Zbornik 27. mednarodne multikonference

INFORMACIJSKA DRUŽBA - IS 2024

Zvezek C

Proceedings of the 27th International Multiconference

INFORMATION SOCIETY – IS 2024

Volume C

Odkrivanje znanja in podatkovna skladišča - SiKDD
Data Mining and Data Warehouses - SiKDD

Uredniki / Editors

Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

7. oktober 2024 / 7–11 October 2024
Ljubljana, Slovenia

Uredniki:

Dunja Mladenić
Department for Artificial Intelligence
Jožef Stefan Institute, Ljubljana

Marko Grobelnik
Department for Artificial Intelligence
Jožef Stefan Institute, Ljubljana

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Mitja Zemljak
Oblikovanje naslovnice: Vesna Lasič

Dostop do e-publikacije:
http://library.ijs.si/Stacks/Proceedings/InformationSociety

Ljubljana, oktober 2024

Informacijska družba
ISSN 2630-371X

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2024

Leto 2024 je hkrati udarno in tradicionalno. Že sedaj, še bolj pa v prihodnost bosta računalništvo, informatika (RI) in umetna inteligenca (UI) igrala ključno vlogo pri oblikovanju napredne in trajnostne družbe. Smo na pragu nove dobe, v kateri generativna umetna inteligenca, kot je ChatGPT, in drugi inovativni pristopi utirajo pot k superinteligenci in singularnosti, ključnim elementom, ki bodo definirali razcvet človeške civilizacije. Naša konferenca je zato hkrati tradicionalna znanstvena, pa tudi povsem akademsko odprta za nove pogumne ideje, inkubator novih pogledov in idej.

Letošnja konferenca ne le da analizira področja RI, temveč prinaša tudi osrednje razprave o perečih temah današnjega časa – ohranjanje okolja, demografski izzivi, zdravstvo in preobrazba družbenih struktur. Razvoj UI ponuja rešitve za skoraj vse izzive, s katerimi se soočamo, kar poudarja pomen sodelovanja med strokovnjaki, raziskovalci in odločevalci, da bi skupaj oblikovali strategije za prihodnost. Zavedamo se, da živimo v času velikih sprememb, kjer je ključno, da s poglobljenim znanjem in inovativnimi pristopi oblikujemo informacijsko družbo, ki bo varna, vključujoča in trajnostna.

Letos smo ponosni, da smo v okviru multikonference združili dvanajst izjemnih konferenc, ki odražajo širino in globino informacijskih ved: CHATMED v zdravstvu, Demografske in družinske analize, Digitalna preobrazba zdravstvene nege, Digitalna vključenost v informacijski družbi – DIGIN 2024, Kognitivna znanost, Konferenca o zdravi dolgoživosti, Legende računalništva in informatike, Mednarodna konferenca o prenosu tehnologij, Miti in resnice o varovanju okolja, Odkrivanje znanja in podatkovna skladišča – SIKDD 2024, Slovenska konferenca o umetni inteligenci, Vzgoja in izobraževanje v RI.

Poleg referatov bodo razprave na okroglih mizah in delavnicah omogočile poglobljeno izmenjavo mnenj, ki bo oblikovala prihodnjo informacijsko družbo. “Legende računalništva in informatike” predstavljajo slovenski “Hall of Fame” za odlične posameznike s tega področja, razširjeni referati, objavljeni v reviji *Informatica* z 48-letno tradicijo odličnosti, in sodelovanje s številnimi akademskimi institucijami in združenji, kot so ACM Slovenija, SLAIS in Inženirska akademija Slovenije, bodo še naprej spodbujali razvoj informacijske družbe. Skupaj bomo gradili temelje za prihodnost, ki bo oblikovana s tehnologijami, osredotočena na človeka in njegove potrebe.

S podelitvijo nagrad, še posebej z nagrado Michie-Turing, se avtonomna RI stroka vsakoletno opredeli do najbolj izstopajočih dosežkov. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejel prof. dr. XXXXX. Priznanje za dosežek leta pripada XXX za XXX. »Informacijsko limono« za najmanj primerno informacijsko tematiko je prejela XXX, »informacijsko jagodo« kot najboljšo potezo pa dobi XXX za XXX. Čestitke nagrajencem!

Naša vizija je jasna: prepoznati, izkoristiti in oblikovati priložnosti, ki jih prinaša digitalna preobrazba, ter ustvariti informacijsko družbo, ki bo koristila vsem njenim članom. Vsem sodelujočim se zahvaljujemo za njihov prispevek k tej viziji in se veselimo prihodnjih dosežkov, ki jih bo oblikovala ta konferenca.

Mojca Ciglarič, predsednica programskega odbora

Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2024

The year 2024 is both ground-breaking and traditional. Now, and even more so in the future, computer science, informatics (CS/I), and artificial intelligence (AI) will play a crucial role in shaping an advanced and sustainable society. We are on the brink of a new era where generative artificial intelligence, such as ChatGPT, and other innovative approaches are paving the way for superintelligence and singularity—key elements that will define the flourishing of human civilization. Our conference is therefore both a traditional scientific gathering and an academically open incubator for bold new ideas and perspectives.

This year's conference analyzes key CS/I areas and brings forward central discussions on pressing contemporary issues—environmental preservation, demographic challenges, healthcare, and the transformation of social structures. AI development offers solutions to nearly all challenges we face, emphasizing the importance of collaboration between experts, researchers, and policymakers to shape future strategies collectively. We recognize that we live in times of significant change, where it is crucial to build an information society that is safe, inclusive, and sustainable, through deep knowledge and innovative approaches.

This year, we are proud to have brought together twelve exceptional conferences within the multiconference framework, reflecting the breadth and depth of information sciences:

- CHATMED in Healthcare
- Demographic and Family Analyses
- Digital Transformation of Healthcare Nursing
- Digital Inclusion in the Information Society – DIGIN 2024
- Cognitive Science
- Conference on Healthy Longevity
- Legends of Computer Science and Informatics
- International Conference on Technology Transfer
- Myths and Facts on Environmental Protection
- Data Mining and Data Warehouses – SIKDD 2024
- Slovenian Conference on Artificial Intelligence
- Education and Training in CS/IS.

In addition to papers, roundtable discussions and workshops will facilitate in-depth exchanges that will help shape the future information society. The “Legends of Computer Science and Informatics” represents Slovenia’s “Hall of Fame” for outstanding individuals in this field, while extended papers published in the Informatica journal, with over 48 years of excellence, and collaboration with numerous academic institutions and associations, such as ACM Slovenia, SLAIS, and the Slovenian Academy of Engineering, will continue to foster the development of the information society. Together, we will build the foundation for a future shaped by technology, yet focused on human needs.

The autonomous CS/IS community annually recognizes the most outstanding achievements through the awards ceremony, especially the Michie-Turing Award. The Michie-Turing Award for an exceptional lifetime contribution to the development and promotion of the information society has been awarded to Prof. Dr. XXXXX. The Achievement of the Year Award goes to XXX for XXX. The "Information Lemon" for the least appropriate information topic was awarded to XXX, while the "Information Strawberry" for the best initiative goes to XXX for XXX. Congratulations to all the award winners!

Our vision is clear: to recognize, seize, and shape the opportunities brought by digital transformation, and to create an information society that benefits all its members. We thank all participants for their contributions and look forward to the future achievements from this conference.

Mojca Ciglarič, Chair of the Program Committee

Matjaž Gams, Chair of the Organizing Committee

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia
Sergio Campos-Cordobes, Spain
Shabnam Farahmand, Finland
Sergio Crovella, Italy

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Mitja Lasič
Blaž Mahnič

Programme Committee

Mojca Ciglarič, chair
Bojan Orel
Franc Solina
Viljan Mahnič
Cene Bavec
Tomaž Kalin
Jozsef Györkös
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Mitja Luštrek
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah
Niko Zimic
Rok Piltaver
Toma Strle
Tine Kolenik
Franci Pivec
Uroš Rajkovič
Borut Batagelj
Tomaž Ogrin
Aleš Ude
Bojan Blažica
Matjaž Kljun
Robert Blatnik
Erik Dovgan
Špela Stres
Anton Gradišek

KAZALO / TABLE OF CONTENTS

Zbornik 27. mednarodne multikonference

INFORMACIJSKA DRUŽBA - IS 2024

Zvezek C

Proceedings of the 27th International Multiconference

INFORMATION SOCIETY – IS 2024

Volume C

Odkrivanje znanja in podatkovna skladišča - SiKDD
Data Mining and Data Warehouses - SiKDD

Uredniki / Editors

Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

7. oktober 2024 / 7–11 October 2024
Ljubljana, Slovenia

PREDGOVOR

Tehnologije, ki se ukvarjajo s podatki so močno napredovale. Iz prve faze, kjer je šlo predvsem za shranjevanje podatkov in kako do njih učinkovito dostopati, se je razvila industrija za izdelavo orodij za delo s podatkovnimi bazami in velikimi količinami podatkov, prišlo je do standardizacije procesov, povpraševalnih jezikov. Ko shranjevanje podatkov ni bil več poseben problem, se je pojavila potreba po bolj urejenih podatkovnih bazah, ki bi služile ne le transakcijskem procesiranju ampak tudi analitskim vpogledom v podatke. Pri avtomatski analizi podatkov sistem sam pove, kaj bi utegnilo biti zanimivo za uporabnika – to prinašajo tehnike odkrivanja znanja v podatkih (knowledge discovery and data mining), ki iz obstoječih podatkov skušajo pridobiti novo znanje in tako uporabniku nudijo novo razumevanje dogajanj zajetih v podatkih. Slovenska KDD konferenca SiKDD, pokriva vsebine, ki se ukvarjajo z analizo podatkov in odkrivanjem znanja v podatkih: pristope, orodja, probleme in rešitve.

FOREWORD

Data driven technologies have significantly progressed. The first phases were mainly focused on storing and efficiently accessing the data, resulted in the development of industry tools for managing large databases, related standards, supporting querying languages, etc. After the initial period, when the data storage was not a primary problem anymore, the development progressed towards analytical functionalities on how to extract added value from the data; i.e., databases started supporting not only transactions but also analytical processing of the data. In automatic data analysis, the system itself tells what might be interesting for the user - this is brought about by knowledge discovery and data mining techniques, which try to obtain new knowledge from existing data and thus provide the user with a new understanding of the events covered in the data. The Slovenian KDD conference SiKDD covers topics dealing with data analysis and discovering knowledge in data: approaches, tools, problems and solutions.

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Janez Brank, Jožef Stefan Institute, Ljubljana

Marko Grobelnik, Jožef Stefan Institute, Ljubljana

Alenka Guček, Jožef Stefan Institute, Ljubljana

Branko Kavšek, University of Primorska, Koper

Dunja Mladenić, Jožef Stefan Institute, Ljubljana

Erik Novak, Jožef Stefan Institute, Ljubljana

Inna Novalija, Jožef Stefan Institute, Ljubljana

Joao Pita Costa, Quintelligence, Ljubljana

Lui Rei, Event Registry, Ljubljana

Jože Rožanec, Jožef Stefan Institute, Ljubljana

Abdul Sitar, Jožef Stefan Institute, Ljubljana

Luka Stopar, SolvesAll, Ljubljana

Swati Swati, Bundeswehr University Munich, Munich

Jan Šturm, Jožef Stefan Institute, Ljubljana

Oleksandra Topal, Jožef Stefan Institute, Ljubljana

Integrating Knowledge Graphs and Large Language Models for Querying in an Industrial Environment

Domen Hočevar
domenhocevar1@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Klemen Kenda
klemen.kenda@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Knowledge graphs have traditionally required the use of specific query languages, such as SPARQL, to retrieve relevant data. In this paper, we present a system capable of performing natural language queries on knowledge graphs by leveraging retrieval-augmented generation (RAG) and large language models (LLMs). Our system can ingest large knowledge graphs and answer queries using two approaches: first, by utilizing LLMs to extract information directly from subgraphs; and second, by generating SPARQL queries with LLMs and using the results to inform further inference, such as counting the number of items.

Keywords

knowledge graph, semantic inference, Industry 4.0, LLM, RAG

1 Introduction

In the context of Industry 4.0, knowledge graphs play a crucial role in mapping and describing the entire production vertical, from supply and demand dynamics to intricate details within the production process. This includes the configuration of shop floors, production lines, machines, and data setups, extending even to specific datasets generated during operations. Knowledge graphs can also include relevant information about the tools required for particular processes, as well as details about personnel, including their skills and roles.

A key standard for representing such data within the Industry 4.0 initiative is the Asset Administration Shell (AAS) [3], which provides a logical representation for a factory asset (can also be a piece of software, etc.). By adopting AAS, industries can ensure interoperability and standardization, enabling more efficient data exchange and integration across various systems, ultimately enhancing the agility and responsiveness of manufacturing processes.

Querying knowledge graphs can be a challenging task for end users, as it often requires expertise in specialized query languages such as SPARQL [8] — a skill that is not widely known among non-experts. Working with SPARQL SELECT queries remains a challenge also for LLMs, with performance varying significantly depending on the specific model and task complexity. While the leading LLMs can reliably address basic syntax errors, generating semantically accurate SPARQL SELECT queries remains difficult in many cases [10]. Similar work has been done on interaction with databases, however even with SQL query generation the results of GPT-4 are still far behind human ability (approx. 55% execution accuracy) [9].

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.5

To overcome these challenges, we propose a system that enables users to interact with knowledge graphs through natural language queries. The system leverages LLMs' capabilities to interpret knowledge graphs while compensating for their limited ability to generate fully syntactically and semantically correct SPARQL queries. Proposed system, depicted in Figure 1, leverages large language models (LLMs) [11] to process natural language inputs and provide responses in natural language. Our approach integrates retrieval-augmented generation (RAG) techniques alongside the automatic generation of SPARQL queries based on natural language input [2].

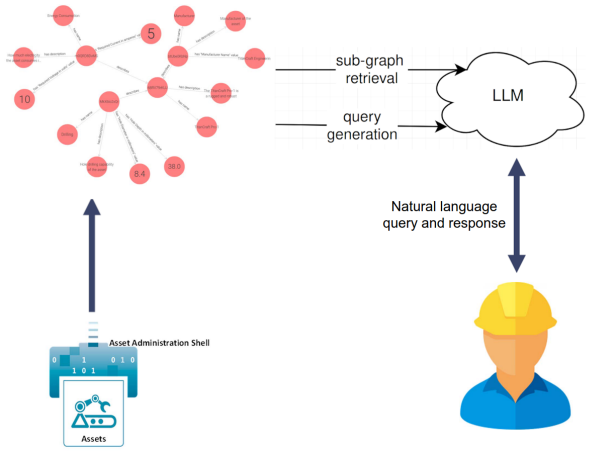


Figure 1: Intended usage of the system: AAS instances are converted into a knowledge graph, enabling natural language queries by the user.

By doing so, our system not only simplifies the querying process but also ensures that the responses are accurate and contextually relevant, making knowledge graphs more accessible and usable for a broader range of users. Additionally, the use of LLMs in combination with SPARQL querying enables the system to handle complex tasks, including those that require logical reasoning, aggregation, or interpretation of data, thus enhancing its utility in real-world applications. For example, our system is able to answer queries such as: "Give me all machines that are capable of drilling a hole with 2cm perimeter".

Finally, question answering with the help of knowledge graphs and language models has been tackled before [16], however, the development of retrieval-augmented generation (RAG) systems has seen significant growth recently. In 2024, several preprints have emerged showcasing the application of the RAG approach to knowledge graphs [12, 13, 14]. This paper contributes to this rapidly evolving field by presenting our own advancements and findings.

2 Data

This study uses a generated dataset representing a hypothetical factory with various machine models, designed to test the capabilities of the developed application. The work is part of the Smart Manufacturing pilot in the EU-funded HumAIne project [7], with the aim of eventually using real-world data from participating factories.

The mock factory includes models of "DRILLERS", "CIRCLE CUTTERS", and "CIRCULAR SAWS", each with unique names, manufacturers, and descriptions. These models are represented using AASs with relevant submodels for energy consumption, manufacturer details, and operation-specific parameters like hole diameter or depth of cut.

We created AASs for 7 drilling machine models, 7 circle cutter models, and 10 circular saw models, along with 1,000 machine instances randomly assigned to these models. Numerical values and availability were populated randomly for testing, reflecting potential real-world variations.

The initial step after acquiring AAS data is to convert it into a knowledge graph. This process involves transforming JSON-serialized AASs into RDF triples, which represent the semantic information of the data. Once the RDF triples are generated, they are stored in a GraphDB[1] repository. To enable semantic data retrieval, we employ a connector that interfaces with the ChatGPT Retrieval Plugin[2], which operates alongside the server application. When new triples are added to the GraphDB repository, the connector triggers the plugin to generate vector embeddings of the text representations of the new nodes. These embeddings are created using a language model and are stored in a separate vector database. The ChatGPT Retrieval Plugin enables interacte to a selection of different vector databases, in our case we employed the Milvus vector database. The system is also designed to maintain consistency; if any triples are removed from the GraphDB repository, the corresponding vector embeddings are automatically deleted from the vector database.

3 Methodology

The system architecture is illustrated in Figure 2. The user interacts with the system through a client application, developed using ReactJS, which serves as the graphical user interface (GUI). This client application communicates with the system's middleware, which is built on the Flask framework. Users have the capability to upload AAS data to construct and enhance the knowledge graph, as well as to issue natural language queries.

The middleware acts as the core of the system, facilitating communication between the client application, the knowledge graph stored in a GraphDB database, and OpenAI's GPT models. The AAS data uploaded by the user is first converted into RDF triples and then stored in the GraphDB repository. The Flask-based middleware also integrates with the ChatGPT Retrieval Plugin, which is responsible for generating vector embeddings of the knowledge graph nodes using OpenAI's `text-embedding-ada-002` model.

These vector embeddings are stored in the Milvus vector database [15]. The ChatGPT Retrieval Plugin allows the system to efficiently retrieve the most relevant embeddings in response to user queries, ensuring that the system can provide accurate and contextually appropriate answers. Additionally, the middleware leverages LlamaIndex[3] to manage sub-graph retrieval and query generation, which are essential for responding to complex queries by the user.

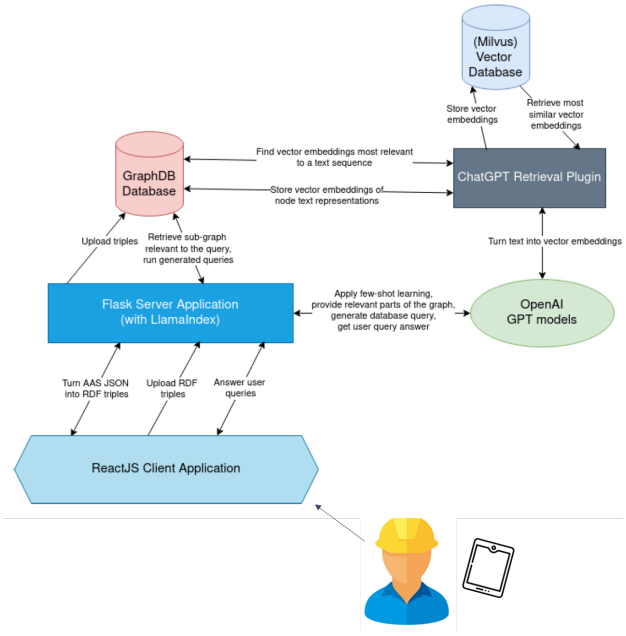


Figure 2: System architecture for retrieval augmented generation with knowledge graphs in Industry 4.0.

In summary, the architecture is designed to streamline the process of building a knowledge graph from AAS data and enables users to query this graph with retrieval-augmented generation (RAG) using natural language, with the system handling the complexities of data storage, retrieval, and natural language processing in the background.

The sequence diagram in Figure 3 illustrates the interaction between system components during query processing. Our system enables two distinct approaches to handle natural language queries, often combining both to generate a comprehensive answer for the user.

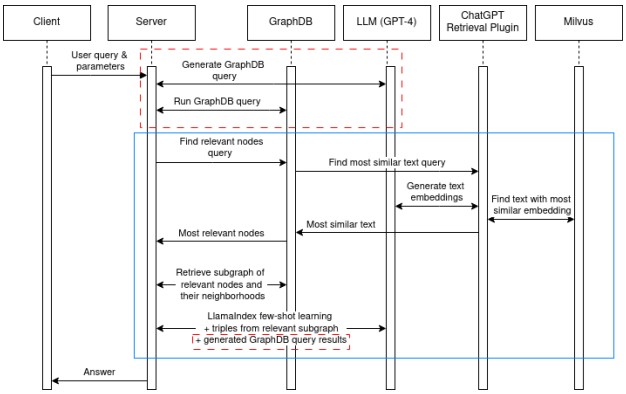


Figure 3: Sequence diagram of different approaches for data extraction. The blue box represents the RAG approach and the red box represents the SPARQL query generation approach. Note that RAG approach utilizes results from SPARQL queries on the knowledge graph.

[1] https://graphdb.ontotext.com/
[2] https://github.com/openai/chatgpt-retrieval-plugin
[3] https://www.llamaindex.ai/

The first approach utilizes a Retrieval-Augmented Generation (RAG) method. Upon receiving a query, the system analyzes the query to identify relevant concepts and generates vector embeddings for these concepts [5]. These embeddings are then matched against the knowledge graph stored in GraphDB to find the most relevant nodes. Once the relevant nodes are identified, a naive neighborhood expansion is performed, capturing additional related nodes to ensure a more complete context. The search is parameterized using parameters: scope, how many nodes from the graph to retrieve; breadth, from how many relevant nodes to start the neighborhood expansion; score weight, how many more nodes are visited from the identified relevant nodes that are deemed more relevant using embedding similarity. This subgraph, along with a few examples for context, is then fed into the Large Language Model (LLM) using a few-shot [1] learning technique to generate a response [4]. The LlamaIndex framework provides a general context query for turning triples into natural language. This method is particularly effective for queries requiring contextual understanding and extraction of complex information from the knowledge graph.

The second approach involves generating a SPARQL query based on the natural language query and the ontology used within the knowledge graph. The system attempts to execute this SPARQL query in the GraphDB database. If the query runs successfully, the resulting data is passed to the LLM to formulate the final answer. This approach is especially beneficial for tasks that involve counting instances or performing specific data aggregation operations, where LLMs alone might struggle. This approach benefits from the first approach as it can use it as backup or to enrich the SPARQL query results with additional context.

4 Results

To thoroughly evaluate the system, we employed three different evaluations: (a) assessing the accuracy of data retrieval based on query parameters (not using query generation), (b) evaluating the system's ability to correctly fetch the number of instances (testing query generation), and (c) conducting a manual assessment of most relevant user queries.

4.1 Accuracy of Data Retrieval

The first approach involved testing the system's ability to accurately retrieve data that met specific query conditions without employing SPARQL query generation. We focused on queries where the user requested a list of machines of a particular type with a voltage requirement less than or equal to a specified value. An example query would be: "Return all drilling machines that consume at most 4 volts and specify their consumption."

We conducted these tests on three types of machines: "drilling machines", "circle cutters", and "circular saws". The voltage values specified in the queries ranged from 0 to 10 volts, inclusive. The evaluation was designed to measure how accurately the system could identify and return the correct set of machines based on these voltage constraints.

For these tests, the following parameters were used (scope: 100, breadth: 1000, score weight: 100, model: `gpt-4-1106-preview`, query generation strategy: **disabled**).

The system's performance was assessed by comparing the retrieved data against the expected results, specifically checking the number of machines that met the voltage criteria and identifying any errors, such as incorrect voltage values or unnecessary machine retrievals. Results are depicted in Figures 4 and 5.

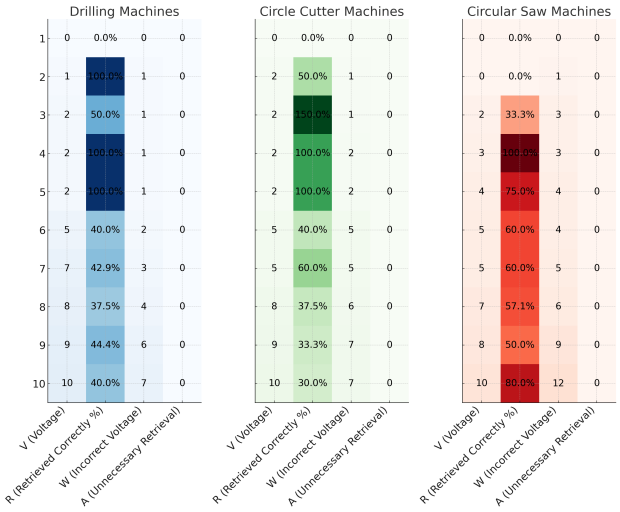


Figure 4: Performance of the system by the type of the machine and query.

In Figure 4, each table contains four columns: "V" (voltage specified in the query), "R" (percentage of correctly retrieved machines), "W" (number of machines with incorrect voltage), and "A" (number of unnecessary machine retrievals). Figure 5 summarizes the results: "Fully Correct Answers" shows the percentage of queries that returned all requested information without errors; "Share of Expected Information Found" indicates the proportion of requested information retrieved; and "Share of Incorrectly Displayed Voltages" represents the percentage of retrieved voltages that were incorrect.

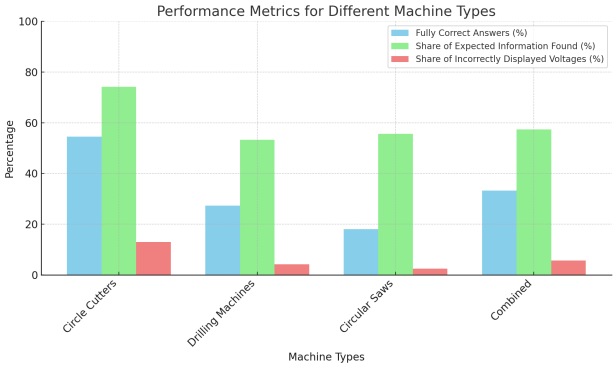


Figure 5: Combined performance.

The results show that sometimes the LLM would incorrectly generate a different voltage requirement for a machine, making it appear to satisfy the query conditions. However, the retrieved machines were always of the correct type. For example, a query like "Name all drilling machines and specify their voltage requirements" correctly retrieves all machines with the right specifications, suggesting the issue may lie with the LLM rather than the knowledge retrieval process.

To address this, users can try adjusting query parameters or rewording the query to verify the information's accuracy. If this

type of query is crucial, incorporating voltage-specific queries into the query generation strategy could improve reliability, although the LLM may struggle with large lists due to its context window limitations. As shown in Figure 5, these types of queries often do not reliably provide all requested information in one answer, so users should run multiple queries to increase the likelihood of retrieving all necessary data.

4.2 Instance Fetching Accuracy

In these tests, we tested query generation strategy. The following parameters were used (scope: 100, breadth: 1000, score weight: 100, model: `gpt-4-1106-preview`, query generation strategy: **enabled**).

The queries asked for the number of available instances for selected machine models, such as "Get the number of available [name of the machine 1], [name of the machine 2] machine instances. Specify the number for each machine type separately.". The query format was picked such that the LLM will benefit from query generation (availability property is specified in the schema supplied for query generation).

A total of 100 queries were run, with 10 queries for each number of specified machine models (ranging from 1 to 10 models). The share of fully correct answers for each query type was between 80 and 100%. The overall accuracy was 96%. This supports our hypothesis that the query generation strategy provides more accurate answers for slightly more complex queries.

4.3 Manual Evaluation of Example Queries

This evaluation was initially performed to identify several shortcomings in our methodologies as mentioned in the previous subsections. By manually evaluating specific queries relevant to end users, we were able to partially address these issues and fine-tune parameters to achieve more accurate results. For instance, while the system's initial results were often incomplete (e. g., query did not return all the machines satisfying certain criteria), increasing the breadth parameter to include a larger subgraph and allowing LLMs to traverse a broader neighborhood improved the results. Additionally, we demonstrated that subgraph retrieval and query generation can complement each other, further enhancing overall performance. All the results are commented in detail in [6].

5 Conclusions

In this paper, we presented a system that bridges the gap between natural language processing and querying knowledge graphs, specifically within the context of Industry 4.0. By leveraging large language models (LLMs) and retrieval-augmented generation (RAG), our system allows users to interact with complex knowledge graphs using natural language queries, thereby simplifying access to detailed manufacturing data.

Our evaluation demonstrated the usability of our system, however with the integration of LLMs for natural language understanding, some challenges remain. These include occasional inaccuracies in data retrieval and the LLM's limited ability to handle large datasets or specific queries. By adjusting subgraph retrieval parameters such as breadth and scope, and by combining it with SPARQL query generation, we were able to significantly enhance the system's accuracy and reliability.

This work highlights the potential of combining knowledge graphs with LLMs to create more intuitive and effective query systems in industrial environments. Future improvements could focus on refining query strategies and further optimizing the balance between subgraph retrieval and SPARQL generation to ensure even more robust and comprehensive query handling.

Acknowledgements

This work was supported by the European Commission under the Horizon Europe project HumAIne, Grant Agreement No. 101120218. We would like to express our gratitude to all project partners for their contributions and collaboration.

References

[1] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[2] Diego Bustamante and Hideaki Takeda. 2024. Sparql generation with entity pre-trained gpt for kg question answering. *arXiv preprint arXiv:2402.00969*.

[3] 2022. Details of the asset administration shell. https://www.plattform-i40.de/IP/Redaktion/EN/Downloads/Publikation/Details_of_the_Asset_Administration_Shell_Part1_V3.pdf?__blob=publicationFile&v=1 (visited on 02/22/2024). (2022).

[4] Chao Feng, Xinyu Zhang, and Zichu Fei. 2023. Knowledge solver: teaching llms to search for domain knowledge from knowledge graphs. *ArXiv*, abs/2309.03118. https://api.semanticscholar.org/CorpusID:261557137.

[5] Luis Gutiérrez and Brian Keith. 2019. A systematic literature review on word embeddings. In *Trends and Applications in Software Engineering: Proceedings of the 7th International Conference on Software Process Improvement (CIMPS 2018) 7*. Springer, 132–141.

[6] Domen Hočevar. 2024. *Integrating Knowledge Graphs and Large Language Models for Querying in an Industrial Environment*. Bachelor's Thesis. University of Ljubljana, Faculty of Computer, Information Science, Faculty of Mathematics, and Physics, Ljubljana, Slovenia, (Aug. 2024). Interdisciplinary University Study Program, First Cycle, Computer Science and Mathematics.

[7] Humaine Horizon. 2024. Humaine horizon. https://humaine-horizon.eu/. Accessed: 2024-08-26. (2024).

[8] Pérez Jorge. 2006. Semantics and complexity of sparql. In *Proc. 5th Int. Semantic Web Conference (ISWC2006)*.

[9] Jinyang Li et al. 2024. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36.

[10] Lars-Peter Meyer, Johannes Frey, Felix Brei, and Natanael Arndt. 2024. Assessing sparql capabilities of large language models. (2024). https://arxiv.org/abs/2409.05925 arXiv: 2409.05925 [cs.DB].

[11] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

[12] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: a roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

[13] Diego Sanmartin. 2024. Kg-rag: bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.

[14] Bhaskarjit Sarmah, Benika Hall, Rohan Rao, Sunil Patel, Stefano Pasquali, and Dhagash Mehta. 2024. Hybridrag: integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. *arXiv preprint arXiv:2408.04948*.

[15] Jianguo Wang et al. 2021. Milvus: a purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, 2614–2627.

[16] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Comparative Analysis of Machine Learning Models for Groundwater Level Forecasting: The Impact of Contextual Data

Rok Klančič
rok.klancic@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Klemen Kenda
klemen.kenda@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

This paper presents a comparative evaluation of three distinct categories of models applied to groundwater level data: traditional batch learning methods, time series deep learning methods, and time series foundation models. By enriching the water level data with weather-related features, we significantly improved the effectiveness of simpler models. The results demonstrate that, despite their state-of-the-art performance on univariate datasets and the corresponding publicity, advanced models without contextual feature support are still surpassed by traditional methods trained on enriched datasets.

Keywords

groundwater level prediction, time series forecasting, deep learning, foundation models, contextual data

1 Introduction

Accurate water level prediction is crucial for mitigating the impacts of climate change on water resources. By forecasting water levels, we can better prepare for potential floods and droughts, and more effectively manage our water supplies. However, predicting water levels presents a significant challenge due to the dynamic nature of the data. As climate change leads to prolonged droughts and increasingly erratic precipitation patterns, the need for reliable forecasting methods becomes even more important [2].

In this paper, we aim to compare the performance of various models in forecasting groundwater levels. Specifically, we focus on the differences between traditional batch learning methods that utilize relevant contextual data and newer univariate time series deep learning and foundation models.

The main contributions of this paper are:

- A comparative analysis of the performance of traditional batch learning methods against state-of-the-art time series deep learning techniques and time series foundation models, particularly in the context of feature vectors enriched with relevant contextual data.
- The application of time series foundation models and deep learning methods to the domain of groundwater level forecasting.

The groundwater dataset used in this study has previously been employed for predictive modeling with traditional batch learning methods [9], where extensive feature engineering was also performed. Our work builds upon and extends this earlier research by incorporating a different set of models.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.6

2 Methods

In our experiments, we employed three categories of methods: traditional batch learning techniques, time series deep learning models, and time series foundation models.

2.1 Traditional Batch Learning Methods

In the context of data-driven modelling of environmental issues, traditional batch learning methods have historically demonstrated significant success [5]. In this study, we employed linear regression alongside two tree-based approaches: random forest and gradient boosting [7] as baselines to evaluate whether the newer, more prominent techniques, which have recently gathered a considerable amount of attention, can perform competitively in this specific setting.

All of the chosen batch learning techniques are regression-based and are valued for their simplicity, speed, and ease of use. However, they often lack the complexity necessary to fully capture intricate patterns in the data. To mitigate this limitation, we incorporated contextual features, such as weather data and forecasts (e.g., precipitation, cloud cover, temperature). While the data fusion problem is solved [8], this approach raises concerns about the availability and relevance of the contextual data.

2.2 Time Series Deep Learning Methods

Time series deep learning models are explicitly designed for forecasting time-dependent data. In our study, we employed N-BEATS [12] and PatchTST [10], both of which have architectures tailored to capture trends and seasonalities inherent in time series data. Despite their advanced capabilities, these models have drawbacks, including longer training and inference times, the necessity for extensive hyperparameter tuning to achieve optimal performance, and limited support for incorporating additional features. Although certain models support multivariate time series, they were not utilized in our experiments.

2.3 Time Series Foundation Models

While deep learning methods require separate training and prediction phases, time series foundation models aim to eliminate the training step. Inspired by large language models, these models are pretrained on extensive time series datasets, enabling zero-shot predictions on new time series without additional training. We used CHRONOS [1], an open source foundation model. The advantages of this approach include ease of use with minimal parameter adjustments and no need for training. However, similar to deep learning models, they lack support for multivariate time series.

Several studies have already evaluated the performance of various deep learning and foundation models for time series forecasting [1] [13]. However, this research extends the application of these forecasting models to groundwater level data, therefore contributing to the better understanding of their effectiveness in this domain.

3 Experiment Setting

The experiments were conducted on a dataset of groundwater levels in Slovenia. Due to the cumulative nature of water levels and to facilitate comparison with the original study [9], predictions were made on daily changes in water levels rather than on absolute values.

3.1 Dataset

The groundwater dataset is a subset of the larger dataset used in the study [9]. It consists of groundwater level measurements taken daily from multiple stations across Slovenia. To apply traditional batch learning methods, we enriched the dataset with weather data, associating each water measurement station with the nearest weather station. Due to the availability of weather data, only data from the years 2010 to 2017 was included in our study. For consistency and ease of comparison with previous study [9], we focused on data from two water measurement stations located in Ljubljana.

In traditional batch learning within the environmental domain, it is essential to not only use the raw data but also to engineer relevant features. Initially, we removed the pressure and dew point features, as they were either unrelated to the target variable or highly correlated with other features [9]. We then created additional features by shifting the data from 1 to 10 days, making historical values available, and by computing the averages of features over a 2- to 10-day window. This process resulted in approximately 2,000 features. Given the excessive number of features, which could degrade model performance, we employed a feature selection algorithm to identify the most informative subset.

We used a genetic feature selection algorithm from scikit-learn, evaluated on 365-day part of training dataset, with the maximum number of features set to 40. The algorithm was executed separately for each model, focusing on one station and a prediction horizon of three days, resulting in distinct feature vectors. Subsequently, weather forecast features with longer offsets were manually added to the selected feature set.

3.2 Evaluation Metrics

The dataset was split into a training set (approx. 2,500 days), a validation set (100 days), and a test set (365 days) for model evaluation. Model performance was evaluated using the R^2 score, averaged across all tested stations. Although alternative metrics such as root-mean-squared error (RMSE), and mean absolute percentage error (MAPE) were considered, they, for this dataset, produce results that are closely related to the R^2. This metric was selected due to its robustness against variations in data offset and amplitude, and for direct comparability with the results in the original study [9]. The R^2 score is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where y_i is the i-th true value, $\hat{y}_i$ is the i-th predicted value and $\bar{y}$ is the average of true values.

3.3 Baseline Methods

The primary objective of our research was to compare the performance of traditional batch learning methods, enriched with relevant contextual features, against that of modern deep learning techniques and foundation models for time series forecasting. Therefore, we selected linear regression, random forest regressor, and gradient boosting regressor as our baseline methods. These models were previously applied to the groundwater dataset [9], necessitating a reproduction of the results as a benchmark.

3.4 Implementation Details

The prediction pipelines varied slightly between the different types of models:

- For **CHRONOS**, we utilized the dataset without weather features, as it only supports univariate time series. Since no hyperparameter tuning was required, the data was divided into training and test sets, omitting the validation set. The model generated the predictions directly from the water level data. We used the chronos-t5-large model from the chronos library.
- For **N-BEATS** and **PatchTST**, the same dataset was used, given the same limitation as mentioned previously. However, a validation set was required for hyperparameter tuning. After selecting appropriate hyperparameters, the models were trained on the training set and evaluated on the test set. Implementations from the NeuralForecast library were used for both models.
- For the **linear regression**, **random forest regressor**, and **gradient boosting regressor** models, we included both water level and weather data. Feature selection was conducted to reduce the number of features, resulting in 42 features for linear regression, 30 for random forest, and 36 for gradient boosting. After feature selection, hyperparameters for the random forest and gradient boosting models were tuned, and the data for linear regression was normalized. The models were then trained on the training set and evaluated on the test set using scikit-learn's implementations.

The hyperparameters used for training are listed in Appendix A, while a description of the selected features is provided in Appendix B.

4 Results

The results for all tested models across various prediction horizons are presented in Table 1. The reported R^2 scores were calculated based on the differences in water levels; if absolute water levels had been used, the R^2 scores would have been significantly higher. For example, in the case of CHRONOS with 1-day ahead predictions, the R^2 score is 0.725 for relative level differences and 0.998 for absolute water levels.

Among the models, linear regression achieved the highest performance, followed by the random forest. In contrast, the more complex methods, including deep learning models and the foundation model, showed generally lower performance, with the exception of the 1-day prediction horizon, where N-BEATS outperformed the tree-based models. Notably, the R^2 scores decrease as the prediction horizon lengthens, with a more pronounced decline observed in the deep learning and the foundation models compared to the traditional batch learning methods.

Figures 2 and 3 display the predictions from CHRONOS, PatchTST, and linear regression compared to the true data for the 1-day and 5-day prediction horizons. It is evident that the predictions from CHRONOS and PatchTST begin to exhibit a rightward shift as the horizon extends. Figure 1 visualizes the R^2 scores for all models across the different prediction horizons.

Table 1: R^2 Scores for Different Prediction Horizons and Models.

Methods	1 day ahead	2 days ahead	3 days ahead	4 days ahead	5 days ahead
Chronos-large	0,725	0,365	0,175	0,04	-0,09
GradientBoostingRegressor	0,640	0,603	0,527	0,556	0,545
RandomForestRegressor	0,726	0,697	0,701	0,706	0,691
N-BEATS	0,742	0,397	0,17	-0,03	-0,143
PatchTST	0,721	0,394	0,215	0,109	-0,02
LinearRegression	**0,792**	**0,781**	**0,785**	**0,784**	**0,780**

The best and second-best results are bolded and underlined respectively.

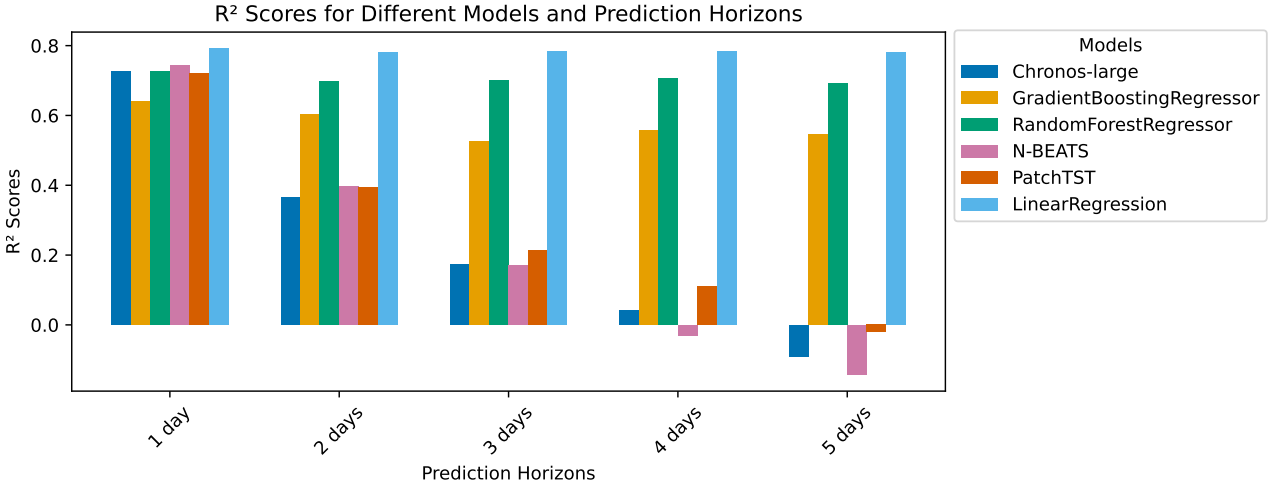


Figure 1: R^2 Scores for All of the Methods and Prediction Horizons.

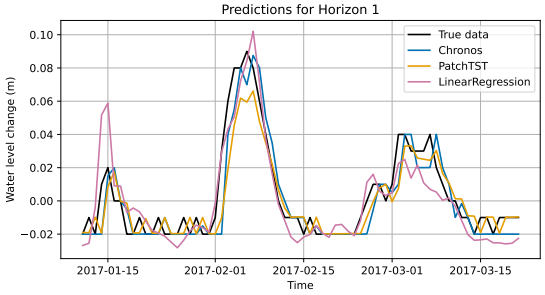


Figure 2: Example Predictions for Three Models for 1-Day Prediction Horizon.

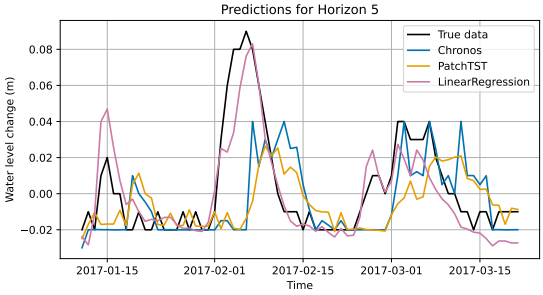


Figure 3: Example Predictions for Three Models for 5-Day Prediction Horizon.

The results indicate that traditional methods, when supplemented with relevant contextual features, outperform more complex models that do not incorporate such data. While the 1-day ahead predictions show comparable performance across all methods, as the prediction horizon extends, the accuracy of CHRONOS, PatchTST, and N-BEATS declines sharply. In contrast, the traditional models, supported by contextual features, maintain their predictive accuracy much more effectively, as shown in Figure 1.

A closer examination of the predictions in Figures 2 and 3 reveals that for 1-day ahead predictions, all models track the true data closely. However, in the 5-day ahead predictions, models lacking contextual data begin to exhibit a rightward shift in their predictions. This likely occurs due to the absence of contextual information, causing these models to lag in capturing the true trajectory of water levels. In contrast, models with access to weather data can predict further ahead by accounting for factors such as the impact of rainfall patterns on water levels.

An unexpected finding is that among the baseline models, linear regression outperforms the more sophisticated methods. For instance, in the article [9], while linear regression produced strong results, it did not surpass the performance of the other two methods.

5 Conclusion and Future Work

After evaluating all models on the groundwater level dataset, we observed that traditional methods, when equipped with relevant features, consistently outperformed newer and more sophisticated techniques, particularly as the prediction horizon lengthened. This suggests that the emphasis on developing the most powerful deep learning or foundation models for time series predictions may be overstated. With thoughtful selection of contextual features, even the simplest models can outperform modern approaches, which is a significant finding for fields with sufficient contextual data, such as data-driven environmental modelling.

To enhance the robustness of our evaluation, future work could involve testing additional methods, expanding the analysis to include more measurement stations and surface water level data, and incorporating deep learning models that support multivariate time series, such as N-BEATSx [11] and N-HiTS [3]. Further insights could be gained by exploring foundation models with multivariate support, such as TimesFM [4], as well as some more univariate models, like TimeGPT-1 [6]. Future research could also compare the inference times of various models and assess performance across different time series lengths.

Acknowledgements

This work was supported by the European Commission under the Horizon Europe project Plooto, Grant Agreement No. 101092008. We would like to express our gratitude to all project partners for their contributions and collaboration.

Furthermore, we would like to thank Erik Novak for his assistance in completing this research.

References

[1] Abdul Fatir Ansari et al. 2024. Chronos: learning the language of time series. *arXiv preprint arXiv:2403.07815*.

[2] ARSO. 2009. Freshwater. Retrieved August 27, 2024 from https://www.arso.gov.si/en/soer/freshwater.html.

[3] Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. 2023. NHiTs: neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* number 6. Vol. 37, 6989–6997.

[4] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2023. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*.

[5] Fan Feng, Hamzeh Ghorbani, and Ahmed E. Radwan. 2024. Predicting groundwater level using traditional and deep machine learning algorithms. *Frontiers in Environmental Science*, 12. DOI: 10.3389/fenvs.2024.1291327.

[6] Azul Garza and Max Mergenthaler-Canseco. 2023. TimeGPT-1. *arXiv preprint arXiv:2310.03589*.

[7] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

[8] Klemen Kenda, Blaž Kažič, Erik Novak, and Dunja Mladenić. 2019. Streaming data fusion for the internet of things. *Sensors*, 19, 8. DOI: 10.3390/s19081955.

[9] Klemen Kenda, Jože Peternelj, Nikos Mellios, Dimitris Kofinas, Matej Čerin, and Jože Rožanec. 2020. Usage of statistical modeling techniques in surface and groundwater level prediction. *Journal of Water Supply: Research and Technology-Aqua*, 69, 3, (Apr. 2020), 248–265. DOI: 10.2166/aqua.2020.143.

[10] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

[11] Kin G. Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. 2023. Neural basis expansion analysis with exogenous variables: forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39, 2, 884–900. DOI: https://doi.org/10.1016/j.ijforecast.2022.03.001.

[12] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2019. N-BEATS: neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.

[13] Hongwei Ye et al. 2024. A transformer-based forecasting model for f10.7 index and its application study on the chinese langfang dataset. *Advances in Space Research*. DOI: https://doi.org/10.1016/j.asr.2024.08.024.

A Hyperparameters

Table 2: Hyperparameters Used for Gradient Boosting Regressor and Random Forest Regressor.

Hyperparameter	GradientBoosting	RandomForest
n_estimators	28	164
max_features	'log2'	0.5
max_depth	10	20

Table 3: Hyperparameters Used for N-BEATS and PatchTST.

Hyperparameter	N-BEATS	PatchTST
loss	HuberLoss	/
n_harmonics	5	/
n_polynomials	5	/
scaler_type	'robust'	/
n_blocks	[3, 3, 1]	/
mlp_units	[[128, 128]]	/
horizon	5	5
input_size	15	71
learning_rate	0.001	0.001
max_steps	25	1323
encoder_layers	/	12
n_heads	/	16
hidden_size	/	64
linear_hidden_size	/	512
dropout	/	0.2
fc_dropout	/	0.1
head_dropout	/	0.1
attn_dropout	/	0.2
patch_len	/	16
stride	/	8
revin	/	True

B Selected Features

Due to the large number of features selected by the feature selection algorithm, we provide a summarized description of the most frequently chosen features. The features that appeared most often include shifts and averages of precipitation, precipitation forecasts, temperature, altitude difference, cloud cover, humidity, and snow accumulation. Notably, the majority of selected features were derived features we generated, with only approximately one original feature being selected per model.

In Table 4, the most common shifts and averages for each individual model are presented. The table indicates that shifts and averages of varying lengths were selected, with a slight preference for shorter ones.

Table 4: Most Frequently Selected Shifts and Averages for Various Methods.

Method	Shifts (days)	Averages (days)
GradientBoostingRegressor	4, 10	2, 6
RandomForestRegressor	2, 6	3, 9
LinearRegression	2, 10	2, 7
Combined	**2, 10**	**2, 3**

Interactive Tool for Tracking Open-source Artificial Intelligence Progress on Hugging Face

Bogdan Šinik
bogdan.sinik@famnit.upr.si
UP FAMNIT
Koper, Slovenia

Domen Vake
domen.vake@famnit.upr.si
UP FAMNIT
Koper, Slovenia

Jernej Vičič
jernej.vicic@upr.si
UP FAMNIT, UP IAM
Koper, Slovenia

Aleksandar Tošić
aleksandar.tosic@upr.si
UP FAMNIT, InnoRenew CoE
Koper, Slovenia

Abstract

Given its increasing importance in our daily lives, Artificial Intelligence has become a prominent subject that needs extensive investigation and understanding. This study presents an analysis of the open-source community in the field of Artificial Intelligence (AI). Various questions arise anytime AI is introduced. open-source AI introduces additional concerns. Should artificial intelligence (AI) be universally accessible, or should it be restricted to private use? Is it worthwhile to offer basic models to the broad user population? We chose the most important data from the primary website in the field, Hugging Face. We have developed a tool that allows for straightforward monitoring of the progress of various open-source AI models using data obtained from their leader board. The platform offers accessible and valuable information about various AI models, including their architectures and the activities of authors. Through performing a quick review with our tool, it becomes evident that the open-source community is becoming large and has an undeniable impact on the AI community.

Keywords

LLM, open-source, AI, Hugging Face

1 Introduction

Artificial intelligence, particularly large language models (LLMs), is an important topic in the computer industry today. Despite the numerous fears and dogmas around it, it is certain that AI has become an integral aspect of our lives. This research has specifically concentrated on the development of a tool for monitoring the impact of the open-source community in the area of artificial intelligence. As implied, these models are accessible to all individuals. There is considerable debate on whether this type of technology should be universally accessible. We wanted to investigate if the open-source community is actively contributing to the development of the field, regardless of one's philosophical convictions. Due to the substantial computational requirements, it was previously impossible to execute Large Language Models on personal computers. As increasingly compact versions with impressive capabilities are being produced, this scenario undergoes a significant transformation. Currently, it is feasible to execute your own model, as long as it is of a modest enough size, on a home computer's graphics processing unit (GPU), even if the GPU is a few years old [9]. The rise in accessibility also enables a larger community to test and develop new solutions and build on top of existing models. We believe that there is a big lack of tools for monitoring the impact of this movement.

Hugging Face [1] has grown into one of the primary platforms for the open-source community. Users are able to download and interact with all significant open-source models. Subsequently, users have the option to publish their models on the platform and compare their performance by adding them to the leaderboard, where all the models are benchmarked and ranked. The open-source community relies heavily on the distribution of models by large corporations, as creating a model from scratch is a hard undertaking [9]. This tool facilitates collaboration among open-source contributors, enabling them to collectively generate social media content, exchange ideas, and even publish concise articles. In addition to the models, they have the ability to generate and upload useful datasets. It represents the most advanced and innovative developments in the field of open-source AI and Machine learning.

An issue that has been observed is the absence of effective visualization tools on Hugging Face, which would enable users to easily see patterns and gain a comprehensive understanding of the open-source AI area. In order to address this issue, we have developed a sophisticated tool that offers users various viewpoints on the data.

2 Literature review

Large Language Models (LLMs) have proven essential in enhancing software engineering (SE) tasks, demonstrating their effectiveness in code comprehension Similar to conventional software engineering tools, open-source cooperation is essential for achieving superior products in this area. [8]

The article authored by Patel et al. [9] emphasizes the significance of the open-source AI community and elucidates its rapid growth in the wake of major industry leaders like Google, Microsoft, and OpenAI. An important milestone in this subject is often emphasized as the day when the LLama model was initially made available to the open-source community. The community promptly recognized the possibilities and potential involved in this release.

Due to its continuous growth, Hugging Face has emerged as the primary platform for exchanging machine learning (ML) models, resulting in an increasing level of complexity. A relational

[1] https://huggingface.co/

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.1

database called HFCommunity was established to facilitate the analysis and resolution of this issue [1].

As previously said, open-source AI models offer an extensive range of possibilities. At the recent conference, the authors [12] demonstrated their effective use of Hugging Face. Due to the significant difficulty in developing a model with broad intelligence, researchers have merged ChatGPT capabilities with models from HuggingFace using agentic architecture to get impressive results in multiple domains. ChatGPT was tasked with creating a plan of action and assigning specific duties to each open-source model based on their own areas of expertise. This is an excellent demonstration of the influence and capabilities of the open-source community, given the familiarity with open models and their capabilities.

The article [6] examines the vulnerabilities associated with open-source AI. A much higher number of repositories with high vulnerabilities has been discovered compared to those with low vulnerabilities, particularly in root repositories. This emphasizes the significance of ensuring the security of technology in order to facilitate its utilization.

In a recent paper [10], authors have analyzed the transparency of Hugging Face pre-trained models regarding database usage and licenses. The analysis revealed that there is often a lack of transparency regarding the training datasets, inherent biases, and licensing details in pre-trained models. Additionally, this research identified numerous potential licensing conflicts involving client projects. 159,132 models were examined. It was found that merely 14% of these models explicitly identify their datasets with specific tags. Furthermore, a detailed examination of a statistically significant sample comprising 389 of the most frequently downloaded models showed that 61% documented their training data in some form.

3 Methodology

We obtained the data by extracting the Open LLM Leaderboard from Hugging Face [2] by saving the data server sent to the client. This data contains information about repositories of models that are currently on the leaderboard and the models that are waiting to be evaluated for the leaderboard. A Python pipeline was developed to clean and enrich this data available on [2]. The leaderboard data includes model architecture and precision as well as the model type and performance on the following benchmarks: ARC[3], HellaSwag[14], MMLU[5], TruthfulQA[7], Winogrande[11] and GSM8K[4]. In addition to the data provided on the leaderboard, additional information on the given models was obtained by using the HF API client. This included data about repository contributors, tags, base models, used datasets, and repo activity. It is important to note that the data is self-reported by the developers and is not enforced by HuggingFace. Additionally, the leaderboard includes duplicates due to developers being able to replace models in the repository with different models under the same name. This means the duplicates have the same repository data but distinct performances. Due to the inability to programmatically determine the current model in the repository, we chose the best-performing model under the repository name as the model representing the repository when removing duplicates. Thus, all datasets were generated for further utilization. The following analysis was conducted using the R programming language. The data was mostly studied via the perspective of time, as our focus was on identifying any obvious trends. The data was categorized using several criteria, such as model type, model architecture, and amount of parameters. The data was initially selected and aggregated to ensure that all crucial components were easily accessible. All models that were categorized as flagged have been excluded from the dataset. In addition, we have collected data on the authors' activities and conducted a study on that particular aspect. Once the data had been cleaned and prepared for visualization, we utilized the R ggplot library to create visual representations of the data. A comprehensive R Shiny app was developed by aggregating all the visuals. We chose to utilize Shiny because it is a great option for constructing interactive data analysis solutions due to several factors. Firstly, it enables the development of web applications that are capable of responding and adapting to real-time changes and user interactions. This simplifies the process of exploring and analyzing data. Shiny easily incorporates with R, utilizing its robust statistical and graphical functionalities to generate complex, interactive visualizations without the need for experience in web technologies such as HTML, CSS, or JavaScript. [13] Finally, our application was deployed to a server, making it accessible online.

4 Results

The outcome of this study is the tool we have developed. The link may be accessed via the following URL. [3]. It has six distinct viewpoints, all conveniently accessible inside its tab. The initial figure, labeled as 1, displays both the count of new models and the distribution of various model types. Hugging Face has identified five distinct categories of models: basic mergers and moerges, fine-tuned on domain-specific datasets, chat models, continuously pretrained models, and pretrained models. If the model did not belong to any of these classes, its type was classified as unknown. The user has the ability to effortlessly choose their preferred categories, along with the desired time frame and unit of aggregate (daily, weekly, or monthly). This allows the viewer to clearly observe the evolution of model types and their popularity over time. It is evident that fine-tuned models are predominantly utilized. This is logical, as users are adapting base models by training them on unique datasets to achieve specialization. Also, we can see that merged models are a relatively recent phenomenon.

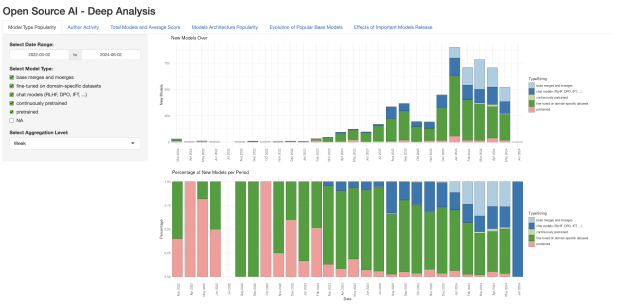


Figure 1: Popularity by model type over time

The second view, referenced as 2, has two interconnected visualizations. The upper section displays the activity of the top 10 authors within a specific range of dates. The display showcases every model they have developed, along with its corresponding type. The lower section presents the average benchmark score for

[2]https://github.com/VakeDomen/HF_analysis

[3]https://oai.dltlt.famnit.upr.si/

each model, organized by author. This visualization enables users to effortlessly monitor the most prominent authors and observe their patterns and accomplishments in model development over time. Users have the ability to effortlessly choose a certain range of dates and also narrow down the list to the top 10 authors according to their preferences. It is evident that leading authors typically do not adhere to trends and consistently provide models of similar type.

Figure 2: Top authors activity over time

The following perspective 3 illustrates two aspects. The first aspect is the alteration in the average benchmark score for each model type as time progresses. The display showcases the top-performing model for each category and time interval (daily, weekly, or monthly). In addition to the dots representing each model, we have incorporated a smooth line to aid the user in seeing the temporal changes for a particular model type. Following the first visualization, we have included a second visualization that displays the total number of models for each model type within the chosen period range. Through these visualizations, users can easily identify the model type that experienced the most improvement and the model types that were mainly produced. We can see the trend, which indicates that open-source AI models are improving, as evidenced by the improvement in average benchmark scores across most of them. The overall number of models is rapidly increasing, indicating a rise in the popularity of open-source AI models.

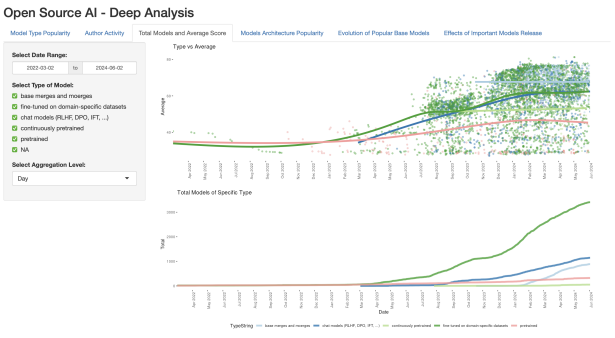


Figure 3: Change of benchmark score and total models per type over time

The fourth perspective, as seen in Figure 4, examines the changing popularity of various model architectures throughout time. The following architectures have been chosen for this specific objective: LLama, Mixtral, Mistral, Qwen2, Gemma, Phi, Opt, GPT2, and GPT2-NeoX. All architectures that did not fit into any one category were classed as "Other". This perspective has two graphics that depict popularity. The first comparison assesses the popularity of a model relative to itself, depending on the number of new models introduced before. The second one compares it to the average number of new models created, taking into account their architecture. Both are depicted by coloring the area, as it is the most convenient way to track. Users may analyze the fluctuation in popularity of well-known model architectures over time and examine how the rising popularity of a particular architecture might impact the popularity of a certain architecture of interest. The lower plot indicates that LLama and Mistral are the predominant models; nonetheless, they have experienced fluctuations throughout time, as visible on the upper plot.

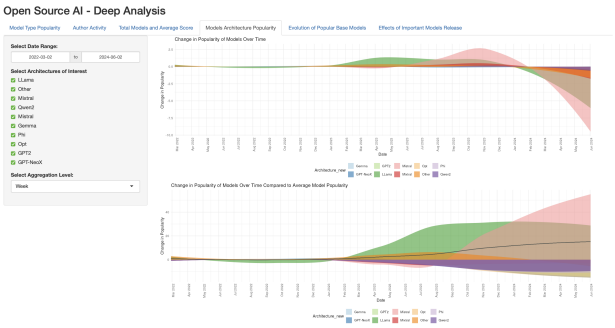


Figure 4: Change of popularity of main architectures over time

The graphic labeled as 5 illustrates the progressive improvement of the key base models developed by famous companies. This was accomplished by isolating each incremental improvement in score over time, using the base model as a reference. In order to fulfill this objective, we have chosen five distinct variations of LLama, Mistral, and Mixtral, as well as three iterations of Phi. The user may easily observe the overall improvement in benchmark scores for each base model. In addition, users have the ability to view the overall duration required for the model to achieve its maximum performance. We have included a feature that enables users to toggle the visibility of model labels, hence enhancing visibility and facilitating more in-depth examination according to their preferences. This allows the user to observe the speed at which specific models reached their peak performance and the extent of their improvement relative to the base models.

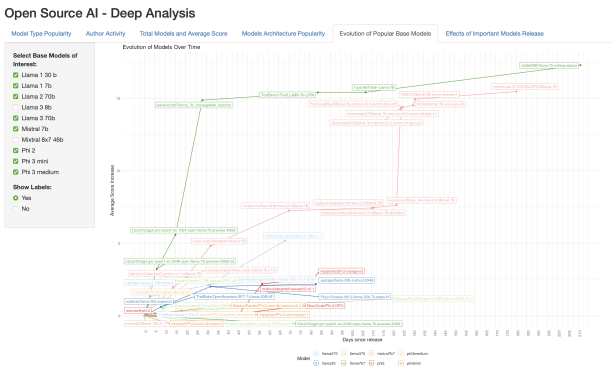


Figure 5: Evolution of famous base models

The final view, as depicted in Figure 6, illustrates the impact of significant releases on the popularity of various model designs. As we have employed identical model designs to those in view four, we have extracted and categorized all significant release dates of these models. The user has the option to choose the time unit for aggregate, which can be either day, week, or month. Users may quickly analyze the impact of significant releases and observe how they influence the popularity and mass creation of specific models. We can observe the evident impact of the recent releases of LLama and Mistral for their popularity.

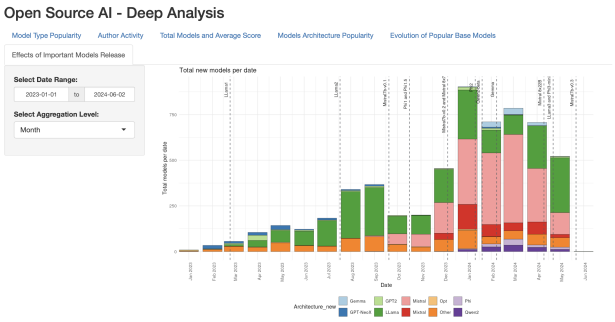


Figure 6: Effect of big releases on architecture of produced models

5 Conclusion and future work

Given the growing importance of Artificial Intelligence in modern culture, it is beneficial to explore the free solutions that are accessible rather than just depending on commercial alternatives. This paper offers valuable insights into a tool designed to simplify the examination of trends in open-source AI in a user-friendly manner. It offers various viewpoints and enables users to acquire knowledge and reach certain conclusions about the subject. Hugging Face has the capability to function as an excellent tool for finding a certain model. As time progresses, open-source AI is expected to provide a growing contribution to the AI community and provide more specific applications for models that could be ignored by big organizations.

We aim to enhance the functionality of our Shiny application by incorporating more perspectives and expanding the range of data interaction options. Our objective is to ensure that the system is as updated as possible. Besides that, we want to conduct a comprehensive analysis of the data to identify patterns and correlations inside this group. We aim to assess the potential of these models and examine their capabilities and potential uses in addressing real-world issues. We would like to analyze the sustained popularity and efficacy of these models over a longer time frame.

References

[1] Adem Ait, Javier Luis Cánovas Izquierdo, and Jordi Cabot. 2023. Hfcommunity: a tool to analyze the hugging face hub community. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 728–732. DOI: 10.1109/SANER56733.2023.00080.

[2] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard. (2023).

[3] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. (2018). arXiv: 1803.05457 [cs.AI].

[4] Karl Cobbe et al. 2021. Training verifiers to solve math word problems. (2021). arXiv: 2110.14168 [cs.CL].

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. (2021). arXiv: 2009.03300 [cs.CY].

[6] Adhishree Kathikar, Aishwarya Nair, Ben Lazarine, Agrim Sachdeva, and Sagar Samtani. 2023. Assessing the vulnerabilities of the open-source artificial intelligence (ai) landscape: a large-scale analysis of the hugging face platform. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1–6. DOI: 10.1109/ISI58743.2023.10297271.

[7] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: measuring how models mimic human falsehoods. (2022). arXiv: 2109.07958 [cs.CL].

[8] Zhihao Lin et al. 2024. Open-source ai-based se tools: opportunities and challenges of collaborative software learning. *arXiv preprint arXiv:2404.06201*.

[9] Dylan Patel and Afzal Ahmad. 2023. Google "we have no moat, and neither does openai.". *SemiAnalysis. May*, 4, 2023.

[10] Federica Pepe, Vittoria Nardone, Antonio Mastropaolo, Gerardo Canfora, Gabriele Bavota, and Massimiliano Di Penta. 2024. How do hugging face models document datasets, bias, and licenses? an empirical study.

[11] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: an adversarial winograd schema challenge at scale. (2019). arXiv: 1907.10641 [cs.CL].

[12] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: solving ai tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems*. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors. Vol. 36. Curran Associates, Inc., 38154–38180. https://proceedings.neurips.cc/paper_files/paper/2023/file/77c33e6a367922d003ff102ffb92b658-Paper-Conference.pdf.

[13] Carson Sievert. 2020. *Interactive web-based data visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

[14] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: can a machine really finish your sentence? (2019). arXiv: 1905.07830 [cs.CL].

Multilingual Hate Speech Modeling by Leveraging Inter-Annotator Disagreement

Patricia-Carla Grigor*
University of Vienna
Vienna, Austria

Bojan Evkoski
evkoski_bojan@phd.ceu.edu
Central European University
Vienna, Austria

Petra Kralj Novak
novakpe@ceu.edu
Central European University
Vienna, Austria
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

As social media usage increases, so does the volume of toxic content on these platforms, motivating the Machine Learning (ML) community to focus on automating hate speech detection. While modern ML algorithms are known to provide nearly human-like results for a variety of downstream Natural Language Processing (NLP) tasks, the classification of hate speech is still an open challenge, partially due to its subjective annotation, which often leads to disagreement between annotators. This paper adopts a perspectivist approach that embraces subjectivity, leveraging conflicting annotations to enhance model performance in real-world scenarios. A state-of-the-art multilingual language model for hate speech detection is introduced, trained, and evaluated using diamond standard data with metrics that consider disagreement. Various strategies for incorporating disagreement are compared in the process. Results demonstrate that the model performs equally or better on all evaluated languages compared to respective monolingual models and drastically outperforms on multilingual data. This highlights the effectiveness of multilingual and perspectivist methods in addressing the complexities of hate speech detection. The presented multilingual hate speech detection model is available at: https://huggingface.co/IMSyPP/hate_speech_multilingual.

Keywords

hate speech detection, inter-annotator disagreement, multilingual language modeling

1 Introduction

The phenomenon of hate speech, which is typically defined as offensive or derogatory language targeting individuals or groups based on characteristics such as race, religion, ethnic origin, sexual orientation, disability, or gender [2], has become a significant problem on social networks in recent years, with communities being increasingly exposed to toxic content as the networks grow and become more interconnected [13, 3]. Consequently, the Machine Learning (ML) and computational linguistics communities have begun developing content moderation strategies using advanced algorithms and Natural Language Processing (NLP) techniques to detect hate speech [10, 11]. However, a key challenge is the subjectivity of hate speech, as annotators often disagree due to diverse backgrounds and perspectives.

To address this challenge, researchers have proposed alternative methodologies to ground-truthing, including the incorporation of diverse perspectives into the training and evaluation pipelines of ML models [1, 14]. One such approach is introduced by [7], who train monolingual hate speech classifiers in several languages directly on datasets that include disagreement. As an alternative to gold-standard data, such data is referred to as diamond standard data, based on the assumption that more than one single truth exists. In terms of evaluation, the researchers focus on the evaluation of models from the perspective of disagreement, with the ultimate goal of estimating the agreement between the annotators themselves, as well as between models and annotators by using the appropriate metrics. Their main findings indicate that disagreement between annotators represents an intrinsic limitation to the performance that can be achieved by automated systems.

This paper aims to explore the potential of training a multilingual hate speech model, as well as further explore the ideas of incorporating inter-annotator disagreement in model training. Therefore, at the basis of this paper lie the following research questions:

- *How does the performance of multilingual hate speech classifiers trained on diamond standard data compare to the performance of monolingual models?*
- *How can inter-annotator disagreement be effectively incorporated into the classifier fine-tuning process?*

In light of these research questions, the expected outcomes are twofold: (1) multilingual classifiers trained on diamond standard data are anticipated to outperform monolingual models, and (2) incorporating inter-annotator disagreement is expected to enhance sensitivity to nuanced hate speech. These findings could benefit computational linguistics research and social media providers by informing the development of more effective content moderation algorithms.

2 Related Work

Several methods exist for incorporating disagreement into ML training pipelines [12, 5], but few focus on hate speech detection. One approach is presented in [7], where monolingual hate speech classifiers were trained for English, Italian, and Slovenian. These classifiers utilized diamond standard datasets sourced from YouTube and Twitter, employing a consistent annotation process for each language. Their main findings indicate that, according to the accuracy scores, the annotators demonstrated a high degree of agreement in approximately 80% of the cases across all three datasets. In terms of Krippendorff's ordinal alpha score, which considers both agreement by chance and the ordering of classes (from least to most severe), the agreement score is approximately

*The first author conducted the research with significant input from the second author, under the supervision and guidance of the third author. All authors contributed to writing the manuscript.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.7

0.6 for all three languages. Furthermore, the evaluation results indicate that the performance of each model aligned with the inter-annotator agreement, both in terms of accuracy and the alpha score. This implies that the performance of models is inherently constrained by the level of agreement among annotators. Consequently, when trained on diamond standard data, it is unlikely that the performance of these models can significantly surpass human performance.

This work was built upon these findings through investigating the potential of multilingual models to enhance hate speech detection, with the aim of broadening their applicability across diverse linguistic contexts. Additionally, strategies for incorporating annotator disagreement were explored, with the goal of improving model performance to approach human-level accuracy and agreement.

3 Method

This section details the methodology for training and evaluating the multilingual hate speech classifier presented in this paper. It begins with a brief overview of the datasets used, followed by an explanation of the chosen pre-trained language model that serves as the foundation for fine-tuning. The section concludes with a description of the methods employed for evaluating the models.

3.1 Datasets

Three monolingual datasets, i.e. the English (Youtube), Italian (Youtube) and Slovenian (Twitter) datasets, introduced in [7] served as the basis for our multilingual model. Each item was annotated by two annotators independently, assigned to one of four available classes: [Appropriate], [Inappropriate], [Offensive], and [Violent]. In the case of conflicting labels, both annotating instances were kept.

To explore strategies for incorporating disagreement, three multilingual datasets were created. First, the *Duplicate All* (DA) dataset, which contains all instances by their respective two annotators from the three monolingual datasets. Second, the *Duplicate Disagreement* (DD) dataset, in which instances where annotators disagreed appear twice with their respective conflicting labels, while instances that they agreed upon appear only once, creating a more balanced training set that reflects both agreement and disagreement, potentially preventing the models from being biased towards instances where annotators agree. And third, the *Remove Disagreement* (RD) dataset, which consists only of instances where annotators agree. Thus, the first two datasets contain diamond standard data, while the third dataset can be considered a gold standard dataset in which disagreement has been explicitly removed.

All instances in these datasets have undergone the same preprocessing steps, such as replacing links and usernames with placeholders. This step was undertaken to mitigate any potential biases associated with certain names, as discussed in [6]. Table 1 presents an overview of the label distribution across the three multilingual training sets. The datasets used for monolingual evaluation are the unmodified evaluation sets presented in [7].

Table 1: Label distribution of the multilingual train sets

Dataset	Acceptable	Inappropriate	Offensive	Violent
DA	191,677	11,005	112,833	7,145
DD	111,324	8,346	72,706	4,992
RD	80,573	2,661	40,255	2,161

3.2 Model Selection and Fine-Tuning

Our proposed multilingual hate speech model builds on the pre-trained XLM-R transformer model [4], chosen for its proven effectiveness in cross-lingual understanding and its ability to handle a wide range of languages. This provides a robust foundation for fine-tuning and optimization, particularly since English, Italian, and Slovenian—the languages used for fine-tuning—were included in XLM-R's pre-training. To explore various strategies for incorporating annotator disagreement during training, three model variants were fine-tuned on the previously presented datasets, referred to in the tables as MDA, MDD, and MRD, respectively.

To address class imbalance and enhance model performance on minority classes, a custom training loop with a weighted cross-entropy loss function was implemented, as proposed in [9]. The class weights were calculated to be inversely proportional to the frequency of each hate speech class within the training data. The hyperparameters for the fine-tuning process included a learning rate of 6×10^{-6}, a batch size of 8, and 3 training epochs. During the training phase, the AdamW optimizer was employed to optimize the model parameters. The fine-tuning process was implemented using PyTorch.

3.3 Model Evaluation

In terms of evaluation, the approach introduced in [7] was replicated in order to compare the performance of the multilingual classifiers to human judgment from the perspective of disagreement. This was achieved by employing identical measures to estimate the agreement between human annotators, as well as the agreement between annotators and models. Accuracy, F1 score and, most notably, Krippendorff's ordinal alpha were used to evaluate all models in this research.

Rarely used in ML applications, Krippendorff's alpha is a robust measure for assessing inter-rater reliability, accounting for agreement beyond what might occur by chance. It is applicable across various data types (nominal, ordinal, interval, and ratio scales) and is particularly effective in dealing with missing data. The value of Krippendorff's alpha ranges from -1 to 1, where 1 indicates perfect agreement and 0 suggests agreement equivalent to chance. Generally, an alpha above 0.80 is considered a strong agreement, while in hate speech datasets, the alpha values range from 0.25 to 0.65. For a detailed discussion, see Krippendorff [8].

4 Results

This section presents the evaluation results on the multilingual model and its variants. It starts with an evaluation from the perspective of inter-annotator and model-annotator agreement. Then, the class specific evaluation results, as well as a model comparison based on the models' average scores are presented. The models are also compared to monolingual baselines fine-tuned on data for their respective languages, including the BERT model for English, AlBERTo for Italian, and CroSloEngual for Slovenian, as presented in [7].

4.1 Inter-Annotator and Model-Annotator Agreement

The inter-annotator agreement was computed on the evaluation sets for each language using Krippendorff's alpha and accuracy. The same measures were also used to compute the agreement between the annotators and the models. The results are presented in Table 2.

Table 2: Inter-Annotator Agreement compared to model-annotator agreement in terms of Krippendorff's ordinal alpha (α) and Accuracy (Acc.) for the models Multilingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), and Multilingual Remove Disagreement (MRD) based on the language-specific evaluation sets

Dataset	Inter-Annotator Agreement		MDA		MDD		MRD	
	α	Acc.	α	Acc.	α	Acc.	α	Acc.
English	58.19	82.91	55.89	79.97	50.18	76.47	**57.90**	**81.41**
Italian	57.00	81.79	**58.29**	82.00	56.15	80.43	57.84	**82.69**
Slovenian	56.62	79.43	55.74	78.60	52.95	76.52	**55.15**	**78.84**

First, in the case of inter-annotator agreement, annotators agree around 80% of the time in terms of accuracy, with an accuracy score between 79% and 82% across all three datasets. However, accuracy does not account for class imbalance, nor the ordering of the classes. A more appropriate estimate of the agreement is computed through Krippendorff's ordinal alpha. Here, the annotators achieve an agreement score alpha in the values between 0.56 and 0.58 across the three languages.

Second, in terms of agreement between annotators and models, the same metrics were applied. The results demonstrate a consistent level of agreement between the models and annotators across all cases. Based on accuracy scores, all models align with at least one annotator approximately 80% of the time, with alpha values comparable to inter-annotator scores. In most instances, the models achieve the upper limit of inter-annotator agreement, and in some cases, even exceed it (e.g., Italian *Multilingual Duplicate All MDA*). This suggests that the models are effectively learning consistent patterns or biases that align well with one or more annotators. Such outcomes are expected in scenarios where annotator disagreement is largely due to subjective interpretation. This should not be construed as the model being inherently superior, but rather as an indication of its efficiency in modeling the predominant patterns present in the training data.

Third, a comparison between the multilingual variants shows that the *Duplicate Disagreement (DD)* strategy consistently shows worse alpha scores, meaning that emphasizing only on disagreement might be detrimental in training. No consistent difference between *Duplicate All (DA)* and *Remove Duplicates (RD)* is evident from the experiments.

4.2 Model Comparison

To evaluate the performance of the models across the four hate speech classes, the F1 score was used. Additionally, the combined (weighted) F1 score was computed for each model to assess their overall performance. To determine the best-performing model, the weighted F1 scores were averaged across all three languages. Table 3 shows the results achieved by each of the models on the English evaluation set. In the case of the English dataset, the results show that the multilingual model outperforms the baseline monolingual English model across all classes except the [Appropriate] class, a case in which it still performs competitively. The variant which achieved the highest score on the minority classes is the *MDA* model, with an F1 score of 39.16 for the [Inappropriate] class and an F1 score of 27.82 for the [Violent] class. This is most likely due to introducing the weighted cross-entropy loss function, which was effective in improving performance on underrepresented classes, a procedure which was not performed in [7].

Similar patterns emerge on the Italian dataset (Table 4). The multilingual model is competitive to the monolingual model while outperforming the Italian baseline on the minority classes. The highest scores on the most important classes [Violent] and [Offensive] were achieved by the *MDA* variant, once again showing the superiority of the *Duplicate All (DA)* strategy.

Table 3: Model evaluation results in terms of class-specific F1 scores on the English dataset. The Total score was calculated using the weighted F1 score. The first three models represent the monolingual baselines. The subsequent models represent the multilingual models

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN	**89.38**	28.95	68.36	24.17	**83.44**
IT	85.25	13.81	0.41	0.00	63.39
SL	88.01	25.17	49.69	2.88	77.71
MDA	86.10	**39.16**	68.24	**27.82**	81.09
MDD	83.33	34.16	65.07	24.52	78.20
MRD	87.43	29.90	**69.02**	27.27	82.18

In the case of the Slovenian dataset, the observed phenomena slightly differ from the previous ones. The evaluation results are presented in Table 5. Here, two of the multilingual variants (*MDA* and *RD*) outperform the Slovenian monolingual model overall, despite predicting worse on the [Appropriate] class. Notably, the monolingual model outperforms all models on the [Violent] class, which has not been the case for the other languages. This could be due to language specifics that the multilingual model fail to capture, or to the specifics of the CroSloEngual BERT which is also heavily pre-trained on Croatian and Slovenian data. Once again, the *DA* disagreement strategy shows slight superiority over *RD*.

Finally, Table 6 shows the average scores of all models, achieved by averaging their combined (weighted) F1 scores across all three languages. Summarizing the multilingual superiority, these final results show how monolingual models drastically falter on unseen languages, while the multilingual models have the capacity to reach the inter-annotator agreement ceiling for all languages.

While overall results show that the *Remove Disagreement (RD)* gold standard strategy for incorporating disagreement is best, one should be cautious when making such conclusions. Class-specific results show that the Duplicate All (DA) strategy outperforms in all the classes most relevant to hate speech detection, except for [Appropriate], which is the least relevant class. Another difference is that the *MDA* model involved training longer on the same data which might have resulted in improvement on minority classes and saturation on the majority class. For a future fairer comparison, the fine-tuning process on gold standard data should be adjusted accordingly. The *MDA* variant of the model is available at: https://huggingface.co/IMSyPP/hate_speech_multilingual.

5 Discussion

In recent years, automated hate speech detection has become crucial for moderating online content and mitigating the negative impact on social dynamics within online communities. This

Table 4: Model evaluation results in terms of class-specific F1 scores on the Italian dataset

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN	86.27	1.28	1.05	0.00	67.42
IT	**91.32**	**58.46**	59.02	40.34	**83.22**
SL	86.23	0.76	3.25	0.00	65.95
MDA	89.77	58.45	**60.42**	**44.97**	82.38
MDD	88.95	56.04	58.31	39.85	81.19
MRD	90.41	55.46	59.49	38.78	82.50

Table 5: Model evaluation in terms of class-specific F1 scores on the Slovenian dataset

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN	79.93	3.98	2.34	0.00	53.84
IT	79.84	3.80	1.24	0.00	53.43
SL	**85.70**	43.69	65.26	**29.12**	78.39
MDA	84.30	**45.22**	**69.69**	24.79	**78.88**
MDD	82.33	43.39	68.59	23.84	77.19
MRD	84.98	38.47	68.40	15.50	78.80

Table 6: Average performance of models based on class-weighted F1 scores across three languages

Model	Avg. Weighted F1 Score (all languages)
EN	68.23
IT	66.68
SL	74.02
MDA	80.78
MDD	78.86
MRD	**81.16**

research proposes a novel multilingual hate speech model to address these challenges on a broader scale. The following discusses the main findings.

First, the inter-annotator agreement and the agreement between annotators and models suggest that inter-annotator agreement sets an intrinsic limit on model performance. Models are limited by the quality and consistency of the annotated data, which directly affects their ability to accurately predict unseen data. However, incorporating areas of disagreement into model development can lead to more robust models capable of handling ambiguous cases by employing one of the several available strategies for incorporating disagreement.

Second, the multilingual model consistently surpassed the monolingual baselines, achieving the inter-annotator agreement ceiling across all languages. This success can be attributed partly to the ability to leverage patterns learned from multiple languages, partly to vast amounts of data incorporated into state-of-the-art pre-trained multilingual models, and partially to the class weighting scheme employed in the fine-tuning. These findings support the first research question, demonstrating that a multilingual hate speech classifier trained on diamond standard data outperforms its monolingual counterparts.

Finally, this research contributes substantially to hate speech classification in a multilingual context by introducing a novel multilingual hate speech detection model and making it available on the Hugging Face platform. Our model underscores the importance of incorporating inter-annotator disagreement into model development, challenging the reliance on gold standard data in subjective tasks, such as hate speech detection.

6 Conclusions

This paper advances automatic hate speech detection by introducing a novel multilingual model fine-tuned on the state-of-the-art XLM-R transformer. By leveraging multilinguality, the model significantly outperforms monolingual baselines, demonstrating its effectiveness across diverse linguistic contexts. This highlights the potential of multilingual approaches in improving hate speech detection, especially in scenarios where content spans multiple languages.

Additionally, this research incorporates inter-annotator disagreement into the fine-tuning process using diamond standard data, offering a valuable alternative to traditional gold-standard models. By embracing rather than ignoring annotator disagreement, the model better reflects the nuances of subjective annotations, enhancing its real-world applicability. However, while this approach shows promise, annotator disagreement continues to present challenges, indicating that further work is needed to fully address its impact on model performance.

Future research could extend this work by evaluating the models on additional languages, exploring alternative baseline models, refining strategies for incorporating annotator disagreement and handling minority classes. As online hate speech extends its impact, developing robust, multilingual content moderation systems is crucial to maintaining safe and inclusive digital environments.

7 Acknowledgments

The authors acknowledge partial financial support from the Slovenian Research Agency (research core funding no. P2-103).

References

[1] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: a closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 45–54.

[2] Alexander Brown. 2017. What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, 36, 561–613.

[3] Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40, 108–118.

[4] Alexis Conneau et al. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR, abs/1911.02116*.

[5] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 2021. Beyond black & white: leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

[6] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115, 16, E3635–E3644.

[7] Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. 2022. Handling disagreement in hate speech modelling. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 681–695.

[8] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

[9] Andraž Pelicon, Syrielle Montariol, and Petra Kralj Novak. 2023. Don't start your data labeling from scratch: opsala-optimized data sampling before labeling. In *International Symposium on Intelligent Data Analysis*. Springer, 353–365.

[10] Juan Manuel Pérez et al. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11, 30575–30590.

[11] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55, 477–523.

[12] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: a survey. *Journal of Artificial Intelligence Research*, 72, 1385–1470.

[13] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, 19–26.

[14] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7, e598.

Predicting Pronunciation Types in the Sloleks Morphological Lexicon of Slovene

Jaka Čibej[1,2]
jaka.cibej@ff.uni-lj.si
jaka.cibej@ijs.si
[1]Faculty of Arts, University of Ljubljana
[2]Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

We present an experiment dealing with the automatic prediction of pronunciation types for lemmas in the *Sloleks Morphological Lexicon of Slovene*. We perform a statistical analysis on a number of mostly *n*-gram-based features and use a set of statistically significant features to train and test several machine learning models to discriminate between lemmas for which a phonetic transcription can be generated automatically using Slovene grapheme-to-phoneme (G2P) conversion rules (e.g. *Novak*), and lemmas with pronunciation that follows other G2P rules (e.g. *Shakespeare*).

Keywords

grapheme-to-phoneme conversion, pronunciation types, morphological lexicon, proper nouns, Slovene

1 Introduction

The *Sloleks Morphological Lexicon of Slovene* [2] is the largest open-access database containing machine-readable information on the morphological properties of Slovene lemmas (e.g. *miza* 'table', noun, common, feminine) and their inflected forms (e.g. *mize*, singular, genitive; *mizo*, singular, accusative). Since version 2.0 [3], each lemma and inflected form also contains accentuated forms (e.g. *míza*) and phonetic transcriptions in the International Phonetic Alphabet (IPA) and its equivalent X-SAMPA (e.g. IPA: /ˈmiːza/, X-SAMPA: /"mi:za/). Both transcriptions were generated automatically from accentuated forms, first in version 2.0 using a rudimentary rule-based system, then again in 3.0 with a greatly improved and linguistically informed rule-based grapheme-to-phoneme (G2P) conversion tool for Slovene.[1]

Rule-based G2P conversion for Slovene (particularly from accentuated forms) yields very good results and leaves only a minority of issues to be resolved manually because in terms of its orthographic depth, Slovene features a shallow orthography ([9]) in which each grapheme in the alphabet generally corresponds to one phoneme (see e.g. [4]) and the spelling-sound correspondence is relatively direct ([1]; [11]): the pronunciation rules allow for words to be pronounced correctly based on their graphemic representation, with some exceptions and several predictable phoneme assimiliations (such as the assimilation of voiceless consonant phonemes to their voiced equivalents *glasba* 'music', IPA: /ˈglaːzba/, or vice-versa, voiced-to-voiceless, *podpreti* 'to support', IPA: /pɔtˈpreːti/).

However, not all entries in Sloleks follow Slovene G2P principles. For a number of words, particularly proper nouns denoting people (*Shakespeare*, *Sharon*), locations (*Sydney*, *Birmingham*), inhabitants (*Newyorčan* 'New Yorker'), etc.; as well as adjectives derived from proper nouns (*aachenski* 'pertaining to Aachen', *Acronijev* 'belonging to Acroni'), the phonetic transcription cannot be generated using Slovene G2P rules. In such cases with foreign orthographic elements that indicate relations between graphemes and phonemes that are unusual for Slovene, Slovene linguistic and lexicographic practice (see e.g. [5]) first requires a transliteration into the closest equivalent using Slovene graphemes, which can then be used to generate the phonetic transcription using Slovene G2P rules (e.g. *Newyorčan* → *njújórčan* → IPA: /ˈnjuːˈjoːrtʃan/).

Because of this, it is necessary to discriminate between different *pronunciation types*: categories of words that follow Slovene G2P rules (*Slovene G2P*) and those that do not (e.g. *Other G2P*; more on this in Section 2). Pronunciation types denote the manner in which the phonetic transcription of the word can be generated. In some cases, assigning the pronunciation type to a lemma is trivial – if the lemma contains a grapheme that is not part of the Slovene alphabet[2] (e.g. *x*, *y*, *w*, *q*), it belongs into the *Other G2P* category (e.g. *Byron*, *Oxford*). There are, however, many exceptions that belong in the *Other G2P* category despite being comprised entirely of Slovene graphemes (e.g. *Matt*, *Sharon*).

In Sloleks 3.0, the first cca. 100,000 lemmas that had been part of version 2.0 were manually annotated with pronunciation types, whereas the 264,000 new entries (added automatically from the *Gigafida 2.0 Corpus of Modern Standard Slovene* [6]) still lack this information. Because manual annotation from scratch is time-consuming, we performed an experiment to determine to what degree the pronunciation type can be predicted automatically by relying on the scarce linguistic and morphosyntactic information that can be extracted from an individual lemma.

The paper is structured as follows: we describe the dataset that was used for the statistical analysis and machine learning experiment (Section 2), as well as the process of feature selection (Section 3). We train several machine-learning models and evaluate their performance using 10-fold cross-validation (Section 4). Finally, we manually evaluate a sample of automatically annotated entries (Section 5) and conclude the paper with our plans for future work (Section 6).

[1]The Slovene G2P tool is part of *Pregibalnik*, a piece of software used for the automatic expansion of the *Sloleks Morphological Lexicon of Slovene*: https://github.com/clarinsi/SloInflector It was developed within the *Development of Slovene in the Digital Environment* project. The Slovene G2P converter is also available as an API-service: https://orodja.cjvt.si/pregibalnik/g2p/docs

[2]Although *ć* and *đ* are not part of the Slovene alphabet, they are phonemically transparent and frequently occur in names of Slovene citizens, so they are not counted as foreign characters for the purposes of this task.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.2

Table 1: Lemmas in Sloleks 3.0 by Pronunciation Type

Pronunciation Type	Frequency	%
-	264,538	72.41%
Slovene G2P	94,750	25.93%
Other G2P	3,066	0.84%
Numeral	1,840	0.50%
Acronym	845	0.23%
Slovene G2P with minor deviation	113	0.03%
Abbreviation	70	0.02%
Ambiguous G2P	69	0.02%
Symbol	49	0.01%
Total	**365,340**	**100.00%**

Table 2: Lemmas in Sloleks 3.0 with *Other G2P* pronunciation type by Morphosyntactic Properties

Morphosyntactic Properties	Frequency	%
Adjective, possessive	1,092	35.62%
Noun, proper, masculine	958	31.25%
Noun, proper, feminine	713	23.26%
Adjective, general	142	4.63%
Noun, common, masculine	127	4.14%
Noun, common, feminine	20	0.65%
Adverb, general	10	0.33%
Noun, common, neuter	2	0.07%
Verb, main, imperfective	2	0.07%
Total	**3,066**	**100.00%**

2 Dataset

Sloleks 3.0 contains a total of 365,340 entries, but only approximately 28% have been manually assigned one of 8 pronunciation types[3] (as shown in Table 1). For the classification task, we focus only on the two most frequent pronunciation types (*Other G2P* and *Slovene G2P*). [4]

In terms of their morphosyntactic features, the *Other G2P* lemmas mostly consist of possessive adjectives and proper nouns, collectively accounting for cca. 90% of the category (as shown in Table 2), but only 15% of the portion of Sloleks annotated with pronunciation types.

The final dataset for statistical analysis and machine learning consisted of 94,863 *Slovene G2P* lemmas (e.g. *dekadentnost*, *Košak*, *prefiltriran*) and 3,066 *Other G2P* lemmas (e.g. *Elizabeth*, *Presley*, *Sinclaire*).

3 Statistical Analysis and Feature Selection

From each lemma, we extracted a series of features that could help discriminate between the two classes: (a) percentage of *Slovene G2P* graphemes within the lemma (i.e. graphemes of the Slovene alphabet as well as *ć* and *đ*); (b) morphosyntactic features (e.g. *noun*, *proper*, *masculine*); (c) relative frequencies[5] of character-level uni-, bi-, and trigrams within the lower-cased lemma (e.g. *Matt* → $f_r(m)$, $f_r(a)$, ..., $f_r(ma)$, $f_r(at)$, ..., $f_r(mat)$, ...); (d) relative frequencies of character-level uni-, bi-, and trigrams from a robust CVC-conversion of the lemma, substituting consonant graphemes with *C* and vowel graphemes with *V* (e.g. *Matt* → *CVCC* → $f_r(C)$, $f_r(V)$, ..., $f_r(CV)$, $f_r(VC)$, ..., $f_r(CVC)$, ...); (e) relative frequencies of character-level uni-, bi-, and trigrams from a finegrained CVC-conversion of the lemma[6] (e.g. *Matt* → *ZVKK* → $f_r(Z)$, $f_r(V)$, ..., $f_r(ZV)$, $f_r(VK)$, ..., $f_r(ZVK)$, ...)

For (c), (d), and (e), the initial and final uni-, bi-, and trigrams of the lemma were extracted separately as well, as in some cases the position of the *n*-gram in the word can be indicative of one class over another.

For general character-level *n*-grams, the first 1,498 with a frequency of at least 500 across all Sloleks 3.0 lemmas were analyzed; these cover cca. 88.34% of all *n*-gram occurrences. For robust CVC and finegrained CVC *n*-grams, all were analyzed. We performed the Kruskal–Wallis H test [7] (k=2, n=97,056) on a total of 6,148 features, out of which 2,490 (40%) were statistically significant.[7] Statistically significant features by categories are shown in Table 3. 1,146 features are more indicative of *Slovene G2P* and 1,344 are more indicative of *Other G2P*. As shown in Table 4, only three of the top 10 general *n*-grams indicative of *Other G2P* actually contain non-*Slovene G2P* characters, confirming that detecting lemmas from the *Other G2P* category is more complex and requires more than simply taking into account non-*Slovene G2P* graphemes.

Table 3: Statistically Significant Features by Category

Feature Category	Number
Percentage of *Slovene G2P* characters	1
Morphosyntactic features	3
General character-level *n*-grams	1,119
Initial character-level *n*-grams	398
Final character-level *n*-grams	468
General robust CVC *n*-grams	66
Initial robust CVC *n*-grams	44
Final robust CVC *n*-grams	39
General finegrained CVC *n*-grams	157
Initial finegrained CVC *n*-grams	102
Final finegrained CVC *n*-grams	93
Total	**2,490**

4 Pronunciation Type Prediction

The identified features (along with several placeholder *n*-grams to take into account any graphemes not covered in the initial dataset) were taken into account to develop a custom vectorizer that converts a given lemma and its lexical features based on the MulText-East v6 (MTE-6) Morphosyntactic Specifications for

[3]It should be noted that all the inflected forms within the entry effectively inherit the pronunciation type.

[4]Symbols in Sloleks are rare, along with entries within the *Ambiguous G2P* category (where an entry can either follow Slovene G2P rules or not, depending on the context – e.g. *Amanda* as a Slovene name: /amˈaːnda/; or as an English name with a pronunciation adjusted to the Slovene set of phonemes: /əmˈɛːnda/). Abbreviations and numerals are easily identifiable, and while acronyms have a separate manner of generating phonetic transcriptions which also depends on their morphological patterns, they are also mostly identifiable with rules. Because of its rarity and similarity to *Slovene G2P*, the *Slovene G2P with minor deviation* category was merged into *Slovene G2P* for the classification task.

[5]Relative frequencies were calculated as $f_r(x_n) = f_a(x_n)/\sum f_a(y_n)$, e.g. the absolute frequency of *n*-gram *x* of length *n* within the lemma divided by the sum of absolute frequencies of each *n*-gram *y* of length *n* within the lemma.

[6]In the finegrained CVC-conversion, consonant graphemes were generalized into more finegrained categories, e.g. graphemes denoting Slovene sonorants (M), voiced (G) and voiceless obstruents (K), foreign consonants (X), etc.

[7]Effect size was calculated as $\eta^2 = (H - k + 1)/(n - k)$, as reported in [10].

Table 4: Top 10 Statistically Significant General Character-Level *n*-Grams by Effect Size (η^2)

n-Gram	H	p	η^2	Means
y	11509.36	$p \leq 0.0001$	0.1186	$\mu_S < \mu_O$
w	9595.25	$p \leq 0.0001$	0.0989	$\mu_S < \mu_O$
ch	7558.60	$p \leq 0.0001$	0.0778	$\mu_S < \mu_O$
ll	6295.96	$p \leq 0.0001$	0.0649	$\mu_S < \mu_O$
ss	3804.26	$p \leq 0.0001$	0.0392	$\mu_S < \mu_O$
nn	3220.65	$p \leq 0.0001$	0.0332	$\mu_S < \mu_O$
th	2973.89	$p \leq 0.0001$	0.0306	$\mu_S < \mu_O$
wa	2761.53	$p \leq 0.0001$	0.0284	$\mu_S < \mu_O$
tt	2745.10	$p \leq 0.0001$	0.0283	$\mu_S < \mu_O$
co	2571.20	$p \leq 0.0001$	0.0265	$\mu_S < \mu_O$

Table 5: Model Performance Based on 10-Fold Cross-Validation

Model	A	BA	P	R	F1	ROC AUC
LinearSVC	**99.08**	**87.87**	**96.36**	**87.87**	**91.64**	**98.89**
Multin. NB	97.38	79.17	78.12	79.17	78.62	96.55
kNN (k=5)	98.25	75.17	93.67	75.17	81.74	91.63
Majority	96.87	-	-	-	-	-

Slovene[8] into a 2,500-dimensional numerical vector. The entire dataset was converted into vectors and split into a training set (80%) and a test set (20%), both stratified by class. Three models[9] (Linear Support Vector Classifier (LinearSVC), Multinomial Naive Bayes Classifier (Multin. NB), and k Nearest Neighbors Classifier (kNN)) were trained and evaluated with 10-fold cross-validation. The results are listed in Table 5[10] and show that LinearSVC outperforms the other two models. All three exhibit above-baseline accuracy compared to the majority classifier, but Multinomial NB and kNN perform much worse in terms of balanced accuracy as well as precision and, in case of kNN, recall. Recall is also somewhat lower with LinearSVC, which is to be expected – some *Other G2P* lemmas might contain no indicative *n*-grams and are thus hard to detect; on the other hand, once identified, the model is very precise in its prediction.

Table 6 shows the confusion matrix for the LinearSVC model tested on the 20% stratified test dataset. The model very rarely misclassifies *Slovene G2P* lemmas, and more frequently errs with *Other G2P* lemmas. A closer inspection of the misclassified *Slovene G2P* examples reveals several errors in the original dataset: *Beethoven*, *Ratzinger*, *Rotterdam*, *Franco*, *Oberstdorf*, and *Keller* were in fact correctly classified as *Other G2P*, but they are miscategorized as *Slovene G2P* in the original dataset. Other misclassifications include examples of foreign proper nouns and possessive adjectives that contain unusual grapheme combinations for Slovene (e.g. *Andreas*, *Aurelio*, *Hilton*, *Simpsonov*), but their pronunciation can still be derived from their graphemic representation (e.g. *Andreas* → IPA: /and'reːas/).

On the other hand, *Other G2P* lemmas misclassified as *Slovene G2P* include *Andersonov*, *Atkinsov*, *Batmanov*, in which the grapheme

[8]MTE-6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html The vectorizer uses Slovene morphosyntactic tags, e.g. *Slz* (S – noun, l – proper, z – feminine).
[9]All models were trained using the Python library *scikit-learn*. [8]
[10]A, BA, P, R, and F1 refer to accuracy, balanced accuracy, macro-precision, macro-recall and macro-F1, respectively.

Table 6: Confusion Matrix for Linear Support Vector Classifier

True → ↓ Predicted	Slovene G2P	Other G2P	∑
Slovene G2P	**18,939**	140	19,079
Other G2P	34	**473**	507
∑	18,973	613	-

Table 7: Confusion Matrix for Manual Evaluation

True → ↓ Predicted	Slovene G2P	Other G2P	∑
Slovene G2P	**86**	9	95
Other G2P	14	**91**	105
∑	100	100	-

'a' is pronounced as /ɛ/, but this cannot be discerned from the graphemic representation itself. Other misclassified examples are more obviously pertaining to *Other G2P*, e.g. *Dorfmeister*, *Faulknerjev*, *Flaubertov*, *Heisenbergov*, *Balfourjev*. This might indicate that not all indicative *n*-grams have been included as features (e.g. 'ei', 'ou'), possibly for lack of evidence in the original dataset or because they are less frequent and have not been included in the initial batch of statistical tests. As the lexicon expands with new entries, the model will be updated with new examples and new features to potentially improve performance.

5 Manual Evaluation

We trained a new instance of the LinearSVC model on the entire dataset and used it to annotate the remaining cca. 264,000 lemmas from Sloleks 3.0 with no pronunciation type, resulting in 86,730 lemmas with *Other G2P* and 177,808 lemmas with *Slovene G2P*.

We performed a preliminary manual evaluation consisting of a random sample of 100 examples from each class. The results are shown in the confusion matrix in Table 7. Although the sample is too small to be representative of the whole, it indicates that the model performs well even on unseen data, achieving an accuracy of 88.50% (P=0.91, R=0.87, F1=0.89) over a majority baseline accuracy of 50.00%.

The misclassifications of *Other G2P* as *Slovene G2P* include examples such as *Mukhamedov*, *Beatli*, *Livenza*, and *Preidler*, with limited indicators that the words belong to the *Other G2P* category. Most graphemes in these examples are pronounced according to *Slovene G2P* criteria, with the exception of individual *n*-grams ('nz', 'ei', 'kh'), some of which have not been included in the set of features. In other examples, only one or two vowel graphemes are indicative of *Other G2P* pronunciation (e.g. *Trendlina*, which is also a lemmatization error; the correct lemma is *Trendline*; and *Sanberg*), and the pronunciation of single vowel graphemes appears harder to predict than consonant graphemes or combinations thereof.

Similarly, the misclassifications of *Slovene G2P* lemmas as *Other G2P* lemmas include examples such as *Doneck*, *Barson*, *Bronson*, *Piersanti*, and *Faustini*. While these are proper nouns of foreign origin, their Slovene pronunciation can either be fully discerned from their graphemic representation (e.g. *Doneck* → IPA: /dɔ'neːtsk/), or it only differs slightly from what Slovene

grapheme-to-phoneme conversion would produce (e.g. *Faustini* → automatically converted IPA: /faus'ti:ni/; correct IPA: /fa͡us'ti:ni/).

6 Conclusion

In the paper, we presented the results of an attempt to automatize the assignment of pronunciation types to lemmas in the *Sloleks Morphological Lexicon of Slovene*. The results show that a model based on a series of mostly *n*-gram features can provide good results when discriminating between *Slovene G2P* and *Other G2P* categories, with the best performance achieved by the Linear Support Vector Classifier. However, there is still room for improvement, particularly in terms of recall – a number of *Other G2P* lemmas from the test set were misclassified as *Slovene G2P*, while those classified correctly were classified with a relatively high precision score. *n*-grams that are statistically significant as indicative of one class have proven to be useful features for model development, but because they are not evenly distributed and occur sporadically in different lemmas, it would make sense to further improve the model by performing the same statistical analysis (as described in Section 3) on the long tail of less frequent *n*-grams to prepare a more comprehensive list of indicative *n*-grams. The current version of the model is very light-weight and additional features should not cause the model to become overencumbered.

There are several possibilities for further development of the model. Firstly, instead of using relative frequencies of *n*-grams as features, it would be useful to test how different measures such as TF–IDF, absolute frequencies, or even Boolean values influence the performance of the model, and potentially also test several other machine learning algorithms (e.g. Random Forest Classifier). Secondly, while the other pronunciation types from Sloleks 3.0 (acronyms, abbreviations, etc.) are relatively easily identifiable (but much less frequent), in the next step, it would be informative to include them in the training set and test out the model's performance on the full set of categories. Thirdly, a statistical analysis should be performed on the probabilities with which the model makes decisions and to what degree they correlate with the percentage of graphemes that differ from the shallow orthographical Slovene G2P rules (e.g. *Anderson*, with arguably only 'a' not following Slovene G2P rules; vs. *Châteaux*, where the majority of graphemes are pronounced completely differently compared to Slovene G2P rules). This would require the preparation of a separate dataset in which graphemes are manually aligned to either the graphemes of their transliterated Slovene graphemic forms (*Newyorčan* → *njújórčan*) or their Slovene IPA transcriptions. By assigning scores that reflect the degree of orthography depth for the individual lemma, it would be possible to use the dataset to train a regression model.

Similarly, *Other G2P* lemmas from Sloleks 3.0 can be manually annotated with their language of origin and transliterated according to the recently published transliteration rules of *Pravopis 8.0*,[11] the new orthographic manual of Slovene, which at the time of writing this paper is still in development. Such a dataset would enable the development of a model for language identification for individual lemmas, and, ultimately, a model for automatizing transliteration of lemmas of foreign origin into their Slovene equivalents. As of now, no such tool yet exists for Slovene, and even the new orthographic manual anticipates that all transliteration should be done manually, which begs the question whether at least part of the work can be automatized. This would be an important step in the development of a modern, digital infrastructure for Slovene orthography, and would facilitate the automatic expansion of modern digital dictionary databases and datasets for automatic speech recognition.

In addition, although our preliminary experiments with LLMs (ChatGPT 3.5 and 4.0) classifying *Slovene G2P* and *Other G2P* lemmas have yielded much worse results than the best performing LinearSVC model, more systematic experiments are warranted.

As part of our future work, we intend to implement the model into *Pregibalnik*,[12] which is used for automatically extending the lexicon and currently assigns no pronunciation type. The model itself is available under the Apache 2.0 license on Github[13], while the pronunciation type annotations will be included in future versions of Sloleks and, eventually, manually validated.

Acknowledgements

The research presented in this paper was conducted within the research project titled *Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language* (J7-4642), the research programme *Language Resources and Technologies for Slovene* (P6-0411), and the CLARIN.SI research infrastructure, all funded by the Slovenian Research and Innovation Agency (ARIS). The author also thanks the anonymous reviewers for their constructive comments.

References

[1] Derek Besner and Marilyn Chapnik Smith. 1992. Chapter 3 basic processes in reading: is the orthographic depth hypothesis sinking? In *Orthography, Phonology, Morphology, and Meaning*. Advances in Psychology. Vol. 94. Ram Frost and Leonard Katz, editors. North-Holland, 45–66. DOI: https://doi.org/10.1016/S0166-4115(08)62788-0.

[2] Jaka Čibej et al. 2022. Morphological lexicon sloleks 3.0. Slovenian language resource repository CLARIN.SI. (2022). http://hdl.handle.net/11356/1745.

[3] Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik, and Marko Robnik-Šikonja. 2019. Morphological lexicon sloleks 2.0. Slovenian language resource repository CLARIN.SI. (2019). http://hdl.handle.net/11356/1230.

[4] Florina Erbeli and Karmen Pižorn. 2012. Reading ability, reading fluency and orthographic skills: the case of l1 slovene english as a foreign language students. English. *Center for Educational Policy Studies Journal*, 2(3), 119–139. https://files.eric.ed.gov/fulltext/EJ1130208.pdf.

[5] Nataša Gliha Komac et al. 2015. Koncept novega razlagalnega slovarja slovenskega knjižnega jezika. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. (2015). https://fran.si/179/novi-slovar-slovenskega-knjiznega-jezika/datoteke/Potrjeni_koncept_NoviSSKJ.pdf.

[6] Simon Krek et al. 2019. Corpus of written standard slovene gigafida 2.0. Slovenian language resource repository CLARIN.SI. (2019). http://hdl.handle.net/11356/1320.

[7] William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 260, 583–621. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1952.10483441. DOI: 10.1080/01621459.1952.10483441.

[8] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[9] Anja Schüppert, Wilbert Heeringa, Jelena Golubovic, and Charlotte Gooskens. 2017. Write as you speak? a cross-linguistic investigation of orthographic transparency in 16 germanic, romance and slavic languages. English. *From semantics to dialectometry*, 32, 303–313. ISBN: 9781848902305.

[10] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 1(21), 19–25.

[11] Antal van den Bosch, Alain Content, Walter Daelemans, and Beatrice de Gelder. 1994. Analysing orthographic depth of different languages using data-oriented algorithms. In *Proceedings of the 2nd International Conference on Quantitative Linguistics*.

[11] *Pravopis 8.0: Pravila novega slovenskega pravopisa za javno razpravo*. https://pravopis8.fran.si/, 9 August 2024

[12] *Pregibalnik*: https://github.com/clarinsi/SloInflector; the entire tool is also available as an API-service: https://orodja.cjvt.si/pregibalnik/docs

[13] GitHub: https://github.com/jakacibej/sikdd2024_predicting_pronunciation_types

Higher-Order Bibliographic Services based on bibliographic networks

Vladimir Batagelj
IMFM
Ljubljana, Slovenia
IAM and FAMNIT, UP
Koper, Slovenia
vladimir.batagelj@fmf.uni-lj.si

Jan Pisanski
Faculty of Arts, UL
Ljubljana, Slovenia
jan.pisanski@ff.uni-lj.si

Tomaž Pisanski
FAMNIT, UP
Koper, Slovenia
IMFM
Ljubljana, Slovenia
tomaz.pisanski@upr.si

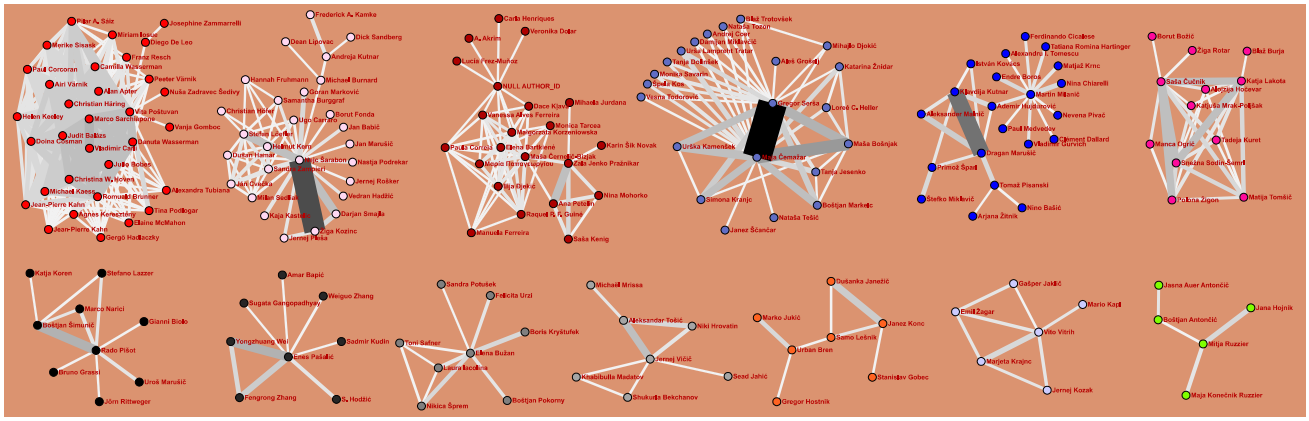


Figure 1: The largest co-author groups at level 10 at the University of Primorska until 2024.

Abstract

Bibliographic databases only provide basic services to users, but they could provide much richer information for specific user needs. The main reason for the delay in developing such higher-order bibliographic services is the limited access to data in proprietary databases. We expect the new open bibliographic databases like OpenAlex will encourage faster development of these services. We describe an approach based on a collection of bibliographic networks as a foundation to support the development of higher-order bibliographic services.

Keywords

bibliographic database, open access, network analysis, higher-order bibliographic service, prototype, OpenAlex

1 Introduction

From special bibliographies (BibTEX, EndNote) and bibliographic databases, it is possible to obtain data about works (papers, books, reports, etc.) on selected topics. A typical work description contains the following data: authors; title; publisher/journal; publication year and pages. In some sources, additional data are available including languages, classification of documents, keywords, authors' institution/country affiliation, lists of references, and the abstract. This data can be transformed into a collection of compatible two-mode networks on selected topics [5]: works × authors; works × keywords; works × countries, and other pairs of characteristics describing works. Besides these networks, we can also get the partition of works by their publication years, the partition of works by journals or publishers, the vector of the number of pages, and, in some cases, the (one-mode works × works) citation network.

When constructing any of these networks, the first task is to specify the nodes and which relations are linking them. In short, the network boundary problem [16] has to be solved. This includes deciding whether a network is one-mode or two-mode and which node properties are important for the intended analyses. For specifying links, this amounts to answering a series of questions:

(1) Are the links directed?
(2) Are there different types of links (relations) to include?
(3) Can a pair of nodes be linked with multiple links?
(4) What are the weights on the links?
(5) Is the network static, or is it changing through time?

Another problem that often occurs when defining the set of nodes is the identification of nodes. The unit corresponding to a node can have different names (synonymy), or the same name can denote different units (homonymy or ambiguity). For example in the BibTEX bibliography from the Computational Geometry Database [14] the same author appears under 7 different names: R.S. Drysdale, Robert L. Drysdale, Robert L. Scot Drysdale, R.L. Drysdale, S. Drysdale, R. Drysdale, and R.L.S. Drysdale. Insider information is needed to decide that Otfried Schwarzkopf and Otfried Cheong are the same person. At the other extreme, there are at least 57 different mathematicians with the name Wang, and Li in the MathSciNet Database [20]. Its editors have tried hard, from 1985, to resolve the identification of the author's problem during the data-entry phase. The significant growth of contributions by Chinese scientists and their full name similarity in

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.12

Roman transcriptions adds additional complexity to the problem. In the future, the problem could be eliminated by implementing initiatives such as using ORCID or resolving the identification problem in bibliographic databases (Scopus, OpenAlex).

2 Higher-Order Bibliographic Services

The data collected in different bibliographic databases can be used to provide higher-order bibliographic and bibliometric services such as what to read (contact/visit)? – a list of relevant articles/books (authors, institutions) on selected topic; where to publish? – a list of journals suitable for the publication of an article, automatic suggestion of keywords; reviewer selection – a list of reviewers suitable for a submitted article; possible partners for research collaboration; a career application – a candidate's activity report draft; etc.) for different types of users (students, researchers, teachers, decision-makers, funding agencies, research institutions, database managers, etc. . To support this goal we have to use high-quality data often obtained by combining data from different databases.

For the development of higher-order bibliographic and bibliometric services, open bibliographic databases such as OpenAlex are particularly welcome, as the developed services can remain open.

3 OpenAlex

The basic type of unit in a bibliographic database is the work. A user searching the database gets a list of works satisfying the query. Usually, some operations with such lists (inspection, filtering, merging, intersection, statistics, etc.) are supported. Only basic services are provided to users.

Some web services also supporting some other types of units (authors, institutions, research fields, conferences, etc.) were developed such as Google Scholar [19], Scholar GPS [12], and DBLP – computer science bibliography [10].

Our approach is based on OpenAlex [18, 9] but this information can be obtained from most bibliographic databases [13, 11]. OpenAlex indexes more than twice as many scholarly works as the leading proprietary products and the entirety of the knowledge graph and its source code are openly licensed and freely available through data snapshots, an easy-to-use API, and a nascent user interface.

OpenAlex is based on 7 types of units (entities): **W**(ork), **A**(uthor), **S**(ource), **I**(nstitution), **C**(oncept), **P**(ublisher), or **F**(under) (and some additional ones such as topics, keywords, countries, continents, languages, etc.). Each unit gets its OpenAlex ID – we assume that the identification problem is solved by the database.

The simplest use of OpenAlex is through its web interface (service) https://openalex.org/ or using a direct URL request in the browser URL line. For example

- Author's name: search the OpenAlex web service
- Known author ID: URL https://openalex.org/A5001676164
- Work with DOI: URL https://api.openalex.org/works/https://doi.org/10.1007/s11192-012-0940-1
- Known work ID: URL https://openalex.org/W2083084326
- Name of the institution: search the Openalex Web service
- Known institution ID: URL https://openalex.org/institutions/I4210106342

This way, the OpenAlex web interface provides basic inspections of the selected unit. For example, by including a link with our OpenAlex author ID on our web page we get a report on our publications. Similarly, we get the report on the publication activity of the selected institution.

3.1 API

An application programming interface (API) is a way for two or more computer programs or components to communicate with each other. It is a type of software interface, offering a service to other pieces of software [21]. In our case, API enables us to use the database data from our programs. An R package supporting the use of OpenAlex is `openalexR` [1].

The OpenAlex API is available at https://api.openalex.org. Its response is returned in JSON format. Here is an R code using the OpenAlex API for the IMFM institution search

```
setwd(wdir <- "C:/work/OpenAlex/API")
library(httr); library(jsonlite)
res <- GET("https://api.openalex.org/institutions",
    query = list(search="imfm"))
str(res)
cont <- fromJSON(rawToChar(res$content))
names(cont); str(cont)
```

The response data are available in the variable `cont`. Similarly, the API can be used also from other programming languages.

The OpenAlex query can be composed of different components. Using **search** we can search for a given search text across titles, abstracts, and full-text. Using a **filter** we can limit our search to units satisfying given conditions. Using **select** we can select data fields that will appear in results. The query can be further controlled by some parameters. For example

```
wd <- GET("https://api.openalex.org/works",
  query = list(
    search="handball",
    filter="publication_year:2015",
    select="id,title",
    page="2", per_page="200"))
names(wd)
wc <- fromJSON(rawToChar(wd$content)); names(wc)
names(wc$meta); wc$meta$count; str(wc$results)
```

returns the second page (with up to 200 entries) on works on handball published in the year 2015. Only information about works ID and title is returned.

The OpenAlex API uses paging – the list data are provided by pages. The **basic paging** (up to 10 000 units) is based on two parameters `page` and `per_page`). The **cursor paging** is a bit more complicated than basic paging, but it allows us to access as many records as we like.

4 A collection of bibliographic networks

We developed an R package `OpenAlex2Pajek` to support the creation of bibliographic networks from OpenAlex [4]. We get a collection of bibliographic networks (citation network **Cite**, authorship network **WA**, sources network **WJ**, keywords network **WK**, countries network **WC**), some partitions and vectors (properties of nodes) (publication year, type of publication, language of publication, cited by count, countries distinct count, referenced works, and additionally two files containing names of works `xyzW.nam` and names of authors `xyzA.nam`. Most acquired networks are 2-mode – they link units of two different types; an ordinary or 1-mode network links units of the same type.

Currently, `OpenAlex2Pajek` contains three main functions `OpenAlex2PajekCite`, `OpenAlex2PajekAll`, and `coAuthorship`.

We split the process of creating the collection of bibliographic networks into two parts:

- determining the set W of relevant works using the **saturation approach** [7, page 506],
- creation of the network collection for the works from W.

The set W is determined iteratively using the function `OpenAlex-2PajekCite` and the collection is finally created using the function `OpenAlex2PajekAll`.

The function `coAuthorship` creates a weighted temporal network describing the co-authorship between world countries in selected time intervals. The weight of an edge is the number of works co-authored by authors from the linked countries.

In an analysis of weighted networks, the 1-neighbor skeleton is often used to get an overall insight into the network's basic structure. In the 1-neighbor skeleton, only its strongest link is kept for each node. The resulting directed network is forest-like. Non-trivial connected components in 1-neighbor skeletons are (usually) directed trees with a pair of nodes linked in both directions with the largest weight in the tree – these two arcs are usually replaced by an edge (undirected link). In Figure 2 the 1-neighbor skeletons for years 1990, 1995, 2000, 2010, 2015, and 2020 are presented. We see that the number of isolated nodes (countries not collaborating with other countries) is decreasing. In all analyzed years the US has a leading (hub) position. In the years 1990, 1995, 2000, and 2010 the edge in the main component links US and GB but in the years 2015 and 2020 GB is replaced by CN. In 1990, stronger secondary hubs were GB, FR, RU, JP, and DE. In the following years, some other countries SE, ES, AU, CN, BR, ZA, and IN (BRICS) became secondary hubs attracting previously non collaborating countries or geographically or linguistically close countries.

Most of the ingredients of basic reports are counters, sorted lists, (weighted) degrees and their distributions obtained from an adequate network. Sometimes also the time is considered producing time series.

An important property of a collection of bibliographic networks is that some of them are compatible – they share a common set (most often the set of works W). This allows us to use network multiplication (defined by the product of network matrices) to compute the corresponding derived network connecting the remaining two sets [5]. For example, in the derived network $\mathbf{AK} = \mathbf{WA}^T \cdot \mathbf{WK}$ its entry $AK[a,k]$ tells us in how many works the author a used the keyword k. Similarly, in the derived network $\mathbf{ACiK} = \mathbf{WA}^T \cdot \mathbf{Cite} \cdot \mathbf{WK}$ its entry $ACiK[a,k]$ tells us how many times the author a cited works described by the keyword k.

A 2-mode network is always compatible with its transpose (on both sets). The corresponding derived networks are called projections – the row projection $\mathrm{row}(\mathbf{WA}) = \mathbf{WA} \cdot \mathbf{WA}^T$ and the column projection $\mathrm{col}(\mathbf{WA}) = \mathbf{WA}^T \cdot \mathbf{WA}$. Both projections are ordinary weighted 1-mode networks that can be analyzed using standard network analysis methods.

For the authorship network **WA** its column projection $\mathbf{Co} = \mathbf{WA}^T \cdot \mathbf{WA}$ is the co-authorship network. Its entry $Co[a,b]$ counts the number of works that authors a and b co-authored. It turns out that a work with k co-authors contributes k^2 links to the co-authorship network – works with a large number of co-authors are overrepresented in it. To treat all authors equally the fractional approach is used [3]. In Figure 1 the largest co-authorship groups at level 10 at the University of Primorska are presented – connected components of the link cut at level 10 in the network **Co**. Each pair of linked authors co-authored at least 10 works

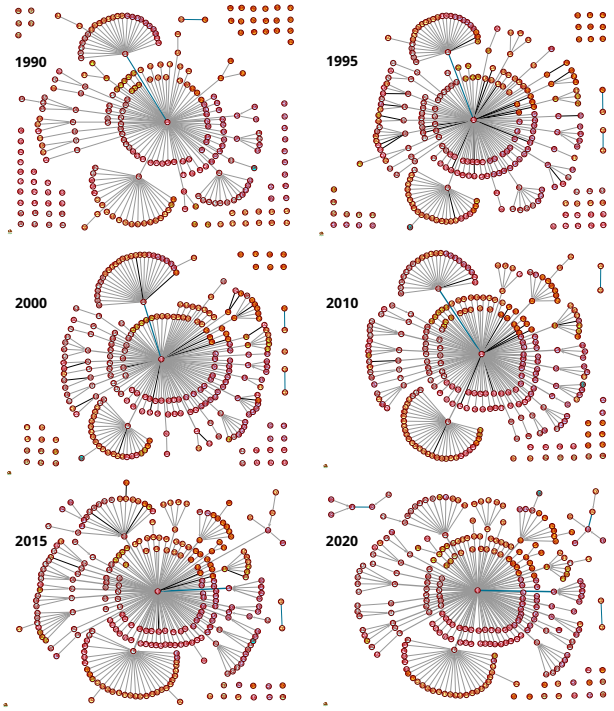


Figure 2: 1-neighbors skeletons of world co-authorship for selected years.

in the bibliography of works with at least one co-author from University of Primorska.

In bibliometric analysis, the citation network **Cite** has a very important role. It collects "votes" about the relevance of previous works for a given work. It is often used for solving the network boundary problem, and also for identifying the most relevant works in the collected bibliography [2, 6]. The derived network $\mathbf{ACiA} = \mathbf{WA}^T \cdot \mathbf{Cite} \cdot \mathbf{WA}$ describes the citations between authors – its entry $ACiA[a,b]$ counts the number of times author a cited author b. The co-citation network is defined as the column projection of the citation network $\mathbf{coCi} = \mathrm{col}(\mathbf{Ci}) = \mathbf{Ci}^T \cdot \mathbf{Ci}$ and the bibliographic coupling network is defined as the row projection of the citation network $\mathbf{biCo} = \mathrm{row}(\mathbf{Ci}) = \mathbf{Ci} \cdot \mathbf{Ci}^T$.

The idea of derived networks can be extended to temporal bibliographic networks [8]. Using derived networks we enlarge the source for different statistics. Additional insight can be gained by analyzing the structure of networks and identifying important subnetworks in them [6].

In the following, we present an overview of typical report ingredients [7, 15]. Because of limited available space, we decided to put examples on Github/bavla.

5 Report ingredients

5.1 Statistics

Because the analyzed networks are often large a complete presentation is not an option. To describe them we use different statistical descriptors.

- sizes of sets (number of nodes, number of links); structural network properties (number of components, size of the largest component, etc.)
- top units – ordered lists of units with the largest values of selected property (degre, weighted degree, link weight,

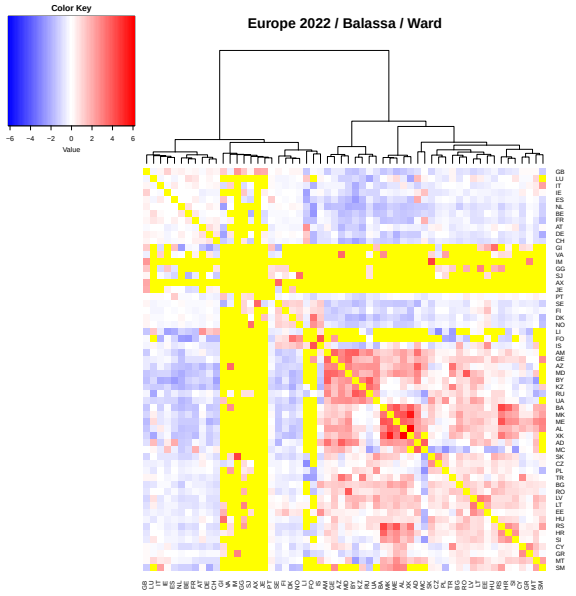


Figure 3: Balassa EU co-authorship for the year 2022.

- distribution of selected property
- time series describing temporal changes of selected properties
- scatter plots showing a possible relationship between two selected properties

Often bibliometric properties of units follow laws such as Zipf (or power) law, Bradford law, Lotka law, lognormal distribution, Hirsch index, etc.

5.2 Network analysis

Derived networks are weighted. To get readable results of reasonable size we usually search for important subnetworks, often a kind of skeleton – from a given network less important elements are removed. There are different types of skeletons (spanning forest, k closest neighbors, cuts, cores, islands, etc. [6]).

A traditional graph-based visualization is used if the obtained result network is not dense. For denser networks, the matrix display is much more readable. In a matrix display, the permutation of nodes (usually obtained by clustering) can create patterns that reveal the network's internal structure.

Figure 3 presents a matrix display of Balassa co-authorship indices between European countries in 2022 (yellow cell – no link, red/blue cell – above/below expectation) [17].

5.3 Special algorithms

Some properties can require special computational procedures and direct access to the bibliographic data. In such cases, open access to the bibliographic database is of crucial importance.

5.4 Reports

The results of analyses can be combined and presented to users in different forms:

- Booklet report (in PDF).
- (Service generated) web pages.
- Dashboards.
- Dataset (JSON, CSV, etc.).

6 Conclusions

We have presented an approach to support higher-order bibliographic services based on networks. Open access to high-quality bibliographic data is crucial for the faster development of such services. The new bibliographic database OpenAlex seems to be a step in the right direction. It needs the support of science policy and also of individual scientists (checking the correctness of their data).

Acknowledgements

The computational work reported in this paper was performed using a collection of R functions OpenAlex2Pajek and the program Pajek for analysis of large networks. Code, data, and figures are available on Github/Bavla/OpenAlex.

VB's work is partly supported by the Slovenian Research Agency ARIS (research program P1-0294, research program CogniCom (0013103) at the University of Primorska, and research projects J1-2481, J5-2557, and J5-4596), and prepared within the framework of the COST action CA21163 (HiTEc). JP's work is partly supported by ARIS (research program P5-0361 and research projects J1-2551 and J5-4596). TP's work is partly supported by ARIS (research program P1-0294 and research projects N1-0140, J1-2481, J5-4596).

References

[1] Massimo Aria, Trang Le, Corrado Cuccurullo, Alessandra Belfiore, and June Choe. 2024. openalexR: an R-tool for collecting bibliometric data from OpenAlex. *The R Journal*, 15, 4, 167–180.

[2] Vladimir Batagelj. 2003. Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*.

[3] Vladimir Batagelj. 2020. On fractional approach to analysis of linked networks. *Scientometrics*, 123, 2, 621–633. DOI: 10.1007/s11192-020-03383-y.

[4] Vladimir Batagelj. 2024. OpenAlex2Pajek. version 4, June 18. (2024). https://github.com/bavla/OpenAlex/tree/main/code.

[5] Vladimir Batagelj and Monika Cerinšek. 2013. On bibliographic networks. *Scientometrics*, 96, 3, 845–864. DOI: 10.1007/s11192-012-0940-1.

[6] Vladimir Batagelj, Patrick Doreian, Anuška Ferligoj, and Nataša Kejžar. 2014. *Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley Series in Computational and Quantitative Social Science*. Wiley, Chichester. ISBN: 978-1-118-91537-0; 978-0-470-71452-2. DOI: 10.1002/9781118915370.

[7] Vladimir Batagelj, Anuška Ferligoj, and Flaminio Squazzoni. 2017. The emergence of a field: a network analysis of research on peer review. *Scientometrics*, 113, 1, 503–532. DOI: 10.1007/s11192-017-2522-8.

[8] Vladimir Batagelj and Daria Maltseva. 2020. Temporal bibliographic networks. *J. Informetr.*, 14, 1, Article No. 101006. DOI: {10.1016/j.joi.2020.101006}.

[9] Dalmeet Singh Chawla. 2022. Massive open index of scholarly papers launches. *Nature*.

[10] DBLP – computer science bibliography. 2024. (2024). https://dblp.org/.

[11] Lorena Delgado-Quirós and José Luis Ortega. 2024. Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 5, 1, 31–49.

[12] Scholar GPS. 2024. (2024). https://scholargps.com/.

[13] Chenyue Jiao, Kai Li, and Zhichao Fang. 2023. How are exclusively data journals indexed in major scholarly databases? an examination of four databases. *Scientific Data*, 10, 1, 737.

[14] Bill Jones. 2002. Computational geometry database. (2002). ftp://ftp.cs.usask.ca/pub/geometry/.

[15] Daria Maltseva and Vladimir Batagelj. 2019. Social network analysis as a field of invasions: Bibliographic approach to study SNA development. *Scientometrics*, 121, 2, 1085–1128. DOI: 10.1007/s11192-019-03193-x.

[16] Peter V. Marsden. 1990. Network data and measurement. *Annu. Rev. Sociol.*, 16, 435–463. DOI: 10.1146/annurev.so.16.080190.002251.

[17] Nataliya Matveeva, Vladimir Batagelj, and Anuška Ferligoj. 2023. Scientific collaboration of post-soviet countries: the effects of different network normalizations. *Scientometrics*, 128, 8, 4219–4242.

[18] Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

[19] Google Scholar. 2024. (2024). https://scholar.google.com/.

[20] Bert TePaske-King and Norman Richert. 2001. The identification of authors in the mathematical reviews database. *Issues Sci. Technol. Librariansh.*, 31. DOI: 10.5062/f4kh0k9m.

[21] Wikipedia. 2024. API. August 22. (2024). https://en.wikipedia.org/wiki/API.

Are papers all that counts? A bibliometric analysis of the Slovenian scientific community

Aymeric Dupuis
Jožef Stefan Institute
Ljubljana, Slovenia
aymeric.dupuis@etu.univ-nantes.fr

Sašo Džeroski
Jožef Stefan Institute
Ljubljana, Slovenia
saso.dzeroski@ijs.si

Boshko Koloski
Jožef Stefan Institute
Ljubljana, Slovenia
boshko.koloski@ijs.si

Matej Martinc
Jožef Stefan Institute
Ljubljana, Slovenia
matej.martinc@ijs.si

Abstract

We conduct a bibliometric analysis of the Slovenian science by scraping the data from Slovenian current research information system (SICRIS) and using it to build a knowledge graph, representing a network of all Slovenian scientific fields and a large majority of Slovenian researchers. By analyzing this network using different graph measures, we obtain valuable insights into the connections between different scientific fields and researchers in Slovenian science. Additionally, we show the importance of graph measures as measures of scientific excellence, since they measure very different aspects of scientific success than the traditional citation metrics.

Keywords

bibliometrics, Slovenian scientific community, knowledge graphs

1 Introduction

With the growth and diversification of the scientific enterprise, obtaining empirical evidence on the research process is crucial for enhancing its efficiency and reliability. Meta-research and bibliometrics are developing scientific disciplines, seeking to analyse, evaluate and refine research practices, and several studies have focused on the analysis of the global scientific endeavour, e.g., identifying most prominent scientists and fields [7]. These studies also focus on the problem of how to properly rank scientific excellence and scientific outputs in general, warning that one should not rely on just a few metrics to obtain a comprehensive picture of the actual impact a specific scientist has [8].

Until now, very few studies have tackled the analysis of scientific ventures at national level, and to our knowledge, there has been no study covering the Slovenian scientific landscape specifically. This kind of research is nevertheless important and could potentially influence policies that would improve scientific production and enable effective distribution of research funds and resources.

In this study, we try to address the identified research gaps by 1.) drawing the map of Slovenian scientific research that would enable proper decision making and policy formulation, and 2.) proposing new metrics of scientific excellence that would allow us to obtain a more complete view of the impact a scientist or a discipline as a whole has. More specifically, our contributions are the following:

- Using the collected data about the Slovenian scientists and their projects, covering different scientific fields and a large majority of researchers working in Slovenian science, we conduct a graph analysis of connections between different fields and researchers. By drawing a comprehensive map of connections between actors and fields, we identify the most important researchers and scientific fields that connect others and play a vital role in the Slovenian scientific ecosystem.
- We created a new ranked list of Slovenian scientists according to graph based metrics, which were not available in any of the previous analyses or databases. We argue that these metrics measure the importance of a role that a specific scientist has in a research community, i.e., their influence which allows them to act as a bridge or a hub connecting scientists from different fields.

2 Related work

Studies in bibliometrics (see [4] for a comprehensive survey of techniques used for measuring scientific excellence) have recently gained traction in parallel with the success of the scientific enterprise, which has grown in both size and diversity, and with the availability of data. According to Ioannidis et al. [7], research on research is becoming important due to the mounting evidence suggesting an alarming drop in reproducibility of research findings, the growing inefficiency of the scientific process, and the fact that the number of false positives in the literature is exceedingly high. To address these problems, they propose a meta-research divided into five main categories that should be studied: methods, reporting, reproducibility, evaluation, and incentives. Studying these five areas would correspondingly allow for five distinct insights into how to perform, communicate, verify, evaluate, and reward research.

Recently, several studies also tackled the problem of how to properly rank scientist and scientific outputs in general. For example, Ioannidis et al. [8] addressed the increasing prevalence of multiauthorship observed in several fields and how this phenomenon affects the effectiveness of the informativeness of citation metrics. They also explored how sensitive the indicators are to self-citation and alphabetic ordering of authors. They concluded that multiple indicators should be used for ranking, as a composite of different metrics gives a more comprehensive picture of the actual impact that a specific scientist has. They also acknowledged that no single or composite citation indicator can be expected to select all the best scientists.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.11

Several studies employed graph-based metrics to enrich the assessment of bibliometric analysis [4, 1]. Network metrics such as degree of centrality, betweenness centrality, eigenvector centrality, closeness centrality, and PageRank were used to pinpoint the relative importance of research constituents (i.e., researchers and institution), which may not necessarily be reflected just through publications. In a large majority of cases, these metrics were calculated on co-authorship graphs.

The studies that would cover Slovenian scientific environment are very scarce. In fact, we are aware of just one, the study by [2], where they claim that research performance is highly dependent on the conditions of (national) research environments. They focus on analyzing research activity in six eastern European countries, namely Croatia, Estonia, Hungary, Latvia, Lithuania, and Slovenia, and try to determine and compare the effectiveness of research in a specific country by obtaining the number of articles belonging to the most cited 10% and the most cited 1% articles in the corresponding subject area and publication year for each country. Their empirical analysis addresses three levels: cross-country, cross-institution, and cross-researcher comparison. The study concludes that Hungary is the country with the highest output, followed by Croatia and then Slovenia, when it comes to the number of influential articles published.

3 Methodology

In this section, we describe our methodology, namely 1.) how we gather the data and 2.) how we analyze these data to obtain a map of the Slovenian scientific community.

3.1 Data Retrieval

Data were retrieved from the Slovenian Current Research Information System (SICRIS) website[1], which lists more than 35,000 researchers working in Slovenian research institutions. Data collection from the SICRIS website proved challenging, as information about a specific researcher can only be obtained by scraping his/her Web page on SICRIS. This required finding a solution to quickly retrieve data from more than 35,000 different pages, and to achieve this, we used the Python Asyncio[2] and BeautifulSoup[3] libraries, which allow the asynchronous connection to several dozen pages simultaneously and extraction of the required data.

Since the script sometimes took several seconds to connect to a specific page, which could quickly accumulate, resulting in considerable overall slowdowns, we optimized the procedure and identified potential slowdowns. Our proposed solution was to implement a strategy that involved canceling the connection and adding the URL to a list whenever a page failed to connect within a 0.5-second time frame. This timeframe was chosen after several trials and was found to be the best compromise. Once all pages had been visited, we repeatedly tried to reconnect to the URLs on this list until it was empty. This change significantly reduced the time required to retrieve all our data. Once all the data was retrieved, we used the Pandas library[2] for data manipulation, which allowed us to export the results into Excel spreadsheets, appropriate for further processing.

From SICRIS, we extracted research areas for each scientist and various bibliometric indicators of their impact, namely A", A', A1/2, citation metrics based on a quantitative assessment of publications in exceptional, high quality and important venues, respectively. We also extracted the A1 metric, which represents a weighted sum of these three metrics, a CI10 metric measuring the number of pure citations of scientific work in the last 10 years, the CImax metric measuring the number of citations in the most cited work, and the h10 metric representing the h-index in the last ten years. Furthermore, we extracted the SICRIS points, a conglomerate metric combing several distinct metrics mentioned above, and the A3 metric, which measures the amount of funds a specific researcher received for his research activity outside of the Slovenian National Research Agency (ARIS).

Finally, the SICRIS database also contains information on projects financed by the Slovenian national research agency in which a specific researcher participated. Scraping this information provided us with an important insight into collaborations between different scientists and fields, allowing us to build collaboration graphs, calculate several graph-based ranking criteria and draw the map of the Slovenian scientific community.

3.2 Methods

Once the data was obtained, we conduct two distinct analysis steps, namely 1.) graph construction and analysis, and 2.) correlation analysis

3.2.1 Graph construction and analysis. To construct the necessary graphs, we used the Python NetworkX library [6]. Using the data from SICRIS, which contain information about project collaboration, we created an undirected graph as follows: all researchers who participated in at least one project are represented by a node, and nodes of researchers who worked together on a project are connected by weighted edges, in which the weights represent the number of shared projects. By removing the isolated nodes, we ended up with a graph with a total of 20,012 nodes and 618,871 edges.

In the next step, we apply several graph statistics and measures in order to obtain several node rankings, each of them measuring a different aspect of the importance a specific node (i.e., a researcher) has in the graph. More specifically, we calculate PageRank (PR), Betweenness centrality (BC), and Eigenvector centrality (EC) measures.

In the context of our graph, the **PageRank** [3] algorithm is applied to evaluate the influence of researchers within the collaboration network. Thus, researchers who are strongly connected to other researchers, who also have many connections (i.e, the so-called hubs in the graph), will have a higher PR score, reflecting their importance and influence in the Slovenian research community. On the other hand, the **Betweenness Centrality** [5] measure evaluates the role of each researcher as an intermediary or a bridge between other researchers. This measure is based on the idea that researchers who are on many collaboration paths between other researchers are considered central and influential in the network. In our contexts, it helps to better understand the structure of the collaboration network among researchers. Researchers with high BC are those who play a crucial role in creating links between different subgroups of researchers and interdisciplinary connections. In practical terms, BC evaluates the number of times a researcher is traversed by the shortest paths connecting other researchers in the network. Thus, researchers who are frequently used as pathways for collaboration among their peers obtain higher BC scores.

[1] https://cris.cobiss.net/ecris/si/en
[2] https://docs.python.org/3/library/asyncio.html
[3] https://www.crummy.com/software/BeautifulSoup
[2] https://pandas.pydata.org/

[1] https://networkx.org/

Another graph centrality measure that we applied to the created graph is the **Eigenvector centrality** [9]. This measure evaluates the influence of a researcher taking into account both the quality and the quantity of connections. EC assigns more weight to connections that include influential researchers. Thus, a researcher connected to influential researchers will be assigned a high score, reflecting potentially greater influence within the network. This measure helps to detect researchers who, even with fewer direct connections, occupy strategic positions in the collaboration network. While this may seem similar to the PR algorithm, there are some differences. Unlike PR, which primarily focuses on the popularity of links, Eigenvector centrality also takes into account the quality of connections. This means that even if a researcher does not have a large number of direct connections, if they are connected to influential researchers, their Eigenvector centrality score can be high. In summary, while both measures aim to evaluate the influence of researchers in a network, they do so through slightly different approaches, thus offering complementary perspectives for analyzing the structure and importance of actors within the collaboration network.

Our second important area of focus in our research is the collaboration between different fields. To build a graph that would represent interdisciplinary collaboration between fields, we grouped all researchers from the same field into a single node, representing an entire field, i.e., we obtain a node for each scientific field found on SICRIS. Similar to the previous graph, edges and their weights represent collaborations on a project between researchers in the linked fields.

3.2.2 Correlation analysis. In order to better understand the metrics from SICRIS and to evaluate the relevance of our scores, we deemed it pertinent to explore the correlation between all our data. This analysis has two main purposes. First, we aim to test the **hypothesis 1** that the new graph ranking we presented, measure different aspects of scientific excellence than the more established measures based on number of citations or publications available on the SICRIS web page. This hypothesis would be deemed correct if one-on-one correlations scores between the newly proposed graph measures and other measures would be low, and incorrect if correlations would be high.

Additionally, we wish to explore the correlation between the established measures available on the SICRIS web page. More specifically, we wish to test the **hypothesis 2** that these measures are strongly correlated, which would indicate that they essentially all measure a very similar aspect of scientific excellence, which is problematic. In order to obtain one-on-one correlations between all measures, we calculate the Spearman correlation coefficient among all of them and then display it through a heatmap.

4 Results

In Table 1, we present some of the results of the graph analysis conducted on the graph of nodes representing researchers, connected by edges representing project collaborations. More specifically, we present 10 best ranked researchers in the SICRIS dataset according to the average between ranks of the three newly proposed graph-based measures, their declared scientific fields, and their ranking (i.e., lower is better) according to the SICRIS points, BC, EC and PR measures.

Note that while the table does contain some highly ranked researchers according to the SICRIS points (e.g., Dr. Sašo Džeroski is ranked as 33rd out of roughly 20K researchers according to this criteria), several researchers in the table are ranked relatively low according to SICRIS points (e.g., the best ranked researcher according to our novel three measures, Dr. Branimir Leskošek, is ranked as 5731th according to the SICRIS points). This finding supports **hypothesis 1** that the proposed new measures measure different aspects of scientific excellence than the more established citation measures. Another important observation is that 7 out of 10 best ranked scientists appear to be active in two fields. This might suggest that they are (or have been) involved in several interdisciplinary projects, which could have a positive influence on the newly proposed graph-based metrics.

In Figure 1, we present the heatmap of the correlations between the different metrics extracted from SICRIS website and the newly proposed graph-based metrics. We observe a strong correlation between PR and BC, 0.7, which might suggest that researchers who collaborate with a wide range of colleagues from different fields are more likely to work with the most important ones.

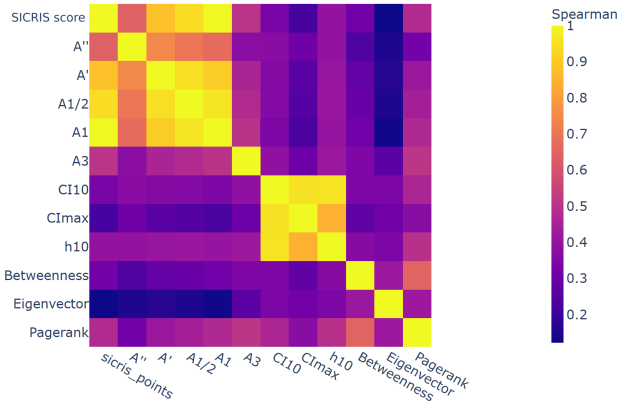


Figure 1: Heatmap of the Spearman correlation among metrics.

We also observe very strong correlations in the top left corner of the heatmap. While a strong correlation was expected, as A", A', A1/2 and A1 are all scores based on the number of publications (in venues of different qualities), the almost perfec correlation between the SICRIS points and A1 (which suggest they measure exactly the same aspect of the scientific impact) is surprising. This finding supports **hypothesis 2** that the current SICRIS measures all measure a very similar aspect of scientific excellence. On the other hand, there is no strong correlation between any of the newly proposed graph-based metrics and metrics extracted from the SICRIS website.

In Table 2, we present the results of our study of interdisciplinary collaboration between different scientific fields. The graph metrics were obtained from a graph of nodes representing fields and edges representing interdisciplinary project collaborations. Note that the field of Computer science and informatics ranks first according to all the criteria. On the other hand, most interdisciplinary collaborations are conducted by the researchers from the field of Chemistry, which ranked as third according to the average (AVG) between the ranks of three graph-based metrics, PG, BC and EV.

5 Conclusions

The graph based bibliometric analysis of the Slovenian scientific community shows that current citations based metrics do not cover some aspects of scientific excellence, such as researcher's

Table 1: 10 best ranked researchers in the SICRIS dataset according to the average between ranks of the three newly proposed measures, BC, EC and PR. We do not show metric scores, but ranks according to scores (i.e., lower value is better).

Researcher	Field 1	Field 2	SICRIS points	BC	EC	PR	AVG
15355 PhD Branimir Leskosek	Public health (occupational safety)	Computer science and informatics	5731	8	4	31	14
06013 PhD Damjana Rozman	Biochemistry and molecular biology	Metabolic and hormonal disorders	704	21	2	33	18
11279 PhD Nives Ogrinc	Control and care of the environment	Animal production	182	7	50	3	20
27733 PhD Tina Kosjek	Control and care of the environment	Pharmacy	809	2	73	9	28
22459 PhD Tadeja Rezen	Neurobiology	Microbiology and immunology	1837	61	3	49	37
22621 PhD Polonca Ferk	Metabolic and hormonal disorders	Pharmacy	5059	13	8	103	41
12688 PhD Kristina Gruden	Biotechnology	/	219	44	139	6	63
08800 PhD Gregor Sersa	Oncology	/	71	3	185	1	63
12315 PhD Ester Heath	Control and care of the environment	Chemistry	208	62	115	23	66
11130 PhD Šašo Dzeroski	Computer science and informatics	/	33	1	195	20	72

Table 2: Scientific fields as defined in the SICRIS database, sorted according to the average (AVG) between the ranks (lower score is better) of three graph-based metrics, PG, BC and EV.

Rank	Field	Collaborations	PR	EC	BC	AVG	Rank	Field	Collaborations	PR	EC	BC	AVG
1	Computer science and informatics	81248	1	1	1	1.0	36	Textile and leather	21080	27	41	39	35.67
2	Materials science and technology	88934	4	3	4	3.67	37	Animal production	34982	29	29	50	36.0
3	Chemistry	101139	2	2	12	5.33	38	Political science	13598	46	37	27	36.67
4	Control and care of the environment	52648	5	8	9	7.33	39	Anthropology	9860	53	36	24	37.67
5	Physics	50010	3	9	14	8.67	40	Ethnology	6698	65	39	11	38.33
6	Plant production	74535	6	6	16	9.33	41	Cardiovascular system	20793	28	43	45	38.67
7	Systems and cybernetics	45584	7	10	23	13.33	42	Telecommunications	14068	41	45	31	39.0
8	Biology	58879	12	7	21	13.33	43	Veterinarian medicine	30954	32	34	60	42.0
9	Civil engineering	36466	22	13	6	13.67	44	Metabolic and hormonal disorders	18518	30	46	55	43.67
10	Biochemistry and molecular biology	79725	11	5	25	13.67	45	Metrology	12978	34	52	47	44.33
11	Neurobiology	45680	14	12	19	15.0	46	Law	7480	54	49	32	45.0
12	Biotechnology	87261	8	4	33	15.0	47	Psychology	8583	51	55	29	45.0
13	Interdisciplinary research	22946	9	33	5	15.67	48	Human reproduction	21535	35	42	58	45.0
14	Public health (occupational safety)	30400	10	25	13	16.0	49	Process engineering	15340	36	47	53	45.33
15	Educational studies	23518	33	15	3	17.0	50	Hydrology	12396	40	53	44	45.67
16	Mathematics	30680	17	20	20	19.0	51	Architecture and Design	4242	58	57	22	45.67
17	Manufacturing technologies and systems	38874	18	14	26	19.33	52	Philosophy	7380	57	44	43	48.0
18	Forestry, wood and paper technology	30620	19	28	15	20.67	53	Sport	10013	43	54	49	48.67
19	Geography	18555	39	23	2	21.33	54	Geodesy	7760	45	56	51	50.67
20	Economics	26891	31	16	18	21.67	55	Electric devices	13633	42	51	59	50.67
21	Microbiology and immunology	54175	16	11	42	23.0	56	Literary sciences	6399	61	50	48	53.0
22	Sociology	19922	44	17	10	23.67	57	Traffic systems	4448	48	60	52	53.33
23	Pharmacy	41125	15	18	41	24.67	58	Culturology	7240	60	48	54	54.0
24	Linguistics	18176	49	19	7	25.0	59	Technology driven physics	6876	47	59	64	56.67
25	Chemical engineering	33753	13	27	38	26.0	60	Communications technology	4388	52	63	56	57.0
26	Energy engineering	32762	23	21	40	28.0	61	Psychiatry	2481	55	65	61	60.33
27	Computer intensive methods and applications	26942	20	32	34	28.67	62	Criminology and social work	2324	66	62	62	63.33
28	Mechanics	26444	24	31	36	30.33	63	Mining and geotechnology	2342	59	68	63	63.33
29	Oncology	37101	21	24	46	30.33	64	Theology	2941	67	58	66	63.67
30	Geology	26961	37	26	28	30.33	65	Ethnic studies	2398	63	61	67	63.67
31	Electronic components and technologies	28858	26	30	37	31.0	66	Art history	1408	70	64	57	63.67
32	Historiography	12390	56	22	17	31.67	67	Archaeology	1177	68	66	65	66.33
33	Urbanism	8669	50	40	8	32.67	68	Information science and librarianship	792	62	70	70	67.33
34	Mechanical design	22352	25	38	35	32.67	69	Stomatology	391	64	71	68	67.67
35	Administrative and organisational sciences	18563	38	35	30	34.33	70	Landscape design	1046	69	67	71	69.0
							71	Musicology	748	71	69	69	69.67

role of connecting a wider research community. Our correlation analysis indicates that existing measures of scientific excellence extracted from the SICRIS web page are strongly correlated. In the future, we plan to expand this analysis to also measure the impact of Slovenian scientists on the global scientific enterprise and conduct additional research to try to find certain patterns across disciplines, or institutions.

6 Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding for the programmes Knowledge Technologies (No. P2-0103).

References

[1] Njål Andersen. 2021. Mapping the expatriate literature: a bibliometric review of the field from 1998 to 2017 and identification of current research fronts. *The International Journal of Human Resource Management*, 32, 22, 4687–4724.

[2] Lutz Bornmann. [n. d.] Research excellence in eastern europe: a bibliometric study focusing on croatia, estonia, hungary, latvia, lithuania, and slovenia.

[3] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30, 1-7, 107–117.

[4] Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. 2021. How to conduct a bibliometric analysis: an overview and guidelines. *Journal of business research*, 133, 285–296.

[5] Linton C. Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry*, 40, 1, 35–41. Retrieved June 27, 2024 from http://www.jstor.org/stable/3033543.

[6] Aric Hagberg, Pieter J Swart, and Daniel A Schult. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 11–15.

[7] John PA Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N Goodman. 2015. Meta-research: evaluation and improvement of research methods and practices. *PLoS biology*, 13, 10, e1002264.

[8] John PA Ioannidis, Richard Klavans, and Kevin W Boyack. 2016. Multiple citation indicators and their composite across scientific disciplines. *PLoS biology*, 14, 7, e1002501.

[9] Paul Turán, editor. 1969. *Publications of edmund landau. Number Theory and Analysis: A Collection of Papers in Honor of Edmund Landau (1877–1938)*. Springer US, Boston, MA, 335–355. ISBN: 978-1-4615-4819-5. DOI: 10.1007/978-1-4615-4819-5_23.

Empowering Open Education Methodologies with AI-based Strategies for the Customization of Education

Tel Amiel
Universidade de Brasilia
Brasilia, Brazil
amiel@unb.br

Antônio J. Moraes Neto
Instituto Federal de Brasilia
Brasilia, Brazil
antonio.neto@ifb.edu.br

Joao Pita Costa
IRCAI, Jozef Stefan Institute
Ljubljana, Slovenia
joao.pitacosta@quintelligence.com

Mitja Jermol, Anja Poljanar
IRCAI, Jozef Stefan Institute
Ljubljana, Slovenia

ABSTRACT

The amount and heterogeneity of data generated in the context of education allied to the rapid progress of scientific research and technological development have created vast amounts of data, much of it open data, but significant challenges to gathering, filtering and making sense of this information. In this paper, we discuss the research outcomes of complementary Artificial Intelligence (AI)-based strategies monitoring and enhancing Open Education, mining online forum interaction student-educator, and empowering mentorship of educators. Firstly, the initial results obtained from the construction of an Observatory focusing Open Education Resources (OERs), contribute to implement 2019 UNESCO OER Recommendation and advance the Education-focused Sustainable Development Goal (SDG) 4. It is acting on five verticals, enriching and treating multilingual data, it displays meaningful information on a dashboard focused on AI and OERs and serving as a collaboration platform focused on existing partnerships within the international research centre on AI under the auspices of UNESCO (IRCAI), the UNESCO Chair in Distance Education and the UNESCO Chair on Open Technologies for Open Educational Resources and Open Learning, mobilizing research collaboration on key AI research challenges relating to generating knowledge about OER. Secondly, we will discuss the recent development of an Educational Recommender System (ERS) that integrates Conversational Analysis (CA) to assess and enhance collaborative learning (CL) in Virtual Learning Environments (VLEs). This novel system was designed to identify collaboration among students and provide tailored recommendations to promote participation and interaction within discussion forums. Finally, we will discuss the development and implementation of AI and OERs in alignment with SDGs, addressing topics of significant social impact over an international online mentoring initiative.

KEYWORDS

Open Education, Machine Learning, Educational Recommender System, Conversational Analysis, Virtual Learning Environment

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.16

1 Introduction

The centralizing piece of the discussions in this paper is an AI-based observatory that allows to explore OER-related topics, particularly those mentioned in the OER Recommendation: promoting OER and acknowledging it's contribution to advancing quality education while providing information on advances focused on the equity and inclusion qualities of OER, as well as on research, activities, projects and news related to OER development, including new initiatives and projects while also promoting public infrastructures for education. The OER Observatory builds on the content made available in UNESCO's OER Dynamic Coalition Portal (oerdynamiccoalition.org) providing the user with access to any of the four proposed views: media; science; policies and training. In each of the views, the user can access interactive data visualisation summarising the sourced data configured to observe the UNESCO OER recommendations. As it is fully based on open data, it allows the user to click on the resources collected and summarized, being taken directly to the source in media, journal, policy or training.

Embracing the intersection of AI and education, which has led to the development of various tools that personalize and enhance learning experiences, we discuss a complementary research based on CA much aligned with the objective of empowering Community interaction at the SDG 4 (Education) Observatory [6]. AI applications in education often focus on providing adaptive feedback, facilitating personalized learning paths, and analyzing student data to improve outcomes. CA is a method that examines the understanding generated through interactions, offering a framework for analyzing how students collaboratively build knowledge. By combining CA with AI, this research aims to develop a system that not only assesses but also actively promotes collaboration in VLEs [10]. The ERS discussed later in this paper, is an example of how IRCAI's SDG4 Observatory gains a complex capability towards the engagement with communities such as in Education. This discussion then expands towards the appropriate mentorship of the professionals that will change the domain's landscape. While initiatives in this context are diverse and disperse, the authors are not aware of existing similar approaches [5].

2 AI-based strategies for the moderation of online forums on education

Entering the age of Big Data, AI is feeding the data-driven digital transformation across industries including Education. CL emphasizes the importance of group tasks and joint participation, wherein students learn by actively engaging in dialogues that facilitate the sharing of ideas and information. Even in remote settings, CL enables students to learn together through virtual platforms. AI offers new opportunities as a pedagogical tool, providing adaptive and personalized environments that can support CL. This research explores the integration of AI into educational contexts, particularly through the development of an Educational Recommender System (ERS) that uses CA to identify and promote collaboration among students in VLEs [1] (see Figure 1).

Figure 1: The ERS forum analysis screen

The research methodology is divided into three key stages: *Conversational Analysis*, applying CA to monitor discussion forums within the Moodle platform, focusing on interactions among students, identifying collaborative behaviors and interaction patterns; *Collaboration Assessment***,** evaluating the level of collaboration among students based on identified interaction patterns; and *Development of ERS*, building a mechanism that provides recommendations to students, teachers, and tutors. These recommendations are aimed at enhancing collaboration and are based on the analysis of forum interactions [15]. The initial dataset comprises 20,976 messages of Moodle discussion forums, with 15,703 posted by students from a vocational education school. The analysis focuses on these messages to develop and validate the ERS's recommendations. The quality of collaboration is measured through various indicators, which are extracted during different stages of CA. *Preprocessing* applies techniques of Natural Language Processing (NLP) to ensure the accuracy of the analysis, preparing data for the *Resource Processing* stage using Social Network Analysis (SNA) to characterize social dynamics and interactions among students. Moreover, the *Message Attribute Identification* is the CA stage that allows identifying characteristics of students' messages, , specifically their questions, and then *Topic Modeling* is employed to identify key terms discussed in the forums [12], using Tomotopy library (bab2min.github.io/tomotopy) The ERS was tested across five experimental cycles in different classes at two Brazilian Federal Institutes , in a Portuguese language context. The results indicated a positive impact on student learning, with 82% of participants acknowledging the relevance of the recommendations. The system motivated increased participation and collaboration, with a notable trend of students writing more and systematically organizing their ideas in forum posts. Additionally, 90% of students engaged in other activities proposed by their teachers, demonstrating the effectiveness of the recommendations. The results also demonstrate the system's effectiveness in fostering collaboration, with positive feedback from students and educators. A dashboard was developed for teachers, containing graphs including one that shows the main terms discussed in the forum by analysis, in which each edge represents a message from the student with two of these terms, and the nodes in blue highlight the new terms that emerged in relation to the previous analysis (see Figure 2).

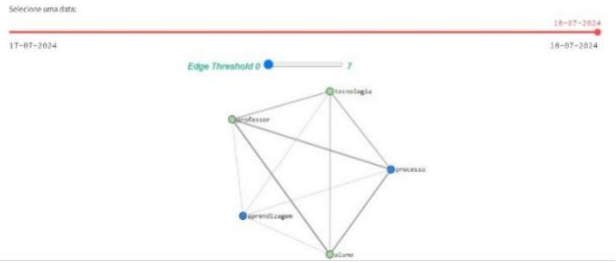


Figure 2: Visual analysis of students' collaboration in a discussion forum where nodes represent actors in the discussion (students/educators) and edges represent interactions.

The development of the ERS represents a significant advancement in promoting collaborative learning in educational settings [6,7]. By integrating CA into the system, the ERS effectively identifies and enhances collaboration among students. The current implementation of this ERS aims to provide personalized recommendations to students, teachers, and tutors, fostering a more interactive and collaborative learning environment [6]. Future work will explore the integration of additional features, such as *wikification* and visualization tools, to further enhance the system's capabilities. Furthermore, the research will benefit from the semi-automatic categorization of educational resources of a range of formats, including videos as in [3].

3 An AI-based Observatory to Assess the Impact of OER Worldwide

Although the abundance of information available online, some of which is labeled as education-related, it is harder and harder to find the appropriate resources that can serve education either at an undergraduate or a professional training level. IRCAI's Open Education Observatory is an initiative dedicated to monitoring, analyzing, and promoting the use of OERs globally. It serves as a hub for research insight and fomenting collaboration, providing valuable insights and data on the

adoption, impact, and trends of OER in education systems worldwide. The observatory supports educators, policymakers, and institutions in leveraging open resources to enhance teaching and learning. It is designed to support government and institutional decision-makers dedicated to promoting the goals of the 2019 UNESCO OER Recommendation, which is centred on OER but generally promotes the ideals of Open Education (see Figure 3).

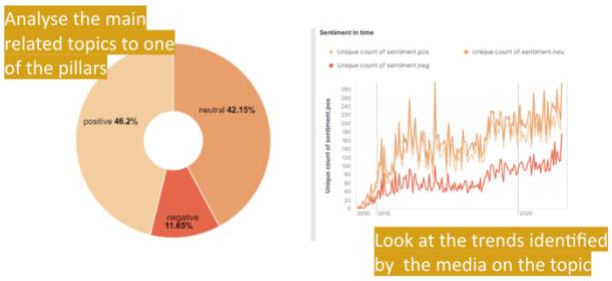


Figure 3: Dashboard of visual modules to analyse the most relevant topics under a certain domain or SDG, and the trends that can direct the education actors preparedness

The Open Education Observatory ingests a range of different data sources with heterogeneous nature and different frequency: (i) worldwide news in almost real-time providing information from a vast catalogue of multilingual world news, captured in more than 60 languages and based on a variety of *wikidata* concepts; (ii) published scientific articles, including journal and conference papers, mostly peer-reviewed, covering over more than 126 million articles with yearly updates; (iii) OER policies from the OER Policy Hub (www.oepolicyhub.org) that needs to be input into the OER DC Portal; subsequent extraction and enrichment of metadata; preparation of dashboard related to dashboard based on filters over the metadata, as well as OECD policies data and metadata on AI and Education with yearly updates; (iv) lectures and videos selected and filtered on content from Videolectures.net [10] resources related to OER; (v) a snippet of worldwide public and private initiatives related to AI and SDG 4 captured by IRCAI's Top100 and related actions; and (iv) a range of worldwide indices with yearly updates on Education-related topics such as the percentage of children out of school, or the literacy rate in youth and adults (see Figure 4).

To ensure that content is readily available for each focus area, materials from the mentioned sources are categorized by relevant keywords and concepts closely associated with the five key areas of the Recommendation. This organization allows users to easily filter and access content based on their specific interests within these areas. By doing so, users can tailor their exploration of resources to match their focus, whether it's capacity building, supportive policy development, leveraging equitable access provided by OER, sustainability models, or international cooperation.

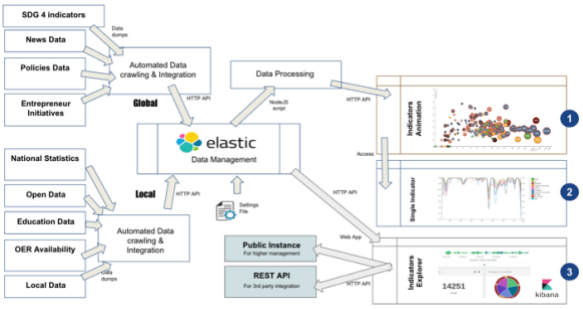


Figure 4: The architecture of the OER Observatory as an Elasticsearch-based system that enables the visualization of heterogeneous data on OERs

For each area, users can filter and find content specific to their domain of interest: up-to-date news and research on OER developments, academic studies related to professional development, and relevant lectures for capacity building; information on OER policy development; resources and research focused on effective, inclusive, and equitable access to quality OER; strategies for developing sustainable OER models; and opportunities for fostering international cooperation through potential new partnerships and shared goals. This organized approach enhances the ability to pinpoint and utilize the most relevant information in each domain. Information generated by the Observatory can be used to aid in the resolution of problems related to the promotion of OER, by identifying trends and major areas of discussion, and to explore successful scenarios through similar challenges and cases. The Observatory provide benefits to a range of stakeholders including: national governments, providing access to a variety of perspectives on OER trends for decision-making; educational and research institutions, facilitating the access to resources and data; civil society, allowing access to information and training materials that explore the knowledge available towards the implementation of the UNESCO recommendations; and the general population, empowering open education.

4 Open Education for a Better World

The Open Education for a Better World (OE4BW) program is an international online mentoring initiative aimed at advancing the development and implementation of open educational resources (OER) that address topics of significant social impact, in alignment with the United Nations Sustainable Development Goals (SDG) [2,14]. As part of the Slo2Svet project, the program received 70 project applications and 87 mentor applications from six continents and 25 different countries (see Figure 5). The program's activities are structured into thematic clusters, focusing on areas such as Artificial Intelligence, Displaced

Persons, Sustainability, Health and Well-being, Renewable Energy, Education, and Youth (specifically targeting developers aged 12-24). Throughout the project development process, progress was closely monitored by a network of mentors and hub coordinators, providing essential guidance and support to OER developers. Additionally, within the scope of the Slo2Svet project, evaluation rubrics for the OER projects were developed and will be utilized during the final conference, where developers will present their completed work.

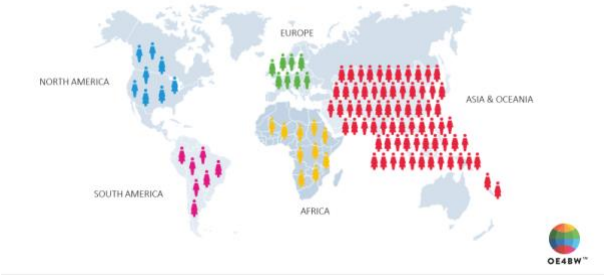


Figure 5: Participants of the OE4BW mentorship in 2023/24.

5 Conclusions and further work

In this paper we discussed the research results and opportunities in Open Education, building on an overall perspective over the OER landscape, the AI-enhanced student-educator interaction, and the mentorship for further progress.

We will be exploring further the potential of the OER observatory, particularly in what regards the appropriate use of LLMs in analyzing the compliance to AI policies in Education. In what regards the future developments of the EduColab, in alignment with IRCAI's *SDG 4 Observatory* and the *Videolectures.net* research agenda and the potential for institutional collaboration, we will focus on: (i) the appropriate *wikification*, incorporating suggestions of *Wikipedia* concepts identified by *Wikifier* and related to the main discussion topics; (ii) integration of interactive data visualization presenting graphical representations of collaboration trajectories, topic evolution, and other key indicators; (iii) extending the system, applying the ERS to other datasets, including public and private message exchange logs, to validate and enhance its applicability; and (iv) personalized recommendations**,** developing a user-based collaborative filtering technique to tailor recommendations more specifically to individual student groups. Moreover, we will explore together the pathways of AI-based citizen science in the context of Open Education and how it can be integrated in the wider scope of the SDG4 Observatory. In the context of the Slo2Svet project, we are conducting a comprehensive analysis of the Open Education for a Better World (OE4BW) mentoring program since its inception, examining outcomes and connections to other initiatives [see for example, 12]. Additionally, we will develop an evaluation framework to assess the impact of the projects produced through the program, mapping project outputs to the five action areas of the 2019 UNESCO OER Recommendation, using insights provided by automatic text analysis and other AI tools. This will allow us to connect the projects produced by OE4BW to the concrete objectives of the Recommendation, providing examples of practice that can be leveraged to advance its goals.

ACKNOWLEDGMENTS

We thank the support of the Slovenian Research Agency (ARIS) and Ministry of Foreign and European Affairs (MZEZ) on the project Slo2Svet - *Connecting cultures, informing and learning through Open Educational Resources and AI* (V2-2363).

REFERENCES

[1] Ahmadian Yazdi, H., Seyyed Mahdavi Chabok, S. J., and Kheirabadi, M. (2022). Dynamic Educational Recommender System Based on Improved Recurrent Neural Networks Using Attention Technique. Applied Artificial Intelligence, 36(1), 2005298.

[2] Drevensek, M., and Urbancic, T. (2022). The Role of Teamwork in the Creation of Open Educational Resources for Closing SDG-Related Knowledge Gaps. Open Praxis, 14(2).

[3] M. Grcar, D. Mladenic., and P. Kese (2009). Semi-automatic categorization of videos on videolectures. net. In Proceedings, of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Springer, pp. 730-733.

[4] Koschmann, T. (2013). Conversation Analysis and Collaborative Learning. In C. Hmelo-Silver, C. Chinn, C. Chan, & A. O'Donnell (Eds.), The Intern. Handbook of Collaborative Learning. Routledge Handbooks, pp. 149–167.

[5] Liu, Q., Huang, J., Wu, L., Zhu, K., and Ba, S. (2019). CBET: Design and evaluation of a domain-specific chatbot for mobile learning. Universal Access in the Information Society.

[6] Moraes Neto, A. J., and Fernandes, M. A. (2019). Chatbot and Conversational Analysis to Promote Collaborative Learning in Distance Education. 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), 2161-377X, pp. 324–326.

[7] Moraes Neto, A. J., Fernandes, M. A. & Amiel, T. (2022). Conversational Analysis to Recommend Collaborative Learning in Distance Education. 14th Intern. Conference on Computer Supported Education, pp. 196–203.

[8] Moraes Neto A. (2024) Sistema de Recomendação Educacional para Diagnosticar e Promover a Colaboração em Ambientes Virtuais de Aprendizagem. Doctoral thesis. Federal University of Uberlandia.

[9] Novak E., Novalija I. (2016) Visual and Statistical Analysis of VideoLectures.NET, Proceedings of the SIKDD'16.

[10] Urbančič, T., Polajnar A., and Jermol, M. 2019. (2019). Open education for a better world: a mentoring programme fostering design and reuse of open educational resources for sustainable development goals. Open Praxis. Open praxis. ISSN 1369-9997, Vol. 11, no. 4; pp. 1-18.

[11] Urbančič, T., Polajnar, A., & Jermol, M. (2019). Open Education for a Better World: A Mentoring Programme Fostering Design and Reuse of Open Educational Resources for Sustainable Develop. Goals. *Open Praxis, 11*(4).

[12] Urbančič, Tanja, et al. (2023) Developing supportive policies and strategies for their implementation: student experience with real-world cases. Open Educational Resources in Higher Education: A Global Perspective. Singapore: Springer Nature Singapore, 2023. pp. 35-53.

[13] Uthus, D. C., & Aha, D. W. (2013). Multiparticipant chat analysis: A survey. Artificial Intelligence, 199–200, 106–121.

[14] Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. Information Systems, 94.

[15] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39.

Addressing Water Sustainability Challenges in North Africa with Artificial Intelligence

Mustafa Zaouini, Maurizio Santamicone, Lee Chana
AI in Africa
Johannesburg, South Africa
mus@fliptin.com

Joao Pita Costa*, Davor Orlic, Mihajela Črnko
IRCAI, Quintelligence
Ljubljana, Slovenia
joao.pitacosta@quintelligence.co

Manal Cherkaoui, Anas Ait Aomar, Ikram Chairi, Karima Echihabi
UM6P
Ben Guerir, Morocco

Hanaa Hachimi, Y. Kaddouri, I. Lirmaqui, A. H. Alaoui, O. Ignammas, H. Rahhou
Ibn Tofail University
Kénitra, Morocco

M. Wahib Abkari, R. Rachidi, W. Laaleg, Z. Hidila, M. Tabaa
Moroccan School of Engineering Sciences, Casablanca, Morocco

K. Gourari, I. Annaki, B. Jearani, S. Trabi, T. Zennouhi, M. Sbaa
UMP Univesity
Oujda, Morocco

J. T. El Azzoiani, M. Ait Essibaa, A. Hamidine, H. Lachheb
Al Akhawayn University
Ifrane, Morocco

ABSTRACT

The topic of water sustainability has been leading the priorities worldwide where Artificial Intelligence (AI) can position research institutions, public & private companies and governments towards evidence-based decision-making in regards to water resources. In this particular domain, the amount and heterogeneity of data generated allied to the rapid progress of scientific research and technological development have created vast amounts of data, but significant challenges to gathering, filtering, and making sense of this information. This paper presents the research outcomes of collaborative effort engaging a total 51 students mentored by 15 professors across 11 research institutions in North Africa, distributed by 14 selected projects focusing the appropriate application of machine learning methods to local and national water sustainability problems. These outcomes were motivated by a youth challenge co-organized during May 2024 between AI Africa and IRCAI with the support of GITEX.

KEYWORDS

Machine learning, text mining, large language models, community engagement, water sustainability, competition

1 Introduction

Building upon common interests, exciting initiatives and existing projects developed by IRCAI and AI in Africa (aiinafrica.org), focused on AI and Sustainability, this activity aimed to build capacity within African youth to advance the Sustainable Development Goals (SDGs) though AI on challenges within their own communities and in the region. The AI Youth Challenge originated in the context of discussions started in GITEX Dubai in 2023 and forwarded to a concrete event in the AI Everything section of the GITEX Africa in the end of May 2024. It was mostly directed to PhD/MsC students and young entrepreneurs working on AI to solve problems for the good of their communities, exploring a wide range of machine learning methodologies (from image recognition on satellite imagery, to text mining on social media, gamification strategies optimizing water consumption, and application of LLM frameworks for RAG and AI Agents in the context of water sustainability), engaging experts from global agencies like, e.g., UNESCO, AI Movement, and UNESCO's Water Education Institute, as well as national companies, research institutions and government. The global challenge of this action, "Water, AI and Sustainability" is one of the MENA priorities, takes into consideration the UN Water Program for 2024-25 [12], and follows the work done by IRCAI with the European Commission (EC) on the NAIADES Water Observatory [9], as well as the recently opened new IRCAI Committee on AI and Water Resource Management [4] focusing on the impact of AI in SDG 6 [11]. This work aligns with UNESCO's interests in taking action to capacitate the Youth towards AI, with focus on the recent activities based from Morocco but with a global scope, including the opening of the new UNESCO AI Centre, the AI Movement (aim.um6p.ma).

Figure 1: Winner of the AI4Water challenge, designed and developed by UM6P students, exposing a water map that pinpoints remote villages with assigned water scores based on satellite imagery and crowdsourced data.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.17

2 Finalist innovative ideas on water sustainability

Attracting the participation of more than 50 PhD and MsC students across 20 teams based in research institutions in Morocco, this initiative was designed to encourage a conversation between the communities, corporate thought leaders, the education visionaries, and the ecosystems builders to have constructive conversations around the shifts and needs of the changing future landscape. The discussions included researchers, start-up communities, technologists, and government representatives to unite and define the future of water sustainability as they see it. The selected AI technologies and methodologies ranged from the use of satellite imagery to the analysis of news and social media, or the input from water-related sensors and the application of Large Language Models (LLMs) to describe good practices. We shall proceed with describing the problems addressed by the finalists of the AI4Water challenge, their prototypes and the value of the innovation they brought with them.

AquaScore. The Rural communities in Morocco's High Atlas Mountains struggle with water management due to limited resources and visibility. Despite needing modest funds, these villages face significant hurdles in accessing support. The challenge lies in objectively quantifying water issues and connecting these communities with potential supporters. AquaScore creates a water map that pinpoints remote villages and assigns them water scores based on satellite imagery and crowdsourced data. This enables ranking villages by water criticality, helping funders and supporters identify where to direct their assistance effectively. The prototype (described in Figure 1) also offers a platform for discussing water solutions, fostering community engagement through gamification features. By increasing the visibility of rural Moroccan villages and providing objective water criticality assessments, AquaScore facilitates efficient resource allocation for donors and experts. This AI-driven approach ensures fair and unbiased assistance to communities in need, promoting water sustainability and improved water management in constrained environments.

AquaScore employs a hybrid approach combining Computer Vision (CV) and Natural Language Processing (NLP). CV algorithms segment satellite images to generate automated baseline water scores, while NLP algorithms extract insights from textual data to enhance score accuracy. This combination allows for objective assessment and continuous improvement of water criticality rankings. The team has already aggregated data on 1,322 High Atlas villages, extracted satellite images, and segmented them using Facebook's Segment Anything model. This process was completed on UM6P servers using 500GB of storage and 80 CPU cores. The system will incorporate user-submitted reports and internet-scraped data to further refine water scores. The uniqueness of AquaScore lies in its data generation and refinement approach. It creates datasets in areas with data scarcity, starting with an automated baseline from satellite imagery and then enriching it through user-generated content. This closed-loop system employs active learning, progressively enhancing accuracy and relevance of water scores.

AquaSense. Water management is a critical issue in many countries, including Morocco. Severe droughts, poor water distribution, and recent natural disasters raise the urgent need for better solutions to manage water resources effectively. AquaSense's prototype (see Figure 2) offers a smart way to handle water resources by predicting future water situations, visualizing key data, and engaging citizens and communities. This helps decision-makers plan better, save resources, and respond quickly to local water issues. AquaSense provides accurate forecasting of water parameters for informed management and answers water-related questions with detailed analysis using the latest data and news. It offers transparent data visualization through interactive charts, allowing users to view and upload data easily. The community & citizens' space features real-time news updates, a water levels map to locate and help regions in need of water, and a tool to easily report local water issues.

Figure 2: Screenshot of the prototype of Aquasense defining parameters, visualizing data and monitoring engagement.

AquaSense combines two distinct parts of AI: DL (LSTM) and Generative AI (RAG and AI Agents). AquaSense uses Multivariate and multistep LSTMs to accurately predict water parameters' levels for the coming years, and Retrieval-Augmented Generation and AI Agents to answer water-related queries with detailed analysis, using the latest data, news, predicted parameters, and documents from sources like UN, UNCCD, and EPA. AquaSense uses Tensorflow and Keras (LSTM model), Pandas and Numpy (data preparation & mgmt.), Langchain (LLM framework for RAG and AI Agents), Chroma (Vector DB), Nomic Embeddings (Open-Source embeddings), GPT3.5-TURBO (LLM model), Streamlit (Web app). Aquasense improves water management by helping stakeholders make informed decisions, enhancing resource allocation, and promoting sustainable practices. Through its innovative features, it bridges the gap between citizens and authorities, which fosters collaboration and reduces water crises over time. Also, AquaSense aligns with several UN Sustainable Development Goals (SDGs) such as SDG 6 (Clean Water and

Sanitation), SDG 13 (Climate Action), and SDG 11 (Sustainable Cities and Communities).

Water Consumption Tracker. This prototype is addressing the global problem of water optimization in the light of the already visible consequences of climate change. That is, the large amount of wasted water due to irresponsible water use by the households. The added value lies in the behavioral approach: the application is designed to make users more aware of their attitude toward water consumption, and to make water conservation a pleasure rather than a responsibility. Introducing a gamification approach as a new strategy should help make water conservation more appealing. It is based on an app that tracks real-time water usage, provides personalized recommendations, and motivates users over a gamification environment, fostering a community focused on sustainable water use.
The use of Machine Learning models such as Random Forest Regressor to find patterns between the households characteristics and their water usage behavior. We plan to add GenAI using LLM model as a chatbot to support our vision by providing custom tips to optimize water usage. The approach was fundamentally based on: (1) collecting data about the households using our application UI; (2) providing optimum water consumption level by the ML model based on the data collected; and (3) monitoring water usage through IoT sensors and the notification system of our App. The data collected is used to optimize the ML model performance. Our approach can potentially reduce household water waste by 20-50% by educating users about their consumption habits through notifications, ranking systems, and feedback mechanisms.

Figure 3: The pitch of one of the top 3 teams – Ghayt – presenting the Water Consumption Tracker at the AI stage of GITEX Africa.

Aquatic Biodiversity. The introduction of non-native species into marine ecosystems presents a significant threat to the fragile equilibrium of these vital environments. Invasive species, often aggressive, can outcompete native organisms, leading to disrupted food chains, altered habitats, and potentially irreversible ecological harm. From coastal areas to the open sea, the swift proliferation of invasive plants, animals, and microorganisms endangers the biodiversity, productivity, and resilience of our marine life. Addressing this escalating global issue requires immediate and decisive action. AI-powered early detection algorithms were prepared to constantly monitor for signs of invasive species, triggering immediate alerts to enable rapid response. Based on species-specific data, the system can precisely deploy the most effective eradication methods, from underwater drones to selective biocides. As invasive species evolve, the AI-driven platform continuously adapts strategies, ensuring that the interventions remain effective and environmentally responsible.

YAZ. High unemployment rates in North Africa often translate into many individuals employed in low-wage jobs, particularly youth from low-income households. Severe water scarcity leading is decreasing exports and rising prices of vegetables and fruits. Challenges meeting the needs of Morocco's population while being a major exporter of produce to global markets. This AI-based agricultural solution is based on Smart Hydroponic Towers designed to efficiently grow crops vertically indoors and outdoors, offering optimal use of available spaces. The adoption of hydroponics in Africa has the potential to create millions of new jobs in the coming years. Integrated with GPT architecture, the technology allows real-time monitoring, pest detection, and yield estimation. YAZ hydroponics are a shift towards a resilient and sustainable Moroccan agriculture.

The tools and technologies presented in this paper that are open source, are available at IRCAI's SDG Observatory GitHub repository (github.com/IRCAI-SDGobservatory).

3 From concept to prototype in a month

AI in Africa in collaboration with IRCAI conducted a gathering of minds which culminated in a 1-day summit around technologies and shifts of the future, hosted by GITEX in the AI Everything section of the GITEX Africa 2024. Between 26th April and 31st May, 55 PhD and MSc students from 11 research institutions took part of a complete program including expert sessions kicked-off at the AI movement, UNESCO's new center for AI in Africa, and engaging experts in water-related topics such as Matjaž Mikoš, UNESCO chair for landslide risk reduction, droughts and floods, discussing our recent research on news mining for extreme weather events [5, 6]; Gerald Corzo Perez, senior researcher at the UN Water Education IHE Delft, discussing our ongoing research on Water, AI and Twitter [7]; and Ignacio Casals, R&D Manager in Aguas de Alicante Spain, providing a industrial perspective on the use of AI to tackle the challenges of wastewater management [8].
The students were followed across 8 stages including: conceptualization; data collection, analysis and visualization; methodology and implementation, prototype building and pitch (see Figure 3). In order to maximize the impact of the programme, the content from the abovementioned opportunities will be organized across UNESCO's most related to the five areas: (1) capacity building; (2) developing

supportive policy; (3) effective, inclusive and equitable access to quality Education; (4) nurturing and creating sustainability models for Water Sustainability; and fostering and facilitating international cooperation.

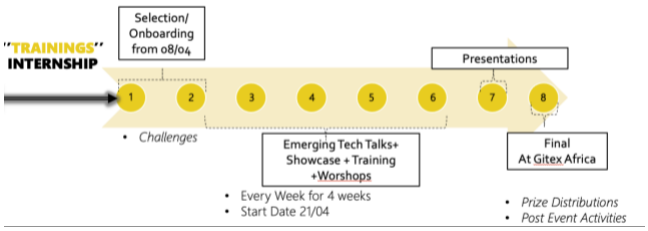


Figure 4: The phases of the training curriculum across 5 weeks.

The training curriculum included weekly seminars open to public, training workshop for participants, showcases and mentoring sessions (see Figure 4). The discussions forming the bae concepts of the participant projects were held in the light of IRCAI's research and research achievements (see Figure 5), aiming at building research collaboration bridges.

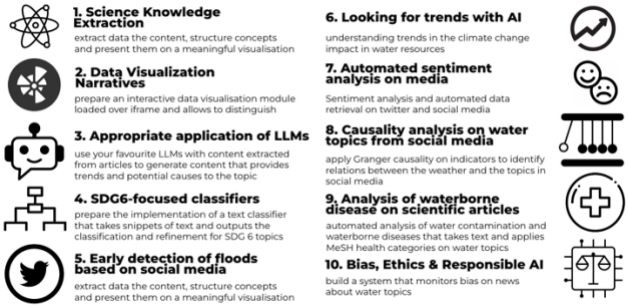


Figure 5: Selected topics from IRCAI's research to motivate challengers in AI and Water research

The data and methods generated by the participants programme can be used by companies, government and research institutions to aid in the resolution of problems related to Water Sustainability, by identifying trends and major areas of discussion, and to explore successful scenarios through similar challenges and cases. IRCAI's SDG 6 Observatory [10] is being built to properly address the challenges of decision makers, using AI. It is benefitting: (i) national governments providing access to a variety of perspectives (including trend and comparative) on a data driven dashboard with information on Water Sustainability trends for decision-making; access to local (e.g. country-level) progress on SDG 6; (ii) educational institutions, offering access to information on current trends on Water Sustainability research and development; (iii) research institutions, sourcing open data over interactive visualisation and research; (iv) the NGO community, easing access to information directly linked to community priorities including citizen science activities; and (v) general population, empowering water education for all.

4 Conclusions and further work

The capacity building to enhance opportunities can benefit from the engagement of the Youth in AI-driven challenges that start in research problems deriving from issues to address in their communities. Problems they know well and data that they often have privilege access to, with promising impact that can ensure the sustainability of the innovation offered. The initiative served us also to collaboratively discuss sustainable solutions that help large scale recovery and define a better and more hopeful inclusive Africa. The winning outcomes of this challenge will integrate a vibrant worldwide Community of researchers and entrepreneurs focusing on AI and SDGs, starting with SDG 6, and supported by initiatives such as IRCAI's Top 100 or the SDG Observatory. Ethical considerations are being addressed in the context of the EC project AI4GOV.

ACKNOWLEDGMENTS

This research was partially funded by the European Commission's Horizon research and innovation program under grant agreement 820985 (NAIADES) and 101120237 (ELIAS).

REFERENCES

[1] Blazhevska V.(2020). United Nations launches framework to speed up progress on water and sanitation goal, United Nations Sustainable Development.

[2] Casale G., Cordeiro Ortigara A.R. (2019) Water in the 2030 Agenda for Sustainable Development: How can Europe act? Water Europe, Brussels. (ISBN 978-90-8277064-3) 36p. https://unesdoc.unesco.org/ark:/48223/pf0000372496

[3] International Water Association and Xylem Inc (2019). Digital Water: Industry leaders chart the transformation journey. [Online] https://iwa-network.org/wp-content/uploads/2015/12/IWA_2019_Digital_Water_Report.pdf

[4] IRCAI Committee Chair on AI and Water Resource Management [online] ircai.org/project/ai-and-water-resources-management/

[5] Mikoš M., Bezak N., Pita Costa J., Nassri M. B., Jermol M., Grobelnik M. (2022). Natural-hazard-related web observatories as a sustainable development tool. In Progress in Landslide Research and Technology, Vol. 1, No. 1, Springer (in print).

[6] Pita Costa J., Rei L., Bezak N., Mikoš M., Massri M.B., Novalija I. and Leban, G. (2024) Towards improved knowledge about water-related extremes based on news media information captured using artificial intelligence. *International Journal of Disaster Risk Reduction*, *100*, p.104172.

[7] Perez, G., Pita Costa J., Novalija I., Rei L., Senožetnik M., Casals del Busto I. C. (2024). Integrating Social Media, News and Machine Learning for Enhanced Hydrological Event Detection and Management. In *15th International Conference on Hydroinformatics* (p. 278)..

[8] Pita Costa J., Massri M. B., Novalija I., Casals del Busto I., et al. (2021). Observing Water-Related Events for Evidence-Based Decision-Making. In: Slovenian Data Mining and Data Warehouses conference (SiKDD2021)

[9] Pita Costa J. (2022). Water Intelligence to Support Decision Making, Operation Management and Water Education - the NAIADES Report. IRCAI Library. [online] https://ircai.org/project/ircais-project-report-on-naiades/

[10] Pita Costa J., Zaouini M., Crnko M., Polzer M., Corzo Perez G., Mikoš M., Orlic D. and Jermol M. (2024) Challenging Water Sustainability in Africa Through AI, Proceedings of the HHAI 2024 workshop on AI in Africa and SDGs.

[11] UN-Sustainable Development [online] The IRCAI Water Observatory - AI in the service of SDG 6 [online] https://sdgs.un.org/partnerships/ircai-water-observatory-ai-service-sdg-6

[12] UN-Water Work Programme 2024-2025 [online] https://www.unwater.org/publications/un-water-work-programme-2024-2025

Predicting poverty using regression

Luka Urbanč
Jožef Štefan Institute
Ljubljana, Slovenija
urbancluka3@gmail.com

Marko Grobelnik
Jožef Stefan Institute
Ljubljana, Slovenija
marko.grobelnik@ijs.si

Joao Pita Costa
IRCAI, Quintelligence
Ljubljana, Slovenija
joao.pitacosta@quintelligence.com

Luis Rei
Jožef Stefan Institute
Ljubljana, Slovenija
luis.rei@ijs.si

Abstract

Poverty reduction is the first Sustainable Development Goal set by the United Nations to be achieved by 2030, but current data indicates that the progress is insufficient. The diverse factors influencing poverty across different nations pose a challenge in developing effective predictive models. This paper evaluates the use of various regression models to predict poverty rates using a comprehensive dataset of 111 variables from sources such as the UN and the World Bank. The data, spanning multiple domains like political stability, education, and economic conditions, was preprocessed and transformed to create auxiliary features and interactions. Among the models, Ridge regression yielded the best results, achieving a Root Mean Square Error (RMSE) of 3.6, indicating high predictive accuracy on a global scale. This study highlights the importance of addressing multicollinearity and incorporating a wide range of features to improve the generalizability of poverty prediction models. Future research should explore more complex methods, such as neural networks, and refine model hyperparameters for enhanced performance.

Keywords

poverty, linear regression, lasso regression, ridge regression, elastic net regression, sustainable development goals

1 Introduction

The need to eradicate poverty has been a long standing issue, which was globally recognized numerous times, most importantly in the United Nations (UN) Sustainable Development Goals (SDGs), being given the number one spot of SDG1: "End poverty in all its forms everywhere", which should be achieved by 2030. The latest UN report on the progress made in achieving SDG1 indicates Poverty has returned to pre-pandemic levels in middle- and high-income countries, with poverty in low income countries still a fraction above those reported in 2019. While the trends seem to be going in the right direction, the UN warns that the current pace of improvement is insufficient to reach the agreed goals before 2030. This raises the question of what impacts poverty rates the most and how countries can most effectively reduce poverty levels.

To fully understand and address the issue of poverty, one must navigate several definitions, which can often lead to confusion. The baseline definition used in this paper is the poverty line as is defined by each country individually, recognizing that different countries have different measures of, e.g., what life conditions and how much income makes an individual reach a "poor" status, as well as how we can normalise this to better compare these relative indicators between countries. We are still missing a clear theory in poverty research, despite the issue existing for a number of decades [2]. With that being said, some authors have already explored the causes of poverty. For instance, corruption, political instability, ineffective local governance, government polices, gender inequality and short-term wage replacement policies, such as maternity leave benefits and sickness pay, impact relative poverty [6, 7]. When assessing what people believe causes poverty some geographical differences emerge. For example, the United States are mostly of the thought that an individuals traits are responsible for poverty, while countries in Europe have a blend of individualistic, fatalistic and structural beliefs such as lack of will, bad luck and social injustice respectively [4].

Machine learning (ML) has also been used in academic research to identify trends and analyze data in most fields, including poverty research. Although a number of papers have already been published on the use of ML to predict poverty [1, 10, 12, 5, 3, 8] (for more see [11]) including combining satellite images and neural networks to help predict poverty in five African countries [5], most take a limited number of variables. Usmanova's literature review found 22 papers published between 2016 and March 2022, with a total of 57 AI methods applied, the most popular being random forest, used in more than half of all papers reviewed. It also found most papers focus only on African and South Asian countries, a finding consistent with our own [11].

In this paper we focus on the following research questions: (i) can regression be useful to identify the most influential features, from a large amount of global indicators; and (ii) can direct and indirect causality relations be identified that signal new indicators relevant to the Poverty-related issues?

2 Data

To address the research questions, we utilized 111 primary variables from sources such as the UN and the World Bank, aggregated through the Our World in Data portal. These variables span diverse domains, including political stability, policies, education, healthcare, economic conditions, and inequality. We prioritized features that prior research has identified as significant, while also incorporating some factors that are less intuitively linked to poverty. The dataset was then used to train various models aimed at predicting poverty rates across countries. This task is particularly challenging because countries respond differently to the same variables. For instance, GDP growth tends to have a more significant impact on poverty reduction in developing nations compared to developed ones. Additionally, many variables

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.20

are strongly correlated, making it difficult for linear regression models to capture their relationships accurately.

As previously mentioned, most of the data used in this paper was sourced from *ourworldindata.com* (OWiD), with some additional data coming from *fao.org*—including variables such as foreign direct investment inflows and outflows, and the added value of agriculture, among others. Data on the transatlantic slave trade and colonial rule was obtained from *www.slavevoyages.org*. All datasets were preprocessed before being merged, following a series of steps.

The first preprocessing step involved light modifications, such as removing irrelevant columns, renaming columns, and excluding data from before 1987 and after 2023 due to gaps and incomplete data. Despite increased reporting in recent years, many countries still omit certain indicators, complicating model training. To address this, missing features with more than n data points for a given country were interpolated, with the edges filled using backward fill (bfill) and forward fill (ffill). Those with less than n data points used the mean of the country's income group for the given year as a filler value. The number n was intuitively chosen to be five and the methods bfill and ffill were chosen to prevent the use of unrealistic data. The World Bank classifies countries into income groups: low (less than 1,045 USD), lower-middle (1,046 USD to 4,095 USD), upper-middle (4,096 USD to 12,695 USD), and high income (12,696 USD or more). However, it is important to note that the data generated using the aforementioned methods somewhat reduces overall robustness.

The next step involved generating auxiliary columns, specifically lagged columns and changes in value for relevant parameters. For instance, the row corresponding to Niger in 2013 would also include the GDP per capita for 2012, 2011, and earlier years, in addition to the value for 2013. This approach reflects the fact that poverty trends often manifest in response to changes over time, rather than immediately. The default number of years for lagged data was set to five. Similarly, we incorporated changes in value over the same five-year period to capture more explicit data on unusual events, such as the onset of wars or significant political changes.

Next each primary parameter was also used as an argument for a number of mathematical functions in an effort to see if any correlations aren't linear but perhaps squared, cubed or another elementary function. The functions used were: x^2, x^3, $\ln x$, $\sin x$, $\cos x$, $\tan x$, $\arcsin x$, $\arccos x$, $\arctan x$ to try and capture any elementary nonlinear dependence within the model.

The last step was to create all possible products with the available primary parameters, as creating all possible products with all auxiliary parameters included would have been computationally inefficient. After all these steps were made, the individual columns were fused together. This method of preprocessing increases the possible variables included, making the model even more general and retaining as many rows of data as possible.

The function responsible for preprocessing, generating and merging the data has a few parameters: *basic_parameters_only*, *combinations* and *math*. *basic_parameters_only* determines, if the model will only contain data obtained from various online databases, or if the model should include generated data: the change in value and values for previous years. *combinations* determines, if the model should create all possible combinations with the primary parameters and *math* determines if mathematical columns are included in an attempt to gain a deeper insight into the features' relationships. The parameters are marked with B, C and M. For instance, B+M would mean the file contains the basic parameters in addition to the mathematically derived columns.

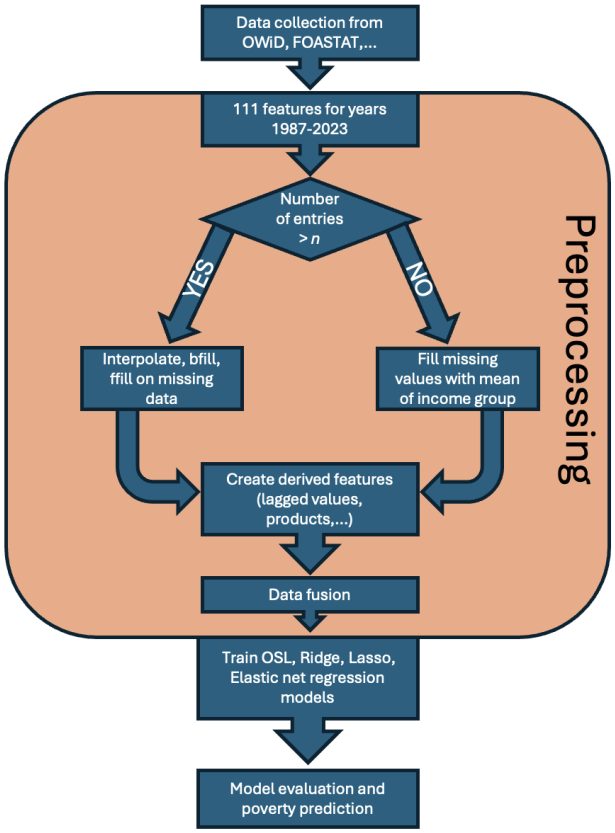


Figure 1: Scheme of adopted methodology

3 Methodology

In order to predict worldwide poverty levels, we have used different linear regression models and compared their accuracies. With this we aimed to ease the interpretability of the models, which is harder to obtain with more complex methods such a neural networks. To perform the research work that is the base of this paper, we have selected ordinary linear regression, lasso regression, ridge regression and elastic net regression as the models to compare. OLS regression struggles with multicollinearity, where predictor variables are highly correlated, leading to unstable estimates of the coefficients. Ridge regression addresses this by adding an L2 regularization term, which penalizes large coefficients and helps to stabilize the estimates in the presence of multicollinearity. By shrinking the coefficients, ridge regression reduces the sensitivity of the model to colinear predictors, ensuring more reliable and generalizable results. Unlike lasso, ridge retains all predictors, making it particularly useful when multicollinearity is a key concern but feature selection is not the goal. We use the implementation of these linear regression algorithms in scikit-learn [9].

The datasets were split into training and test sets using the sklearn function `train_test_split`, with 80% for training and 20% for testing. The training set was used to train four regression variants (`LinearRegression`, `Lasso`, `Ridge`, `ElasticNet`), all with a random state seed of 42. while the test set was used to

determine the mean squared error (MSE) and R^2 value using the functions `mean_squared_error` and `r2_score` from [9], both common metrics used to assess models accuracy. All models except OLS regression also had the data standardized before training. The hyperparameter α for the models was sensibly chosen as 0,1. The results, seen in Table 1 are color coded: red for poor performance, yellow for intermediate, and green for the best. The variation in the number of rows is due to the exclusion of rows with insufficient yearly data, which were dropped when calculating differences from previous years.

After identifying the most successful model, we proceeded to compare its performance between high-income and low-income countries. This comparison aimed to assess how the accuracy and frequency of reported data influence the model's performance. These two income groups were chosen because low-income countries typically report less data with lower accuracy, while high-income countries provide more precise reports. We selected all high- and low-income countries from the dataset that were not used during the model's training. From the 20% of data reserved for evaluation, 444 rows (30%) belonged to high-income countries, and 368 rows (24%) belonged to low-income countries.

We used the trained model to predict poverty levels for these groups and evaluated its performance using the MSE metric to analyze differences between income groups. Additionally, we calculated the maximum error to determine if the average performance was skewed by outliers. A similar evaluation was conducted on the data from Slovenia and Somalia, which were part of the split. Slovenia had 8 rows of data, and Somalia had 6, allowing us to explore how missing data impacts the model's performance, as Somalia had significantly fewer data points overall.

4 Main Results

The file configuration plays a critical role in the model's performance. The results show that C+M, C, and B+C are the best configurations. The C+M file includes all basic features, lagged values, changes in value, mathematical columns, and all possible combinations of basic parameters, totaling 8,236 parameters. Configuration C contains all basic features, combinations, and lagged and difference columns. Lastly, B+C includes only the basic parameters and their combinations. All top-performing models were trained on these datasets.

The results in Table 1 show considerable variation. Models trained with ordinary least squares regression performed poorly, with the best model reaching an RMSE just under 10.15 and an R^2 of 0.50. In contrast, lasso and elastic net regression achieved better results, with RMSEs around 7 and R^2 values close to 0.80. Ridge regression also struggled, except for configuration B+C, which provided the best results with an RMSE of 3.6 and an R^2 of 0.94. However, caution is advised when interpreting models using configuration C+M or C, due to the high number of features relative to the dataset size, which could affect their real-world reliability.

The model weights reveal that only products are present among the top ten most important factors. These products include data on population, population density, agriculture, equality, healthcare, and education. The largest weights show the biggest differences, gradually decreasing in magnitude. The top ten weights range from just over 10 to 7, with the highest weights involving combinations such as population and population density, meadows and pastures with the global peace index, and

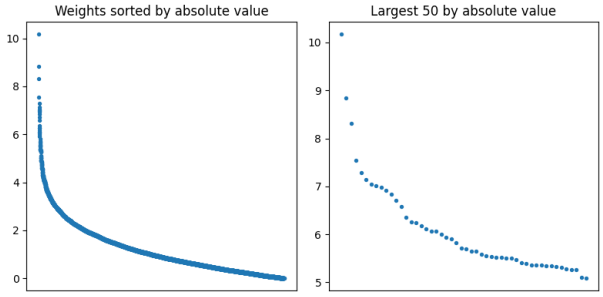


Figure 2: Visual representation of model weights

population with urban and rural population share. Other notable combinations include secondary school completion with women's civil liberties, internet usage with sanitation access, and military spending with wealth distribution. The weights also reflect factors like infant mortality, years colonized, and agricultural employment. Figure 2 further illustrates the decline in the absolute value of these weights.

The model performed better on high-income countries, with an MSE of 6.60, significantly below the overall MSE. In contrast, the MSE for low-income countries was 20.68. The maximum error was also lower for high-income countries (22.1) compared to low-income ones (34.4).

The difference in the model's performance on Slovenia and Somalia was notable. For Slovenia, the MSE was 0.78 with a maximum error of 1.54, far below the overall metrics. Somalia, however, had a much higher MSE of 95.7 and a maximum error of 18.7, likely due to less reliable and extreme poverty data, which skews the model's performance on extreme cases.

5 Discussion

Firstly, the fact that ordinary least squares linear regression couldn't produce an accurate model confirms the fact that the parameters are indeed correlated. This is probably also the reason why the ridge regression model performed the best: ridge regression is used to address the issue of multicollinearity and the features included are mostly strongly correlated, as stated in the introduction. Furthermore, the correlation between parameters is obviously drastically increased by generating all possible products of basic parameters.

Secondly, the impact of mathematical columns needs to be considered. Of the first four models, two have mathematical columns and two don't. Of the eight models generated, three of them perform worse if mathematical data is present, while 5 performed better with mathematical data included. This might indicate some deeper connection, which would be interesting to try and understand. Furthermore, lasso regression handles mathematical columns much better compared to the other models used due to its ability to exclude features.

The impact of product combinations of basic features stands out, with all better-performing models having the *combinations* parameter set to `True`, suggesting deeper relationships between variables. Exploring these connections further, perhaps by training a neural network on the basic parameters and comparing it to linear regression models, could be insightful. If the neural network performs better, further investigation into these correlations would be needed.

Structure	Linear MSE	Linear R^2	Lasso MSE	Lasso R^2	Ridge MSE	Ridge R^2	Elastic net MSE	Elastic net R^2	Shape of X
M	203	0.031	74	0.65	-	-	-	-	(7653, 2131)
None	-	-	109	0.48	163	0.22	108	0.49	(7653, 1221)
C+M	198	0.054	45	0.78	-	-	40	0.81	(7653, 8236)
C	-	-	50	0.76	-	-	45	0.79	(7653, 7326)
B	103	0.50	110	0.47	103	0.50	111	0.46	(7661, 111)
B+C	-	-	48	0.77	13.3	0.94	43	0.79	(7661, 6216)

Table 1: MSE and R-squared values for different regression models and dataset configurations. The presence of B, C or M signals the presence of basic basic parameters only (B), combinations (C) and mathematically (M) derived columns in the dataset. A dash is used to label non-converging models with a negative R-squared value.

The dataset used spans from 1987 to 2023, which is relatively short, given that poverty often has deep historical roots. Although data becomes scarcer in earlier years, those points could still be crucial for improving model accuracy. Moreover, most hyper parameters in this paper were chosen sensibly due to time and computational constraints. Different values for the number of lagged years, years of differences, hyperparameters in the training of models and the minimum number of data points required to interpolate missing data could all lead to interesting discoveries and improvements of the generated models. Our result here shows it is possible to achieve this degree of accuracy, but it doesn't limit what the best model could be. The elastic net, especially, should benefit from such a tuning.

As stated in [11], the recent literature mostly uses the random forest model and, in fact, ordinary linear regression wasn't even in the top ten most common methods. An interesting thing to explore would also be the performance of random forest using the best configuration, B+C. The models may struggle to capture correlations between variables due to differing impacts across countries, as mentioned in the introduction. A potential solution is to split the countries into k groups and train separate models for each group. While this could improve predictions, it raises two challenges: how to split countries without bias and how to ensure enough data for training.

The weights in the model further emphasize the issue of multicollinearity among the parameters, with only product terms emerging as the most influential. However, this does not reveal the true importance of individual parameters, as they may enhance the impact of another factor within the product term. Additional research is needed to better determine the true significance of these parameters and gain a clearer understanding of what drives poverty rates up or down. It can be seen in Figure 2, the models weights occupy a wide range. It is clear that some features are more important, based on their weights and further work is being done to understand which features stand out and why.

The model also performed better in predicting poverty levels in high-income countries compared to low-income countries. This discrepancy can likely be attributed to the fact that high-income countries report more data with greater accuracy, allowing the model to identify underlying patterns more effectively. In contrast, much of the data for low-income countries had to be interpolated, which reduced variability between countries and negatively impacted the model's performance.

6 Conclusion

In this paper, we have shown that a general model exists, based on linear regression methodologies, which can predict poverty with a relatively high accuracy (RMSE of 3.6). This was achieved through testing of numerous linear regression models using open data, with the best model being created by using ridge linear regression trained on data which also included all possible combinations of the basic features included in the dataset. The basic parameters included consist of 111 different parameters describing countries across 36 years. Better models could possibly be generated using more complex methods such as neural nets or random forest, gaining in accuracy but compromising the explainability of the model. The models could also benefit from hyperparameter tuning during the whole process to improve results and find the optimal values. We will be addressing this in further research.

7 Acknowledgements

This research was partially funded by the Future of Life Institute under the project "An AI-driven Observatory Against Poverty", and the European Commission's projects under grant agreement 101135800 (RAIDO) and 101120237 (ELIAS).

References

[1] Gianni Betti, Antonella D'Agostino, and Laura Neri. 2002. Panel regression models for measuring multidimensional poverty dynamics. *Statistical methods and applications*, 11, 359–369.

[2] David Brady. 2019. Theories of the causes of poverty. *Annual Review of Sociology*, 45, 1, 155–175.

[3] Muse A.H. Hassan A.A. and Chesneau C. 2024. Machine learning study using 2020 sdhs data to determine poverty determinants in somalia. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 14, 1, 5956.

[4] Dariush Hayati and Ezatollah Karami. 2005. Typology of causes of poverty: the perception of iranian farmers. *Journal of Economic psychology*, 26, 6, 884–901.

[5] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353, 6301, 790–794.

[6] AH Ng, Abdul Ghani Farinda, Fock Kui Kan, Ai Ling Lim, and Teo Ming Ting. 2013. Poverty: its causes and solutions. *International Journal of Humanities and Social Sciences*, 7, 8, 2471–2479.

[7] Rense Nieuwenhuis, Teresa Munzi, Jörg Neugschwender, Heba Omar, and Flaviana Palmisano. 2019. Gender equality and poverty are intrinsically linked: A contribution to the continued monitoring of selected sustainable development goals. Tech. rep. LIS Working Paper Series.

[8] Shah O. and Tallam K. 2023. Novel machine learning approach for predicting poverty using temperature and remote sensing data in ethiopia. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5, 6, 2302.14835.

[9] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[10] Mubaraq Dele Sulaimon. 2020. Multidimensional poverty and its determinants: empirical evidence from nigeria.

[11] Aziza Usmanova, Ahmed Aziz, Dilshodjon Rakhmonov, and Walid Osamy. 2022. Utilities of artificial intelligence in poverty prediction: a review. *Sustainability*, 14, 21, 14238.

[12] Huang Zixi. 2021. Poverty prediction through machine learning. In *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*. IEEE, 314–324.

Fact Manipulation in News: LLM-Driven Synthesis and Evaluation of Fake News Annotation

Luka Golob
lukag26@gmail.com
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Abdul Sittar
abdul.sittar@ijs.si
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Abstract

Advancements in artificial intelligence and increased internet accessibility have made it simpler to create and disseminate fake news with customized content. However, they also improved the ability to analyze and identify such misinformation. To effectively train high-performance models, we require high-quality, up-to-date training datasets. This article delves into the potential for generating fake news through factual modifications of articles. This is facilitated by prompt-based content generated by large language models (LLMs), which can identify and manipulate facts. We intend to outline our methodology, highlighting both the capabilities and limitations of this approach. Additionally, this effort has resulted in new quality synthetic data that can be incorporated into the standard FAK-ES dataset.

Keywords

fake news, synthetic data, fact extraction, fact verification, large language models

1 Introduction

Synthetic data refers to artificially generated data that is not obtained by direct measurement or observation of real-world events. Instead, it is created using algorithms and simulations. The primary purpose of synthetic data is to provide a realistic alternative to real data for various use cases, such as training machine learning models, testing systems, ensuring data privacy, and more.

We will generate synthetic data from news articles. By making sure, that the information in the news is changed we can safely call it fake news. In our article, fake news will denote articles that are *intentionally* and *verifiably* false [4]. Synthetic data enhances model training by providing additional examples to supplement scarce labeled datasets and allows for privacy-conscious testing without real content manipulation. It enables adaptability to evolving fake news tactics by simulating diverse scenarios from the newest data, thereby improving the robustness and resilience of detection algorithms [3].

Large language models (LLMs) made a huge difference in the world of news. Fake news is now much easier and cheaper to construct, but we also have additional methods to help us tackle its spread. Numerous articles appeared trying to partake in this effort. The following are the main scientific contributions of this paper:

(1) A methodology to create synthetic data for fake news using LLMs.
(2) We then use this methodology, to adapt the FA-KES dataset with 100 additional synthetic fake news [1].

In Section 2, we discuss work that is closely related to our task. Section 3 then outlines the methodology for generating synthetic fake news, culminating in Section 4, where we present the results and introduce some modifications to the methodology. Finally, in Chapter 5, challenges, capabilities, and potential improvements are considered.

2 Related Work

A wide range of approaches to generate fake synthetic news with LLM has been developed. In [8] authors generated huge amounts of fake news and categorized them into multiple categories. LLMs can generate fake news by altering the style to mimic credible sources or using sensationalism to influence perception. They can subtly manipulate content to be perceived as true, blend real and fabricated information to exploit cognitive biases, or create convincing fictional narratives.

In general, when making a dataset we want a diverse distribution of fake datasets. In our case, we will focus on one way of data change, which comes under the umbrella of *Content Manipulation*. Similar news manipulations can be seen in [7] where the authors use two main techniques. The first one extracts the summary from the original text, which preserves the main content, which is then changed to produce a fake article. The second one asks a question about the article and changes the content of its answer, to construct a new article. Our approach is in nature similar to the Question-Answer framework.

Many articles provide fake news detection models made using synthetic data. Most popular are deep neural networks such as BERT [1]. But there are other fact-based approaches for fake news labeling as in [3]. In [2] they used GPT4-turbo for prompt-driven fake news detection.

3 Methodology

The methodology is divided into four conceptual steps: Data collection, Characterization of facts, Fact extraction, and Fact manipulation as presented in Table 1.

3.1 Data Collection

The publicly available FA-KES dataset [5], focused on the Syrian war, addresses the deficiency of manually labeled datasets in this domain of news data. It comprises 804 articles sourced from various media outlets. We used 426 articles that were manually labeled as authentic news, but we could just as well use the other (fake) articles.

[1] https://github.com/golobluka/Fake-news-generation-from-FA-KES-dataset

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.13

1. Data collection	2. Characterization of facts
• Should have textual and statistical facts	1.Name of casualty 2.Gender or age group 3.Cause of death 4.Type 5.Actor 6.Place of death 7.Date of death
3. Fact Extraction	**4. Fact manipulation**
Name of casualty: Civilians **Gender or age group**: e.g., child, adult, senior **Cause of death:** shooting, shelling, weapons, etc. **Type:** military personnel **Actor:** rebels, forces **Place of death:** Airbase **Date of death:** April 7, 2017	**Name of casualty**: Manipulated fact **Gender or age group**: Manipulated fact **Cause of death:** Manipulated fact **Type:** Manipulated fact **Actor:** Manipulated fact **Place of death:** Manipulated fact **Date of death:** Manipulated fact

Figure 1: A methodology to generate synthetic data for fake news detection

3.2 Characterization of Facts

While making the FA-KES dataset, its authors created seven factual categories:

(1) `Name of casualty or group`,
(2) `Gender or age group`,
(3) `Cause of death`,
(4) `Type`,
(5) `Actor`,
(6) `Place of death`,
(7) `Date of death`.

It is crucial to note that all articles have a similar structure, describing war incidents. This allows us to establish a consistent framework of facts, such as actor and casualty details. We stick to those facts, but generate them differently, employing LLMs capabilities with faster and cheaper execution, albeit with a slight reduction in reliability.

3.3 Fact Extraction

We extract facts by constructing prompts for LLMs. First approach was a few-shot prompt, which gives some examples of output. Later we constructed an additional approach: Say we are extracting the fact `Place of death` with this second technique. We give a detailed description of what should be extracted and then LLM reads the article and performs the task solely on this basis. This description is usually longer and contains more context. The issues with fact extraction in general are:

- Some articles lack certain facts or merely imply them. LLMs can identify this, outputting responses such as "No information."
- Longer articles may contain multiple events, each with distinct data such as dates or casualties. This can be managed by creating separate tables for each event or consolidating all events into a single table with various facts.

3.4 Fact Manipulation and Synthetic News Generation

The objective is to modify relevant information without altering the writing style or topic of the article. For this transformation, we used a chain of thought prompt, which for a given fact: 1) changes the fact to another with a different meaning, 2) generates a new article based on the altered facts. By changing one fact at a time, quality is improved compared to altering multiple facts simultaneously, as one fact creates a clearer chain of instructions. LLMs such as `Llama3.1:8B` often struggle with precise changes in the article, such as modifying implicit references or incorporating new facts. Quality can be improved by carefully adjusting the prompt content.

LLMs are also exceptional in summarization and paraphrasing. Both are used simultaneously with changing the facts. The problem is that we aim to maintain the extracted facts when summarizing. But this is not crucial, as it usually has better results as article generation.

3.5 Fake News Annotation and Fact verification

After we have generated the fake articles, we can label that data as "fake" or "non-fake", based on comparison with extracted facts. We performed this labeling with various models and compared the performance of labeling,to get the best model. In this experiment we decided for `Llama3.1`. To do the labeling, we are performing fact verification [4]. The fact verification task in general is making a decision as to whether a claim is correct, based on the explicitly-available evidence, such as Wikipedia articles or research papers. We have the extracted fact, which will be compared to the article content. The question thus becomes: Do these facts appear in the given article? This approach emphasizes factual content rather than the overall sentiment of the article.

There are two primary types of prompts: 1) Direct prompts that present the article and a table of facts, asking if the facts relate to the article, 2) Structured prompts that inquire about the correspondence of one fact at a time with the article. The question is: Does this fact correspond to the content of the article? This method combines individual results into an aggregated score. Say the `Place of death` is characterized as `Idlib and Daraa provinces`. Then the question posed to LLM is of the form: `Read the article and understand its places of death. Do Idlib and Daraa provinces "really correspond" to places of death in the article?`

We are not as interested in labeling, as we are interested in the quality of produced synthetic fake news. For this purpose, we will also use fact verification in a slightly different way. We are asking the LLM: Were the factual changes in fake news really made, as they were supposed to? A similar method is used in the article [7].

4 Experimentation and Results

4.1 Experimental settings

We selected 426 articles labeled as authentic news from FA-KES dataset. Then facts were extracted and transformed, as described in the previous section. At first two basic approaches were used to randomly choose 70 news articles and transform them. Afterward, we used the labeling procedure to compare performance, resulting in the table 1. Based on the results we then composed the final algorithm, which would be manually evaluated.

4.2 Evaluation

For every experiment, we first manually checked a minimum 10 percent of random examples to get an overview of how well the LLM was able to do the job. It is quite useful to print text that represents the procedure of decision-making that LLM undertakes, when challenged with the task. It was even helpful to see LLMs generated thinking procedure, as this gives valuable insight, into what is going on "under the hood". We believe that manual fact-checking is the first and most crucial step in generating good prompts. Based on fallacies one can then adjust prompts content. To shed some light on this procedure we have made the following overview.

4.3 Fact Extraction Results

Name of casualty or group:	Members of Nusra Front
Gender or age group:	Adults (no specific age mentioned)
Cause of death:	Explosion at a mosque
Type:	Non-civilian (militants)
Actor:	Unknown (no group claimed responsibility, but supporters blamed ISIS)
Place of death:	Ariha, Idlib province, Syria
Date of death:	Not specified in the article

Figure 2: Example of fact extraction.

LLMs are capable of recognizing different topics and extracting words that correspond to this topic, and also noting if the fact is not mentioned. At first, we extracted short words as represented in Figure 2.

The issue begins with nuances. For example, in many articles the `Actor` is only suspected but not known. In some cases, actor and causality are not precisely distinguished. This usually leaves LLM to some kind of arbitrariness. For this purpose, We also added a longer description that better captures the nuanced subtleties related to facts. This can also be captured in Table 1. There we see the results for short (normal) or detailed extracted facts. The recall is far worse in the case of short prompts. This likely means that there is an abundance of false negatives, which result from the fact, that labeling does not manage to match true articles and their corresponding short facts.

The shorter extracted facts are often not comprehensive. For example, under the label `Type` (which classifies civilian or non-civilian) it writes only civilians, even though, contextual understanding also includes some non-civilian casualties.

Overall the most important insight remains: fact extraction has better quality than article generation.

4.4 Quality and coherence of synthetically generated fake news

The LLM can detect (for example) the `Actor` of some attack in the news, and then it is mostly able to change every occurrence of this `Actor` with another `Actor`. But if we would like to preserve all the coherence of the article much more would need to be done.

News usually contains background information, that provides context for the accident. Our algorithms failed to properly adjust this context, leaving it unchanged in most cases. Our fake news fails to preserve enough coherence to be trusted by a skeptical reader, who tries to connect background material to the event in the article.

Table 1: Comparison of fake synthetic data.

Type of data	Number of facts manipulated	Precision	Recall	F1	Accuracy
Summarization	2/7	0.74	0.63	0.68	0.71
Detailed facts	2/7	0.70	0.80	0.75	0.73

Generating false text, while maintaining coherency, is challenging for LLM. In this task, we have changed one fact: for example, the `Place of death` may be changed to another city or neighborhood. Then this fact must be changed in the article while maintaining other factual information. Here are the main issues:

- In the beginning some facts did not get changed, or the facts were altogether just removed from the article. We managed to reduce this error by adjusting the prompt. It is difficult to adjust all occurrences of the fact, especially if it is only implied and not explicitly stated. We managed to minimize this problem, by a method yet to be shown in section 4.5.
- What remains is the problem of a wider context, Suppose we change the town of the incident, then we must change the name of the neighborhood accordingly. LLM usually fails in this, leaving our article inconsistent, which is a widespread problem.
- LLM does not want to output the content because of harmful content or does not want to produce articles that could be used with illegal intent. This was quite a common problem, which is also reasonable, based on the violent content of articles and the possible abuse of LLM-generated content. The best thing to prevent this error is to use uncensored LLM. In other cases, one can adjust the prompts by removing suspicious words like "fake news".
- The Generated article was shorter, skipping the original text which was not linked to extracted facts. This problem was reduced but still exists in long articles.
- If the fact is not present in the article, then it is hard for LLM to incorporate a new fictitious fact into the text. Mainly it just adds the information in separate sentences.
- When we change facts, traces of the old facts still persist. This is especially common in complicated articles with diverse structures.
- Sometimes the change does not bring about any additional meaning. For example, LLM might change previously unknown casualties and designate them as civilians. They were implied to be civilians all along, and this makes only a minor change and is not really fake.

4.5 Fact verification with LLMs

Remember that in this task, the prompt asks: Does this fact "really correspond" to the content of the article? Performance largely depends on how the program takes the word "really correspond". Words have many nuances: different words can have different meanings, which can complicate labeling. To simplify: we can be stricter, in the sense that words must be the same in the literal sense, or we can count on the similarity of meaning [6]. Based on our goal of creating fake news it is best to focus on meaning and not concrete words.

Here are some common problems:

- Sometimes the fact is changed, but LLM skeptically assumes, that those two names refer to the same group.
- In longer articles, where there are many events, the names get changed only in some events (usually at the beginning of the article). In this case, the LLM can make unwanted predictions, labeling the fact as true rather than false.

Manual checking shows that labeling is more accurate than generation of fake news. This leads us to use labeling as a means to improve article generation.

Table 1 was used to compare different ways to generate fake news. It shows two of the best datasets, which contain true articles and their false twins, generated in two ways:

(1) Fake news generated by "standard" fact extraction and with additional summarization.
(2) Fake news generated by "detailed" fact extraction and with an additional paraphrasing of the article.

In this experiment, instead of merely categorizing the articles as true or false, the results shown in Table 1 reflect how well the generation process aligns with fact verification.

Low precision in the row with Detailed facts led us to detect articles that were not changed. We implemented a strategy where labeling was applied after generating the fake articles to assess the quality of the generation. LLMs often provide incomplete responses and struggle to correct them directly. By introducing an additional verification step, we were able to enhance the overall accuracy of the results.

4.6 Final Dataset Description

In the end, we constructed 100 fake-news based on a prior experiment, which can be found on GitHub [2]. In every article we randomly chose three facts and changed them. Afterward, we carefully went through 10 examples, which are also present on Git Hub, while here we present only the main points:

- Fact verification improved quality by making sure, that the synthetic fake article really incorporated new information. More than 90% new facts really got incorporated in the article. Sometimes new information is only added as additional text(and does not seriously change the main topic).
- Fact is not always incorporated in all places where it is referenced, which leads to inconsistencies. The new article is then a blend of old and new information.
- There are problems with ¨detailed" prompts. Containing more information results in contradictions as we change only one fact at a time.

5 Conclusion

In this article, we focused on exploring the potential of LLMs in fact extraction and generation of fake news. Our motivation was primarily to understand how accurate are LLMs in fact extraction and how reliably LLMs generate synthetic news by altering facts. As a result of our experiment, we have generated 100 synthetic news by randomly transforming there out of seven facts and have performed a manual evaluation, to observe the quality of the generated news dataset.

5.1 Problems, Capabilities and Possible Improvements

- In this stage, LLMs like *Lamma3.1:8B* are not able to coherently change certain facts of news articles. Changing facts can distort the article content, which appears to be extremely hard to manage. This normally does not happen for manageable data as dates (changing the time of some event), but for much more involved actors of the attack in the article. Even so, the synthetic fake news provides valuable information.
- We did not use the model, which has additional information about the news content. Providing additional context would likely have a beneficial effect on all the processes.
- In our case facts were largely dependent on each other. For example `Gender or age group` is an extraction of `Name of casualty or group`. We think it is best if such dependencies are removed because they bring to inconsistencies when changing facts. An additional solution would also be to change `Gander or age group` whenever `Name of casualty or group` is changed.
- Fact extraction is close to human-like quality. The issue is, that besides manual checking, it is hard to find a good measure of the quality of extracted facts.
- Detection of changed facts is in quality similar to extraction of facts (this is not surprising, since they are based on the same skill). Because of the diversity of meanings in language, it is hard to specify the exact reasoning procedure of LLMs and many mistakes come from this kind of miscommunication.

6 Acknowledgments

This work was supported by the European Union through AI4Gov (101094905) and TWON (101095095) EU HE projects and the Slovenian National grant (CRP V2-2272).

References

[1] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Francesco David Nota. 2023. Content-based fake news detection with machine and deep learning: a systematic review. *Neurocomputing*, 530, 91–103. DOI: https://doi.org/10.1016/j.neucom.2023.02.005.

[2] Fredrik Jurgell and Theodor Borgman. 2024. Fake news detection : using a large language model for accessible solutions. (2024).

[3] Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024. Detect, investigate, judge and determine: a novel llm-based framework for few-shot fake news detection. (2024). https://arxiv.org/abs/2407.08952 arXiv: 2407.08952 [cs.CL].

[4] Taichi Murayama. 2021. Dataset of fake news detection and fact verification: a survey. (2021). https://arxiv.org/abs/2111.03299 arXiv: 2111.03299 [cs.LG].

[5] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. 2019. Fa-kes: a fake news dataset around the syrian war. In *Proceedings of the international AAAI conference on web and social media*. Vol. 13, 573–582.

[6] Abdul Sittar, Dunja Mladenic, and Tomaž Erjavec. 2020. A dataset for information spreading over the news. In *Proceedings of the 23th International Multiconference Information Society SiKDD*. Vol. 100, 5–8.

[7] Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: a study of real-world detection challenges. (2024). https://arxiv.org/abs/2403.18249 arXiv: 2403.18249 [cs.CL].

[8] Lionel Z. Wang, Yiming Ma, Renfei Gao, Beichen Guo, Zhuoran Li, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Ng. 2024. Megafake: a theory-driven dataset of fake news generated by large language models. (2024). https://arxiv.org/abs/2408.11871 arXiv: 2408.11871 [cs.CL].

[2] https://github.com/golobluka/Fake-news-generation-from-FA-KES-dataset

Borrowing Words: Transfer Learning for Reported Speech Detection in Slovenian News Texts

Zoran Fijavž
Jožef Stefan Postgraduate International School
Peace Institute
Slovenia, Ljubljana
zoran.fijavz@mirovni-institut.si

Abstract

This paper describes the development of a reported speech classifier for Slovenian news texts using transfer learning. Due to a lack of Slovenian training data, multilingual models were trained on English and German reported speech datasets, reaching an F-score of 66.8 on a small manually annotated Slovenian news dataset and a manual error analysis was performed. While the developed model captures many aspects of reported speech, further refinement and annotated data would be needed to reliably predict less frequent instances, such as indirect speech and nominalizations.

Keywords

reported speech, natural language processing, transfer learning, news analysis

1 Introduction

Reported speech, ubiquitous in literary and news texts, has clear lexical and syntactic patterns which may be reliably modeled via natural language processing (NLP) and may be useful for downstream tasks by drawing a distinction between source and background information. The paper applies transfer learning to extend reported speech classification to Slovenian news texts and provides a provisional classification model. A manual error analysis reveals the model's strengths and weaknesses, highlighting possible steps for further improvements.

2 Related Work

2.1 Role of Reported Speech

Reported speech is common in news texts, generally expressed as direct or indirect speech, with the former repeating the original utterance verbatim and the latter embedding it in a that-clause [18] (e.g., *Jimmy said: "Another systematic review would be great!"* and *Jimmy said that another systematic review would be great.*). More complex forms include mixed speech (*City officials rebuffed the accusations as "groundless and blatantly false".*) and reportative nominalizations with an analogous function as reported speech (*The speaker particularly emphasized the pressures on the media*) [7]. Around 50% of sentences in newspaper corpora may be attributed to a source in the text, predominantly through direct and indirect speech [17]. Verbs cue 96% of reported speech, followed by prepositional phrases (3%) [13]. Reported speech lends objectivity to statements [9], summarizes source statements [16], and is used in discourse analysis and communication studies to explore speaker representation by gender [1], institutional affiliations [8], and topic stances [15], or to distinguish between journalists' and sources' voices [11].

2.2 Existing Datasets and Modelling Approaches

Datasets with reported speech annotations mostly cotain literary or news texts. Key corpora include RiQuA [12], SLäNDa 2.0 [19], Redewiedergabe [3], QUAC [14], PolNeAR [10], Quotebank [21], and STOP [22]. RiQuA and Redewiedergabe are the largest annotated corpora, covering English and German 19th century texts. QUAC contains 212 annotated articles from the Portuguese newspaper Público, while Quotebank spans 162 million news articles with automatic annotations. PolNeAR, consisting of 1,028 news articles, includes attribution annotations, which include and exceed the definition of reported speech. A summary of the datasets is provided in Table 1.

The corpora differ in annotation complexity and size. They are mostly monolingual, warranting the used cross-lingual transfer learning for low-resource languages by employing multilingual models such as mBERT [6] and XLM-R [4]. Narrower multilingual models, such as CroSloEngual BERT, often outperform broader ones [20]. Reported speech modeling may be operationalized as speaker or quotation detection tasks [23, 17]. Simplifying the task to sentence-level classification is warranted by the fact news (unlike literary texts) rarely mix statements by sources and authors in the same sentence and can improve classification reliability at the expense of detailed aspects of reported speech [17] and simplify the annotation structure. Missing fine-grained outputs, such as speakers and boundaries of reported and reporting clauses, may thus be an acceptable trade-off for NLP-based content analysis in news texts. A systematic review of such approaches points to the limits resulting from a low number of features with no guarantee of reliable (joint) prediction, which preclude drawing rich conclusions expected from the method's manual counterpart [2].

3 Experimental Setting

3.1 Task Overview

We treated reported speech as a sentence-level classification task. Sentence splitters were applied to existing datasets, and binary labels were assigned by matching annotated spans with the split sentences. Reported speech sub-types were unified under a single label, joining the annotation schemes of individual datasets. A Slovenian dataset of 10 news texts was manually annotated at the sentence level. The datasets were split into training, evaluation, and test sets to train multilingual pretrained models. For CroSloEngual BERT, preprocessing also involved machine translating the German training data into English. The model outputs were binary labels indicating reported speech, used to calculate F-scores on the test data. A manual error analysis was performed on

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.21

Table 1: Summary of Datasets' Characteristics.

Corpus	Type	Annotations	Language	Sentence No.	Role	Positive Class
RiQua	fiction	direct and indirect speech, cues, speakers, addressees	English	38,610	72% train, 18% development, 10% test	48%
Redewiedergabe	fiction, news	direct, indirect, free indirect and reported speech, speaker, cues	German	24,033	76% train, 16% development, 9% test	33%
Quotebank (manual)	news	speaker, direct speech	English	9,071	test	30%
QUAC	news	speaker, direct speech	Portuguese	11,007	test	11%
PolNeAR	news	speaker, cues, attributions	English	34,153	test	59%
Slovenian parliamentary news	news	sentence-level binary labels	Slovenian	744	test	43%

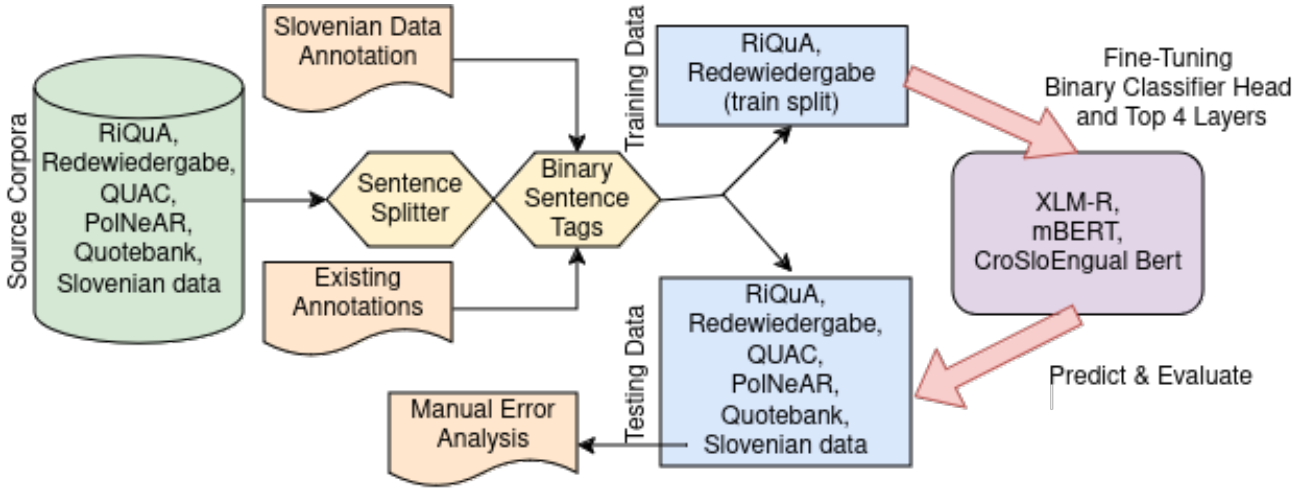


Figure 1: Flowchart of Data Preprocessing, Model Training and Evaluation Processes for Sentence-Level Reported Speech Classification.

the best model's outputs for Slovenian. Preprocessing, training, and evaluation steps are visualized in 1.

3.2 Training and Test Data

Our experiments were based on existing annotated reported speech datasets and a small Slovenian dataset. The training data included sections from RiQuA and Redewiedergabe, both large datasets with labels for direct and indirect speech. For CroSloEngual BERT training, the Redewiedergabe data was machine translated into English. Testing was conducted on the test sections of RiQuA, Redewiedergabe, the entire Portuguese corpus QUAC, and the manually annotated portion of the English Quotebank corpus. Additionally, we manually annotated 10 Slovenian news articles from RTV Slovenia. The datasets are summarized in Table 1.

The Slovenian dataset comprised 10 parliamentary news texts, covering various reporting strategies. Retrieved articles were split into sentences and annotated. Sentences were considered reported speech if they included direct or indirect speech cued by a reporting clause or prepositional phrase. We excluded nominalizations and phrasal quotes (e.g., *They emphasized the pressures on the media and the "illegal non-funding of the Press Agency."*) as well as implied quotes (e.g., *There will be more than 300,000 recipients, he emphasized. 169 million euros will have to be paid out.*).

3.3 Evaluation Procedure

The models' performance on the test datasets was calculated with an F-score. A baseline of assigning a positive label to all examples was calculated for all test datasets. The models' results on the test datasets were compared with a Friedman's test as suggested in the literature [5].

The best Slovenian model's predictions were reviewed with close reading. The error typology consisted of direct speech, indirect speech, speech fragments, annotation errors, annotation errors and *unrelated* and *other* tags. *Direct speech fragments* were sentences part of multi-sentence direct speech quotations. *Annotation errors* were examples with annotations inconsistent with the definition described in *Section 3.2.* For *unrelated* examples, close reading revealed no clear misclassification cause. *Other* was used for examples that did not fit any of the mentioned categories.

3.4 Training Settings

XLM-R and mBERT were used as base models with the default training settings from the *transformers* library with the exception of using 16 gradient accumulation steps and freezing the bottom 8 layers of all models. The latter reduces the training time without significant performance drops (Kovaleva idr., 2019; Merchant idr., 2020). Additionally, a Slovenian-Croatian-English BERT model was trained on English machine-translated data from Redewiedergabe.

4 Results

4.1 Model Results

The model performance varies based on the congruence between the language and precise task definitions in each dataset. The differences between model predictions were not statistically significant (χ^2_F = 9.66; df = 5; n = 8; p = 0.14) so post-hoc tests were not performed. As Table 2 demonstrates, the XLM-R model trained on both RiQuA and Redewiedergabe performed well across the datasets with an F-score of 80.5 and 77.6 on the Redewiedergabe and RiQuA test set, respectively. The high results from training on combined data suggests the RiQuA and Redewiedergabe datasets may benefit from additional or complementary data, at least when using cross-lingual transfer learning. The most successful strategy for Slovenian data was training on RiQuA and English machine-translated Redewiedergabe data using the CroSloEngual BERT model, reaching a F-score of 66.8. We did not evaluate the impact of using translated training data with mBERT and XLM-R.

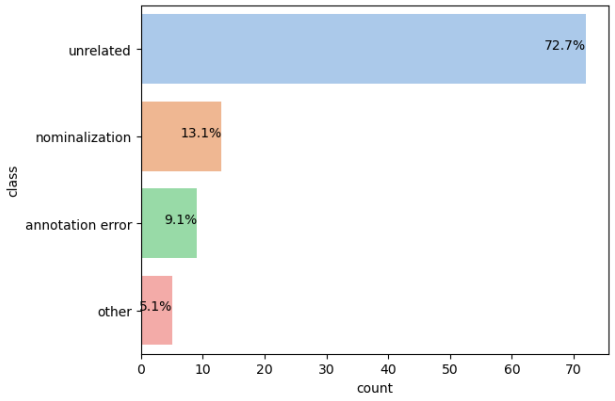


Figure 2: False Positives from the CroSloEngual BERT Classifier.

4.2 Error Analysis Results

The results from CroSloEngual BERT on Slovenian data were analyzed further. False positives were more common than false negatives, representing 23.4% and 9.8% of all examples (n = 744), respectively. Close reading of a sample of 100 false positives did not show a definite pattern for most (72.9%) of them. These examples were clearly unrelated to reported speech, although some did include words lexically related to reporting verbs (e.g. *The proposed law is still under discussion*). The second category of false positives were nominalizations of reported statements (13.1%) not included in our annotation schema. The final source of false positives were annotation errors consisting of wrongly

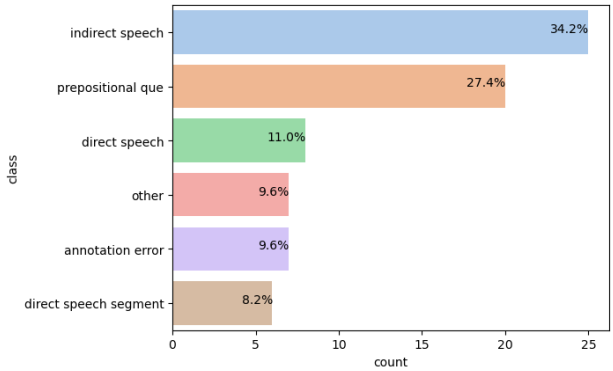


Figure 3: False Negatives from the CroSloEngual BERT Classifier.

unmarked examples of direct or indirect speech (9.1%). The distribution of categories identified in the sample of false positives are illustrated in Figure 2. The most common errors in the 73 false negative examples were instances of indirect speech (34.2% of false negatives) and prepositional queing of reported speech (27.4%). The remainder were instances of direct speech, direct speech fragments and annotation errors representing 11%, 8.2% and 9.6% of the false negatives, respectively. The annotation errors included nominalizations and statements reported as adjective complements (*The speaker was happy that the provisions were accepted)* not included in our annotation schema. Figure 3 summarizes the identified false negative categories .

5 Discussion

This paper presents the development of a reported speech classifier, tested through a small annotated Slovenian dataset and manual error analysis. Cross-lingual transfer learning from the annotated RiQuA and Redewiedergabe datasets achieved an F-score of 66.8 on a small manually annotated dataset of Slovenian news of parliamentary sessions using the base CroSloEngual model with RiQuA and English machine-translated Redewiedergabe training data [1]. This these results corroborate the observation that language models trained on a limited number of languages may outperform less specialized ones such as mBERT and XLM-R [20]. The major source of errors were false positives (23.4% of all sentences) for which no systematic pattern was discernible in the majority (72.9%) of examples. Instances of indirect speech and prepositional queing of statements were overrepresented in the false negatives, accounting for 61.6% of false negatives. Although rare, nominalizations were present in both false positives and false negatives and should be considered in future annotation guidelines. These obeservations indicate reported speech classifiers may benefit form approaches for addressing imbalanced classes.

6 Conclusion

This study developed a sentence-level reported speech classifier for Slovenian news texts using cross-lingual transfer learning. By leveraging existing multilingual models (mBERT, XLM-R, and CroSloEngual BERT) with the English and German datasets RiQuA and Redewiedergabe, we demonstrated that sentence-level

[1]The fine-tuned CSE model is available on the Hugging Face Hub under the name *zo-fi/rep-sp-CSE-rwg-riq*.

Table 2: Model Performances across Datasets (F-scores).

	Redewiedergabe	RiQuA	PolNeAR	QUAC	Quotebank	Slovenian dataset
Positive by default	52.1	60.6	74.2	19.5	45.8	60.3
mBERT+Both	77.5	77.4	73.1	40.5	53.5	63.2
mBERT+RiQuA	68.2	76.9	72.6	31.1	52.6	39.1
mBERT+RWG	78.4	70.4	65.5	43.4	49.1	63.2
XLM-R+Both	80.5	77.6	70	38.8	57.7	63.2
XLM-R+RiQuA	66.6	76.7	73.6	25.5	53.7	60.3
XLM-R+RWG	80.9	70.7	66.4	43.9	50	63.2
CroSloEngBERT+Both+MT	54	76.6	73	24	52.5	66.8

classification can detect some aspects of reported speech in Slovenian. However, the performance estimates are limited due to the small size of the Slovenian testing set and the limited definition used for the annotations. Future research should focus on developing a Slovenian annotated dataset, refining the annotation schema for multiple use cases, and exploring additional modeling features such as encoding broader sentence contexts. This work contributes a provisional tool for computational discourse analysis of Slovenian media texts. Further development is necessary for its application in more nuanced tasks.

Acknowledgements

This work was supported by the Slovenian Research Agency grants via the core research programs Equality and Human Rights in the Times of Global Governance (P5-0413) and Hate Speech in Contemporary Conceptualizations of Nationalism, Racism, Gender and Migration (J5-3102).

References

[1] Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. *PLOS ONE*, 16, 1, (Jan. 29, 2021), e0245533. DOI: 10.1371/journal.pone.0245533.

[2] Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2022. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16, 1, (Jan. 2, 2022), 1–18. DOI: 10.1080/19312458.2021.2015574.

[3] Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus REDEWIEDERGABE. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Nicoletta Calzolari et al., editors. European Language Resources Association, 803–812. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.100.

[4] Alexis Conneau et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors. Association for Computational Linguistics, 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

[5] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, 7, (Dec. 1, 2006), 1–30.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT 2019. Jill Burstein, Christy Doran, and Thamar Solorio, editors. Association for Computational Linguistics, 4171–4186. DOI: 10.18653/v1/N19-1423.

[7] Gabriel Dvoskin. 2020. Reported speech and ideological positions: the social distribution of knowledge and power in media discourse. *Bakhtiniana: Revista de Estudos do Discurso*, 15, 193–213.

[8] Zoran Fijavž and Darja Fišer. 2021. Citatnost in reprezentacija v spletnem migracijskem diskurzu. In *Sociolingvistično iskrenje*. Maja Bitenc, Marko Stabej, and Žejn Andrejka, editors. Založba Univerze v Ljubljani. Retrieved Apr. 3, 2024 from https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/259/370/6011.

[9] Elizabeth Holt. 1996. Reporting on Talk: The Use of Direct Reported Speech in Conversation. *Research on Language and Social Interaction*, 29, 3, (July 1, 1996), 219–245. DOI: 10.1207/s15327973rlsi2903_2.

[10] Edward Newell, Drew Margolin, and Derek Ruths. 2018. An Attribution Relations Corpus for Political News. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018. Nicoletta Calzolari et al., editors. European Language Resources Association (ELRA). Retrieved Apr. 10, 2024 from https://aclanthology.org/L18-1524.

[11] Mojca Pajnik and Marko Ribać. 2021. Medijski populizem in afektivno novinarstvo: časopisni komentar o »begunski krizi«. *Javnost - The Public*, (Dec. 14, 2021). Retrieved Apr. 24, 2024 from https://www.tandfonline.com/doi/abs/10.1080/13183222.2021.2012943.

[12] Sean Papay and Sebastian Padó. 2020. RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Nicoletta Calzolari et al., editors. European Language Resources Association, 835–841. ISBN: 979-10-95546-34-4. Retrieved Apr. 21, 2024 from https://aclanthology.org/2020.lrec-1.104.

[13] Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2013. David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors. Association for Computational Linguistics, 989–999. Retrieved Apr. 17, 2024 from https://aclanthology.org/D13-1101.

[14] Marta Ercília Mota Pereira Quintão. 2014. Quotation A ttribution for Portuguese News Corpora. In Retrieved Apr. 21, 2024 from https://www.semanticscholar.org/paper/Quotation-A-ttribution-for-Portuguese-News-Corpora-Quint%C3%A3o/69fea7d030d5e71b973ec67aa897a7c9aadadac2.

[15] Masaki Shibata. 2023. Dialogic Positioning on Pro-Whaling Stance: A Case Study of Reported Speech in Japanese Whaling News. *Japanese Studies*, 43, 1, (Jan. 2, 2023), 71–90. DOI: 10.1080/10371397.2023.2191839.

[16] Michael Short. 1988. Speech presentation, the novel and the press. In *The Taming of the Text*. Willie Van Peer, editor. Routledge. ISBN: 978-1-315-54452-6.

[17] Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2023. Identifying Informational Sources in News Articles. Version 1. DOI: 10.48550/ARXIV.2305.14904.

[18] Stef Spronck and Daniela Casartelli. 2021. In a manner of speaking: how reported speech may have shaped grammar. *Frontiers in Communication*, 6, 624486.

[19] Sara Stymne and Carin Östman. 2022. SLäNDa version 2.0: Improved and Extended Annotation of Narrative and Dialogue in Swedish Literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. LREC 2022. Nicoletta Calzolari et al., editors. European Language Resources Association, 5324–5333. Retrieved Apr. 21, 2024 from https://aclanthology.org/2022.lrec-1.570.

[20] Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *Text, Speech, and Dialogue* (Lecture Notes in Computer Science). Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák, editors. Springer International Publishing, Cham, 104–111. ISBN: 978-3-030-58323-1. DOI: 10.1007/978-3-030-58323-1_11.

[21] Timoté Vaucher, Andreas Spitz, Michele Catasta, and Robert West. 2021. Quotebank: A Corpus of Quotations from a Decade of News. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. WSDM '21: The Fourteenth ACM International Conference on Web Search and Data Mining. ACM, 328–336. ISBN: 978-1-4503-8297-7. DOI: 10.1145/3437963.3441760.

[22] M. Wynne. 1996. Speech, Thought and Writing Presentation Corpus. Retrieved Apr. 21, 2024 from https://ora.ox.ac.uk/objects/uuid:6caa73c1-d283-4d51-a78f-55df69bae986.

[23] Dian Yu, Ben Zhou, and Dong Yu. 2022. End-to-End Chinese Speaker Identification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors. Association for Computational Linguistics, 2274–2285. DOI: 10.18653/v1/2022.naacl-main.165.

What kind of ESG is profitable? Connecting company performance to ESG terms in financial reports

Luka Andrenšek
trovato@corporation.com
Jožef Stefan Institute
Ljubljana, Slovenia

Katarina Sitar Šuštar
University of Ljubljana
Ljubljana, Slovenia
katarina.sitar@ef.uni-lj.si

Senja Pollak
Jožef Stefan Institute
Ljubljana, Slovenia
senja.pollak@ijs.si

Matthew Purver
Jožef Stefan Institute
Ljubljana, Slovenia
matthew.purver@ijs.si

ABSTRACT

In this paper, we examine the relationship between the discussion of Environmental, Social and Governance (ESG) in companies' annual financial reports and their financial performance. Specifically, we analyse the companies' use of specific ESG terms alongside the performance metric, sector-normalized Return on Assets (ROA). Our motivation is to determine whether companies frequently mentioning terms such as "gender", "equality", "talent", and "innovation" in their reports demonstrate a higher annual ROA compared to those that rarely used these terms. To explore this, we used existing datasets with reports and performance metrics from 348 companies, covering the years from 2009 to 2021. In order to better examine differences, we then selected companies whose ROA significantly differed from the average (either higher or lower), allowing for a more pronounced examination of the impact of ESG term usage on financial performance. The filtered dataset consisted of 107 companies, with a total of 427 reports; split into two sections representing higher and lower performing companies. We then used an existing list of ESG terms derived from a range of separate data sources, and applied a basic statistical n-gram language model to extract the probabilities of each ESG term's occurrence in each of the higher- and lower-performing dataset sections. Results show that while certain sets of ESG concepts correlate with higher financial performance, others do the opposite, and give some initial interpretation into the light this sheds on company reporting behaviour.

KEYWORDS

financial report analysis, language modelling, environmental, social and governance reporting

1 INTRODUCTION & RELATED WORK

There is increasing interest in the behaviour of companies in the area of Environmental, Social and Governance (ESG) criteria, including a company's environmental impact (Environmental), relationships with the community including employees, suppliers and customers (Social), and leadership structures including executive pay and shareholder rights (Governance). Although until recently, ESG analyses were almost entirely performed manually by experts (see e.g. [10]), there has been a large amount of work in the last few years on applying computational machine learning and statistical methods to ESG analysis (see e.g. the recent review by Lim [9]).

However, much of this analysis examines numerical company performance data and categorical metadata; our interest is in developing and applying natural language processing (NLP) technologies to not only help automate analyses, but allow understanding of how human actors discuss and understand the important and meaning of ESG aspects.

Application of NLP in finance is not new: for example, topic modelling has been used to predict company performance and investigate strategies [14, 7]. Recent work also includes application to ESG aspects: Nugent et al. [12] automatically extract news about ESG controversies, and Lee et al. [8] analyse sentiment on ESG issues. Closer to our interests, Purver et al. [13] investigated how the use of ESG terms by companies has changed over time. By analysing and annotating a set of existing resources, they defined a set of 93 ESG terms categorised into 5 core ESG areas. They then showed how these terms can be used to analyse changes in reporting, by analysing a collection of company annual reports, collated over a period of 8 years, using language modelling and distributional methods to reveal changes in the frequency and in the usage of the ESG terms.

Here, we are interested not in changes in ESG discussion over time, but in whether and how the reporting of ESG aspects is connected to financial performance. We take Purver et al. [13]'s resources and methods as a starting point, but augment the financial report text data with available metadata on financial performance, allowing us to compare how ESG reporting varies between more and less well-performing companies.

2 DATA AND METHODS

2.1 Hypotheses

In general, we expect increased probability of appearance of ESG terms in the annual reports from the more profitable firms, based on a number of factors. In general, overall high ESG performing companies exhibit high financial performance [1, 5]; although we note that the link between high ESG score performance and mention of ESG terms is not guaranteed to be straightforward. More specifically, during the period between 2010-2020 analysed here, there was a growing emphasis on corporate social responsibility (CSR) and sustainability. Investors, consumers, and other stakeholders increasingly prioritised companies that demonstrated a commitment to innovation, diversity, and environmental sustainability [11, 2]. Busru and Shanmugasundaram [3] find that firms closely engaging in fostering innovation, attracting top talent,

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.3

Year	# Reports	# Words
2012	178	12.5M
2013	181	14.0M
2014	184	15.0M
2015	196	16.3M
2016	198	17.5M
2017	200	18.4M
2018	200	19.6M
2019	202	21.2M
total	1539	134.6M

Table 1: Number of annual reports available by year

promoting gender and diversity initiatives, could confer a competitive advantage over the industry peers. Furthermore, some policy and regulatory changes (e.g. the 2018 UK Corporate Governance Code, the 2014 EU Directive on Non-Financial Reporting, Carbon Disclosure Project (CDP)) directly or indirectly encouraged companies to address issues related to diversity, gender equality, and environmental sustainability.

2.2 Data and pre-processing

To test this hypothesis, we build on the resources and methods of Purver et al. [13], who provide a dataset of annual reports from FTSE350 companies over the years 2012-2019, based on the FTSE350 list as of 25th April 2020 and obtained from the publicly accessible collection at www.annualreports.com. The reports are already converted to plain text, and we use their publicly available tools to tokenize the collection into words and build ngrams of length 1-4 padded with sentence start and end symbols; the dataset size is reported in Table 1 below (taken from [13]). We use their set of ESG terms, defined via a process of extracting candidate terms from a set of public ESG definitions and taxonomies, asking financial expert annotators to label them as to their representativeness as ESG terms and their ESG subcategory, and keeping the terms with high inter-annotator agreement (see [13] for details).

2.3 Financial performance analysis

The reports were then linked to financial indicators for the respective year and company. The data on company fundamentals was obtained from the Refinitiv EIKON Datastream.[1] Each entry contained annual financial indicators, as well as the companies' industry and sector codes. The main variable of interest was normalized, averaged *return on assets (ROA)* as defined below:[2]

$$\frac{\begin{array}{l} NetIncome - BottomLine \\ \quad +((InterestExpenseOnDebt - InterestCapitalized) \\ \qquad \times(1 - TaxRate)) \end{array}}{AverageOfLastYear'sAndCurrentYear'sTotalAssets}$$

After extracting financial reports with available ROA data, we categorized the financial reports into two groups, in order to examine differences in the associated reports' use of ESG terms. The distribution of ROA shows a heavy concentration around the mean, so in order to derive two distinctive groups we took the two extremes and excluded the central group around the mean. The 'negative' group comprised reports with a yearly ROA less than -0.2, indicating very poor performance. Conversely, the 'positive' group included reports with an ROA of at least 0.2, reflecting very good yearly performance.

Subsequently, we employed a statistical n-gram language model (using NLTK[3]) to analyze the occurrence of each ESG term. For each term, we calculated the probability of its occurrence in positive reports (p_+) and in negative reports (p_-), and the difference ($p_+ - p_-$). Terms with a large difference in these probabilities are more strongly associated with positive reports than with negative ones, and vice versa: terms with a large negative difference are common in negative reports but rare in positive ones. We conducted this analysis for both unigrams and bigrams.

3 RESULTS AND DISCUSSION

The results for 1- and 2-grams are shown in Figures 1 and 2 below (3- and 4-grams showed no clear interpretable associations).[4] As hypothesized, many ESG terms show a strong association with positive performance, with many of these being core terms associated with human resources (*innovation*, *talent*), with social aspects (*gender*, *diversity*), environmental aspects (*renewable*, *carbon footprint*, *environmental impact*) and overall ESG descriptors (*ethical*). However, many terms are conversely (and contrary to our general hypothesis) associated with negative performance, including, again, terms across various ESG categories including environmental (*carbon emissions*, *energy efficiency*, *greehouse*), human resources (*mental health*, *wellbeing*) and general ESG descriptors (*governance*).

However, by combining these terms with recent work in clustering and describing ESG terms [4], we can shed more light on which categories seem to be more positive and which more negative. Ferjancic et al. [4], using the same dataset and ESG term list [13], perform a further topic analysis using BERTopic [6], in which they derive 30 ESG-related topics and 6 higher-level clusters of ESG concepts; they then examine the correlations between these ESG topics and company ESG scores as obtained from external analysts. We align our ESG terms with Ferjancic et al. [4]'s 30 topics by matching against the words most associated with each topic (if a term appears in the top 10 words associated with a topic, we take the term and topic as aligned); we can then compare our positive/negative associations with Ferjancic et al. [4]'s correlations with company ESG scores. Table 2 shows this alignment for our most positive and negative bigram terms here, with the topic labels and an indication of the strength and direction of correlation with overall company ESG scores, as given by [4].

Given this, we see some systematic groupings. *Climate change*, as part of the 'climate risk and policy' topic, as well as *supply chain* and *human trafficking* as part of the 'human rights' topic, represent the themes that appear to be, across different industries, related to high company ESG scores. A similar observation holds for *gender balance*, *gender pay* and *environmental impact*, which all fall in a group of topics which are strongly and significantly correlated with high ESG scores throughout different industries. Overall high ESG performing companies exhibit high financial performance [1, 5], therefore our results for terms such as *climate change*, *supply chain* and *human trafficking* are not surprising: as indicators of topics associated with high ESG, they are good terms for tracking these ESG aspects associated with high financial performance.

[1] https://www.refinitiv.com
[2] We use this normalization and averaging to smooth and remove one-off effects.
[3] https://www.nltk.org/
[4] Note that these figures show differences in absolute probabilities: magnitudes are comparable within 1-grams, and within 2-grams, but not between 1- and 2-grams.

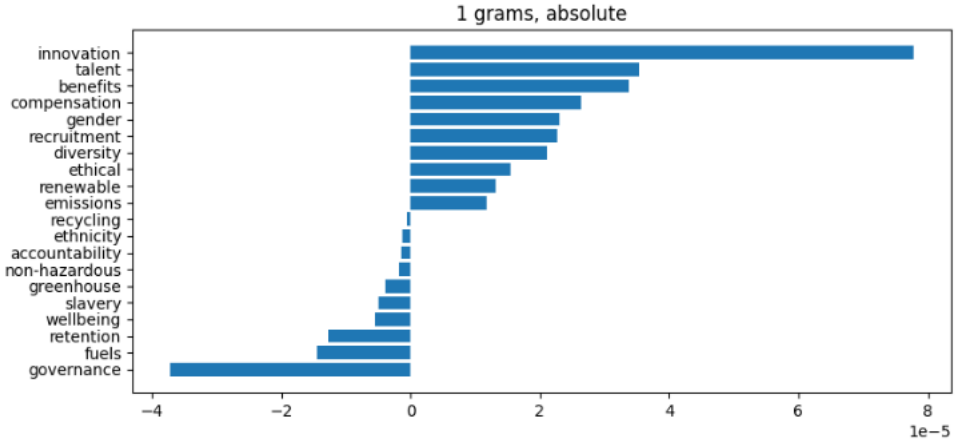


Figure 1: Difference in probability between positive and negative reports $p_+ - p_-$ for the most positive and negative unigram ESG terms.

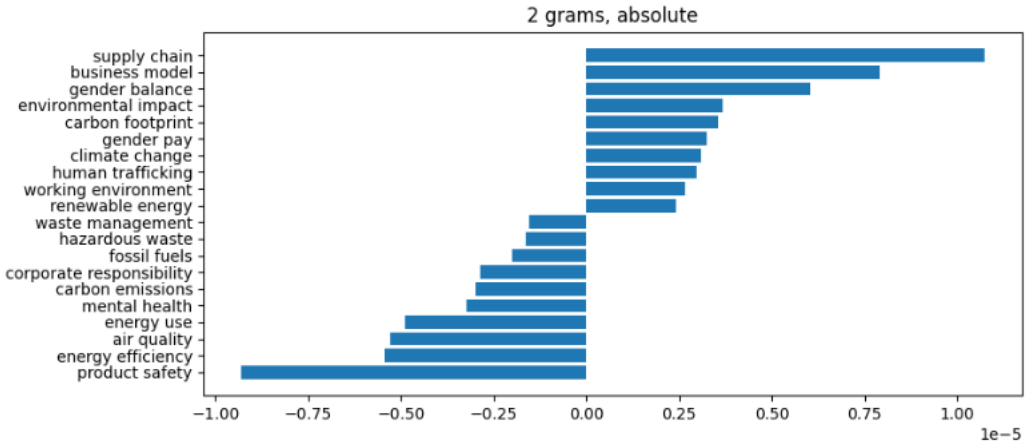


Figure 2: Difference in probability between positive and negative reports $p_+ - p_-$ for the most positive and negative bigram ESG terms.

Looking at the terms with low values which are associated with low RoA, *waste management* and *corporate responsibility* are associated with topics, for which in some industries proportion of these correlate with ESG scores significantly positively and in other industries this correlation is significantly negative. Based on overall correlation between ESG scores and topic proportions across different industries, these two topics are among the third of the topics for which negative correlation between the topic proportion and ESG score prevails. Due to the aforementioned correlation between ESG and financial performance it is therefore understandable that these terms are associated with mention in annual reports of companies with low RoA. Overly extensive discussion on specific topics (such as 'waste management' and 'corporate responsibility') can negatively impact ESG score (see [4]) which can by analogy of ESG and financial performance [1, 5] hold for companies with low RoA.

There is a surprising number of bigrams in both the high RoA and low RoA groups which seem to be associated with the same topic, namely 'climate footprint and energy management'. For companies with high RoA, these terms are *carbon footprint* and *renewable energy*, and for companies with low RoA, the terms are *fossil fuels*, *carbon emissions*, *energy use*, *air quality* and *energy efficiency*. It seems that better performing companies use *carbon footprint* instead of *carbon emissions*, and discuss more on the use of *renewable energy* than on *energy use*, *energy efficiency* and/or *fossil fuels*. In future work, we plan to analyse the use of these terms in more depth, including analysis of the lexical and topical contexts in which they appear, and adding techniques such as sentiment and topic analysis to shed more light on these distinctions.

ACKNOWLEDGEMENTS

The authors thank the reviewers for helpful suggestions, and acknowledge financial support from the Slovenian Research Agency for research core funding (No. P2-0103), as well as for funding of the research project *Quantitative and qualitative analysis of the unregulated corporate financial reporting* (No. J5-2554).

REFERENCES

[1] Nisar Ahmad, Asma Mobarek, and Naheed Nawazesh Roni. 2021. Revisiting the impact of ESG on financial performance of FTSE350 UK firms: static and dynamic panel data analysis. *Accounting, Corporate Governance & Business Ethics*. DOI: 10.1080/23311975.2021.1900500.

2 grams	Term/ROA correlation	Topic	Topic/ESG correlation	score
Supply chain	+	Human rights	++	
Business model	+	Customer services, People and culture	+; -	
Gender balance	+	Diversity and inclusion	++	
Environmental impact	+	General ESG	+	
Carbon footprint	+	Climate footprint and energy management	=	
Gender pay	+	Diversity and inclusion	++	
Climate change	+	Climate risk and policy	++	
Human trafficking	+	None directly related, in broader context in Human rights	++	
Working environment	+	People and culture	-	
Renewable energy	+	Climate footprint and energy management	=	
Waste management	-	Waste management	–	
Fossil fuels	-	No explicit match; contextually appears in Climate footprint and energy management	=	
Corporate responsibility	-	Corporate governance	–	
Carbon emissions	-	Climate footprint and energy management	=	
Mental health	-	Health and safety	+	
Energy use	-	Climate footprint and energy management	-	
Air quality	-	No explicit match; contextually appears in Climate footprint and energy management	-	
Energy efficiency	-	Climate footprint and energy management	-	
Product safety	-	Health and safety	=	

Table 2: Selected ESG terms with their ROA correlation direction (+/−), topic according to [4], and topic/ESG score correlation strength (+ + /+/= /−/−−) as calculated by [4].

[2] A. C. Amason and H. J. Sapienza. 2012. The effects of top management team size and interaction norms on cognitive and affective conflict. *Journal of Management*, 23, 495–516.

[3] S. A. Busru and G. Shanmugasundaram. 2017. Effects of innovation investment on profitability and moderating role of corporate governance: empirical study of indian listed firms. *Indian Journal of Corporate Governance*, 10, 97–117, 2. https://doi.org/10.1177/0974686217730938.

[4] Ursa Ferjancic et al. forthcoming. Textual analysis of corporate sustainability reporting and corporate ESG scores. under review. (Forthcoming).

[5] Gunnar Friede, Timo Busch, and Alexander Bassen. 2015. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5, 4, 210–233. DOI: 10.1080/20430795.2015.1118917.

[6] Maarten Grootendorst. 2022. BERTopic: neural topic modeling with a class-based TF-IDF procedure. (2022). https://arxiv.org/abs/2203.05794 arXiv: 2203.05794 [cs.CL].

[7] M. Jagannathan, D. Roy, and V. S. K. Delhi. 2022. Application of NLP-based topic modeling to analyse unstructured text data in annual reports of construction contracting companies. *CSI Transactions on ICT*, 10, 2, 97–106.

[8] H. Lee, S. H. Lee, K. R. Lee, and J. H. Kim. 2023. Esg discourse analysis through bertopic: comparing news articles and academic papers. *Computers, Materials & Continua*, 75, 3, 6023–6037.

[9] Tristan Lim. 2024. Environmental, social, and governance (esg) and artificial intelligence in finance: state-of-the-art and research takeaways. *Artificial Intelligence Review*, 57, 76. DOI: 10.1007/s10462-024-10708-3.

[10] Steve Lydenberg, Jean Rogers, and David Wood. 2010. From Transparency to Performance: Industry-Based Sustainability Reporting on Key Issues. Tech. rep. Available from https://iri.hks.harvard.edu/links/transparency-performance-industry-based-sustainability-reporting-key-issues. Hauser Center for Nonprofit Organizations at Harvard University.

[11] M. Marzook and B. Al Ahmady. 2022. Linking organisational performance and corporate social responsibility. *European Jnl. of Business and Management Research*, 7, 335–343, 3. https://doi.org/10.24018/ejbmr.2022.7.3.1466.

[12] T. Nugent, N. Stelea, and J. L. Leidner. 2020. Detecting ESG topics using domain-specific language models and data augmentation approaches. (2020). http://arxiv.org/abs/2010.08319.

[13] Matthew Purver, Matej Martinc, Riste Ichev, Igor Lončarski, Katarina Sitar Šuštar, Aljoša Valentinčič, and Senja Pollak. 2022. Tracking changes in ESG representation: initial investigations in UK annual reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*. Mingyu Wan and Chu-Ren Huang, editors. Marseille, France, (June 2022), 9–14. https://aclanthology.org/2022.csrnlp-1.2.

[14] W. Xu and K. Eguchi. 2021. Topic embedding regression model and its application to financial texts. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, 15–21.

Classification of Patents Into Knowledge Fields: Using a Proposed Knowledge Mapping Taxonomy (KnowMap)

Elham Motamedi
elham.motamedi@upr.si
University of Primorska
Koper, Slovenia

Inna Novalija
inna.koval@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Luis Rei
luis.rei@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Various platforms, including patent systems and repositories like GitHub and arXiv, support knowledge dissemination across domains. As knowledge increasingly spans multiple disciplines, there is a need to track innovations that intersect various fields. Despite available data, a comprehensive knowledge taxonomy for effectively tracking innovations across domains is lacking. Developing such a taxonomy and employing automated classification methods will enhance the ability to track shared knowledge.

In this work, we first developed a knowledge taxonomy based on the CPC schema. We formulated the classification of textual data into defined knowledge fields as a multi-label problem. Then, we evaluated the effectiveness of the classification models by fine-tuning pre-trained transformer language models. The multi-label framework enables the tracking of knowledge trends at the intersection of various disciplines.

Keywords

Knowledge Taxonomy, Knowledge Tracking, Patent Classification, Hierarchical Classification, Multi-label Classification

1 Introduction

According to the World Intellectual Property Organisation (WIPO), a patent is an exclusive right granted for an invention, providing legal protection to the inventor while simultaneously benefiting society by making the invention publicly accessible [1]. Each year, patent offices receive numerous patent applications that need to be processed [13].To ensure the novelty of patent applications, inventors should also be able to search existing patents. Organising patents with unique codes in a hierarchical structure aids efficient retrieval and aligns with natural human navigation, starting from broad categories and narrowing down to specifics[21]. Among these hierarchical structures, the CPC system is widely recognised [6]. The CPC codes are organised as a taxonomy, meaning that each entity in the lower level is the detail group of the parent. A patent can be assigned to one or more labels by the experts in patent offices [8, 18]. In the first level of the CPC hierarchy, there are nine sections, which are divided into classes, subclasses, groups, and subgroups. Each level of this hierarchy can have several codes ending in approximately 250,000 classification labels [11]. An example of the hierarchical structure of CPC code is provided in Tab. 1.

The CPC schema's top level has only nine sections, but the number of groups increases substantially at lower levels. In this study, we created a knowledge field taxonomy by merging CPC's detailed classes into a more abstract representation. This taxonomy not only serves as a framework for knowledge representation but also offers a benchmark for patent classification systems. While some studies address the issue of numerous class labels by excluding less-represented classes or truncating hierarchies [24], a consistent benchmark taxonomy has been lacking. Since our proposed knowledge taxonomy aligns with the CPC schema, it is able to provide a benchmark for future studies, facilitating the comparison of different models.

In summary, our paper's contribution is the proposal of a knowledge field taxonomy, KnowMap, which aligns with the widely used CPC schema. The KnowMap merged several class labels within the CPC schema based on the scope of the knowledge field and the number of patents associated with each class. The KnowMap taxonomy is available online [2]. In this study, we also performed a classification task to categorise patents into the fine-grained classes defined by our proposed taxonomy.

Table 1: Example of a sequence of codes across different levels of the CPC hierarchy

CPC	Code	Title
Section	H	Electricity
Class	H03	Electronic circuitry
Subclass	H03C	Modulation
Group	H03C3/00	Angle modulation
Subgroup	H03C3/005	Circuits for asymmetric modulation

2 Related Work

Patent documents contain various types of information, including text, diagrams, plots, and references to other patents or scientific publications [20]. The textual content of a patent is divided into several sections, such as the title, abstract, claim, and description [11]. The title and abstract are shorter than the description but still provide relevant information for classification. Li et al. [15] evaluated various lengths of the abstract and title, finding that using the first 100 words of title and abstract resulted in the best classification performance in their study.

Various classification systems exist for organising patents [6]. In this work, we focus on the CPC schema. The hierarchical representations help organise patents and facilitate efficient searching. Kamateri et al. [11] discussed several potential challenges that artificial intelligence technologies face in patent classification. One such challenge is the extensive number of class labels. As an example, the IPC contains approximately 86,000 classes, while the CPC has around 250,000.

Patent classification is a multi-label classification problem since every patent can belong to several knowledge fields [18,

[1] https://www.wipo.int/portal/en/

[2] https://github.com/elmotamedi/KnowMap-Taxonomy

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.19

10]. Given the large number of classes at the lowest level of the taxonomy tree, the performance of automatic models in predicting such granular categories is limited. Various models have been used to classify patents in a multi-label setting, ranging from classical machine learning models to deep learning models [15, 5, 8]. Several previous studies have focused on higher levels of the hierarchy, limiting classification to broader categories such as sections, classes, or subclasses within the taxonomy [3]. Bekamiri et al. [3] fine-tuned the SBERT model to predict labels at the subclass level (i.e., 663 class labels) using a multi-label formulation. They achieved F1-score of 66%, outperforming previous studies that used the same datasets. Aroyehun et al. [1] similarly truncated the IPC hierarchy at the subclass level and predicted these labels by transferring knowledge from two higher levels (section and class) to the lower level (subclass), achieving a precision score of 0.53. While it remains valuable for patent office experts to use an automatic model that can narrow down applications to higher levels of the taxonomy tree, this approach has limitations and challenges. One such challenge is that the choice of target class labels does not depend on the scope of the knowledge area. More established and expansive areas may benefit from directing experts to detailed groups, while less developed areas may be adequately served by broader classifications.

3 Methods and Materials

In this work, we developed a knowledge taxonomy and classified patents into fine-grained classes by fine-tuning pre-trained models. Below, we outline the methods and materials used.

3.1 Patent Collection and Preprocessing

The dataset used in our experiments is the Google Patents Public Datasets on BigQuery [3]. Each patent has several pieces of information, including the publication number, application number, CPC code, title, abstract, and detailed description. We have expanded the dataset to include the titles associated with each CPC code from Espacenet. [4]. In this study, we focused on the textual data. We generated the input text by concatenating the title, followed by the abstract, and then the description. We included only those documents where the concatenated text is at least 100 words long. Previous studies have examined various lengths of textual data and found that using the first 100 words often results in higher performance for classification tasks [15].

To create a hierarchical structure where we have enough documents among leaf-node labels (i.e., avoiding scenarios where one group contains only a few hundred documents while others contain hundreds of thousands as an example), we needed to count the number of documents which fall into the defined categories. As a preprocessing step before counting, we performed de-duplication, which involved removing duplicate and near-duplicate textual data [4, 12, 14].

Due to the large size of the dataset, we employed MinhHash Locality Sensitive Hashing (LSH) as a deduplication method to efficiently identify similar documents [7, 9, 22]. Specifically, we used MinHash to approximate the Jaccard similarities between sets of n-grams within the documents. MinHash is particularly advantageous for large datasets because it supports parallel computation, enhancing scalability [2]. We set the similarity threshold at 0.9, meaning that documents with a Jaccard similarity of 90% or higher were considered duplicates. To generate the hash signatures in MinHash, we used 128 permutations. For the n-gram representation, we used a range of 1 to 3, incorporating 1-grams, 2-grams, and 3-grams.

3.2 Refining Hierarchical Structure Through Group Merging

The hierarchical structure of the CPC groups was refined at each level of the tree. We started with nine sections at the top level (i.e., *level 1*), which were preserved. At subsequent levels (i.e., *level 2* to *level 4*), groups were merged by manual analysis based on shared knowledge and the number of documents. Groups with relatively few documents (i.e., groups with fewer than 40,000 for *level 2*, 20,000 for *level 3*, and 9,000 for *level 4*) were combined with other groups at the same level that shared similar knowledge. As an example, at the subclass level of the CPC hierarchy, "A01B" (i.e., Soil working) and "A01C" (i.e., Planting, Sowing, Fertilising) represent related steps in agricultural practices, as both are foundational processes in land preparation and management. We merged them into a single group labelled "Soil working and planting," resulting in 162,567 patents in this category. The refinement continued until the fine-grained classes contained at least 9,000 documents.

3.3 Text Classification

We formulated the classification problem as a multi-label problem, in which each document can be assigned to multiple knowledge fields. In this study, we aimed to classify the patents into the fine-grained classes in the lowest level of the proposed taxonomy (i.e., 83 classes). To balance performance and computational cost given the large size of the dataset, We used the pre-trained language models *distilroberta-base*, a distilled version of RoBERTa [16, 19], and *all-MiniLM-L6-v2*, a version of MiniLM fine-tuned for semantic similarity [22, 17]. The pre-trained models were fine-tuned for the downstream task by adding a classification head. The classification head takes the hidden state of the first token from the model and processes it through a fully connected dense linear layer, followed by a dropout layer for regularisation and a tanh activation function for non-linearity. Since our task is multi-label classification, the output logits for each class are converted into probabilities using a sigmoid function.

For model training, we used a learning rate of 4e-5 with a linear scheduler and a weight decay of 0.1. To prevent overfitting, the best checkpoint was selected based on evaluation metrics on the validation set. We trained the model for up to 5 epochs with early stopping criteria based on validation accuracy. The dataset, consisting of 1,092,991 samples randomly selected after deduplication, was split into training, validation, and test sets with ratios of 0.8, 0.1, and 0.1, respectively. To preserve the ratio of samples per class in training, validation, and test sets, we used stratified splitting [5].

3.4 Classification Evaluation

The F1-score is a common metric for classification tasks. We report both Micro-F1, averaged across all instances, and Macro-F1, averaged across all classes.

4 Results and Analysis

In this section, the results are presented in two parts. First, we present our proposed KnowMap taxonomy. Then, we report the

[3]https://github.com/google/patents-public-data
[4]https://worldwide.espacenet.com/
[5]https://github.com/trent-b/iterative-stratification?tab=readme-ov-file#multilabelstratifiedkfold

performance of classifiers in categorising patents into the fine-grained classes of this taxonomy.

4.1 The Proposed Knowledge Mapping Taxonomy (KnowMap)

The taxonomy, along with the associated CPC sections, classes, subclasses, groups, and subgroups are provided in the shared online source. An example of detailing the knowledge field of *soil working and planting* within the broader knowledge field of *human necessities* is illustrated in Fig. 1.

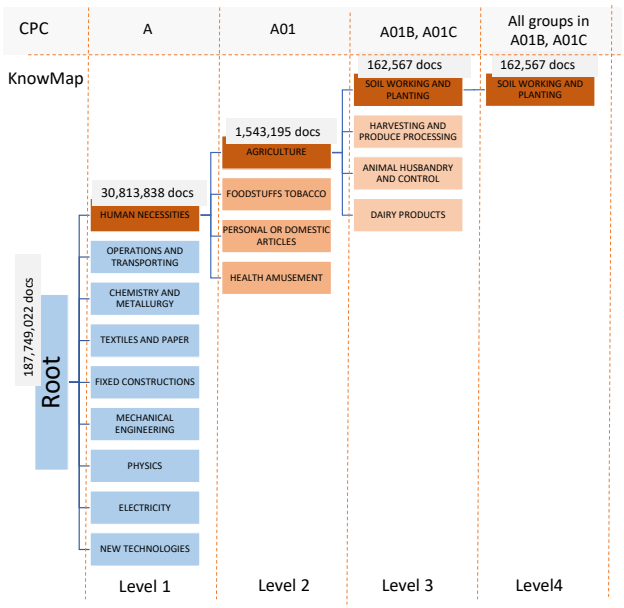


Figure 1: An example of a branch extension in KnowMap from the root to the lowest level, showing the association of KnowMap classes with corresponding CPC classes at each level.

4.2 Classification Results

The classification task in this study was to classify patents into 83 fine-grained classes within our proposed KnowMap taxonomy. The dataset comprised 1,092,991 documents, which were split into the train, validation, and test sets with a ratio of 0.8, 0.1, and 0.1 respectively. We preserved the ratio of samples per class in all three sets with stratified splitting. The average number of documents in the train set, validation set, and test sets are presented in Tab. 2.

Table 2: Overview of sample metrics: total number of samples, average number of samples per class, and normalised average number of samples per class across training, validation, and test sets.

Set	Total	Avg/ class	Normalised Avg
Train	1,092,991	132,202	0.012
Val	874,372	16,476	0.012
Test	218,619	15,543	0.012

Table 3: Classification Results

Metric	RoBERTa	SBERT
Micro-F1 (Val)	0.76	0.76
Macro-F1 (Val)	0.86	0.86
Micro-F1 (Test)	0.77	0.76
Macro-F1 (Test)	0.90	0.90

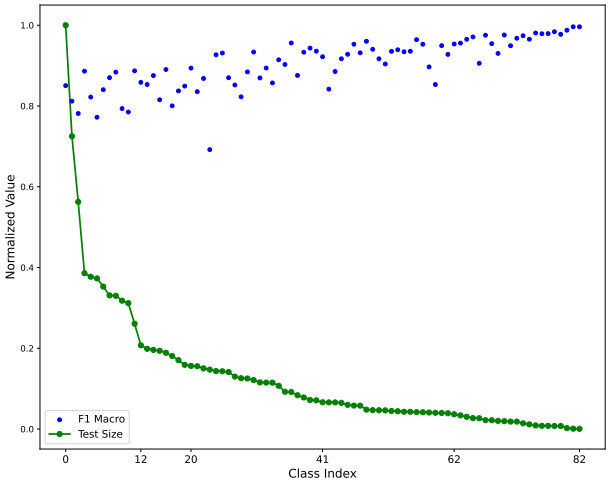


Figure 2: Normalised test size along with F1 Macro scores for each class. The x-axis represents class indices. The y-axis shows normalised values for test size and F1 Macro scores (blue dots).

We demonstrated the experimental results on the two classification models *RoBERTa* and *SBERT* in Tab. 3.

As observed from the results, the Macro-F1 score is higher than the Micro-F1 score, which may indicate that the model performs better for minority classes compared to majority classes. To gain more insights into these results, we generated a plot (see Fig.2), showing the F1 scores along with the normalised number of documents for each class in the test set. We used normalised values to allow both F1 scores and class sizes to be displayed in a single figure, facilitating better comparison.

The plot shows that the Macro-F1 score is higher for minority classes than for majority classes, also indicating that random sampling led to an unbalanced dataset. The imbalanced sample likely caused the higher Macro-F1 score relative to Micro-F1, reflecting poorer performance in the majority classes. Future work will focus on using balancing techniques when sampling to address this issue and enhance model performance.

When looking more closely at the lowest F1-Macro scores, we found that the bottom 10 classes were all leaves under the *chemistry and metallurgy* section. Moreover, the highest F1-Macro scores (0.996) were achieved by the two classes in the *textiles and paper* section, followed by all 17 leaves from the *physics* section. We suspect this performance difference may be due to greater variation in the textual data of *chemistry and metallurgy* class compared to *physics* and *textiles and paper*, leading to more variation between the training and test sets. Analysing this variation in detail remains a task for future work. Additionally, we believe future work could benefit from adapting the classifier to a hierarchical structure, prioritising correct predictions at higher

levels before refining predictions at the leaf level. In our current approach, the classifier does not account for the hierarchy and predicts all leaves directly.

5 Discussion and Conclusions

In this work, we proposed a knowledge field taxonomy, KnowMap, which aligns with the widely used CPC schema. The taxonomy consists of 83 groups at the lowest level, with fine-grained classes containing a minimum of 9,000 samples from the original Google Patents Public Dataset after preprocessing. KnowMap serves as a benchmark taxonomy, addressing a gap in the existing literature.

From the preprocessed original dataset, we randomly selected 1,093,151 samples to fine-tune pre-trained RoBERTa and SBERT models for downstream tasks. However, the random sampling resulted in an unbalanced dataset, which contributed to higher Macro-F1 scores compared to Micro-F1 scores. To enhance classification results, we plan to create a balanced dataset from the original data. Additionally, we aim to use larger models than those used in this study to further improve the fine-tuning process.

6 Future Work

Several knowledge platforms, such as news sites and GitHub, host various types of information shared online. In future work, we aim to incorporate these sources to extend and enhance the knowledge taxonomy's coverage. For example, the All Science Journal Classification (ASJC), which organises research publications by subject area, can be used to identify alignments with the existing taxonomy. This taxonomy alignment can then be further analysed to determine whether to merge or split classes at various levels. Beyond patents, we plan to evaluate the classifier on other data, using domain adaptation methods to transfer knowledge from the labelled patent domain to those with limited or no labels. Large language models (LLMs) could further aid in evaluating the classifier's performance across different domains. Recent research has shown the potential of LLMs to augment or even replace human-labeled training data with labels generated by these models [23].

Moreover, we plan to enhance the classification task by balancing the dataset using balancing techniques for multi-label problems and leveraging larger pre-trained models. we will also closely examine the different knowledge fields to better understand the variations in classifier performance across them.

Acknowledgements

This work was supported by the Slovenian Research and Innovation Agency under grant agreements CRP V2-2272, V5-2264, CRP V2-2146 and the European Union through enrichMyData EU HORIZON-IA project under grant agreement No 101070284.

References

[1] Segun Taofeek Aroyehun, Jason Angel, Navonil Majumder, Alexander Gelbukh, and Amir Hussain. 2021. Leveraging label hierarchy using transfer and multi-task learning: A case study on patent classification. *Neurocomputing*, 464, 421–431. DOI: 10.1016/j.neucom.2021.07.057.

[2] Mehmet Aydar and Serkan Ayvaz. 2019. An improved method of locality-sensitive hashing for scalable instance matching. *Knowledge and Information Systems*, 58, 2, 275–294. ISBN: 1011501811995. DOI: 10.1007/s10115-018-1199-5.

[3] Hamid Bekamiri, Daniel S. Hain, and Roman Jurowetzki. 2024. PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT. *Technological Forecasting and Social Change*, 206, June, 123536. DOI: 10.1016/j.techfore.2024.123536.

[4] Gianni Costa, Alfredo Cuzzocrea, Giuseppe Manco, and Riccardo Ortale. 2011. Data De-duplication : A Review Data De-duplication : A Review. *Learning structure and schemas from documents*, January. ISBN: 9783642229138. DOI: 10.1007/978-3-642-22913-8.

[5] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. 2003. Automated categorization in the international patent classification. *ACM SIGIR Forum*, 37, 1, 10–25. DOI: 10.1145/945546.945547.

[6] Juan Carlos Gomez and Marie Francine Moens. 2014. A survey of automated hierarchical classification of patents. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8830, 215–249. DOI: 10.1007/978-3-319-12511-4_11.

[7] Bikash Gyawali, Lucas Anastasiou, and Petr Knoth. 2020. Deduplication of scholarly documents using locality sensitive hashing and word embeddings. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association, 894–903.

[8] Arousha Haghighian Roudsari, Jafar Afshar, Wookey Lee, and Suan Lee. 2022. PatentNet: multi-label classification of patent documents using deep learning based language understanding. *Scientometrics*, 127, 1, 207–231. DOI: 10.1007/s11192-021-04179-4.

[9] Omid Jafari, Preeti Maurya, Parth Nagarkar, Khandker Mushfiqul Islam, and Chidambaram Crushev. 2021. A Survey on Locality Sensitive Hashing Algorithms and their Applications. *ACM Computing Surveys*. eprint: 2102.08942.

[10] Guik Jung, Junghoon Shin, and Sangjun Lee. 2023. Impact of preprocessing and word embedding on extreme multi-label patent classification tasks. *Applied Intelligence*, 53, 4, 4047–4062. DOI: 10.1007/s10489-022-03655-5.

[11] Eleni Kamateri, Michail Salampasis, and Eduardo Perez-Molina. 2024. Will AI solve the patent classification problem? *World Patent Information*, 78, June, 102294. DOI: 10.1016/j.wpi.2024.102294.

[12] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating Training Data Mitigates Privacy Risks in Language Models. In *International Conference on Machine Learning, Baltimore* number 1. Vol. 162, 10697–10707.

[13] Jong Wook Lee, Won Kyung Lee, and So Young Sohn. 2021. Patenting trends in biometric technology of the Big Five patent offices. *World Patent Information*, 65, March, 102040. DOI: 10.1016/j.wpi.2021.102040.

[14] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 8424–8445. eprint: 2107.06499. DOI: 10.18653/v1/2022.acl-long.577.

[15] Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117, 2, 721–744. ISBN: 1119201829. DOI: 10.1007/s11192-018-2905-5.

[16] Yinhan Liu et al. 2019. Roberta: a robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[18] Arousha Haghighian Roudsari, Jafar Afshar, Charles Cheolgi Lee, and Wookey Lee. 2020. Multi-label patent classification using attention-aware deep learning model. In *Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020*, 558–559. ISBN: 9781728160344. eprint: arXiv:1910.01108. DOI: 10.1109/BigComp48618.2020.000-2.

[19] Victor Sanh, L Debut, J Chaumond, and T Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arxiv 2019. *arXiv preprint arXiv:1910.01108*.

[20] Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, and Stuart M. Shieber. 2023. The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks* number NeurIPS, 1–39. eprint: 2207.04043.

[21] Christoph Trattner, Philipp Singer, Denis Helic, and Markus Strohmaier. 2012. Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *ACM International Conference Proceeding Series*, 0–7. ISBN: 9781450312424. DOI: 10.1145/2362456.2362474.

[22] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33, 5776–5788.

[23] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI '24) Article 303. Association for Computing Machinery, Honolulu, HI, USA, 21 pages. ISBN: 9798400703300. DOI: 10.1145/3613904.3641960.

[24] Junghwan Yun and Youngjung Geum. 2020. Automated classification of patents: A topic modeling approach. *Computers and Industrial Engineering*, 147, July, 106636. DOI: 10.1016/j.cie.2020.106636.

Enhancing causal graphs with domain knowledge: matching ontology concepts between ontologies and raw text data

Jernej Stegnar
Jožef Stefan Institute
Ljubljana, Slovenia
jernej.stegnar@gmail.com

Jože M. Rožanec
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
joze.rozanec@ijs.si

Gregor Leban
Event Registry d.o.o.
Ljubljana, Slovenia
gregor@eventregistry.org

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

When building a causal graph from textual sources, such as media reports, a key task is to provide an accurate semantic understanding of the causal variables encoded as nodes and to link them to existing ontologies with at least two purposes: (i) expand the knowledge with the domain knowledge captured in such ontologies and (ii) provide accurate and different levels of abstraction of the extracted causal variables. This article describes how we used OntoGPT, a tool for matching raw text to ontology concepts initially designed for the medical domain, to match concepts from media events to relevant ontologies. We build upon our previous work on extracting causal variables and enrich the extraction pipeline by matching causal variables to concepts from specific domain ontologies. In particular, we describe our work regarding the GEO ontology. Future work will focus on expanding OntoGPT's capabilities by utilizing a wider selection of ontologies. Addressing its limitations, such as dealing with multiple instances of the same class, will also be crucial for improving its utility. These improvements will allow the tool to better support strategic foresight applications by providing more detailed insights across a multitude of different sectors, further enriching causal graphs and facilitating more accurate predictive modeling.

KEYWORDS

strategic foresight, ontology matching, artificial intelligence

1 INTRODUCTION

Strategic foresight is a discipline concerned with anticipating future trends, uncertainties, and disruptions to inform decision-making and enable the creation of resilient, long-term strategies. As such, it is valuable to governments, organizations, and enterprises, who can use it to remain competitive and adaptable in a rapidly changing world [4].

The pace of technological advancement, shifting geopolitical landscapes, environmental crises, and unpredictable market trends make it essential to react quickly to change. Traditionally, foresight has been based on trend analysis, expert opinion, and qualitative insights. Such approaches lack the agility required to scan real-world events in near-real time and produce strategic foresight outcomes at such a pace. Nevertheless, this would be possible with the use of artificial intelligence.

AI enhances strategic foresight by automating the analysis of data and detecting patterns that may go unnoticed by human experts [1]. Machine learning algorithms can continuously monitor emerging trends, geopolitical shifts, and market fluctuations in near-real time, offering dynamic insights into potential future scenarios. Natural language processing (NLP) enables AI to sift through massive amounts of text, extracting relevant information from reports, news, and social media, thus accelerating the forecasting process. By integrating AI into strategic foresight, organizations can adapt more swiftly and make more informed, data-driven decisions in the face of uncertainty.

Ontologies provide structured knowledge informing the relationships between concepts within a specific domain. Furthermore, they describe those concepts through properties and can link such classes to specific instances observed in the real world. As such, they are of key importance when building a causality graph, given they can augment our understanding of the causal relationships between variables with a better understanding of the context and the variable implications [3]. For example, if the causal relationship reports about the ceasing of an armed conflict, knowing whether a causal variable relates to a country, the location of that country, the neighboring countries, and international organizations it is involved in would help to understand the magnitude of that event and contextualize other likely outcomes (refugee repatriation, impacts on investments, and others).

In the scope of the graph massive project, ontology matching is being used to link the extracted causal relationships from text to concepts inside the ontologies, allowing for a more detailed understanding of the concepts that appear in causal relationships and their interconnectivity.

2 ENRICHING CAUSAL GRAPHS WITH DOMAIN KNOWLEDGE

We consider ontologies a framework (an organized and structured system for representing knowledge) used to represent knowledge within a specific domain by defining the relationships between concepts. They consist of classes (concepts), properties (attributes), and relationships that connect different concepts. This structure provides a standardized way to organize and interpret data, ensuring consistent understanding across systems. For example, in a medical ontology, concepts like "disease" might be linked to "symptoms," "treatments," and "causes," each with its own defined properties. By formalizing these relationships, ontologies allow AI systems to better interpret and reason about

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.25

complex information, leading to more accurate data processing and decision-making.

Ontologies enhance causality graphs by providing domain-specific knowledge that improves the accuracy and depth of relationships represented. When extracting causal relationships from large datasets, such as media reports, the data can often be ambiguous or incomplete. Ontologies address this by offering structured knowledge that defines concepts and their relationships within a specific domain, linking extracted causal relationships to well-defined entities in the ontology. This enriches the causality graph, uncovering implicit connections and non-obvious relationships that may otherwise be missed. In strategic foresight, for example, ontology-based enrichment helps capture a broader range of potential future scenarios by incorporating knowledge beyond the immediate dataset. This leads to more reliable predictions, especially when the training data is limited or domain-specific. Ultimately, ontologies are expected to enable the system to generalize better, predict outcomes with higher accuracy, and improve the overall reliability of causality graphs.

The causality graph pipeline in the Graph Massivizer strategic foresight project is designed to automate the extraction, organization, and analysis of causal relationships from large datasets, particularly news articles. The Figure 1 showcases the structure of our causality graph's data pipeline. The process begins with extracting these relationships from news articles, which are then organized into a causality graph that maps the interactions between various factors and events. The goal is to develop link prediction models that estimate the likelihood of future events based on observed patterns. For instance, one use case involves predicting oil price trends by analyzing factors that influence pricing.

Ontology matching is then integrated into the pipeline to link extracted causal relationships with concepts from structured ontologies. This enrichment adds layers of context and enables the discovery of connections that may not be evident from raw data alone. By incorporating ontologies, the pipeline transcends the limitations of its training data, identifying causal relationships that may be implied by broader knowledge contained in the ontologies. This not only enhances the accuracy of the graph but also allows it to capture more complex and non-direct relationships, improving its predictive capabilities.

As shown in Fig. 1B, the process of ontology linking in our pipeline consisted of creating ontology matching templates, then linking the concepts in text to ontologies, using the information to add additional data to existing causalities, all with the purpose of finding extra implicit connections based on the information provided by the ontologies.

The main problem that needed solving for that purpose was, how to link ontologies to raw text data. In our case that was done using OntoGPT [2], which is a tool used for ontology linking. Another key challenge is inter-ontology matching, which involves linking multiple ontologies through shared concepts. This process expands the knowledge framework, making it even more valuable for our purposes. The challenge of inter-ontology matching hasn't been addressed yet and remains a matter of future work.

3 ONTOGPT: A BRIEF OVERVIEW

OntoGPT is an advanced tool that integrates large language models (LLMs) with ontologies to improve knowledge extraction and organization across various domains. Ontologies provide a

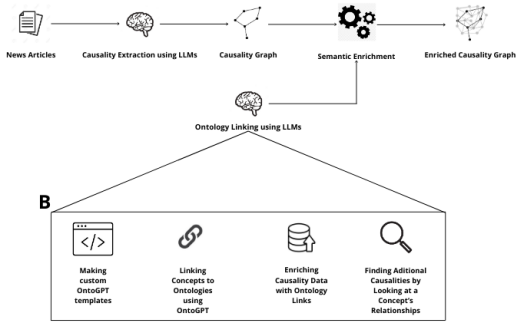


Figure 1: The figure showcases our pipeline for building a causality graph. The sub-figure B showcases how the process of ontology linking was executed as a part of our pipeline

consistent and accurate representation of complex information by defining structured relationships between concepts.

The primary purpose of OntoGPT is to enhance AI systems' understanding, processing, and categorization of data by linking extracted information to predefined concepts and relationships within an ontology. This structured approach ensures greater accuracy and reliability compared to traditional AI systems that rely on unstructured data.

OntoGPT works by connecting data from sources such as text or reports to specific concepts in an ontology, allowing for more informed and contextually accurate connections. For example, in healthcare, OntoGPT can link symptoms from patient records to diseases and treatments outlined in medical ontologies, helping to suggest possible diagnoses or treatment plans.

By combining the language-processing capabilities of LLMs with the structured knowledge available in ontologies, OntoGPT enables AI systems to go beyond keyword matching and consider the relationships between terms. This leads to more intelligent data interpretation and improved decision-making.

OntoGPT is widely used in fields where structured knowledge is critical for high accuracy, such as healthcare, biology, and pharmaceutical research. In medical research, for instance, OntoGPT links clinical trial data, medical records, and scientific literature to medical ontologies, supporting better analysis and decision-making.

The key advantage of OntoGPT lies in its ability to ground AI outputs in domain-specific, structured knowledge, reducing the likelihood of errors and improving the relevance of insights. This grounding ensures that AI responses are not just based on patterns but also on well-defined concepts and their relationships.

In summary, OntoGPT bridges the gap between the raw data-processing power of LLMs and the structured knowledge in ontologies. By leveraging both, it provides a more accurate and reliable approach to extracting and linking data across various domains, particularly when working with large, complex datasets.

3.1 OntoGPT's role

At a lower level, OntoGPT operates using YAML templates that define how data should be extracted from text and linked to ontological concepts. These templates serve as blueprints, specifying which types of entities, relationships, and properties to look for in the input text. The templates guide the large language model by mapping textual data to predefined concepts and relationships

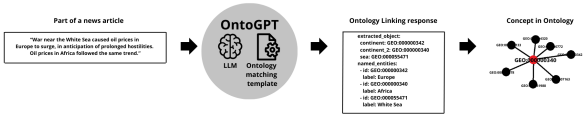


Figure 2: A Showcase of the function of OntoGPT

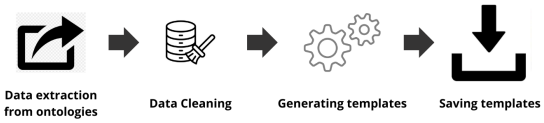


Figure 3: The Process of Templates Generation

from the ontology, ensuring that the extracted information is both relevant and structured. The figure 2 shows the process of ontology linking for an example of a simple sentence. Each YAML template contains detailed instructions on how to identify key terms, their corresponding ontology classes, and the relationships between them. This allows OntoGPT to recognize when a piece of text, such as a sentence from a media article, contains a concept that aligns with an entity or event in the ontology. Once identified, the tool links the extracted data to these ontology entries, enabling richer and more meaningful connections in the data, as it is now grounded in an established knowledge framework.

The approach described in this article uses an ontology file as input to create such templates for data extraction and linking. This enables for a broader range of ontology linking, as the templates can be created on demand.

4 TEMPLATES AND PYTHON CODE GENERATION

The approach works by using the information defined inside the ontology, to generate the YAML templates. The Figure 3 showcases the process of how this is done.

First the class information, for each class inside the ontology, is extracted. This is done by using the "owlready2" python library to parse the ontology into an object, and then extract the relevant information from the new object.

Every class inside the ontology is used to create a corresponding template class, which is optimal, as it covers all parts of the ontology that could potentially be linked. A small portion of the data extraction process is ontology-specific and was custom-tailored to the individual ontology, as some information (like class descriptions) is saved in different parts.

Secondly the data extracted from the ontology is processed and used to create custom YAML templates. This is done by simply using the extracted information to fill in a "general template" we used for generation. Specifically the class names and descriptions are used, to do so. This gives OntoGPT the names of the classes inside the ontology, that we are trying to link the text data to, and their descriptions, which assists OntoGPT in more accurately identifying these classes inside the text. The YAML file also contains the information of "annotators" which tells OntoGPT, which ontology to ground the responses to. The generated YAML templates are saved into a separate file after generation, which makes them ready for use.

The python code that is used by OntoGPT in the process of ontology linking, is similarly generated by using the extracted information to fill in the "general template" and is then saved to a separate file.

5 LIMITATIONS

5.1 Multiple Same-Class Concepts

OntoGPT has problems trying to link two or more concepts to a place in the ontology if the concepts are of the same class. This happens because both concepts suit the description and similar criteria that OntoGPT extracts the information based on. This causes OntoGPT to merge both concepts into a single string and then try to locate the said string inside the ontology, which fails because there is no individual inside the ontology class with such a name. An example of such a response is shown in Listing 1:

Listing 1: Example of a bad response

```
extracted_object:
  continent: AUTO:Europe%2C%20Africa
named_entities:
  - id: AUTO:Europe%2C%20Africa
    label: Europe, Africa
```

If OntoGPT managed to locate the concept inside the text in the ontology, it returns its id (an example of this is "sea: GEO:000055471" and "id: GEO:000055471 : White Sea") If the concept suits the class criteria, but couldn't be located inside the ontology, it returns it as a "AUTO" detection. For the purpose of ontology linking this is not optimal as it does not give us access to the additional information that is stored inside the ontology's individual information. The ontology's individual information is a set of predefined relationships and properties, that an individual concept has. For example, if the individual "Africa" is defined inside the ontology, the individual's data would include its size, countries on the continent, population, and climates, among others. This information gives us reliable information about a certain concept, allowing for more contextual understanding.

To solve this problem, the approach of creating "buffer" classes was taken, where a certain class from ontology would be used to generate three classes describing the different occurrences of the ontology class and a description that would provide sufficient context to OntoGPT to separate the same class concepts into different entities. The corrected response is showcased in Listing 2:

Listing 2: Example of a corrected response

```
extracted_object:
  continent: GEO:000000340
  continent_2: GEO:000000342
named_entities:
  - id: GEO:000000340
    label: Africa
  - id: GEO:000000342
    label: Europe
```

While this approach deals with a high percentage of this type problem, it does not cover the cases where more than three same-class concepts are inside the piece of text being analyzed.

6 CONCLUSIONS

Using OntoGPT in the Graph Massivizer strategic foresight project will prove valuable for enriching causal graphs with linked ontology data, aiming to improve predictive accuracy in predicting future events. Despite OntoGPT's initial focus on medical data, some custom adaptations were successfully implemented to suit a portion of different domains. However, limitations persist in distinguishing between multiple instances of the same concept class. These challenges highlight the need for further development to enhance the tool's versatility across a broader array of applications and ontologies.

ACKNOWLEDGMENTS

The Slovenian Research Agency supported this work. This research was developed as part of the Graph-Massivizer project funded under the Horizon Europe research and innovation program of the European Union under grant agreement 101093202.

REFERENCES

[1] Patrick Brandtner and Marius Mates. 2021. Artificial intelligence in strategic foresight–Current practices and future application potentials: current practices and future application potentials. In *Proceedings of the 2021 12th International Conference on E-business, Management and Economics*. 75–81.

[2] J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, et al. 2024. Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *Bioinformatics* 40, 3 (2024), btae104.

[3] Fatma Özcan, Chuan Lei, Abdul Quamar, and Vasilis Efthymiou. 2021. Semantic enrichment of data for AI applications. In *Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning*. 1–7.

[4] David Sarpong and Nicholas O'Regan. 2014. The Organizing Dimensions of Strategic Foresight in High-Velocity Environments. *Strategic Change* 23, 3-4 (2014), 125–132.

Measuring and Modeling CO_2 Emissions in Machine Learning Processes

Ivo Hrib
Jožef Stefan Institute
Ljubljana, Slovenia
ivo.hrib@gmail.com

Oleksandra Topal
Jožef Stefan Institute
Ljubljana, Slovenia
Oleksandra.Topal@ijs.si

Jan Šturm
Jožef Stefan Institute
Ljubljana, Slovenia
jan.sturm@ijs.si

Maja Škrjanc
Jožef Stefan Institute
Ljubljana, Slovenia
maja.skrjanc@ijs.si

Abstract

With the rapid expansion of the computing industry, efficient energy utilization and reduction of CO_2 emissions are critically important. This research develops analytical tools to predict CO_2 emissions from various machine learning processes. We present a novel methodology for data acquisition and analysis of CO_2 emissions during model training and testing. Our results demonstrate the environmental impact of different algorithms and provide insights into optimizing energy consumption in artificial intelligence applications.

Keywords

CO_2 Emissions, Machine Learning, Energy Consumption, Environmental Impact, AI Model Optimization, Green AI, Sustainable Computing, Carbon Footprint

1 Introduction

The global computing industry significantly contributes to CO_2 emissions, with data centers accounting for 2.5 to 3.7 percent of global greenhouse gas emissions [1]. These emissions exceed those of the aviation industry due to continuous operations and heavy reliance on fossil fuels [11]. Given the growing demand for artificial intelligence (AI) applications, there is an urgent need for CO_2-conscious solutions.

This research aims to develop tools for predicting CO_2 emissions associated with machine learning processes, thus enabling the reduction of the environmental impact of AI models. In collaboration with Eviden (Spain) and under the FAME EU project, we have developed a CO_2 emissions analysis system using tools like CodeCarbon [2] and eco2AI [3].

1.1 Research Goals

The primary goal of this research is to develop a service that predicts CO_2 emissions and power consumption of different machine learning models during both training and evaluation phases with emphasis on hyperparameter dependency. The CO_2 emissions are measured in kilograms per second ($\frac{kg}{s}$) , while the power consumption is measured in kilowatt-hours (kWh).

While existing services, such as CodeCarbon [2] or eco2AI [3], provide real-time measurement of emissions, they do not offer insights into a model's emissions before its construction or use. The service we aim to provide addresses this gap by offering an estimation of emissions and power consumption for different models before they are selected for specific use cases. This forward-looking approach allows for more informed decisions when choosing models, potentially reducing their environmental footprint.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.23

2 Related Work

The environmental impact of machine learning models has been a growing concern in recent years. Several studies have focused on quantifying and reducing the carbon footprint of artificial intelligence (AI) processes. For instance, [12] highlighted the energy consumption of training large neural models and suggested methods for minimizing emissions. Similarly, tools like CodeCarbon [2] and eco2AI [3] have emerged to measure real-time CO_2 emissions from computational tasks. However, these tools often lack predictive capabilities for assessing emissions before model selection, as pointed out by. Our work builds on these existing methodologies, concretely on the work of eco2AI[3], by providing a forward-looking approach that estimates emissions during the model selection phase, thus complementing real-time monitoring tools. This is achieved through heavy dependency on eco2AI[3] measuring systems for data collection, later used for modeling based on the collected data and registered hyperparameters.

2.1 Research Gap and Contribution

Despite the growing availability of tools like CodeCarbon [2] and eco2AI [3], a significant gap remains in the preemptive evaluation of environmental impact during the machine learning (ML) model selection phase. The mentioned tools are valuable for post hoc analyses but do not assist ML practitioners in making **informed decisions upfront**—before model development—on the environmental footprint of different model architectures or hyperparameters.

This gap is crucial, as the model selection phase often involves trial-and-error across multiple models and configurations, potentially leading to unnecessary resource consumption. Without predictive capabilities, practitioners have limited insight into which models will have the lowest environmental impact before engaging in resource-intensive training.

Our research aims to fill this gap by introducing a **predictive service** that estimates the environmental footprint of different ML models before they are trained or used. This service leverages the data collected from existing tools like eco2AI [3], incorporating key features such as hyperparameters, and model architecturre into predictive models. By doing so, we enable developers

to make **more sustainable choices** at the model selection stage, reducing carbon emissions from the start of the ML lifecycle.

The table 1 below presents a feature matrix comparing our proposed service with current tools, showing how our approach addresses unmet needs:

3 Methodology

Due to the lack of suitable data on CO_2 emissions of machine learning models, we began by developing an infrastructure for data collection. This infrastructure is composed of the following steps:

- **Dataset Generation:** Creating synthetic datasets using random data generation methods.
- **Data Preprocessing:** Cleaning and preparing the data for analysis.
- **CO_2 Emission Measurement:** Recording CO_2 emissions during both training and testing phases using different machine learning algorithms.
- **Feature Extraction:** Extracting relevant features such as project ID, experiment details, epoch duration, power consumption, and hardware configurations.
- **Adding Hyperparameters to Final Dataset:** Documenting hyperparameters used in each experiment to assess their impact on emissions.
- **Containerization:** Utilizing Docker for containerization to ensure reproducibility and scalability of the experiments.
- **Data Storage:** Storing all datasets, features, and emission records systematically in a database for further analysis.
- **Modeling:** Developing and training machine learning models to predict CO_2 emissions and power consumption.

The software implementation uses Python, with dependencies including pandas [7], scikit-learn [10], matplotlib [5], eco2AI [3], TensorFlow [**abadi2016tensorflow**], Keras [**chollet2015keras**], and Docker for containerization [**merkel2014docker**].

3.1 Dataset Generation

In this step, we created a synthetic dataset by generating random data points using tools like `sklearn.datasets.make_regression` or `make_classification`. The primary objective here is not to reflect real-world data scenarios but to produce a controlled environment where the focus is on measuring CO_2 emissions and power consumption during model training and evaluation. Datasets generated vary in size from ranges of 250 to 15000 samples and 5 to 2000 features. In classification cases additionally the number of classes ranges from 2 to 50. These parameter ranges were selected to mitigate the risk of computational overload, ensuring that the experiments remain feasible within the available computational resources while maintaining the integrity of the analysis.

3.2 Data Preprocessing

Before analysis, the dataset must be cleaned and prepared. This includes handling missing values, normalizing or standardizing data, encoding categorical variables, and splitting the data into training and testing sets. Proper preprocessing ensures that the data is in the optimal format for the models to learn from and minimizes biases that may affect model performance and emission measurements.

3.3 CO_2 Emission Measurement

We measure CO_2 emissions produced during both the training and testing phases of the machine learning models. This involves using tools like eco2AI [3] to track energy consumption and convert it into equivalent CO_2 emissions. The measurements are taken for various models, such as Decision Trees, Random Forests, Logistic Regression, and Neural Networks, to assess their environmental impact under different computational loads.

3.4 Feature Extraction

To gain deeper insights, we extract various features that could impact CO_2 emissions and energy consumption. These features include project identifiers, detailed descriptions of each experiment, the duration of each training epoch, power consumption metrics, hardware configurations (such as the type of CPU/GPU used), and hyperparameters. The project identifiers refer to unique alphanumeric codes assigned to each machine learning experiment upon execution. These identifiers help differentiate between various model configurations and experimental setups. They are generated and stored automatically by our system during the dataset generation process to ensure traceability and reproducibility of the experiments.

3.5 Adding Hyperparameters to Final Dataset

We document the hyperparameters used in each machine learning experiment, such as learning rates, batch sizes, and the number of layers in neural networks. This allows us to evaluate how these hyperparameters influence CO_2 emissions and energy consumption.

3.6 Containerization

To ensure reproducibility and scalability of our experiments, we employ Docker for containerization. This approach encapsulates the code, dependencies, and environment settings, allowing the experiments to be easily replicated and deployed across different platforms.

3.7 Data Storage

All datasets, extracted features, hyperparameter configurations, and CO_2 emission records are systematically stored in a database. This central repository facilitates efficient querying, retrieval, and analysis of data to support ongoing and future research.

3.8 Modeling

In this step, we develop and train machine learning models to predict CO_2 emissions and power consumption based on various features, such as the type of algorithm used, hardware configuration, and model parameters. This modeling allows us to estimate emissions for different machine learning workflows before their actual deployment. The models help identify the most efficient algorithms and configurations, thus guiding the selection of environmentally friendly AI solutions.

The general pipeline for the previously mentioned steps can be seen below (see Figure 1).

A more thorough view of the workings of this can be seen as shown below for running a single measurement (see Figure 2).

4 Model Architecture

In this section, we explain the architecture of the model used for predicting CO_2 emissions and power consumption based on

Tool/Technology	Platform Compatibility	Model Coverage	Metric Granularity	Carbon Metrics	Energy Metrics	Additional Features	Real-time measurement	Forward-looking Prediction
CodeCarbon	Cloud, On-Premise	All ML models	Per training session	CO_2 emissions (kg)	Energy consumption (kWh)	Dashboard Visualization	Yes	No
eco2AI	Cloud, On-Premise	All ML models	Per training session	CO_2 emissions (kg)	Energy consumption (kWh)	Not RAPL based	Yes	No
Proposed Service	On-Premise	Specific models (mentioned bellow)	Per model, per selection phase	CO_2 emissions $\frac{kg}{s}$	Energy consumption (kWh)	Predictive modeling	No	Yes

Table 1: Feature comparison of existing tools and the proposed service

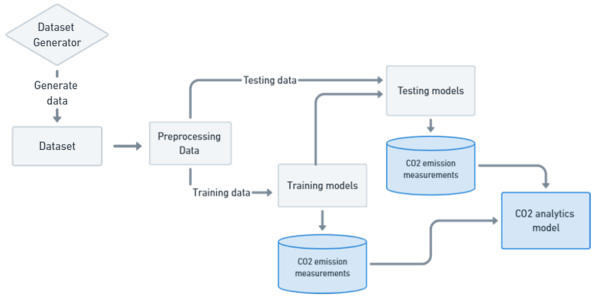


Figure 1: General Measurement Pipeline

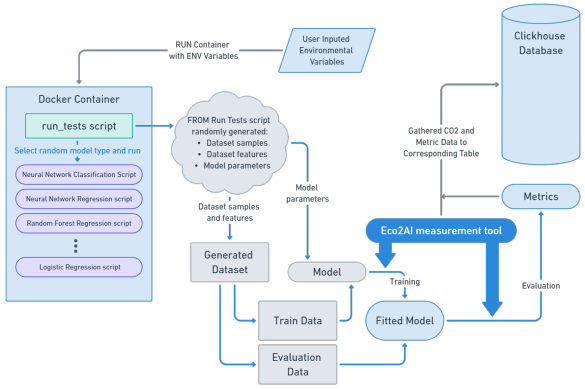


Figure 2: Single Model Measurement Pipeline

various features such as CPU type, GPU type, region, and other experiment-specific details. The model implementation is encapsulated within a Python class named `MultiModel`, which is responsible for managing the entire process from data preprocessing to training and prediction.

The model employs two separate neural networks for predicting CO_2 emissions and power consumption. The architecture for each neural network is as follows:

- **Input Layer:** Receives the scaled and encoded features.
- **Hidden Layers:** Consist of multiple Dense layers with ReLU activation functions. The CO_2 emissions model includes three hidden layers with 128, 64, and 128 neurons, respectively, while the power consumption model has three hidden layers with 64, 64, and 128 neurons.
- **Output Layer:** A single neuron that outputs the predicted value for either CO_2 emissions or power consumption.

4.1 Model Training

The model is compiled using the Adam optimizer [6] and the Mean Squared Error (MSE) loss function. Seeing as we were unable to gather adequate real-time environmental data of factors that may influence our predictions (e.g. Distribution of energy sources, real time CO_2 per kWh), our model relies on static yearly averages of these values[8] [9] . Our model uses the aforementioned features for the purpose of regression with the goal of predicting power consumption and CO_2 emissions gathered by previously mentioned random tests. Each model is trained for 25 epochs using the preprocessed data. After training, the models, along with their respective scalers and encoders, are saved to disk for later use.

4.2 Prediction

Once trained, the model can predict CO_2 emissions and power consumption for new data points by loading the appropriate model, scaler, and one-hot encoder. The input data is preprocessed in the same manner as during training, and the predictions are obtained by applying the trained models.

This modular approach allows for easy extension to additional models or data sources and provides a scalable solution for analyzing the environmental impact of machine learning processes.

5 Web Application Interface for CO_2 Emissions and Power Consumption Prediction

In addition to the backend model developed for predicting CO_2 emissions and power consumption of various AI models, a web application was created to provide a user-friendly interface for real-time predictions. The web app, as shown in Figure 3, allows users to select different machine learning models and configure parameters to estimate the associated environmental impacts.

5.1 Key Features of the Web Application

The web application interface is designed with simplicity and functionality in mind. It includes several key components:

- **Model Selection:** Users can choose the type of machine learning model they are interested in evaluating (e.g., Logistic Regression (abbr. LogR), Decision Tree Classifier (abbr. DTC), Decision Tree Regression (abbr. DTR), Neural

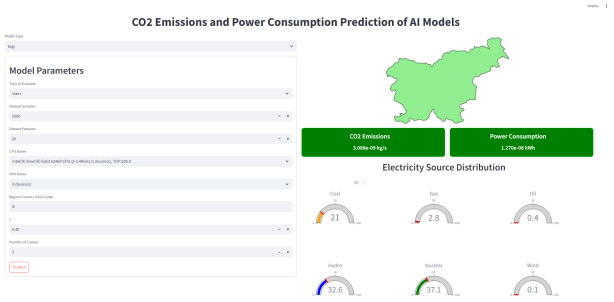


Figure 3: Web App Interface

Network Classifier (abbr. NNC), Neural Network Regression (abbr. NNR), Linear Regression (abbr. LinR), Random Forest Classifier (abbr. RFC) and Random Forest Regression (abbr. RFR)). The dropdown menu in the upper-left corner of the interface provides a list of available models.

- **Model Parameters Configuration:** A section labeled "Model Parameters" allows users to specify various inputs:
  - **Train or Evaluate:** Users can choose whether to estimate emissions for the training or evaluation phase of the model.
  - **Dataset Samples and Features:** Input fields are provided for users to define the size of the dataset in terms of the number of samples and features.
  - **CPU and GPU Specifications:** The app allows the selection of the CPU and GPU type, reflecting different hardware configurations, such as "Intel(R) Xeon(R) Gold 6246R CPU @ 3.40GHz/1 device(s), TDP:205.0" or "AMD Ryzen 7 4800H with Radeon Graphics/1 device(s), TDP:45.0".
  - **Region/Country Selection:** A dropdown to select the geographic location where the model is being executed, which influences the CO_2 emissions based on local energy sources.
- **Real-Time Predictions:** Once all parameters are configured, the application dynamically calculates and displays:
  - **CO_2 Emissions:** The predicted emissions are shown in kilograms per second (kg/s).
  - **Power Consumption:** The power consumption is provided in kilowatt-hours (kWh).
- **Electricity Source Distribution:** A graphical representation is provided for the distribution of electricity sources, such as coal, gas, and oil, in the selected region. This information is crucial for understanding the environmental impact of power consumption based on the local energy mix.

5.2 User Experience and Accessibility

The web application is developed with accessibility in mind, ensuring that users, regardless of technical background, can interact with the model's predictive capabilities. By offering a clear and intuitive interface, it aims to make the process of estimating CO_2 emissions and power consumption transparent and straightforward.

Figure 3 illustrates the application's main screen, where the model type, parameters, and results are all visible at a glance. This real-time feedback loop allows users to make informed decisions based on the predicted environmental impact.

6 Results

6.1 Model Error

To evaluate the performance and accuracy of the models, we conducted a 10-fold cross-validation to estimate the errors in predicting CO_2 emissions and power consumption. The results are presented in Table 2. The errors for both CO_2 emissions and power consumption were computed for both training and evaluation phases of each model type.

Note: In this context, "Train." refers not to the error on the training set, but rather to the error made by our model in predicting the CO_2 emissions / Power Consumption during the training phase of the listed model. Similarly, "Eval." refers not to the error on the evaluation set, but rather to the error made by our model in predicting the CO_2 emissions / Power Consumption when the listed model makes predictions. This distinction is crucial to understanding the results accurately.

Table 2: Model Scaled Error Estimates from 10-Fold Cross-Validation

Model	Phase	CO_2 Error	Power Error
DTC	Eval.	0.0036	0.0043
DTC	Train.	0.0631	0.0649
DTR	Eval.	0.0032	0.0034
DTR	Train.	0.0133	0.0517
RFC	Eval.	0.0094	0.0098
RFC	Train.	0.3242	0.3582
RFR	Eval.	0.0087	0.0081
RFR	Train.	0.2565	0.2779
LogR	Eval.	0.0063	0.0057
LogR	Train.	0.0055	0.0043
LinR	Eval.	0.0099	0.0105
LinR	Train.	0.0104	0.0095
NNC	Eval.	0.0018	0.0030
NNC	Train.	0.1083	0.1216
NNR	Eval.	0.0045	0.0112
NNR	Train.	0.1051	0.1008

Based on the results obtained through the 10-fold cross-validation, it is evident that the model performance varies significantly across different algorithms and phases. One notable observation is that the errors in predicting CO_2 emissions and power consumption are relatively higher during the training phases, particularly for more complex models like Neural Networks and Random Forests [4].

This discrepancy in model performance can be attributed to the sparsity of the data collected during the measurement phase. The limited data points lead to substantial gaps in the attribute space covered by the models, resulting in erratic behavior when predicting outside these ranges. Consequently, the models show diminished accuracy and reliability when confronted with input configurations that fall beyond the scope of the original data.

Future research should focus on enhancing the robustness of these models by expanding the dataset to include a broader range of scenarios and conditions. This would help mitigate the effects of sparsity and improve the model's generalizability, ensuring more reliable predictions across diverse settings.

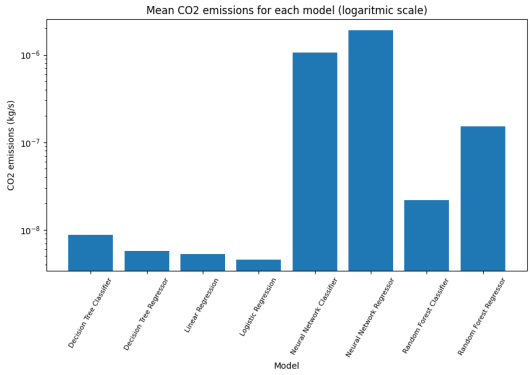


Figure 4: Logarithmically scaled mean emissions across different models

6.2 CO_2 Emission Analysis Across Different Models

Figure 4 provides a comparative analysis of the mean CO_2 emissions generated by different machine learning models during their operation, represented on a logarithmic scale to accommodate the wide range of emission values.

The chart highlights significant variations in CO_2 emissions among models, with the Neural Network Classifier and Neural Network Regressor exhibiting the highest emissions by a considerable margin. This is expected due to the intensive computational requirements and numerous parameters these models necessitate, resulting in elevated power consumption and consequently higher CO_2 output.

In contrast, simpler models like Logistic Regression, Linear Regression, and Decision Tree models show substantially lower CO_2 emissions, reflecting their reduced computational complexity and lower resource demand.

Interestingly, the Random Forest models, particularly the Regressor, present moderate emissions, illustrating that even ensemble methods, which typically involve training multiple decision trees, can maintain reasonable emission levels depending on their configuration.

This analysis underscores the importance of model selection not only for performance but also for minimizing environmental impact, particularly when scaling up operations or deploying in resource-constrained settings.

7 Discussion

The results highlight the significant environmental impact of training complex AI models, particularly neural networks. The variability in emissions suggests that optimizing model hyperparameters and selecting appropriate hardware configurations can reduce CO_2 output.

Future research should focus on model improvement for better and more accurate prediction, expanding the range of algorithms studied, as well as intensive data collection to accommodate gaps in training data.

8 Limitations

This study presents several limitations, particularly regarding the data, model evaluation, and hardware configurations, which must be considered when interpreting the results.

8.1 Training Duration and Model Learning

The models were trained for a fixed number of epochs (e.g., 10 or 20), prioritizing computational cost over learning performance. The focus was on estimating CO_2 emissions rather than model accuracy or convergence, meaning the models may not have fully captured patterns in the data. As such, the reported emissions reflect standardized training durations (with an upper limit for computational efficiency), not optimized learning outcomes.

8.2 Lack of Meaningful Learning Objective

The use of randomly generated data limits the evaluation of model learning. Since the data lacked inherent structure, the models' ability to learn was not assessed. Instead, the models were primarily evaluated on their resource consumption during training, reducing the focus on generalization or predictive power.

8.3 Hardware and Software Considerations

The experiments were conducted on specific hardware (e.g., GPU/CPU configurations), and variations in hardware were not examined. Different hardware setups, especially energy-efficient systems, could significantly impact CO_2 emissions and energy consumption. Therefore, the findings may not generalize across all hardware environments. However, we would like to point out that this was due to lack of infrastructure for broader experimentation.

9 Future Work

Future research should incorporate real-world datasets, optimize hyperparameters, and evaluate diverse hardware configurations to extend these findings to broader machine learning scenarios. The exploration of more complex architectures and learning objectives will provide a deeper understanding of the trade-offs between performance and environmental impact.

10 Conclusion

Our study presents a methodology for monitoring and analyzing CO_2 emissions during machine learning processes. The findings demonstrate that different machine learning models exhibit significant variability in their energy consumption and CO_2 emissions, with complex models like neural networks having a higher environmental impact. By providing predictive insights into these emissions, our approach enables more informed decision-making during model selection, thus contributing to the broader goal of reducing the carbon footprint of AI applications.

Future work will focus on expanding the dataset to include more diverse models and configurations. Additionally, we plan to integrate real-time monitoring tools to compare predictions with actual emissions, further refining our predictive capabilities. Moreover, optimizing model hyperparameters and exploring alternative, more sustainable hardware configurations will be key areas of investigation for minimizing the environmental impact of machine learning workflows.

Acknowledgements

This work was supported by the FAME project, funded by the European Union's Horizon 2023 Research and Innovation Programme under grant agreement nº 101092639.

References

[1] Climatiq. 2023. Climatiq: emissions intelligence platform. Provides data on the carbon emissions of various activities, including computing. (2023). https://www.climatiq.io/.

[2] CodeCarbon Development Team. 2023. Codecarbon: an open source tool for tracking the carbon emissions of machine learning experiments. An open-source tool for tracking and reducing carbon emissions in machine learning models. (2023). https://github.com/mlco2/codecarbon.

[3] Eco2AI Development Team. 2023. Eco2ai: real-time co2 emission tracking for machine learning. A tool for real-time tracking of CO2 emissions during machine learning processes. (2023). https://github.com/sb-ai-lab/Eco2AI.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. A comprehensive resource on machine learning algorithms, including neural networks. MIT Press. http://www.deeplearningbook.org.

[5] John D Hunter. 2007. Matplotlib: a 2d graphics environment. *Computing in science & engineering*, 9, 3, 90–95.

[6] Diederik P Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[7] Wes McKinney. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*. Vol. 445, 51–56.

[8] OurWorldInData. [n. d.] (). https://ourworldindata.org/grapher/carbon-intensity-electricity?tab=table.

[9] OurWorldInData. [n. d.] (). https://ourworldindata.org/electricity-mix.

[10] Fabian Pedregosa et al. 2011. Scikit-learn: machine learning in python. *Journal of machine learning research*, 12, 2825–2830.

[11] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *arXiv preprint arXiv:1907.10597*. 12 pages. DOI: 10.48550/arXiv.1907.10597.

[12] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 3645–3650.

Enhancing Ontology Engineering with LLMs: From Search to Active Learning Extensions

Ganna Kholmska
Jožef Stefan Institute
Ljubljana, Slovenia
anna.kholmska@gmail.com

Klemen Kenda
Jožef Stefan Institute
Ljubljana, Slovenia
klemen.kenda@ijs.si

Joze Rozanec
Jožef Stefan Institute
Ljubljana, Slovenia
joze.rozanec@ijs.si

Abstract

This paper explores the use of LLMs in ontology engineering within the HumAIne project, focusing on the discovery, analysis, and extension of ontologies in Data Mining, Machine Learning, and manufacturing. The methodology leverages fine-tuned prompts and combines LLMs with traditional tools like Protege for validation. A multi-LLM approach improved domain-specific concept coverage and reduced errors, though challenges remain in addressing deep domain-specific gaps and ensuring logical consistency.

Keywords

LLMs, Ontology Engineering, Active Learning, Data Mining, Machine Learning, Ontology Selection, Ontology Extension

1 Introduction

The HumAIne project, funded by the European Commission under the Horizon Europe program, aims to develop a platform integrating advanced AI paradigms such as Active Learning (AL), Neuro-Symbolic AI, Swarm Learning, and Explainable AI. This platform is designed to enhance human-AI collaboration in dynamic, unstructured environments, with applications spanning healthcare, manufacturing, finance, energy grids, and smart cities. Its primary goal is to support decision-making by combining human expertise with AI capabilities.

One of the project's key challenges is developing multiple ontologies that provide a structured framework for integrating domain-specific knowledge. This framework is essential for enhancing the clarity and reliability of AI-driven decisions, while ensuring adaptability across diverse applications. To construct these ontologies, we first explored publicly available ontologies relevant to the project's scope, then extended selected ones with concepts from HumAIne's AI paradigms, starting with Active Learning

However, manual ontology construction is a complex, resource-intensive process that requires expertise across multiple domains, collaboration among stakeholders. Ensuring modularity, reusability, and scalability adds to this complexity.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.28

Recent studies show that leveraging Large Language Models (LLMs) can streamline ontology construction by reducing manual effort and improving consistency and quality. For instance, [1] demonstrates semi-automatic knowledge graph construction using open-source LLMs, while [2] proposes methods for automatic concept hierarchy generation through LLM queries. Building on this research, this paper contributes a methodology that integrates LLMs with traditional tools like Protege to streamline the discovery, analysis, and extension of ontologies. By employing a multi-LLM approach, we address challenges in domain-specific concept identification and ensure more consistent, accurate results in ontology development for fields like Data Mining, Machine Learning, and manufacturing.

2 LLM-Assisted Search and Analysis of Domain Ontologies

Our experimentation with methodologies and tools for efficient web search and ontology analysis in Data Mining (DM), Machine Learning (ML), and manufacturing domains led to the development of the LLM-leveraging algorithm shown in Fig. 1. This algorithm uses carefully crafted prompts to guide LLMs in generating accurate, targeted queries. Before each step, the initial prompt is optimized through several iterations in a dialogue with the LLM to improve accuracy and relevance. Further details on the iterative query refinement process are provided in the Discussion section.

Step 1: Define the Search Objective. At this stage, LLMs like Bing Chat, Google's Bard, or ChatGPT with Web Browsing are employed to iteratively refine the search objectives initially formulated by the researcher, along with relevant keywords, phrases, and terms describing the ontologies or concepts of interest. For instance, our initial search objective for DM and ML ontologies was to "Find ontologies that offer up-to-date, detailed descriptions of the DM and ML domains, following best practices in ontology engineering." Keywords included "Active Learning" and "CRISP-DM standard."

Step 2: Formulate Search Queries Using LLMs. Based on the refined search objectives and keywords, and using a carefully crafted prompt, LLMs generate targeted search queries. These queries are fine-tuned through feedback or early search results to maximize relevance and accuracy. For example, for a DM ontology, the LLM generated queries such as "Data Mining ontology for semi-supervised machine learning," which were further refined before finalizing the query.

Step 3: Conduct Web Search. This step involves real-time browsing tools like Copilot in Microsoft Edge (GPT-4) and Perplexity AI to execute searches and identify relevant sources. Our study prioritized high-quality sources like ontology

repositories (e.g., BioPortal, OBO Foundry) and academic databases (Google Scholar, IEEE Xplore, ACM Digital Library).

It is important to acknowledge that LLM-driven web searches are generally confined to public repositories and a limited range of academic databases. As a result, proprietary or lesser-indexed ontologies may require manual exploration to ensure a more thorough search.

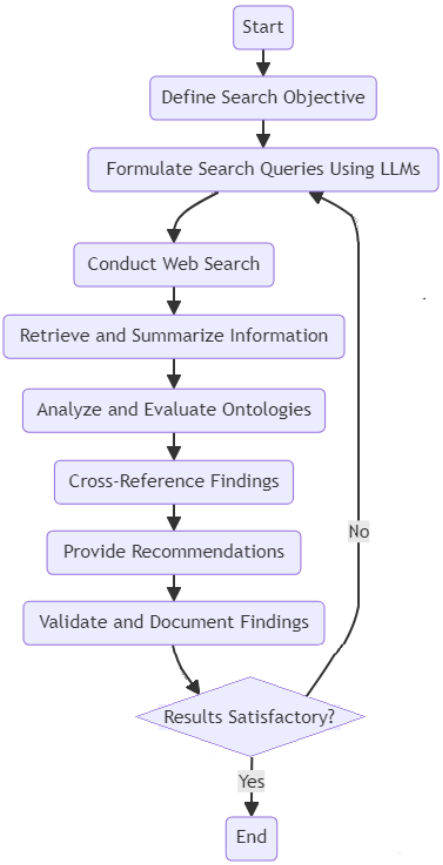


Figure 1: Key steps of LLM-leveraging algorithm

Step 4: Retrieve and Summarize Information. LLMs (Google Bard, Copilot (GPT-4), Perplexity AI) were employed to extract and summarize key information from ontology descriptions found in publications, technical papers, and repository documentation identified during the search. Using a specifically tuned prompt, LLMs extracted 11 characteristics for each of the 34 identified DM and ML ontologies. These characteristics included purpose, availability, ontology metrics, reused ontologies, software editors, representation language, and evaluation methodologies. This structured data, organized in table format, provided valuable insights into each ontology's scope, quality, and reusability. From these results, we selected 6 ontologies for further exploration, prioritizing comprehensive coverage of DM and ML concepts, adherence to ontology engineering best practices, and alignment with established standards in these domains.

Step 5: Analyze and evaluate ontologies. LLMs were further utilized to assess the relevance, content, and structure of the selected ontologies. In our study of DM and ML ontologies, LLMs such as GPT-4, which can process, explain, and generate OWL and RDF code, were used alongside ontology tools like Protege. This combination ensured that the ontologies addressed relevant concepts and aligned with frameworks like CRISP-DM. GPT-4 helped significantly in bridging the gap between textual descriptions and formal ontology representations.

Step 6: Cross-Reference and Compare Findings. LLMs with contextual understanding were employed to integrate and refine information from multiple sources. For this task, we used. Additionally, ChatGPT (GPT-4) categorized 65 manufacturing ontologies categorized 65 manufacturing ontologies, assessing them for relevance to process planning, standardization, industry adoption, interoperability, and support for advanced manufacturing concepts. Further exploration of the top 8 LLM-scored ontologies showed strong alignment with expert evaluations, but domain-specific tasks required carefully crafted prompts and human oversight for effectiveness.

Step 7: Provide Recommendations for Further Exploration. LLMs generated recommendations for the most suitable ontologies or areas for additional research based on the previous step's results. This includes identifying underexplored concepts and areas needing further investigation.

Step 8: Validate and Document Findings. The findings were manually validated for accuracy and relevance, then systematically documented. ChatGPT (GPT-4) was used to summarize and structure the documentation.

Step 9: Iterate and Refine Search (if needed). When results were too broad or irrelevant (e.g., Active Learning misinterpreted as an educational method), we refined the search prompt by adding more context.

By using this LLM-based algorithm, we conducted comprehensive web searches and extracted relevant information to identify the most suitable ontologies for the HumAIne project. In the DM and ML domains, we selected the OntoDM suite (OntoDM-Core, Onto-KDD, and OntoDT). For the manufacturing domain, we identified the Industrial Ontologies Foundry Core (IOF Core) as the best fit.

3 LLM-Assisted Ontology Extension with Active Learning Concepts

Integrating Active Learning (AL) into an ontology requires extending it with new classes, properties, and relationships representing key AL concepts. While traditional methods of building and extending ontologies are well-documented, we leveraged GPT-4 for this task using iteratively refined prompts (see Discussion section). This section outlines how LLMs, particularly GPT-4, were used to extend the IOF Core ontology with AL concepts.

Step 1: Define the Problem and Objectives. Through iteratively refined prompts, LLMs formulated clear objectives, specifying the domain (e.g., manufacturing) and key concepts (e.g., Active Learning). These outputs were used to guide further steps, with LLMs leveraging contextual understanding, knowledge synthesis, and language generation to suggest relevant AL applications such as adaptive scheduling. Queries like "How can Active Learning improve adaptive scheduling in manufacturing?" generated valuable insights into potential use cases. where AL would be most beneficial.

Step 2: Analyze the Ontology to be Extended. By combining Protege's visualization and navigation tools with GPT-4's ability

to process textual and machine-readable data (e.g., OWL/RDF), we thoroughly examined the IOF Core ontology structure and identified areas for introducing AL concepts. For example, GPT-4 helped uncover key classes like “Process,” “Resource,” and “PerformanceMetric” within IOF Core, highlighting relevant properties for AL integration. Queries such as "What aspects of IOF Core can benefit from AL integration?" and "What key concepts are missing from the IOF Core ontology for integrating Active Learning in manufacturing?" guided us in identifying areas for improvement, including handling uncertainty and adjusting dynamic processes.

Step 3: Identify Active Learning Concepts. The main tasks of this step and the role of LLMs in supporting each task are summarized in the Table 1:

Table 1: LLMs applications for Identifying AL Concepts

Task	LLM Application	Example Output
1. Identify fundamental AL concepts	Use LLMs to generate a list of core AL strategies and techniques	Concepts like“Uncertainty sampling,” “Query-by-committee”
2. Extract domain-specific AL concepts	Query LLMs about AL in specific industrial contexts	Concept like "Query Efficiency" in decision-making for manufacturing
3. Mine AL concepts from literature	Process academic papers, reports to extract relevant AL terms	Concepts like “Stream-base selective sampling” from papers on AL in manufacturing
4. Assign properties to new classes	Generate properties for AL ontology classes	QueryStrategy class properties: “hasuncertaintySampling” “queryByCommittee”
5. Refine and validate terminology	Ensure definitions, resolve overlaps	Refined and validated terms based on domain-specific standards

By prompting, LLMs generated nearly 200 fundamental AL concepts, structuring them into a hierarchy by leveraging their vast training data. Additionally, LLMs helped generate definitions, assisting in verifying and refining concepts. However, after a point, LLMs began repeating concepts or producing less relevant terms. LLMs were also effective in generating domain-specific concepts through targeted queries. For instance, querying AL in manufacturing led to concepts like "uncertainty management" and "query efficiency." More specialized concepts required extraction from academic papers, which were cross-referenced with existing standards in DM, ML, and manufacturing (e.g., CRISP-DM, IEEE 7000 Series, ISA-95, ISO 15531). Ontology learning tools like Text2Onto and OntoLearn were combined with LLMs for cross-verification.

*Step 4: Develop Ontology Extensions.*LLMs helped create AL-related classes, properties, and relationships based on the identified concepts, using OWL-compliant syntax (see Fig. 2). By combining GPT-4’s knowledge synthesis with Protege’s structural reasoning and consistency checking, we improved the efficiency and accuracy of reviewing, debugging, and validating OWL code.

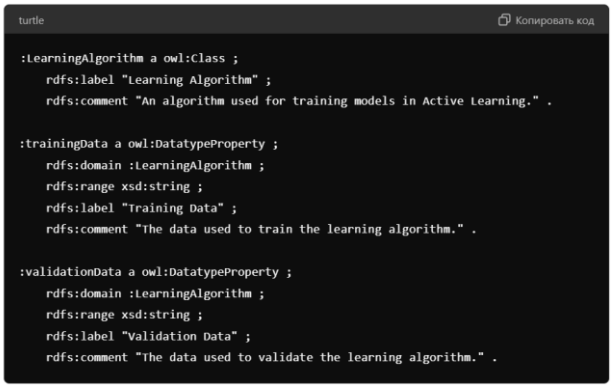


Figure 2: Screenshot of LLM-generated code defining the “LearningAlgorithm” class with properties “trainingData” and “validationData”

Step 5: Ensure Semantic Consistency. LLMs, such as GPT-4, assisted in ensuring semantic consistency by reviewing new and existing ontology elements and suggesting how new concepts could align with the existing framework. For example, an LLM suggested how an AL “QueryStrategy” class fits within the IOF Core ontology.

Example Prompt: " Review the new QueryStrategy class and suggest how it can align with the existing classes in IOF Core."

LLM Output: The QueryStrategy class aligns with decision-making aspects of the Process concept. Strategies such as “UncertaintySampling,” “QueryByCommittee,” “ExpectedModelChange,” and “ExpectedErrorReduction” can be viewed as specialized decision-making processes within the broader process framework of IOF Core.

However, LLMs cannot guarantee logical consistency and face limitations in handling complex relationships, making it necessary to use ontology reasoners, such as Protege or HermiT, to perform consistency checks.

Step 6: Map to Existing Ontologies. LLMs, such as GPT-4, assist in generating initial mapping suggestions by analyzing similarities in definitions, relationships, and properties between new and existing concepts. This involves creating explicit relationships like “owl:sameAs,” “owl:equivalentClass”, and “owl:equivalentProperty”.

Example LLM Output:

:FeedbackMechanism a owl:Class ;
owl:equivalentClass :ControlSystem ;
rdfs:label "Feedback Mechanism" ;
rdfs:comment "Mechanisms that provide feedback in Active Learning to control systems."

While LLMs are effective in identifying high-level similarities, they may face challenges with complex or domain-specific relationships, requiring further refinement. Although we didn’t encounter these issues during our initial work extending IOF Core with AL concepts, we used Protege’s alignment plug-ins to refine LLM-generated mappings. For more complex mappings, tools like AgreementMaker or COMA can further refine the suggestions.

Step 7: Prototype and Test. LLMs, such as GPT-4, were prompted to generate validation scenarios, competency questions, and SPARQL queries based on the integrated AL concepts. For instance, a prompt like "Suggest validation scenarios for adaptive scheduling with Active Learning" helped us produce realistic test cases, including prototype code, descriptions of initial setup, process flows, validation steps, and queries based on newly integrated concepts.

SPARQL queries generated by LLMs were executed using Protege with SPARQL plugins to assess the ontology's ability to retrieve relevant information and answer competency questions.

However, some LLM-generated scenarios revealed limitations in domain-specific knowledge, resulting in generic outputs that required refinement. Additionally, LLMs struggled with modeling intricate relationships or complex data retrieval conditions, making human oversight essential for ensuring accuracy and thorough testing.

Step 8: Iterative Refinement. Following initial prototyping and testing, we gathered feedback from domain experts and users to further refine the ontology. Validation reports were uploaded to AskPDF Research Assistant (GPT-4), where LLMs reviewed the reports, extracted key improvement suggestions, and refined task lists. The LLM provided insights into areas where ontology relationships or properties required adjustments and identified additional concepts that might have been overlooked.

Step 9: Document and Disseminate. LLMs like ChatGPT or Bard were instrumental in generating comprehensive documentation, including details on the ontology extensions. Additionally, LLMs contributed to drafting technical reports and research papers.

Using this methodology, we successfully extended the IOF Core ontology with Active Learning (AL) concepts. Future stages of the HumAIne project will focus on further validation and refinement, particularly during pilot case implementations.

4 Discussion

This study highlights LLMs' potential in ontology engineering by reducing manual effort and increasing efficiency. LLMs rapidly identified key ontologies like OntoDM and IOF Core and generated structured classes, properties, and relationships, reducing the need for manual OWL/RDF code generation and concept mapping. However, LLMs face challenges in domain-specific precision, requiring human oversight to refine outputs and address nuances in specialized fields. While tools like Protege excel at ensuring logical consistency, LLMs offer dynamic capabilities for generating new concepts and relationships. Despite these advantages, traditional tools like AgreementMaker and COMA are still necessary to refine and validate LLM-generated mappings.

One strategy to mitigate LLM limitations was iterative prompt engineering. We refined prompts for ontology search and extension tasks through multiple cycles of improvement. These cycles, with LLMs like GPT-4, involved clarifying questions, refining queries, and generating more focused outputs. Initial prompt for starting the cycle can be the following:

"Your role is my Prompt Creator. Your goal is to craft the best possible prompt for my needs. The prompt will be used by you, [LLM's name]. I want to write about: [keyword/topic]. Based on my input, you will now generate 3 sections. a) Revised prompt (clear, concise, and easily understood by you), b) Suggestions (on what details to include in the prompt to improve it), and c) Questions (ask any relevant questions to improve the prompt). We will continue this iterative process with me providing additional information to you and you updating the prompt until it's complete."

After 4-5 cycles, the prompts were highly optimized, ensuring relevant outputs. This refinement process reduced inconsistencies and improved LLM-generated content across both search and extension phases.

We integrated multiple LLMs, including Bing Chat (GPT-4), Google's Bard, and Perplexity AI, to cross-validate outputs, reducing errors and refining results. This ensured consistency in LLM-generated ontologies and mappings.

To evaluate this multi-LLM approach, we propose the following metrics: Inter-Model Consistency (measures alignment between LLM outputs). Error Rate Reduction (Tracks how often one LLM corrects another's errors),.Coverage of Relevant Concepts (assesses LLMs' ability to capture domain-specific concepts). Although these metrics provide a framework, formal measurements are yet to be implemented.

Future stages will involve applying these metrics to validate outputs and testing extended ontologies in real-world applications. This hybrid method combines LLMs and traditional tools, ensuring both efficiency and accuracy in scalable ontology development.

5 Conclusions

This study demonstrates how LLMs can streamline ontology engineering by automating the search, analysis, and extension of domain-specific ontologies. Leveraging multiple LLMs, we successfully identified and extended key ontologies, including OntoDM and IOF Core, for the HumAIne project, improving efficiency in generating classes, properties, and relationships.

While LLMs significantly enhance the process, they face challenges in domain-specific precision and require human oversight, particularly for complex relationships. Traditional tools like Protege and ontology reasoners remain critical for ensuring logical consistency and validation.

Future work will focus on refining these extended ontologies through real-world pilot tests and applying evaluation metrics to LLM-generated outputs. This hybrid approach, combining LLM automation with traditional validation tools, offers a scalable solution that balances efficiency with the need for human expertise.

Acknowledgments

This work was supported by the European Commission under the Horizon Europe project HumAIne, Grant Agreement No. 101120218.

References

[1] Kommineni, Vamsi Krishna, Birgitta König-Ries and Sheeba Samuel. "From human experts to machines: An LLM supported approach to ontology and knowledge graph construction." *ArXiv* abs/2403.08345 (2024): n. pag. DOI: https://doi.org/10.48550/arXiv.2403.08345

[2] Funk, Maurice, Simon Hosemann, Jean Christoph Jung and Carsten Lutz. "Towards Ontology Construction with Language Models." ArXiv abs/2309.09898 (2023): DOI: https://doi.org/10.48550/arXiv.2309.09898

On the Brazilian Observatory for Artificial Intelligence

Rafael Meira Silva, Cristina Godoy Oliveira
CIAAM, C4AI, Univ. of São Paulo
São Paulo, Brazil
rafael@meirasilva.com.br
cristinagodoy@usp.br

Luiz Costa, Alexandre Barbosa
CETIC, OBIA
São Paulo, Brazil
tuca@nic.br
alexandre@nic.br

Joao Paulo Candia Vieira
CIAAM, C4AI, Univ. of São Paulo
São Paulo, Brazil
candia@usp.br

Joao Pita Costa*
IRCAI, Quintelligence
Ljubljana, Slovenia
Joao.pitacosta@quintelligence.com

ABSTRACT

Artificial Artificial Intelligence (AI) is rapidly transforming industries and economies worldwide, with Brazil and South America emerging as significant players in this global shift. The fundamental need to monitor the impact of artificial intelligence (AI) in the verticals for sustainable development, government engagement, investment and society at large motivated the Brazilian Artificial Intelligence Observatory (OBIA). It is also an integral part of the Brazilian Artificial Intelligence Plan (PBIA), and a former objective of the Brazilian Strategy of AI aims to become the leading platform for monitoring the uses of AI in the country. OBIA is part of Axis 5 of the PBIA focused on supporting the regulatory and governance process of AI. This research paper explores the current state, challenges, and potential of AI development in the region, examining how technological advancements are influencing economic growth, societal change, and policy-making across South America, with a particular focus on Brazil as a leading hub of innovation. It is also investigating common aspects of the research agendas as with IRCAI's SDG Observatory, particularly in what regards machine learning workflows and approaches complementing traditional and crowdsourced heterogeneous data collection and analysis.

KEYWORDS

Artificial Intelligence, Observatory, Survey Data Analysis, Complex Data Visualization, Multidisciplinary Collaboration.

1 Introduction

AI is increasingly shaping the economic landscape and societal dynamics across Brazil and South America, positioning the region as a growing hub for technological innovation. Despite challenges such as uneven infrastructure and regulatory hurdles, Brazil is making significant strides in AI research and development, contributing to the regulation and better understanding of the impact of AI in South America. OBIA [5] is answering this need, serving as a platform to support the strategy and other government actions with data on the uses and impacts of AI (see Figure 1).

Figure 1: Screenshot of OBIA showing some results on the preparedness of Brazilian industry to adopt AI workflows.

The objectives of OBIA include compiling, recording, and providing information related to Artificial Intelligence in Brazil, enabling analyses of its adoption and its main impacts on society. It also has the mission of consolidating and disseminating knowledge about the repercussions of this technology, providing support to guide policies, strategies, and actions in promoting development and responsible use of AI.

The observatory gathers Brazilian data on the use and adoption of Artificial Intelligence by different sectors, such as education, business, government, health, and others (see Figure 2).

The currently available indicators rely mostly on traditional data sources for analysis, such as surveys and data sets made available for the team. The first product of OBIA is the book "Artificial Intelligence in Healthcare - Potentialities, Risks and Perspectives", published in July 2024. In a second line of action, it functions as a repository of guiding documents in the area, originating from all parts of the world. In a third line, it acts as an "information exchange point" between AI centers operating in Brazil: the IAX. All indicators collected will be public and can be accessed on the OBIA portal [4].

The Center for Artificial Intelligence (C4AI) at the University of São Paulo, funded by FAPESP (the public agency for research funding in the State of São Paulo) and IBM, participates in the OBIA through its Humanities area. C4AI will contribute with qualitative research in the horizontal axes of "Legislation, Regulation, and Ethical Use" and "AI Governance," while also conducting studies across various vertical axes to be monitored. The research group dedicated to this effort comprises scholars from the fields of law, computer science, electrical engineering, sociology, and political science, allowing for an interdisciplinary analysis of the key topics monitored by

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.18

OBIA. This interdisciplinary approach will provide a comprehensive view of the current state of AI development and implementation in Brazil. Various reports, articles, and data will be provided to support OBIA in fulfilling its mission.

In addition to the participation of professionals from various NIC.br departments, the Observatory has a network of external partners, including the Center for Management and Strategic Studies (CGEE), the São Paulo State System Data Analysis Foundation (SEADE), C4AI, CIAAM (Center of Artificial Intelligence and Machine Learning) and others. The following will explore how C4AI contributes to OBIA through a complementary approach, focusing on the qualitative analysis of decisions by the São Paulo Court of Justice related to AI.

2. Data and Methodology

2.1. Legislation, Regulation, and Ethical Use: A Qualitative Analysis

The research presented in this paper is the base of an action contributing to implement the PBIA strategy [7], responsible for monitoring AI regulation and legislation. It has divided its research into three main areas: the Executive, the Judiciary, and the Legislative branch, combining traditional and modern data collection methods. Regarding the Executive branch, monitoring is being conducted through data scraping of government transparency websites based on a curated and continuously updated list of AI-related terms developed by the group. This monitoring aims to understand what AI systems are being purchased or contracted by public authorities. For the Judiciary, we have been analyzing court decisions from the São Paulo Court of Appeal (TJSP) related to AI, to understand judicial interpretations and rulings in the absence of specific AI legislation [2]. As of the latest data scraping in August 2024, more than 13.000 relevant decisions have been identified. Lastly, in relation to the Legislative Branch, the group is closely following the progress of discussions on Bill 2338/2023, which focuses on AI regulation, by participating in public hearings and issuing technical notes to guide legislators. The goal is to expand this research to monitor AI-related legislation at the state and municipal levels, as many municipalities are legislating on the matter to prepare their cities to assume roles of “smart cities”.

2.2. Monitoring and exploring the local data

To effectively monitor developments in AI, it is essential to establish a comprehensive list of AI-related terms that can guide data collection efforts. This list is derived from multiple sources, including scientific articles, standards like [3], and reports such as OECD's [1]. The monitoring process involves monthly web scraping of court rulings, based on the AI-related terms list, from TJSP (Judiciary Power) and data from the Brazilian Transparency Portal (Executive Power), which occurs on the 15th of each month. For the Judiciary Power, the scrapes and data treatment are performed with scripts developed in Python and R programming languages, based on the TJSP API, by Jesus Filho (github.com/jjesusfilho/tjsp). For the Executive Power, a script was developed to scrape data from the Data Download section of the Brazilian Transparency Portal (portaldatransparencia.gov.br/download-de-dados).

Currently, we are developing an automation tool, based on NLP techniques, to enhance the qualitative analysis of these court rulings, allowing for more efficient identification and categorization of data relevant to AI research. The first approach for this automation tool is using a NER (Named Entity Recognition) model, to automate the identification of relevant entities, including litigants and court judgments. The next step would be to apply a classification model, yet to be chosen, to filter out noise data. The process of constructing the terms for web scraping is a critical step to ensure the relevance and accuracy of the data collected for AI research. This process begins with the development of a comprehensive list of AI-related terms, which is built using multiple authoritative sources. One primary source is the OECD's report "Identifying and Measuring Developments in Artificial Intelligence," which offers a foundation of 226 AI-related terms identified through extensive analysis of scientific articles, open-source systems, and patents. Another source is the ISO/IEC 22989:2022 standard [3], which provides a framework for AI concepts and terminologies. These terms are carefully selected, refined, and translated into Portuguese by experts working within the Brazilian Technical Standards Association (ABNT) to ensure that only those terms that are highly relevant and specific to AI are included. Terms that are too general or contextually irrelevant—such as "transparency," which could result in unrelated hits concerning Brazil's Access to Information Law—are excluded to avoid false positives in the scraping process. The final list of terms, consisting of 103 terms in both English and Portuguese, is used to guide the web scraping data collection processes, allowing a focused and efficient retrieval of information that aligns with specific research objectives.

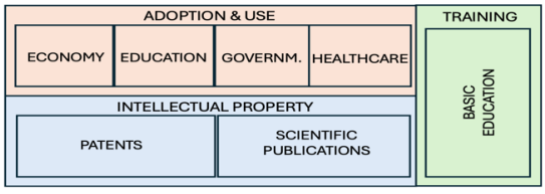


Figure 2: Current Dimensions of OBIA’s monitoring topics

2.3. How to implement and classify repositories with reference documents and statistics?

As part of the data collection and structuring process for qualitative analysis, we are implementing and classifying repositories containing reference documents and statistics. These repositories will focus on key thematic areas, such as "Legislation, Regulation, and Ethical Use" and "AI Governance," and will be populated with data from sources like TJSP, the

Transparency Portal, and other relevant databases. By combining different methods, data retrieval becomes more efficient and targeted, ensuring the collection of relevant information. Web scraping supplements this process by capturing data unavailable through APIs, ensuring comprehensive coverage. The data is regularly updated, with documents classified by relevance to AI terms, creating a dynamic and organized repository (see Fig 3) described in [6].

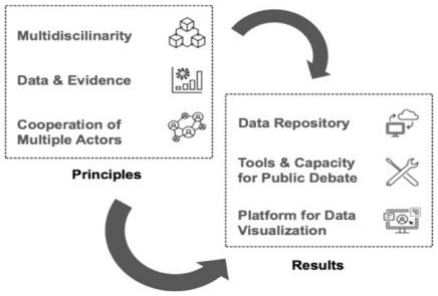


Figure 3: OBIA's guiding principles and expected results [6]

2.4. How to establish and maintain cooperation networks?

Establishing and maintaining cooperation networks requires fostering collaboration among interdisciplinary researchers from fields such as law, computer science, engineering, sociology, and political science. These networks are essential for sharing insights and methodologies related to AI monitoring. Using APIs and web scraping tools enables access to current data, supporting continuous knowledge exchange. Regular workshops, webinars, and joint research projects help keep participants engaged. Publishing reports, articles, and datasets strengthens the network and supports OBIAs mission to monitor AI developments comprehensively.

3 Discussion of initial results

As of June 28, 2024, a total of 13,064 decisions were scraped from the São Paulo State Court of Justice based on AI-related terms. Out of 103 terms searched, 45 returned at least one result. Graph 1 shows the monthly distribution of all results, while Figure 5 (logarithmic scale) displays the distribution of results by AI term. Both Portuguese and English terms were used for scraping. The top 15 terms with the most occurrences were analyzed over time, and Figure 6 presents the temporal evolution of these results by publication date. A qualitative review of 597 decisions from the São Paulo Court of Justice (TJSP) using a detailed list of AI-related terms, focused on terms like "Facial Recognition" and "Facial Biometrics," showing they are often used in various legal contexts, sometimes diverging from their technological meanings.

Terms like "Facial Expression Recognition" and "Learning Agent" were often interpreted in psychological or social contexts rather than purely technological ones. The analysis used analytical, comparative, and monographic methods, with the latter focusing on case-specific factors to draw broader generalizations. From TJSP's website, 597 rulings were reviewed: "Facial Recognition" (1), "Facial Expression Recognition" (1), "Machine Learning" (7), "Artificial Intelligence" (163), "Artificial Intelligence" in English (4), "Machine Learning" in Portuguese (3), "Learning Agent" (1), and "Facial Biometrics" (417).

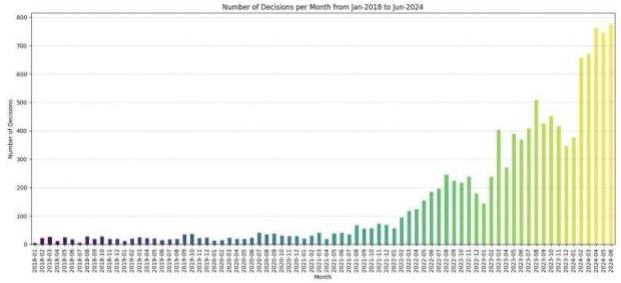


Figure 4: Nr. of Decisions per Month from Jan 2018 to Jun 2024.

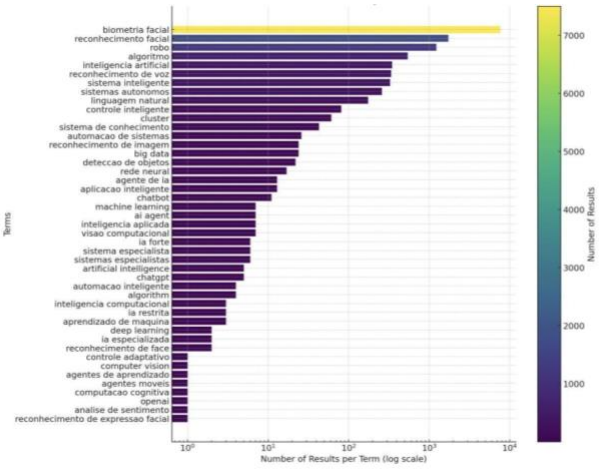


Figure 5: Number of results per AI term.

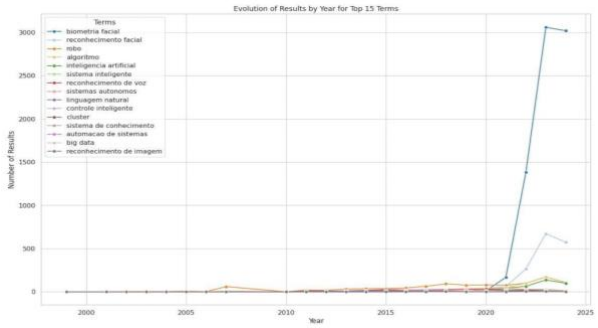


Figure 6: Evolution of results by year for top 15 terms.

The rulings followed a structured format, and the analysis included 14 categories, such as case number, appeal type, judge, district, and the context of term usage. Key findings highlighted the use of "Artificial Intelligence" and "Machine Learning" in commercial disputes and credit issues rather than solely technological matters. The rulings analyzed represent

decisions, rendered by collegiate bodies composed of multiple magistrates. Each ruling follows a structured format: Description and Qualification, covering aspects such as appeal, case number, judicial district, presiding judge, and parties involved; Summary of the ruling; Report, offering a brief description of the facts; Majority Opinion; and Dissenting Opinion (if applicable). The analysis was conducted with each of the 14 subcategories corresponding to columns in a single row: case number; type of appeal; reporting judge; district; judicial body; subject matter; judgment date; publication date; summary; parties; reasoning; final decision; context of term usage in the full text; and relevant jurisprudence. While the first nine categories were predefined based on the complete jurisprudence search, the remaining five were more subjective, created to enhance the understanding of the rulings' content and improve data visualization. Significant findings were noted in cases involving "Artificial Intelligence" and "Machine Learning," where the terms were often associated with commercial disputes, service contracts, or credit-related issues rather than purely technological applications. A recurrent theme in cases involving "Facial Biometrics" was the legality and validity of loan contracts signed through biometric recognition. The majority of decisions upheld the legality of such contracts, highlighting issues of consent and the technical reliability of biometric systems [1]. However, inconsistencies in judicial reasoning were identified, where similar cases had varying outcomes depending on the presiding judge. Overall, the analysis highlighted several gaps and challenges in the legal treatment of AI-related technologies, particularly concerning transparency, fairness, and consumer protection. The study underlined the need for more consistent legal standards and better understanding among judges of the technical nuances involved in AI applications to ensure fair and equitable rulings.

4 Conclusions and further work

The qualitative research findings from the analysis of court decisions related to AI reveal several key conclusions. AI-related terms such as "Facial Recognition," "Voice Recognition," and "Autonomous Systems" are frequently used in judicial contexts that extend beyond their traditional technological meanings, intersecting with areas like consumer protection, contract law, and fraud. The inconsistency in judicial reasoning and varying outcomes in similar cases highlight the need for clearer legal frameworks and a deeper understanding of AI's technological implications among judges. Moving forward, the incorporation of NLP techniques into the analysis will help extract key arguments from judicial decisions, providing deeper insights into the legal discourse on AI. This will enhance the robustness of future research on AI regulation and its implications for public policy.

Furthermore, a preliminary analysis of news using the NLP capabilities of the *Eventregistry.org* system (see Figure 7) show how this source can provide complementary results to the study when, e.g., capturing the attention of media on the terms "criminal law" and "AI" in "Brazil" in the past 12 months, where 1.4% exhibits discussions on Human Rights, and terms like "democracy" and "discrimination" are within the top 30. When performing sentiment analysis over these results we can see large variations after the summer of 2022 with a predominantly negative sentiment regarding this search topic.

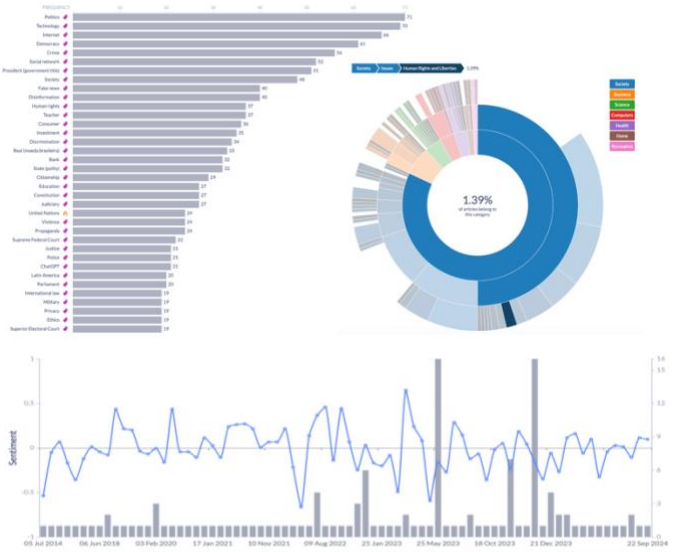


Figure 7: Significance of criminal law and AI in the news.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the Center for Artificial Intelligence (C4AI) at the University of São Paulo (USP), supported by FAPESP and IBM, for their invaluable support to the AI Observatory team. We thank to the CIAAM for their continued collaboration and contributions to this research. We thank the support of the European Commission project ELIAS - Lighthouse of AI for Sustainability (10080425).

REFERENCES

[1] Baruffaldi, Stefano, et al. (2020) Identifying and measuring developments in artificial intelligence: Making the impossible possible. OECD.
[2] Cristina Godoy B. de Oliveira, Otávio de Paula Albuquerque, Emily Liene Belotti, Isabella Ferreira Lopes, Rodrigo Brandão de A. Silva, Glauco Arbix. Intelligent Systems: 12th Brazilian Conference, BRACIS 2023, Belo Horizonte, Brazil, September 25–29, 2023, Proceedings, Part I, pp 18 – 32.
[3] ISO (2022) Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. ISO/IEC 22989:2022. [Online]. Available: https://www.iso.org/standard/74296.html/ [27 8 2024]
[4] Luiz Costa et al. (2024) The Brazilian Artificial Intelligence Observatory (OBIA). [Online]. Available: https://www.obia.nic.br/ [27 8 2024]
[5] MCTI (2021). Brazilian Strategy of Artificial Intelligence. [Online]. Available: ebia-documento_referencia_4-979_2021.pdf (www.gov.br) [07 9 2024]
[6] MCTI (2023). OBIA: Observatório Brasileiro de Inteligência Artificial,. Available: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/1_ebia-reuniao-ro_7_24_05_2023_anexo_2_eixo2-pdf.pdf [27 8 2024]
[7] PBIA (2024). Brazilian Artificial Intelligence Plan . [Online]. Available: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/07/plano-brasileiro-de-ia-tera-supercomputador-e-investimento-de-r-23-bilhoes-em-quatro-anos/ia_para_o_bem_de_todos.pdf/view [07 09 2024]

Pojavljanje incidentov ob uporabi Umetne Inteligence

Marko Grobelnik
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
marko.grobelnik@ijs.si

Besher M. Massri
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
m.besher.massri@gmail.com

Alenka Guček
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
alenka.gucek@ijs.si

Dunja Mladenić
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
dunja.mladenic@ijs.si

Povzetek

Prispevek predstavi prve rezultate ob uporabi sistema, ki je bil zasnovan in razvit v sodelovanju z OECD za spremljanje incidentov, povezanih z umetno inteligenco. Glavna motivacija teh prizadevanj je podpora zakonodaji, povezani z umetno inteligenco, in učinkovitemu oblikovanju politik, saj sistem zagotavlja vpoglede na podlagi zbranih podatkov. OECD AI Incidents Monitor za spremljanje incidentov, povezanih z umetno inteligenco, dokumentira incidente in nevarnosti v zvezi z umetno inteligenco, da bi oblikovalcem politik, strokovnjakom za umetno inteligenco in vsem zainteresiranim stranem po vsem svetu pomagal pridobiti dragocen vpogled v tveganja in škodo, ki jo povzročajo sistemi umetne inteligence. Ideja je, da bo sistem sčasoma pomagal povečati ozaveščenost javnosti in vzpostaviti skupno razumevanje incidentov in nevarnosti umetne inteligence, in tako prispeval k zaupanja vredni umetni inteligenci.

Ključne besede

umetna inteligenca, analiza podatkov, oblikovanje politik,incidenti

Abstract

This paper presents a system designed and developed in collaboration with OECD for monitoring of AI-related incidents. The main motivation behind the efforts is in supporting AI-related legislation and effective policymaking, as the system provides evidence based on the collected data. The OECD AI Incidents Monitor documents AI incidents and hazards to help policymakers, AI practitioners, and all stakeholders worldwide gain valuable insights into the risks and harms of AI systems. The idea is that over time the system will help to raise awareness and establish a collective understanding of AI incidents and hazards contributing to trustworthy AI.

Keywords

Artificial Intelligence, data analysis, policy making, AI incidents

1 Uvod

Ob vse širši uporabi umetne inteligence (UI) prihaja tudi do incidentov ob njeni uporabi. Spremljanje teh incidentov je nujno za zagotavljanje preglednosti, nadzora in razvoj politik, ki lahko takšne incidente preprečujejo ali vsaj zmanjšujejo. Predstavljeni sistem deluje kot orodje, ki pomaga uporabniku, ki si prizadeva v realnem času slediti dejanskim incidentom, povezanim z umetno inteligenco, ter zagotavljati dokazno osnovo za oblikovanje okvira poročanja o incidentih in povezanih političnih razpravah o UI. Z zbiranjem podrobnih vpogledov v vsak incident omogoča učenje iz preteklih napak ter spodbuja varnejši in bolj odgovoren razvoj ter uporabo umetne inteligence. Koristi skupnosti, ki se ukvarja z umetno inteligenco, saj izpostavlja trende in področja, ki potrebujejo pozornost ali regulativni poseg.

Prednost sistema je, da je zbiranje podatkov avtomatizirano, kar je prednost v primerjavi s podobnimi repozitoriji, ki so urednikovani ročno, kot je na primer AIAAIC Repository [2]. Repozitorij je prosto dostopen in namenjen tako oblikovalcem politik, kot razvijalcem UI, raziskovalcem, pravnikom in javnim organizacijam.

V nadaljevanju predstavimo metodologijo za spremljanje incidentov, prikažemo delovanje sistema na nekaj realnih primerih, predstavimo deležnike in nekaj zaključkov.

2 Metodologija

Metodologija OECD za spremljanje AI incidentov se osredotoča na identifikacijo in klasifikacijo incidentov, s čimer zagotavlja vpogled v realno dogajanje in podpira razvoj okvira za poročanje o incidentih. Začetna točka je identifikacija in klasifikacija incidentov, ki so poročani v uglednih mednarodnih medijih, s pomočjo modelov strojnega učenja, kar omogoča gradnjo zanesljive baze podatkov (incidenti so zajeti od 2014 naprej).

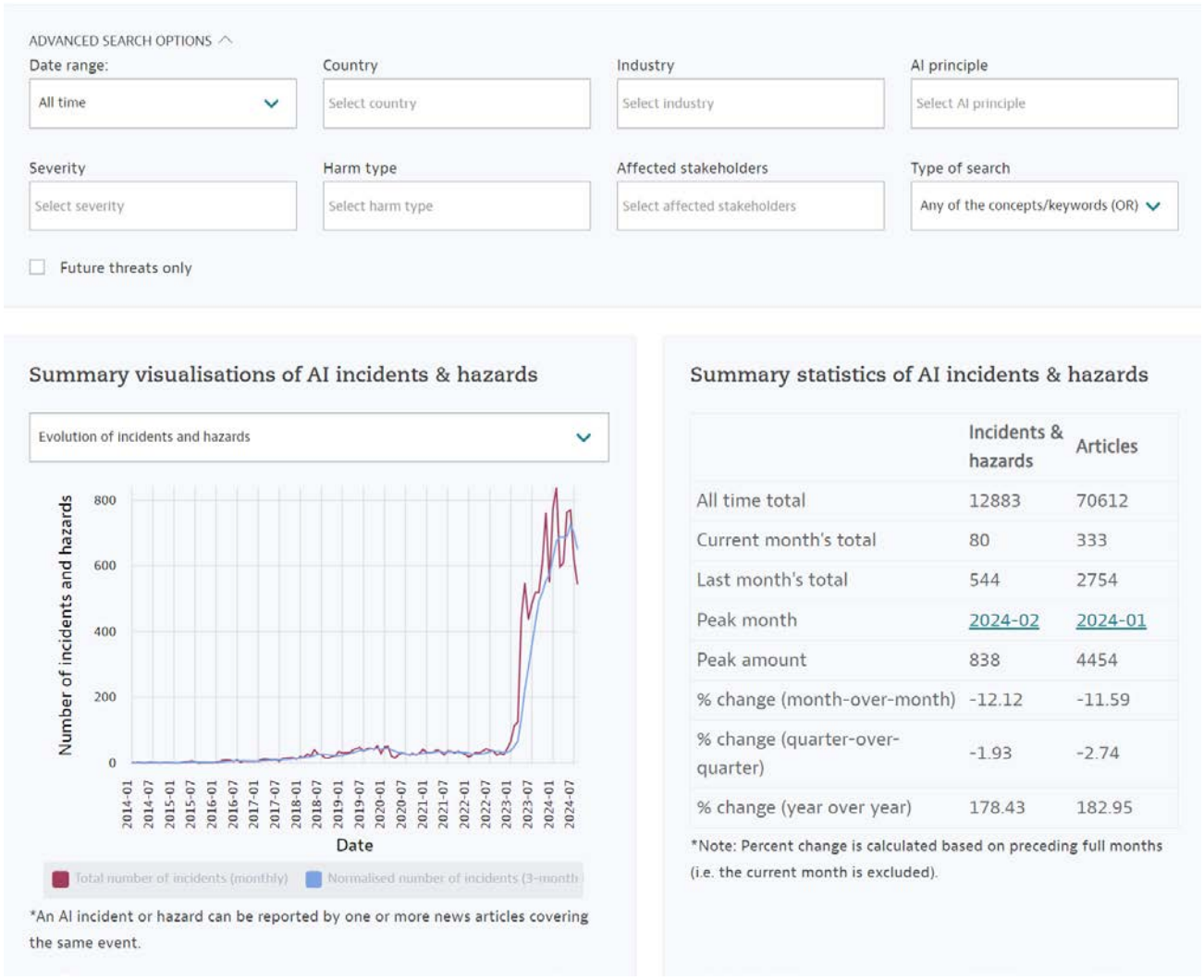
Kljub prizadevanjem, ti incidenti predstavljajo le podmnožico vseh globalnih AI incidentov. Incidenti so razvrščeni glede na resnost, industrijo, povezane AI principe (OECD AI Principles [3]), vrste škode in prizadete deležnike. Analiza temelji na naslovih, povzetkih in prvih odstavkih novinarskih člankov, pri čemer se pridobljeni podatki uporabljajo za izgradnjo zanesljive, objektivne in kakovostne baze podatkov o incidentih, povezanih z AI. Kot vir novic služi sistem Event Registry [4].

Razvoj sistema, h kateremu smo prispevali, nadgrajuje delo mednarodne skupine strokovnjakov (OECD Expert group), ki razvija teoretično ogrodje za poročanje o incidentih, definira pojem AI incidenta in oblikuje povezano terminologijo, kot je AI nevarnosti in njene potencialne posledice. Podrobna metodologija in definicije so razložene na spletni strani OECD: https://oecd.ai/en/incidents-methodology.

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.26



Slika 1 Prikaz začetne strani OECD monitorja AI incidentov (https://oecd.ai/en/incidents) prikazuje vmesnik za iskanje po konceptih, vizualizacijo incidentov v času (spodaj levo; y os: število incidentov; x os: čas (2014-danes) in statistični povzetek (spodaj

3 AI Incidents Monitor

AI Incidents Monitor je do konca avgusta 2024 zaznal preko 12 000 incidentov in nevarnosti v zvezi z UI, Kot je razvidno iz Slike 1. Sistem je popolnoma avtomatski in zaznava incidente s skeniranjem številnih podatkov objavljenih v novicah, ter nato s pomočjo UI določa kaj se označi kot incident ali nevarnost. Na naslovni strani (Slika 1) je prikazan črtni diagram naraščanja incidentov v času (levo) in pripadajoča statistika (desno). Uporabnik lahko izbira med absolutnim prikazom incidentov (kot na Sliki 1) ali v ustreznem meniju izbere pod-področja. Če se poglobimo v prikaz na Sliki 1, vidimo z različnimi barvami označeni kumulativni incidenti (vijolična) oz. njihovo trimesečno povprečje (modra). Statistika na desni prikazuje absolutno število incidentov glede na izbrano področje (12 883 incidenti in nevarnosti o katerih je poročalo 70612 novičarskih člankov), statistiko za zadnji mesec in mesece z največjimi vrednostmi (februar 2024). Iz statistik o spremembi glede na mesec, na četrtletje in na leto, vidimo padec števila incidentov in nevarnosti o katerih so mediji poročali v zadnjem mesecu glede na prejšnji mesec oz. prejšnje četrtletje.

3.1 Primer analize pojavitve incidentov UI

Sistem omogoča napredno filtriranje po incidentih UI za sledeče kategorije: čas, država, industrija, princip UI, resnost, tip škode, oškodovanci, tip iskanja po vsebini (glej Sliko 1). Tako so na primer možne vrednosti za resnost: smrt, poškodba, nevarnost, nefizična nevarnost, možni tipi škode pa so: fizična, psihološka, ekonomska, ugled, javni interes, človekove pravice, neznana.

Sistem omogoča napredno iskanje po konceptih, recimo za primer generativne UI, sistem poroča statistike, ki kažejo 2302 incidenta in nevarnosti, en od primerov incidentov, ki jih je sistem zaznal pa se nanaša na Apple in razvoj »AI personality«, ki naj bi nadomestil obstoječi Applov Siri.

Poleg konceptov uporabnik lahko nadalje izbere tudi napredno iskanje za natančnejšo identifikacijo želene podskupine incidentov, ki ga zanimajo. Tako lahko recimo izbere državo, ki je povezana s poročanjem o incidentih in nevarnostih UI. Na Sliki 2 je tako prikazan primer iskanja po kategoriji države, za Slovenijo. Sistem najde dva incidenta, ki sta bila povezana s Slovenijo. Prvi incident se nanaša na Microsoftov povečan prispevek k emisiji CO2. Na prvi pogled ni očitna povezava s Slovenijo, ko pa pogledamo povezane novice naletimo na omembo Slovenije: *»…But the tech giant's electricity consumption last year rivaled that of a small European country—beating Slovenia easily.«* [6]. Vsak primer je tudi semantično označen. Tako je na Sliki 2 za prvi primer označena povezanost s principi UI učinkovitost, trajnostni razvoj. Microsoft s tem lahko prizadene več deležnikov: splošno javnost, podjetja, delavce, vlade (Affected Stakeholders, Slika 2). Poleg tega predstavlja nevarnost za okolje, javne interese in človekove pravice (Harm type, Slika 2). Klasificirano je kot ne-fizična nevarnost (Severity, Slika 2).

Iz podrobnih analiz, ki so zbrane v nedavnem poročilu *»Observatory of the social and ethical impact of artificial intelligence«* [5], je razvidno, da večina incidentov (96%) spada pod kategorijo ne-fizično nevarnih, a imajo lahko zelo resne psihološke in finančne posledice, vključujoč nadlegovanja, odvisnosti in škodo ugledu tako posameznikom kot tudi inštitucijam.

4 Deležniki

OECD-jev monitor incidentov AI (AIM) je dragoceno orodje, zasnovano za različne deležnike, ki sodelujejo pri razvoju, regulaciji in uporabi umetne inteligence. Potencialni uporabniki tega orodja vključujejo oblikovalce politik, razvijalce AI, raziskovalce, pravne strokovnjake in javne organizacije.

Oblikovalci politik lahko AIM uporabljajo za sledenje in analizo podatkov v realnem času o incidentih, povezanih z AI, po vsem svetu, kar jim pomaga pri oblikovanju informiranih in na dokazih temelječih predpisov. Zmožnost orodja za kategorizacijo incidentov glede na resnost, industrijo in vrste škode je ključna za razumevanje širših posledic tehnologij umetne inteligence in oblikovanje politik, ki zmanjšujejo tveganja.

Razvijalci AI in raziskovalci lahko koristijo AIM, da prepoznajo pogoste težave, povezane s sistemi umetne inteligence. S preučevanjem incidentov, zabeleženih v AIM, lahko izboljšajo svoje modele, da bi se izognili podobnim težavam in povečali varnost ter zanesljivost aplikacij umetne inteligence.

Pravni strokovnjaki lahko uporabljajo AIM za pridobitev vpogledov v spreminjajočo se pokrajino tveganj, povezanih z umetno inteligenco, kar bi lahko bilo koristno v pravnih primerih ali ocenah skladnosti. Razumevanje preteklih incidentov in njihovih pravnih posledic lahko usmerja razvoj robustnih okvirov upravljanja AI.

Nazadnje lahko javne organizacije in zagovorniške skupine uporabljajo AIM za spremljanje družbenih vplivov umetne inteligence, s čimer zagotavljajo, da so interesi javnosti zaščiteni. To lahko vključuje analizo vzorcev incidentov z umetno inteligenco za zagovarjanje boljše zaščite potrošnikov in etičnih standardov pri uvajanju AI.

5 Diskusija

V prispevku smo predstavili OECD-jev monitor incidentov umetne inteligence, pri razvoju katerega smo sodelovali. Sistem služi kot dober vir za širok nabor uporabnikov, ki želijo razumeti in upravljati tveganja, povezana s tehnologijami UI. Sistem se nadgrajuje z dodatnimi podatkovnimi viri.

V prihodnosti je predvideno, da bo omogočen odprt postopek oddaje podatkov, ki bo dopolnil informacije o incidentih, pridobljene iz trenutnih virov. Nadaljnje delo zajema tudi avtomatsko analizo podatkov o incidentih za namen bolj celovitega vpogleda. To vključuje avtomatsko odkrivanja vzorcev, kot so verižne reakcije ali učinki na več industrij hkrati. Za potrebe preverjanja resničnosti poročanih incidentov, bi lahko vključili kombiniranje informacij iz več neodvisnih virov in uporabljal algoritme za odkrivanje lažnih novic, kot tudi ročno preverjanje.

Zahvala

Delo, opisano v tem prispevku, so podprli OECD in številni mednarodni eksperti, Ministrstvo za digitalno preobrazbo in Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru CRP V2-2272 in V5-2264.

Acknolwedgements

The described work was supported by OECD and many os its international experts, Slovenian Ministry of Digital Transformation and Slovenian Research and Innovation Agency under CRP V2-2272 and V5-2264.

Literatura

[1] OECD AI Incidents Monitor (AIM), https://oecd.ai/en/incidents. August 2024

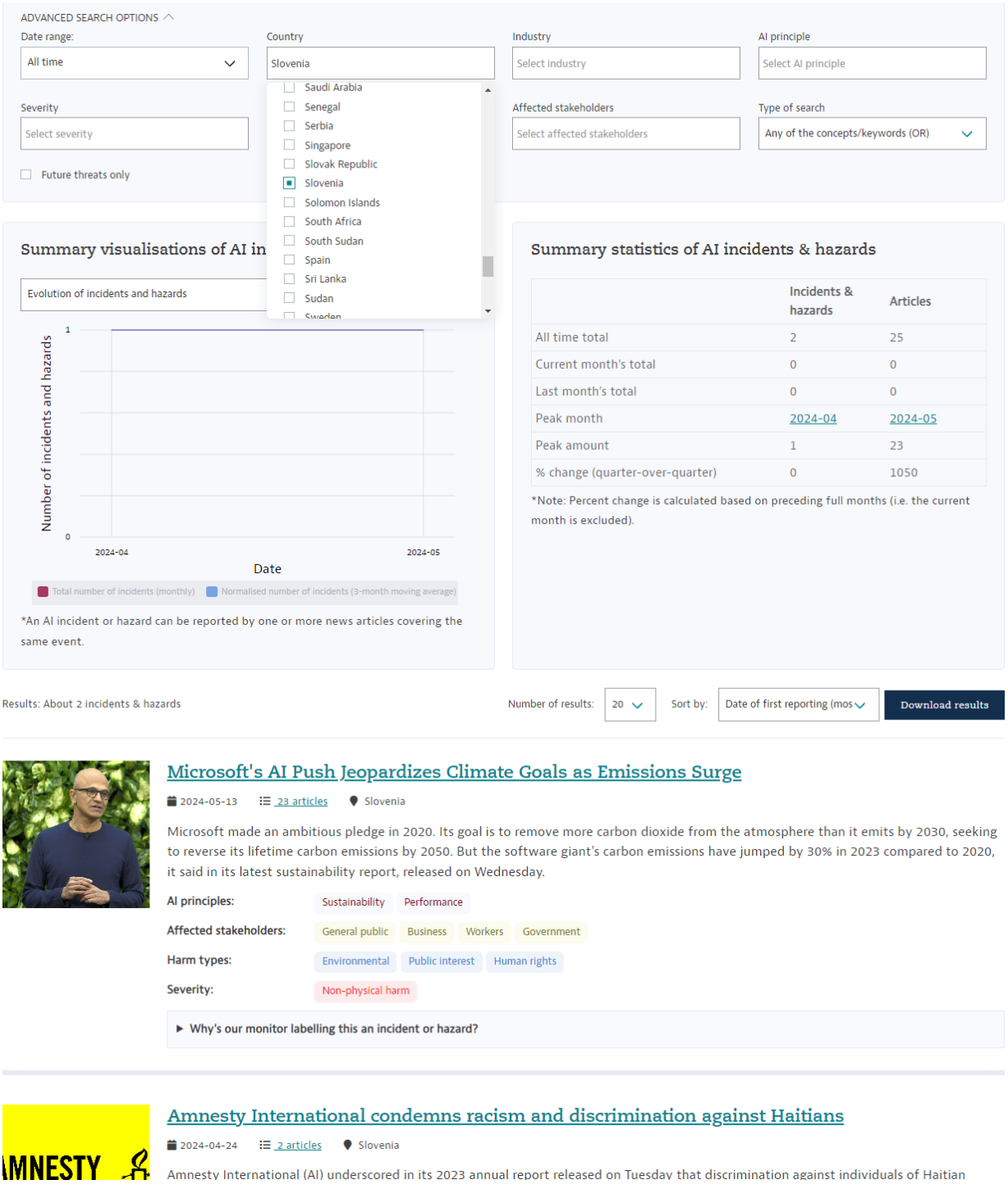
[2] AIAAIC Repositoryhttps://www.aiaaic.org/aiaaic-repository. August 2024

[3] OECD AI Principles for trustworthy AI https://oecd.ai/en/ai-principles August 2024

[4] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In Proceedings of the 23rd International Conference on World Wide Web, 107–110.

[5] Richard Benjamins, Another Inconvenient Truth: The Societal Emergency of AI Incidents - We Should Do Something About It https://www.odiseia.org/post/another-inconvenient-truth-the-societal-emergency-of-ai-incidents-we-should-do-something-about-it

[6] Microsoft's AI Push Imperils Climate Goal as Carbon Emissions Jump 30% https://tanaka-preciousmetals.com/en/elements/news-cred-20240821/



Slika 2 Prikaz naprednega iskanja na OECD monitorju AI incidentov (https://oecd.ai/en/incidents) filtrirano po državi za Slovenijo. Podane so statistike dveh incidentov o katerih je poročalo 25 novičarskih člankov, in spodaj sta prikazana oba incidenta.

Perception of AI in Slovenia

Abdul Sittar
abdul.sittar@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Alenka Guček
alenka.gucek@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Abstract

This paper introduces the AI News Monitor system developed for real-time monitoring and analysis of artificial intelligence (AI) perception in global and local news media. Leveraging data from the Event Registry platform, the AI News Monitor tracks AI-related news articles across multiple dimensions, providing insights through three key views: a global historical overview, current global trends, and local trends specific to Slovenian media. The system facilitates both passive observation of AI discourse and active exploration of specific AI-related events. Our illustrative analysis reveals significant global trends, including heightened media focus on deep learning, generative AI, and robotics, and examines the implications of these trends on public trust in AI. Additionally, the paper discusses the practical applications of the AI News Monitor for stakeholders such as policymakers, journalists, business leaders, and researchers. We conclude with a discussion on the impact of media coverage on public perception of AI and propose possible future enhancements of the system, including broader language and source coverage.

Keywords

datasets, artificial intelligence, media monitoring, perception

1 Introduction

Artificial Intelligence (AI) is increasingly becoming an integral part of society, influencing various aspects of daily life and industries [4]. As AI continues to evolve, so does its portrayal in the media, which plays a critical role in shaping public perception and trust. Understanding how AI is perceived globally and locally is essential for policymakers, businesses, and researchers to ensure that AI technologies are developed and deployed in ways that are socially acceptable and trustworthy [3, 4].

In response to this need, we have developed the AI News Monitor system designed for real-time monitoring and exploratory analysis of AI-related news coverage. The AI News Monitor offers a comprehensive view of how AI is discussed in the media, capturing data from the Event Registry platform on a monthly basis [7].

The AI News Monitor system is structured around three main views: a global overview that presents historical data from the past year, global trends that highlight recent AI-related events, and local trends focusing on mentions of AI by Slovenian news sources. These views allow the users to either passively monitor ongoing developments in AI or actively explore specific events and trends that may influence public opinion.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.14

Following are the main scientific contributions of this paper:
(1) We present a methodology to understand public perception about AI in news.
(2) We analyse some trends in AI's Perception.
The remainder of the paper is structured as follows. Section 2 describes the methodology to collect historical data, AI news categories and gaining insights in public perception about AI in news. Section 3 presents the analysis of trends in AI's Perception. We present different user scenarios and possible applications of AI News Monitoring in Section 4 and discussion in Section 5. Section 6 concludes the paper and outlines possible areas of future work.

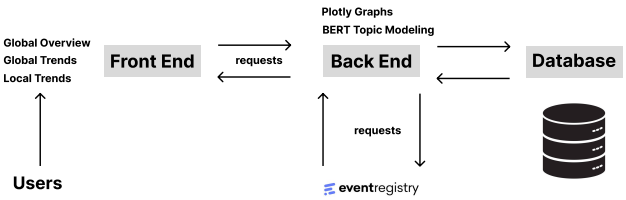


Figure 1: Architecture for Real-Time AI News Monitoring and Visualization based on Event Registry and implemented using Flask and Plotly.

2 Methodology

The proposed approach to creating a web service to analyze public perception involves two key steps: 1) identifying AI-related categories and gathering news within these categories, and 2) developing a web service that displays trends across these categories, news publishers, and highlights current trends among both global and local (Slovenian) news sources (see Figure 1).

Firstly, we selected AI-related categories based on the Slovenian AI observatory[1] and Wikipedia[2]. The key categories associated with Artificial Intelligence include 'Generative AI', 'Artificial Intelligence', 'NLP', 'Chat-GPT', 'Deep Learning', 'Robotics', 'Computer Vision', 'Neural Networks', 'Graph Neural Networks', 'Self-supervised Learning', and 'Zero-shot Learning'.

Next, we collected news articles from the last year related to these categories. These articles were classified into the appropriate categories based on Wikipedia concepts, and we also obtained sentiment data from Event Registry. The portrayal of AI-related news significantly impacts public perception, with the emphasis on risks, benefits, or ethical concerns shaping public opinion and driving narratives that can either build trust or instill fear[8],[12], [1].

To understand global trends, we retrieved news events published globally in the last month. For local trends, we focused on news

[1] http://siai.ijs.si/dashboards/Main/SlovenianObservatoryIntro?globalCountry=SVN
[2] http://country-dashboards.ijs.si/dashboards/Main/Index?

Figure 2: Time series of the number of news articles by specific areas (in colors, at the top). Detailed view upon precise exploration (middle) and corresponding sentiment of news from specific areas (at the bottom).

articles published by the top 50 Slovenian news publishers. Finally, we employed topic models to analyze the corpus of news articles and extract underlying themes [9], [2].

3 Analysis of trends of AI's Perception

3.1 Global Overview

The global overview provides a historical review of global AI-related news (see Figure 2). Users can explore the number of news articles across 13 AI fields (Generative AI, Chat-GPT, Deep Learning, Robotics, Computer Vision, Neural Networks, Graph Neural Networks, Artificial Intelligence, Federated Learning, Few-shot Learning, Meta Learning, Self-supervised Learning, and Zero-shot Learning) or by news providers and have an overview of the sentiment of the news.

Global trends allow for the review and exploration of global AI-related trends based on captured events from the last month. Figure 3 shows a detailed view of the Global Trends, where a written report of the number of news articles and events, a histogram of the number of AI-related news articles over time, and the ability to explore the last 10 events in a selected field.

3.2 Local Trends

Local trends allow for the review of news from Slovenian news providers for the last month. The local trends show the detailed view, where a written report of the number of news articles and events, a histogram of the number of AI-related news articles over time, and the ability to explore (see Figure 4).

3.3 EXAMPLES OF TRENDS

Figure 3: A detailed view of Global Trends, showing the option to select news events based on chosen AI fields.

3.3.1 Global Overview. In the historical overview of AI trends in March 2024 (Figure 2), there was a significant increase in the number of news articles and interest in deep learning, generative AI, and robotics. Specifically, on March 18th, there were 1,800 news articles about generative AI, 970 about robotics, and 274 about deep learning. This spike in news highlights several key events: one of the standout stories was the launch of Gen-2 by Runway, a generative video model capable of creating high-quality short clips. An important topic was the use of AI in political campaigns, particularly the creation of deepfakes and misinformation. This raised concerns about AI's impact on elections and voter trust. In the field of robotics, researchers were inspired by advancements in generative AI to develop more versatile robots. These new robots can perform various tasks using a single, comprehensive

model, demonstrating significant progress in robotic capabilities. Overall, the sentiment in March 2024 was positive (as seen from the sentiment analysis), reflecting enthusiasm and optimism regarding this technological progress. The increased media attention highlights the rapid development and growing importance of AI in various fields.

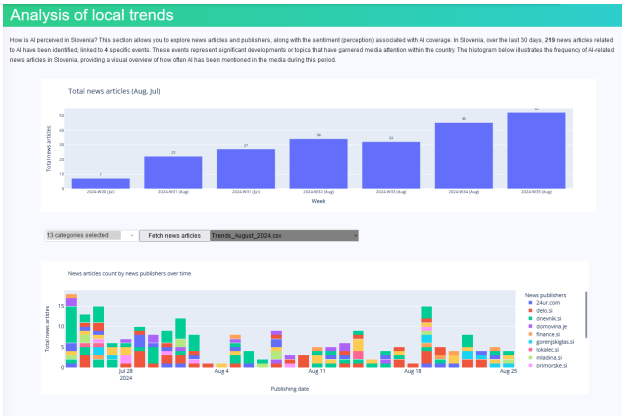


Figure 4: A detailed view of Local Trends, showing the option to select news events based on chosen AI fields.

3.3.2 Global Trends. In our examination of global trends, we selected the news story "AI and heat waves pose dual threats to the power grid" and found that two specific newspapers published more articles on this topic compared to others. The sentiment of these articles, as shown in the middle graph (Figure 4), fluctuates between positive and neutral. Upon delving into the content of these publications, we found that Forbes focused on the issue of fake news generated by AI, while Lexology explored future AI applications in various fields.

3.3.3 Local Trends. In the last month (at the time of writing the report, this was June 2024), there was an increase in AI-related news from Slovenian news providers, particularly from Delo.si and Sta.si (Figure 5). When analyzing the sentiment of these articles, most were neutral, with a few expressing positive opinions about AI. Delo.si focused on the growing adoption of AI by companies in Slovenia, highlighting discussions on the potential of quantum computing and recent advancements in AI technology. This coverage indicates a balanced view of AI's impact and potential. Sta.si reported on the construction of a state-of-the-art data center in Maribor, which will also house a supercomputer. This event represents a major development in Slovenia's technological infrastructure. Additionally, Sta.si wrote about AI trends that benefit semiconductor manufacturers, reflecting a positive outlook on the economic impact.

4 User Scenarios and Applications

The AI News Monitor can cater to a range of stakeholders with varying use case objectives [10], [6], [5]. Policy makers can utilize the developed system to track global and local trends in AI-related topics, enabling them to craft data-driven policies that balance innovation with societal concerns. Journalists can leverage the system to gather comprehensive insights into public sentiment and media coverage, enriching their reporting with accurate and timely information [11]. Detailed scenarios for both policy makers and journalists are explained below, illustrating how the AI News Monitor can support their specific goals. Other potential stakeholders are business executives, NGOs, researchers and educators.

Policy Makers: Scenario: A policymaker uses AI News Monitor to track trends in robotics.
Background: Jure, a decision-maker at a government agency for technology and innovation, is tasked with drafting new guidelines for the development and implementation of robotics in Slovenia. To understand the broader context and local trends, he needs to explore the global perception of robotics and compare it with local perspectives.
Steps: Step 1: Searching for a Global Overview: Jure logs into AI News Monitor and searches for "robotics" under the global overview section. The system displays a line chart showing how robotics has been mentioned over time, along with a sentiment graph for the past year. He finds that robotics is globally discussed with mostly positive sentiment, particularly in Asia and North America. Step 2: Global Trends: Jure selects "robotics" among the topics and reviews recent events on this subject. He chooses an event focusing on robotics in the EU and examines the sentiment of the publications and the main themes. In his browser, he looks at the specific articles and discovers that discussions predominantly revolve around automation and industrial robotics. Step 3: Local Trends in Slovenia: Next, Jure is interested in a review for Slovenia to understand how robotics is perceived at the local level. The dashboard for the selected topic displays an analysis of recent articles from Slovenian media. By using the browser, he discovers that discussions mainly focus on the impact of robotics on employment and the potential use of robots in healthcare. Jure finds that local concerns are more focused on social and economic impacts. He includes these insights in his preparatory documents for the new guidelines. Step 4: Compiling the Report and Recommendations: Finally, Jure exports key data, including sentiment graphs and media summaries, from AI News Monitor. He compiles a report that summarizes global trends and local concerns and proposes balanced guidelines that promote innovation in robotics while addressing social impacts.

Journalists: Scenario: A policymaker uses AI News Monitor to track trends in robotics.
Background: Ana, a journalist at a technology magazine, is tasked with writing an article on the growing trend of using generative AI to create videos. She needs to explore both global trends and local perspectives in Slovenia to provide a comprehensive overview.
Steps: Step 1: Searching for a Global Overview: Ana searches for "generative AI" under the global overview section. The system displays a line chart showing that this topic is on the rise, identifies the media outlets reporting on generative AI, and provides a sentiment graph for the past year. Step 2: Global Trends: Ana selects "generative AI" and reviews recent events on this topic. She focuses on deepfake video generation, checking who has written about it and what the main themes are. She then looks up these articles in her browser. Step 3: Local Trends in Slovenia: Ana shifts her focus to Slovenia to understand local views. The dashboard reveals that Slovenian media coverage is largely positive, particularly for certain providers. However, Ana realizes the need to include concerns about authenticity and misinformation to provide a balanced perspective. Step 4: Compiling and Writing: Ana exports key data, including sentiment graphs and media summaries, from AI News Monitor. She drafts her article, starting with global trends and then delving into specific concerns in Slovenia, enriched with visual data.

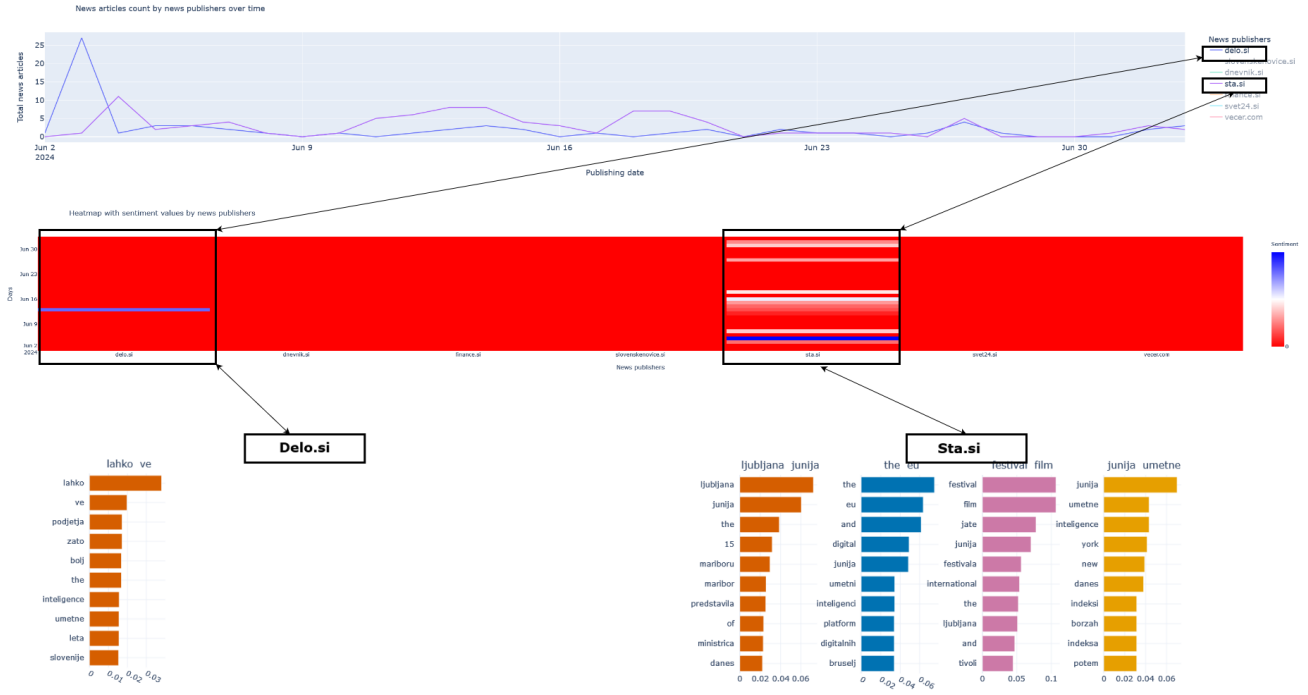


Figure 5: Time series of the number of news articles by news provider in Slovenia (at the top). Sentiment analysis (in the middle) and frequency of topics for this period (at the bottom).

5 Discussion

Services like AI News Monitor can play a role in fostering greater transparency around AI by offering detailed insights into how AI is being discussed across various media platforms. By tracking public sentiment and highlighting both positive and negative trends, it helps to ensure that the development and deployment of AI technologies are aligned with public concerns and expectations.

While AI News Monitor offers valuable insights, it has limitations, such as its reliance on media reporting, which may not capture the full spectrum of public opinion. Additionally, potential biases in media sources or the algorithms used for sentiment analysis could skew the results, presenting challenges in ensuring a fully accurate and balanced representation of public perception.

6 Conclusions

AI News Monitor was developed to understand and track public sentiment around AI, offering policymakers, journalists, and other stakeholders the insights needed to make informed decisions. AI perceptions can be monitored globally and locally, for the context of Slovenia. However, there are opportunities for future work to enhance its capabilities. Expanding the its coverage to include more languages and diverse sources would provide a more global perspective, while refining sentiment analysis techniques could improve accuracy and reduce potential biases.

7 Acknowledgments

This work was supported by the European Union through AI4Gov (101094905) and TWON (101095095) EU HE projects and Ministry of Digital Transformation and Slovenian Research and Innovation Agency under CRP V2-2272.

References

[1] Iyad AlAgha. 2021. Topic modeling and sentiment analysis of twitter discussions on covid-19 from spatial and temporal perspectives. *Journal of Information Science Theory and Practice*, 9, 1, 35–53.

[2] David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: aggregating tweets by conversations. In *Tenth international AAAI conference on web and social media*.

[3] Stephen Cave, Kate Coughlan, and Kanta Dihal. 2019. " scary robots" examining public responses to ai. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 331–337.

[4] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence* number 1. Vol. 31.

[5] Fabian Gilson, Matthias Galster, and François Georis. 2020. Generating use case scenarios from user stories. In *Proceedings of the international conference on software and system processes*, 31–40.

[6] Debasish Kundu and Debasis Samanta. 2007. A novel approach of prioritizing use case scenarios. In *14th Asia-Pacific Software Engineering Conference (APSEC'07)*. IEEE, 542–549.

[7] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, 107–110.

[8] Kalle Lyytinen, Heikki Topi, and Jing Tang. 2021. Information systems curriculum analysis for the macude project. *Communications of the Association for Information Systems*, 49, 1, 38.

[9] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 889–892.

[10] Frank Moisiadis. 2000. Prioritising use cases and scenarios. In *Proceedings 37th International Conference on Technology of Object-Oriented Languages and Systems. TOOLS-Pacific 2000*. IEEE, 108–119.

[11] Abdul Sittar, Daniela Major, Caio Mello, Dunja Mladenić, and Marko Grobelnik. 2022. Political and economic patterns in covid-19 news: from lockdown to vaccination. *IEEE Access*, 10, 40036–40050.

[12] Abdul Sittar, Dunja Mladenić, and Marko Grobelnik. 2022. Analysis of information cascading and propagation barriers across distinctive news events. *Journal of Intelligent Information Systems*, 58, 1, 119–152.

What will happen tomorrow? Predicting future event types for businesses

Tesia Šker
Jožef Stefan Institute
Ljubljana, Slovenia
tesia.sker@gmail.com

Jože M. Rožanec
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
joze.rozanec@ijs.si

Gregor Leban
Event Registry d.o.o.
Ljubljana, Slovenia
gregor@eventregistry.org

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

Strategic foresight helps organizations anticipate future challenges and opportunities, allowing them to handle uncertainty better. While strategic foresight is becoming more widely adopted across organizations, the process still heavily relies on expert knowledge, and little of it has been automated through artificial intelligence. In this research, we explore how media news events can be analyzed to forecast event types that will take place in the near future. In particular, we consider it a supervised machine learning problem with a well-defined set of event types and leverage graph representation of the media news events to create graph embeddings, train a classifier, and predict event types that will likely occur one day ahead. We validated our approach on a real-world dataset of an American multinational conglomerate operating in industry, worker safety, healthcare, and consumer goods.

KEYWORDS

strategic foresight, event prediction, machine learning, graphs

1 INTRODUCTION

Strategic foresight helps organizations anticipate future challenges and opportunities, allowing them to handle uncertainty better [9]. Therefore, predicting future event types as a part of strategic foresight became necessary for businesses to manage their operations without significant losses. Various events on a major scale, such as floods, earthquakes, internet failures, or pandemics, as we are witnessing recently, or on a minor scale, such as road closures due to sports events or promotions at fairs, can have a major impact on business operations. By predicting the next event type, businesses can adjust prices, reschedule staff, manage stocks, reschedule transportation routes to avoid delays, and more, and thus reduce losses or increase their sales and profits.

There is currently a massive number of articles written on Future Event Predictions. Based on Zhao [11], the event prediction methods can be classified in terms of goals into time prediction, location prediction, semantics prediction, and a combination of these. Each goal is divided into subgoals for which various techniques can be applied. According to the classification provided by Zhao, our technique can be classified as a semantic prediction.

In this research, we explore how graphs can be used to model media news events and to forecast event types in the near future. By doing so, we provide a valuable tool for decision-makers, offering them a clearer view of potential outcomes. Specifically, our research focuses on using a JSON dataset containing a variety of articles about a particular business company. We create a graph representation of the articles and use Graph2Vec to create embeddings that can be used downstream to fit other machine-learning models. Using this information, we apply a Random Forest Classifier to predict the categories of articles about the company for the following day.

In particular, we expect this to be useful to give organizations a competitive advantage in fast-changing markets [5]. While human expertise is valuable, it varies from person to person, leading to inconsistent predictions. Manually analyzing large datasets is also time-consuming and prone to errors. AI, however, can process vast amounts of data, spot patterns, and predict future event types more accurately.

This work is structured as follows. Section 2 presents related work that is relevant for this paper. Section 3 describes the data in the dataset, and the data extraction process. Section 4 introduces a new approach to predict future event types. Section 5 presents the results of this research. Section 6 concludes this work and proposes future improvements.

2 RELATED WORK

In recent decades there has been an increasing interest in strategic foresight in the academic field. According to Fergnani (2020) [2] this is because by "using corporate foresight, organisations can reconfigure their strategy based on the analysis of business opportunities suggested by future possibilities". Even in academia "one of the domains heavily impacted by Artificial Intelligence is innovation management and in this context especially the area of Strategic Foresight (SF)" as per Brandtner et. al (2021) [1].

However it seems that strategic foresight methods related to AI only end up being used by bigger companies with a larger number of resources. As noted by Kim and Seo (2023) [6], "except for AI start-ups and players in the consumer electronics and information and communication industry, small- and medium-sized enterprises (hereafter SMEs) in other industries do not demonstrate competence in AI." Therefore, effective implementation of AI solutions for strategic foresight in smaller and medium sized

Jože M. Rožanec and Tesia Šker are co-first authors with equal contribution and importance.
Corresponding author: Jože M. Rožanec: joze.rozanec@ijs.si.

Information Society 2024, 7–11 October 2023, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.24

companies would be one of the topics to be explored in future research.

In this research however, we focus more on the general implementation of strategic foresight by means of next event prediction. Exploring similar fields, we found that there was already some research exploring the field of event predictions, which rather than focusing on businesses focused on other domains. In the field of sequential event prediction, several researchers are exploring diverse methods. Although the methods share some conceptual similarities with our research, they differ significantly in methodology and focus. Letham, Rudin, and Madigan (2013) [7] developed a model that predicts the next event using an ERM-based approach with logistic regression, focusing on the presence of events rather than their order. On the other hand, our work uses labeled article databases and considers the sequence of past events, using techniques like graph construction, random walks, and random forests. Yeon, Kim, and Jang (2015) [10] focus on predicting event flow through visual analytics, using LDA for topic extraction and emphasizing specific keywords, while our approach is entirely text-based and relies on graphs. On the other hand, Hu et al. (2017) [4] use LSTM networks for predicting future subevents, which offers an alternative method to our non-LSTM-based text analysis.

Although these studies provide useful insights and have offered significant improvement in sequential event prediction, they may face certain challenges. For instance, Letham, Rudin, and Madigan (2013) [7] emphasize event presence over sequence, potentially missing key temporal relationships, while Yeon, Kim, and Jang (2015) [10] depend heavily on keywords, overlooking broader context. Additionally, LSTM-based models like those used by Hu et al. (2017) [4] are powerful however they require significant computational power. In contrast, our work addresses these limitations by employing a graph-based approach that prioritizes event sequences and leverages standardized data from sources like DMOZ and Wikipedia. This enables us to make more accurate and efficient predictions, offering a practical and scalable solution that enhances predictive accuracy.

3 DATASET

3.1 Data Extraction Pipeline

The event detection pipeline processes about 300.000 English news articles per day. Each news article is first annotated using tools like entity linking, topic classification and sentiment detection. Each article is then split into sentences where each sentence retains it's annotations and other meta-data. For each pair of the entities in the sentence, an event classifier then determines if there is a particular relation of interest expressed in the sentence between the two entities. The predefined taxonomy currently includes 133 event types of interest, ranging from security, environment, natural disasters, accidents, politics, and other areas. To classify the events, a neural network transformer architecture with a pretrained encoder is used. The entire network, including the encoder, is trained on our supervised dataset using best practices like online hard example mining, class balancing, dropout, and consistency regularization. The sentences for which the classifier finds that it mentions a relation of interest are then stored in a database, together with the pair of associated entities and other available meta-data.

data = {"uri": ["2024-07-423118842-0"], "date": ["2024-07-16"], ..., "eventType": ["et/business/equity-actions/fundraising"], ..., "slots": ["http://en.wikipedia.org/wiki/Czech_Republic"]}

Figure 1: Sample of relevant data considered when parsing an event type to build the dataset.

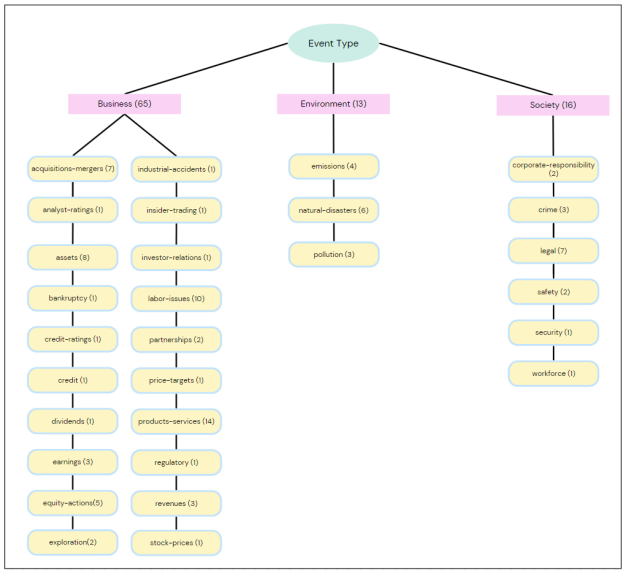


Figure 2: Event Type Taxonomy

3.2 Data Description

For our research, we used a dataset of events provided by Event Registry, with media events encoded in JSON format. Specifically, we analyzed 4,216 events related to the company 3M, recorded between June 23, 2021, and July 23, 2024. We used a URI to classify each event, drawing from DMOZ and Wikipedia categories (Fig. 1). These were selected because they provide standardized descriptions of the events being reported, which makes the data consistent and reliable. The events are categorized into 94 distinct types, which are further grouped into three primary domains: business, environment, and society. The business domain makes up the largest proportion of events, accounting for 65 types (69% of the total), while the environment and society domains contain 13 types (14%) and 16 types (16%), respectively. Within these domains, the event types are further divided into smaller subdomains, which can be aggregated into larger subdomain units as demonstrated in the event type taxonomy (Fig. 2).

4 METHODOLOGY

This study uses graph-based techniques to predict future event types from news articles about a specific company. The process starts by building a graph that maps relationships between event types and concepts from Wikipedia and DMOZ. Random walks are then performed on this graph to extract key information such as URIs, dates, and event types, which are then transformed into embeddings using Graph2Vec [8]. Next, the event types are encoded and adjusted through a process called target shifting. This step aligns the features to better forecast future outcomes based on previous data. The predictions are made using a Random Forest classifier, which is then validated through stratified k-fold

a) b) c)

Figure 3: Event Type Graphs

cross-validation for higher accuracy. The following sections will present each step of this process in more detail (see Fig. 4).

4.1 Graph Construction

For each article in the JSON dataset, a detailed graph G is generated using the NetworkX library [3]. The graph construction process starts by extracting key information such as the article's URI(unique identifier), as well as the date associated with the article and the event types, which are represented by specific URIs. In addition to these elements, each article also includes two important lists: 'slots' and 'categories'. The 'slots' list contains wiki and dmoz addresses that are directly related to the event described in the article, while the 'categories' list includes various classifications of the event. To complete the graph, labels are created by extracting URIs from the 'slots' list and filtering the 'categories' to focus on those with the "dmoz" prefix.

4.2 Random Walks for Feature Extraction

Once the graphs for each article are constructed, random walks are performed, starting at a given node (event type) and moving to adjacent nodes based on specific probabilities. Several random walks are generated for each node, forming the foundation for feature extraction processes. A single random walk begins by initializing the path with the starting node and iterating over a specified path length. At each step, a random number is compared with a probability p. If the number is less than p, the walker stays at the current node, otherwise it moves to a random neighbor. If no neighbors are available, the walk ends.

Generating multiple random walks for every node follows a similar approach, using p as the probability of staying at the current node (set at 0.05). The process involves creating an empty list to store all random walks and iterating through each node in the graph. For each node, the specified number of random walks is generated, and each walk is appended to the list.

4.3 Embedding Generation Using Graph2Vec

The random walks from the graphs are processed similarly to word sequences in a document. The 'embedding_data' function generates vector embeddings for graph data using the Doc2Vec model. It begins by converting each random walk into a TaggedDocument, storing these in 'documents_gensim'. The Doc2Vec model, with a vector size of 5 is trained on these documents, creating a vector space where similar sequences are positioned close together.

The function then processes each graph in the graphs dictionary, extracting uri, date, and event type, and generating additional random walks. These walks are converted into embeddings using the 'infer_vector' method, and the resulting vectors are averaged into one final embedding. This embedding is stored in a dictionary across 'embedding1' to 'embedding5', alongside the graph's metadata.

4.4 One Hot Encoding & Target Shifting

To transform the categorical event types into binary vectors, One hot encoding is applied. This allows the model to treat each event type as a separate class. After extracting relevant column names, the encoded target data is concatenated with the feature embeddings, creating a dataset for model training and evaluation. The dataset is then aggregated by averaging out the embeddings and calculating the maximum value of the encoded target columns for a given day. Finally the 'target' data is shifted by one day, which allows the embeddings to forecast the event types for the following day.

4.5 Random Forest Classification & Stratified K-Fold Cross Validation

To ensure an effective classification and prediction of the data, A Random Forest classifier is created. When employing this method, embeddings are used as features and the one-hot encoded event types are used as labels. The data itself is split into testing and training sets, followed by the incorporation of the Stratified K-Fold cross validation. This technique splits the data into 10 folds, while ensuring that the event type proportion in each fold remains equal. The model is then trained on 9 folds, with the remaining fold being used for validation. This ensures balanced representation of each class across the folds resulting in a more effective performance.

5 RESULTS

As mentioned above, the model was trained on a training set, and then evaluated on a test set. The training set included approximately 508 samples for each fold, and the test set included about 10% of the whole set, which amounted to 56 samples per fold. Using this, the model then predicted the probabilities for event types for each set. When training the model for each class, we noticed certain classes did not have enough occurrences to have at least one entry of such a class per dataset fold and were skipped. We, therefore, trained the model and predicted for a total of 45 classes.

To evaluate the discriminative performance of the model, the ROC AUC score was used. The results produced showed us how well the model distinguishes between different classes, as well as the model's ability to predict future event types. The ROC AUC score showed us that the average performance of the model was around 0.5674, and the median was close to it, with an AUC ROC score of 0.5559, with the highest score reaching 0.8194 and the lowest reaching a value of 0.3338. While the best scores demonstrate we can effectively forecast event types ahead of time, further work is required to enhance results, which in most cases remain close to 0.5.

6 CONCLUSIONS

This study was used to develop a graph-based approach to predicting event types in articles. In the process, we utilized random walks for feature extraction and Doc2Vec for embedding generation. Then, we trained the resulting model on a Random Forest classifier and evaluated it with a Stratified K-Fold Cross Validation. The model demonstrated solid performance with an average ROC AUC score of around 0.5674, reaching a peak at approximately 0.8194. This indicates the model's effectiveness in capturing relationships within the data and predicting future event types.

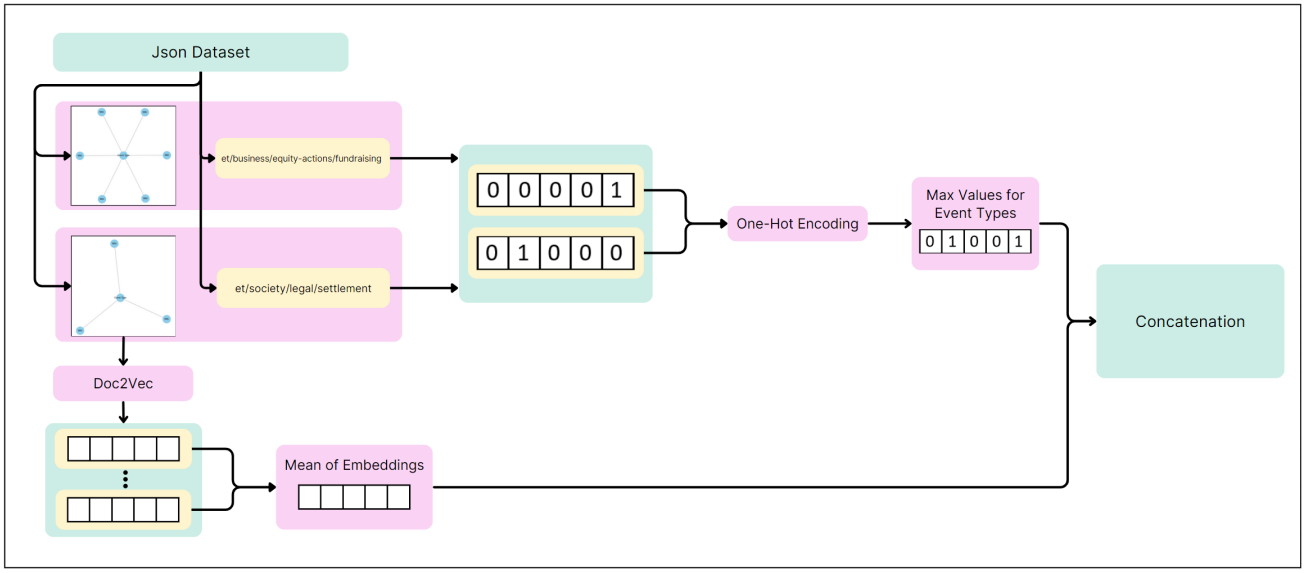


Figure 4: Data Extraction Pipeline

However, while the model performed well overall, occasional fluctuations in accuracy suggest space for further improvement. We are currently striving to find ways to make graphs more informative. In future work we could refine the feature extraction process by incorporating larger datasets, with a wider variety samples and a larger number of companies.

ACKNOWLEDGMENTS

The Slovenian Research Agency supported this work. This research was developed as part of the Graph-Massivizer project funded under the Horizon Europe research and innovation program of the European Union under grant agreement 101093202.

REFERENCES

[1] Patrick Brandtner and Marius Mates. 2021. Artificial intelligence in strategic foresight–Current practices and future application potentials: current practices and future application potentials. In *Proceedings of the 2021 12th International Conference on E-business, Management and Economics*. 75–81.

[2] Alex Fergnani, Andy Hines, Alessandro Lanteri, and Mark Esposito. 2020. Corporate foresight in an ever-turbulent era. *European business review* 25 (2020), 26–33.

[3] Aric Hagberg, Pieter J Swart, and Daniel A Schult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

[4] Linmei Hu, Juanzi Li, Liqiang Nie, Xiao-Li Li, and Chao Shao. 2017. What happens next? Future subevent prediction using contextual hierarchical LSTM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[5] Jon Iden, Leif B. Methlie, and Gunnar E. Christensen. 2017. The nature of strategic foresight research: A systematic literature review. *Technological Forecasting and Social Change* 116 (2017), 87–97. https://www.sciencedirect.com/science/article/pii/S0040162516306035

[6] Jong-Seok Kim and Dongsu Seo. 2023. Foresight and strategic decision-making framework from artificial intelligence technology development to utilization activities in small-and-medium-sized enterprises. *foresight* 25, 6 (2023), 769–787.

[7] Benjamin Letham, Cynthia Rudin, and David Madigan. 2013. Sequential event prediction. *Machine learning* 93 (2013), 357–380.

[8] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005* (2017).

[9] Freija van Duijne and Peter Bishop. 2018. Introduction to strategic foresight. *Future* 1 (2018), 67.

[10] Hanbyul Yeon, Seokyeon Kim, and Yun Jang. 2015. Visual Analytics using Topic Composition for Predicting Event Flow. *KIISE Transactions on Computing Practices* 21, 12 (2015), 768–773.

[11] Liang Zhao. 2021. Event Prediction in the Big Data Era. *Comput. Surveys* 54, 5 (2021), 1–37.

Generating Non-English Synthetic Medical Data Sets

Lenart Dolinar
University College London
London, United Kingdom

Erik Calcina
Jožef Stefan International Postgraduate School
Jožef Stefan Institute
Ljubljana, Slovenia

Erik Novak
Jožef Stefan International Postgraduate School
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Using synthetic datasets to train medicine-focused machine learning models has been shown to enhance their performance, however, most research focuses on English texts. In this paper, we explore generating non-English synthetic medical texts. We propose a methodology for generating medical synthetic data, showcasing it by generating Greeklish medical texts relating to hypertension. We evaluate our approach with seven different language models and assess the quality of the datasets by training a classifier to distinguish between original and synthetic examples. We find that the `Llama-3` performs best for our task.

Keywords

Synthetic data, healthcare data, multilingual data, large language models, classification

1 Introduction

The healthcare domain produces a lot of medical data that can be used to train machine-learning models to help medical personnel. For example, a machine-learning model designed to perform Named Entity Recognition (NER) on electronic health records (EHRs) needs extensive labeled datasets to accurately identify medical terms like diseases, treatments, and patient details. However, the data contains a lot of personal information, and hospitals cannot share it freely due to data protection. In addition, there are not enough examples to train the models for some problems, such as those relating to rare diseases. Because of this, synthetic data is being used as a substitute to train the models.

Most synthetic data generation approaches focus on generating English texts. These usually utilize large language models trained on predominantly English documents retrieved from the web. However, there are few examples of using them to generate non-English texts. Furthermore, the language models have difficulties generating texts that do not reflect the distributions found in the training sample. This includes medical texts, which are usually not accessible to the general public.

This paper proposes a methodology for generating medical synthetic data using open-source large language models. We apply the methodology to a medical data set written in Greeklish, a combination of Greek and English scripts. We test it with seven large language models and assess performance by training a classifier to distinguish original examples from synthetic ones. Using the same prompt, we find that the open-source Llama-3 model best generates synthetic data that reflects the original data set.

The remainder of the paper is as follows: Section 2 presents the related work on generating synthetic data using large language models. Next, the proposed methodology is described in Section 3. The experiment setting is presented in Section 4, followed by the experiment results in Section 5. We discuss the results in Section 6 and conclude the paper in Section 7.

Information Society 2024, 10–14 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.4

2 Related Work

This section describes the related work, focusing on large language models and methods for generating synthetic data.

2.1 Large language models

Large Language Models (LLMs) are models that were trained to generate human-like texts based on an extensive process of training on vast amounts of data. Models, such as Llama 3 [2], GPT-4 [9], Aya 23 [3] and Mistral [7], are often easy to work with by providing an input textual prompt, based on which the models respond. The LLMs are helpful in specialized fields, such as medicine, since they can be fine-tuned on extensive data sets containing medical terms and concepts. This enables them to perform well in tasks such as medical synthetic data generation [12]. Despite that, they are sometimes unable to follow the instructions in the prompt accurately, leading them to hallucinate, i.e. confidently produce wrong responses [5].

In our experiments, we investigate the LLMs' performance in generating synthetic medical data given specific constraints and detailed prompts to simulate the original data set as best as possible.

2.2 Synthetic medical data generation

Recently, synthetic medical data, generated using LLMs, has been used to enhance the performance of models for solving different natural language processing tasks in medicine.

One work focuses on generating a synthetic dataset of electronic health records of Alzheimer's Disease (AD) patients based on a label that is provided [8]. They find that the performance of their system for detecting AD-related signs and symptoms from EHRs improves vastly when trained on synthetic and original data sets as opposed to training the system only on the original one. Another work investigated using LLMs for extracting structured information from unstructured healthcare text [13]. By generating synthetic data using LLMs and fine-tuning the model, they significantly improved the models' performance for medical-named entity extraction and relation extraction tasks.

Most related works focus on English synthetic data due to scarce non-English training data and the dominance of English in medical terminology [6]. This paper focuses on generating non-English texts, specifically medical texts written in Greeklish about hypertension.

3 Methodology

This section outlines our research methodology. We first present the pre-processing of the data set, followed by describing the synthetic data generation process. Finally, we present the description of synthetic dataset evaluation using a classifier. Figure 1 shows the diagram overviewing the proposed methodology.

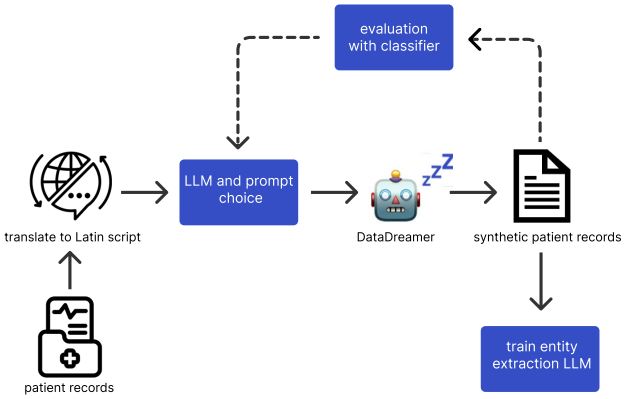


Figure 1: An overview of the methodology. The image was designed using resources from flaticon.

3.1 Data pre-processing

The data set used consisted of 1,299 examples of medical history in Greeklish, where the Latin and Greek scripts were used interchangeably. It also contained 1,495 labels, most of which were in English. The labels consisted of drugs, medical events, and measurements.

To translate the labels into Greek, we used the `NLLB-200` [14] translation model[1]. Since LLMs were predominantly trained on texts written in Latin script, we decided to transliterate both the labels and examples from Greek to Latin script. This allowed the LLMs to generate longer tokens with richer information.

We split the original data set into two subsets to ensure no data leakage. The first one, consisting of 930 examples, was used for synthetic data generation. The second one, containing the remaining 369 examples, was used for evaluation.

3.2 Synthetic data generation

We utilized the datadreamer library [10] to generate the synthetic data set. The library enables open-source models to create synthetic data sets and was developed to work in research settings, supporting prompt templates and few-shot learning.

We developed a prompt containing the instructions and restrictions on generating the examples. To better showcase the structure of the generated text, we also provided five random examples from the original data set as few-shot examples. Next, using datadreamer, we sent the prompt to the chosen LLM. We experimented with multiple LLMs, and about 800 examples were generated for each used LLM. When experimenting with LLMs that required calling an external provider (e.g., OpenAI), we provided five static few-shot examples that did not include any patient personal data due to data privacy concerns.

To ensure the quality of generated data, we implemented a post-processing step. This included formatting the generated text into one line and excluding examples where the length was too long or where the model started repeating words meaninglessly. This ensured that all generated examples followed the same format and could be used for evaluation.

Table 2 presents generated examples for the label "OSTEOPOROSH". Similarities in the examples highlight the need for rigorous methods to evaluate how closely they resemble the original data set. The methods are explained in Section 4.1.

[1]https://huggingface.co/facebook/nllb-200-distilled-600M

3.3 Technical details

In this section, we describe the models and the parameters used in the experiment. All models used are available via the HuggingFace's transformer library [15].

We tested five open-source models to generate the synthetic data sets, all of which can be run on a 32GB GPU: `Llama-3` [2] only has support for the English language but has been fine-tuned to understand user prompts, which is a feature we expected would help a lot with the synthetic data generation.[2] `Aya-23` [3] is a multilingual language model and offers support for 23 languages, including Greek.[3] `Mistral` [7] supports a variety of languages but omits Greek[4]. The models `Gemma-2` [4] and `Phi-3` [1] were also tested and compared in the experiments.[5][6] In addition, we experimented with `GPT-4o` [9] and `GPT-3.5-Turbo`, which are accessible via the OpenAI API.

All models were given the same prompt containing instructions that included (1) generating Greek texts written in Latin script and (2) containing a label randomly selected from the original data set, (3) examples are supposed to be at most 6 words long, (4) should provide concise responses, (5) structured format (all text must be in a single line, must use // and commas as separators, and must be similar in format as the provided few-shot examples). To stress some more important instructions, some instructions were given in capital letters and were also repeated.

4 Experiment Setting

This section describes the experiment setting, which consists of the evaluation process and the metrics used to measure the approach's performance.

4.1 Evaluation approach

The quality of the generated synthetic data was measured in two parts. The first consisted of statistical measurements, such as calculating the average length of the generated examples and finding the proportion of examples that included the required labels. These statistics were then compared to the original data set.

The second part consisted of training a classifier to discern if the input text was from the original or from the synthetic data set. The data set used to train and evaluate the classifier involved 369 randomly selected synthetic examples and 369 examples from the original data set, transliterated into Latin script. We chose 5-fold validation as our classification procedure and calculated the mean performance across all trials.

The classifier was trained using the BERT [11] language model, specifically the `bert-base-multilingual-cased` variant[7]. The classifier was trained using the following parameters: batch size = 16, epochs = 3, and learning rate = 2e-5. The same parameters were used for all synthetic data sets.

4.2 Metrics

To assess the quality of the generated synthetic data sets, we used the F1 score as our main metric for evaluating the classifier's performance. The target value was 0.5; if the performance is greater than 0.5, the classifier can discern the original from the synthetic examples. Hence, the synthetic data does not reflect the original data set. If the performance is less than 0.5, the

[2]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[3]https://huggingface.co/CohereForAI/aya-23-8B
[4]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[5]https://huggingface.co/google/gemma-2-9b-it
[6]https://huggingface.co/microsoft/Phi-3-medium-4k-instruct
[7]https://huggingface.co/google-bert/bert-base-multilingual-cased

classifier has difficulties separating the synthetic from the original data, which can be because the synthetic data contains copies of the original examples. In addition to the F1 score, we measured the classifier's accuracy, precision, and recall, which are also reported.

5 Results

In this section, we present the results of our experiment. We first present the statistical results, followed by the classifier's evaluation.

5.1 Statistical analysis

Table 1 compares the synthetic data sets and the original one regarding label occurrence and average example length. The label occurrence is 1.000 in the original data set, as all examples from the original data set are assumed to include relevant labels and information.

The most aligned synthetic data set regarding label occurrence was generated using `GPT-4o`, followed by `Llama-3`. However, in terms of average example length, the data set generated using `Gemma-2` performed the best, followed by `Llama-3`.

The worst-performing models, in terms of label occurrence, were `Mistral` and `Phi-3`, which in about 25% did not include the selected label. The data set generated using the `Aya-23` had the largest difference in terms of average example length, on average generating examples with three extra words.

Table 1: Statistical comparison between the original and synthetic data sets. The bold and underlined values represent the best and second-best statistics, respectively.

LLM	Label occurrence	Avg example length
original dataset	1.000	4.682
`Llama-3`	<u>0.990</u>	<u>5.330</u> (+0.648)
`Aya-23`	0.949	8.040 (+3.358)
`Mistral`	0.740	6.376 (+1.694)
`Gemma-2`	0.988	**4.207** (-0.475)
`Phi-3`	0.782	6.071 (+1.389)
`GPT-4o`	**0.996**	3.691 (-0.991)
`GPT-3.5-Turbo`	0.867	6.764 (+2.082)

Looking at both statistics, we can conclude that `Llama-3` had the best alignment to the original data set in terms of label occurrence and example length, closely followed by `GPT-4o`.

To better imagine the differences between the generated examples, we handpicked an example from each synthetic data set related to the label "OSTEOPOROSH", shown in Table 2.

5.2 The classifier evaluation

Table 3 shows the F1, Precision, Recall, and Accuracy performances of the trained classifier on different synthetic data sets. The best performance was achieved by `Mistral` with approximately 0.85 scores in all four metrics, followed by `Llama-3`, with approximately 0.88 scores in all metrics. The worst performances were on data sets generated by the `Aya-23` and `GPT-3.5-Turbo` models. Surprisingly, the `Aya-23` is a language model supporting Greek; thus, it was expected to generate better examples.

6 Discussion

This section discusses the synthetic data generation performance, outlines our methodology's limitations and drawbacks, and proposes potential improvements to the approach.

6.1 LLM performance

Results in Table 1 show significant quality differences among synthetic datasets from different LLMs, with label occurrence ranging from 0.740 for `Mistral` to 0.996 for `GPT-4o`, and average example length from 3.691 for `GPT-4o` to 8.040 for `Aya-23`.

However, Table 3 indicates no significant performance differences within a single synthetic dataset, with a maximal standard deviation of the metrics being 0.021 for the `Llama-3` dataset.

We can also notice that the F1 and accuracy scores are very close for all synthetic data sets. This means the classifier was likely performing relatively similarly on both classes (synthetic and original datasets) without significant bias to either class.

We can observe much better performance on the `Llama-3` data set, which is primarily trained on English data, than on the `Aya-23` data set, which is also trained on Greek data. This shows that a model does not need to be extensively trained in Greek texts to generate this type of synthetic medical data well.

6.2 Limitations

Due to limited computing power, only one GPU with 32GB of space was available, restricting the testing of larger LLMs. To address these challenges, using cloud-based resources or distributed computing could help run larger models and improve the variety of synthetic data generated.

Due to privacy concerns, when using `GPT-4o` and `GPT-3.5-Turbo` models, which are not locally-run models, we had to use five fixed examples when generating synthetic data instead of a larger variety. This potentially led to larger similarities of the `GPT` synthetic datasets to the examples instead of the original dataset and, consequently, worse performance.

6.3 Potential improvements

The prompt was the same for all seven LLMs and was primarily tested on `Llama-3`. Hence, the performance might be biased towards the model. The method could be improved by tailoring the prompts to each model individually.

The evaluation of synthetic datasets could be further extended by checking for repeating examples in the synthetic dataset or by checking how different the generated example is from the five provided examples. The evaluation could also be improved by checking for overfitting to the original data set.

7 Conclusion and Future Work

This paper presents a method for generating Greek synthetic medical data sets. To synthetically create datasets similar to the original, we carefully craft a prompt and perform pre-processing and post-processing of the data to increase performance and eliminate the effect of hallucinations.

Using a classifier and considering the inclusion of labels and generated text length, we conclude that `Llama-3` is best for generating examples that most closely resemble the original dataset. In the future, we plan to explore the underlying architectures of the models to understand their performance differences in multilingual contexts. This will allow us to further refine our methods and create more accurate data sets.

Furthermore, we intend to use the synthetic dataset to train a named entity recognition (NER) system to recognize medical labels from medical history examples. Measuring the performance of the NER trained on synthetic datasets will give us another way of evaluating their quality. We also intend to create a more general pipeline enabling the code to generate synthetic medical data in a wider variety of languages and formats.

Table 2: Generated examples for label "OSTEOPOROSH".

LLM	Examples
original dataset	APO 2O ETON YPERTASH ME AGOGI// OSTEOPOROSH // YPOTHYROIDISMOS
Llama-3	YPOTHYROEIDISMOS, OSTEOPOROSH, APO//
Aya-23	CA ORTHOU, ANEYRISMA KOILAKHS AORTHOU, OSTEOPOROSH.
Mistral	OSTEOPOROSH, APO 60 ETOS, APO 2 MHNES KAI APO 10 GRAMM
Gemma-2	OSTEOPOROSH, ARTHROSITIS, ETOVIR
Phi-3	OSTEOPOROSH, XAROSTHROMA, ALPHA-BISFIOVITINI, 2018, DIATHRHSH, DIA
gpt-4o	OSTEOPOROSH, ANEMIA
gpt-3.5-Turbo	OSTEOPOROSH, GASTREKTOMH, EMFISIMA, YDRONERFOSI, PSIXROS.

Table 3: Mean performance metrics of the classifier for synthetic data sets, with standard deviation. Performances that are closer to 0.5 are considered better. The bold and underlined values represent the best and second-best performances, respectively.

LLM	F1	Precision	Recall	Accuracy
Llama-3	<u>0.875 ± 0.021</u>	<u>0.881 ± 0.020</u>	<u>0.875 ± 0.020</u>	<u>0.875 ± 0.020</u>
Aya-23	0.945 ± 0.005	0.947 ± 0.004	0.945 ± 0.005	0.945 ± 0.005
Mistral	**0.848 ± 0.012**	**0.856 ± 0.001**	**0.849 ± 0.011**	**0.849 ± 0.011**
Gemma-2	0.928 ± 0.005	0.930 ± 0.005	0.928 ± 0.005	0.928 ± 0.005
Phi-3	0.927 ± 0.009	0.932 ± 0.008	0.927 ± 0.009	0.927 ± 0.009
GPT-4o	0.906 ± 0.014	0.912 ± 0.012	0.907 ± 0.014	0.907 ± 0.014
GPT-3.5-Turbo	0.940 ± 0.013	0.944 ± 0.011	0.940 ± 0.013	0.940 ± 0.013

Acknowledgments

This work was supported by the Slovenian Research Agency. Funded by the European Union. UK participants in Horizon Europe Project PREPARE are supported by UKRI grant number 10086219 (Trilateral Research). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA) or UKRI. Neither the European Union nor the granting authority nor UKRI can be held responsible for them. Grant Agreement 101080288 PREPARE HORIZON-HLTH-2022-TOOL-12-01.

References

[1] Marah Abdin et al. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. 2024. arXiv: 2404.14219 [cs.CL]. URL: https://arxiv.org/abs/2404.14219.

[2] AI@Meta. "Llama 3 Model Card". In: (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[3] Viraat Aryabumi et al. *Aya 23: Open Weight Releases to Further Multilingual Progress*. 2024. arXiv: 2405.15032 [cs.CL].

[4] Google DeepMind Gemma Team. *Gemma 2: Improving Open Language Models at a Practical Size*. 2024. URL: https://storage.googleapis.com/deepmind-media/gemma/gemma-2-report.pdf.

[5] Xu Guo and Yiqiang Chen. *Generative AI for Synthetic Data Generation: Methods, Challenges and the Future*. 2024. arXiv: 2403.04190 [cs.LG]. URL: https://arxiv.org/abs/2403.04190.

[6] Rainer Hamel. "The dominance of English in the international scientific periodical literature and the future of language use in science". In: *AILA Review* 20 (Dec. 2007), pp. 53–71. DOI: 10.1075/aila.20.06ham.

[7] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[8] Rumeng Li, Xun Wang, and Hong Yu. "Two Directions for Clinical Data Generation with Large Language Models: Data-to-Label and Label-to-Data". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 7129–7143. DOI: 10.18653/v1/2023.findings-emnlp.474.

[9] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[10] Ajay Patel, Colin Raffel, and Chris Callison-Burch. *DataDreamer: A Tool for Synthetic Data Generation and Reproducible LLM Workflows*. 2024. arXiv: 2402.10379 [cs.CL]. URL: https://arxiv.org/abs/2402.10379.

[11] Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493.

[12] Karan Singhal et al. "Large language models encode clinical knowledge". In: *Nature* 620 (2023), pp. 172–180. DOI: 10.1038/s41586-023-06291-2.

[13] Ruixiang Tang et al. *Does Synthetic Data Generation of LLMs Help Clinical Text Mining?* 2023. arXiv: 2303.04360 [cs.CL]. URL: https://arxiv.org/abs/2303.04360.

[14] NLLB Team et al. *No Language Left Behind: Scaling Human-Centered Machine Translation*. 2022. arXiv: 2207.04672 [cs.CL]. URL: https://arxiv.org/abs/2207.04672.

[15] Thomas Wolf et al. "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

LLNewsBias: A Multilingual News Dataset for Lifelong Learning

Swati Swati
swati.swati@unibw.de
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute and
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract

The rise of digital media enhances information accessibility but also introduces challenges related to the quality and impartiality of news reporting, particularly regarding biases that influence public perception during key global events. In response, this study introduces *LLNewsBias*, a dataset designed to detect and analyze political bias in multilingual news headlines, covering four major events from 2019 to 2022 — Brexit, COVID-19, the 2020 U.S. election, and the Ukraine-Russia war. With over 350,000 headlines in 17 languages, annotated with bias labels, this dataset is compiled using Media Bias/Fact Check and Event Registry. Our contributions include a structured framework for data collection and organization, enabling event-wise and year-wise analysis while supporting lifelong learning. We also highlight potential use cases that demonstrate the dataset's utility in advancing bias prediction models, multilingual adaptation, and model robustness. Additionally, we discuss the dataset's limitations, addressing potential biases, sample size constraints, and contextual factors. This work provides a valuable resource for improving bias detection in dynamic, multilingual news environments, contributing to the development of more accurate and adaptable models in natural language processing and media studies. For code and additional insights, visit: https://github.com/Swati17293/LLNewsBias

Keywords

Dataset, News, Bias, Multilingual, Headline, Low-resource, Media Bias, News Bias, Continual Learning, Lifelong Learning

1 Introduction

The rapid growth of digital media has greatly enhanced the accessibility of information, but it has also introduced significant challenges concerning the quality and impartiality of news reporting. Political bias in news content is particularly concerning, as it has the potential to influence public perception and shape societal narratives, especially around key global events. Understanding and predicting such biases, particularly in multilingual contexts where biases can manifest differently across cultural and linguistic boundaries, is essential for promoting fair and balanced journalism. Traditional approaches to bias detection often rely on monolingual datasets and static models that may not effectively capture the evolving nature of news content [6]. These limitations underscore the need for more robust datasets and methodologies that can adapt to the dynamic and multilingual landscape of modern news reporting.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.8

In this study, we address these challenges by introducing a novel dataset *LLNewsBias* specifically designed for the detection and analysis of political bias in multilingual news headlines. Our dataset spans four major global events from 2019 to 2022: Brexit, COVID-19, the 2020 U.S. election, and the Ukraine-Russia war, capturing a wide range of political discourse across 17 languages. To collect this dataset, we use Media Bias/Fact Check for the assignment of bias labels, and Event Registry [2] for the extraction of relevant headlines and metadata. The resulting dataset is not only comprehensive in its linguistic diversity but also structured to support both event-wise and year-wise analyses, with an emphasis on lifelong learning.

1.1 Contributions

Our study makes the following contributions:

- **Multilingual bias-annotated dataset:** We introduce a multilingual bias-annotated dataset containing over 350,000 news headlines in 17 languages, each annotated with political bias labels.
- **Data collection and organization framework:** We present a structured framework for data collection and organization, enabling event-wise and year-wise analysis while ensuring adaptability for lifelong learning.
- **Potential use-cases:** We outline several potential applications of our dataset, highlight its potential for advancing lifelong learning models, particularly in bias prediction, multilingual adaptation, and model robustness.
- **Discussion of limitations:** We identify and discuss the dataset's limitations, such as biases in data collection, sample size constraints, and contextual influences, offering a transparent assessment of its applicability.

In summary, our paper introduces a comprehensive dataset and a framework for the study of political bias in multilingual news headlines. By focusing on key global events and providing support for lifelong learning, our study contributes to the ongoing effort to develop more accurate and adaptable models for bias detection in diverse linguistic and cultural contexts.

2 Related Work

Several datasets focus on news articles and political bias [5], but there is a notable scarcity of multilingual, bias-annotated datasets designed for lifelong learning [4]. While resources like the *media bias chart* by Ad Fontes Media and PolitiFact provide insights into bias, they are often limited to English-language sources or specific fact-checked claims, lacking the continuous, event-centric data necessary for broader analysis. GDELT [3], a large-scale event-oriented news dataset, covers multiple languages but focuses on location, network, and temporal attributes rather than political bias or the event-outlet relationship. Existing multilingual datasets are often domain-specific [1], limiting

their utility for general bias analysis. In contrast, LLNewsBias dataset fills these gaps by offering a generalized, multilingual, and bias-annotated data designed for event-wise and year-wise analyses, particularly suited for lifelong learning models.

3 Dataset Description

In this section, we introduce our dataset *LLNewsBias* and describe the framework used for its collection and organization. We begin by detailing the primary data sources that form the foundation of this dataset. Following this, we present a comprehensive overview of the data collection process, with a focus on the methodologies employed to ensure robustness and reliability. Finally, we provide an in-depth overview of the dataset's structure, including its directory organization, file contents, and the various ordering methods applied to facilitate detailed analysis. Our dataset is documented in accordance with the FAIR Data Principles.

3.1 Primary Data Sources

In this section, we outline the two primary data sources used in our study: Media Bias/Fact Check (MBFC) and Event Registry (ER). MBFC serves as the bias rating portal, providing bias labels for selected media outlets, while ER is used to extract the headlines and corresponding metadata from articles published by these outlets.

3.1.1 ***Media Bias/Fact Check.*** For bias labeling in this study, we utilized Media Bias/Fact Check (MBFC), a well-established platform known for its comprehensive coverage and frequent updates. Although other platforms like allsides.com and adfontes-media.com also provide bias ratings, MBFC was selected for its reliability and particular focus on low-resource languages. MBFC assigns bias labels based on political orientation and evaluates outlets for credibility and factual accuracy. These labels are determined by a team of contractors and volunteers who follow a standardized methodology, ensuring that the ratings are both consistent and dependable for our analysis.

3.1.2 ***Event Registry.*** In this study, we use Event Registry [2] platform as the primary source for collecting multilingual news headlines. It aggregates content from over 150,000 news sources across more than 60 languages, making it an ideal resource for analyzing bias in diverse and low-resource languages. Apart from the headlines, it allows access to numerous metadata such as publication date, news category, and political bias. By leveraging its Python API, we efficiently filtered and extracted headlines relevant to our study. This ensured a comprehensive dataset that supports the analysis of bias in a lifelong learning setup, exploring how emerging events and domain shifts influence the performance of bias prediction models over time.

3.2 Data Collection Framework

Our data collection framework as depicted in Figure 1, is designed to support both event-wise and year-wise analyses, with the additional capability of facilitating lifelong learning.

For data collection, we begin by defining two sets: a set of significant global events ($E = \{e_1, e_2, ..., e_n\}$), and a set of years ($Y = \{y_1, y_2, ..., y_m\}$), where n and m represent the total number of events and years, respectively. We then use the Media Bias/Fact Check (MBFC) platform to select media outlets ($O = \{o_1, o_2, ..., o_p\}$) and determine their respective political bias, with p as the total number of outlets. To maintain data reliability, we exclude outlets labeled as questionable and assign each remaining outlet $o_i \in O$ a bias label $b_i \in B$, where $B = \{b_1, b_2, ..., b_q\}$ represents the set of bias labels, with q representing the number of distinct bias labels.

Next, we define a temporal query Q_t to extract article headlines ($H = \{h_1, h_2, ..., h_r\}$), where r represents the total number of headlines retrieved from the Event Registry (ER). The query Q_t is formulated as:

$$Q_t = \{Q_e, Q_o, Q_{cat}, Q_{dt}\} \tag{1}$$

where Q_e, Q_o, Q_{cat} specify the event, media outlet, and news categories (limited to those classified as 'news' by ER $Q_{cat} = \{$'*politics*', '*business*', '*sports*', '*arts and entertainment*', '*science*', '*technology*', '*health*', '*environment*'$\}$), respectively. The time constraint is represented as $Q_{dt} = [Q_{sd}, Q_{ed}]$, where Q_{sd} and Q_{ed} denote the start and end dates. To scrape all the article headlines (H), we utilize Q_t to query ER.

We then associate the extracted headlines H with the corresponding bias labels in B and structure the dataset according to two classification types: event-wise and year-wise. To organize the data, we define an event-based order O_{event} and a year-based order O_{year} as follows:

$$O_{event} = \{e_1 \rightarrow e_2 \rightarrow ... \rightarrow e_n\} \tag{2}$$

$$O_{year} = \{y_1 \rightarrow y_2 \rightarrow ... \rightarrow y_m\} \tag{3}$$

For lifelong learning, we design the dataset to be extendable, allowing for the integration of new events and years as they emerge, denoted by $E' \subseteq E$ and $Y' \subseteq Y$, where E' and Y' represent the sets of newly added events and years.

We designed the dataset with a flexible framework that allows for the seamless integration of new events and years as they emerge, represented as $E' \subseteq E$ and $Y' \subseteq Y$, where E' and Y' denote the newly added events and years. This structured approach ensures scalability for continuous learning without requiring major restructuring and supports the training of adaptive models capable of integrating new information effectively. Unlike standard multi-year datasets, our dataset includes annotations that facilitate contextual understanding, enabling models to learn from historical data while adapting to evolving trends and patterns in news reporting. This ensures that the models remain relevant as new information becomes available.

Finally, we split the dataset into training and test sets using a stratified sampling approach to ensure the preservation of bias label distributions across both events and years. We perform this step as it is critical for maintaining the integrity of the model training process in a lifelong learning context.

3.3 Data Synopsis and Structure

In this section, we present an overview of the data and explain how it is systematically organized, making it easier to understand both the content and format of our dataset.

3.3.1 ***Data Synopsis.*** The dataset features 356,060 headlines on four major events from 2019 to 2022: *Brexit*, *COVID-19*, *the election*, and *the Ukraine-Russia war*. These headlines, sourced from 45 unique news outlets in 17 different languages, are annotated with 3 political bias labels: *Left Centre*, *Least Biased*, and *Right Centre* covering diverse topics such as *politics*, *business*, *arts and entertainment*, *sports*, *science*, *technology*, *health*, and *environment*. The dataset is structured into 7 distinct columns within .csv files. Table 1 presents a comprehensive summary of the dataset statistics.

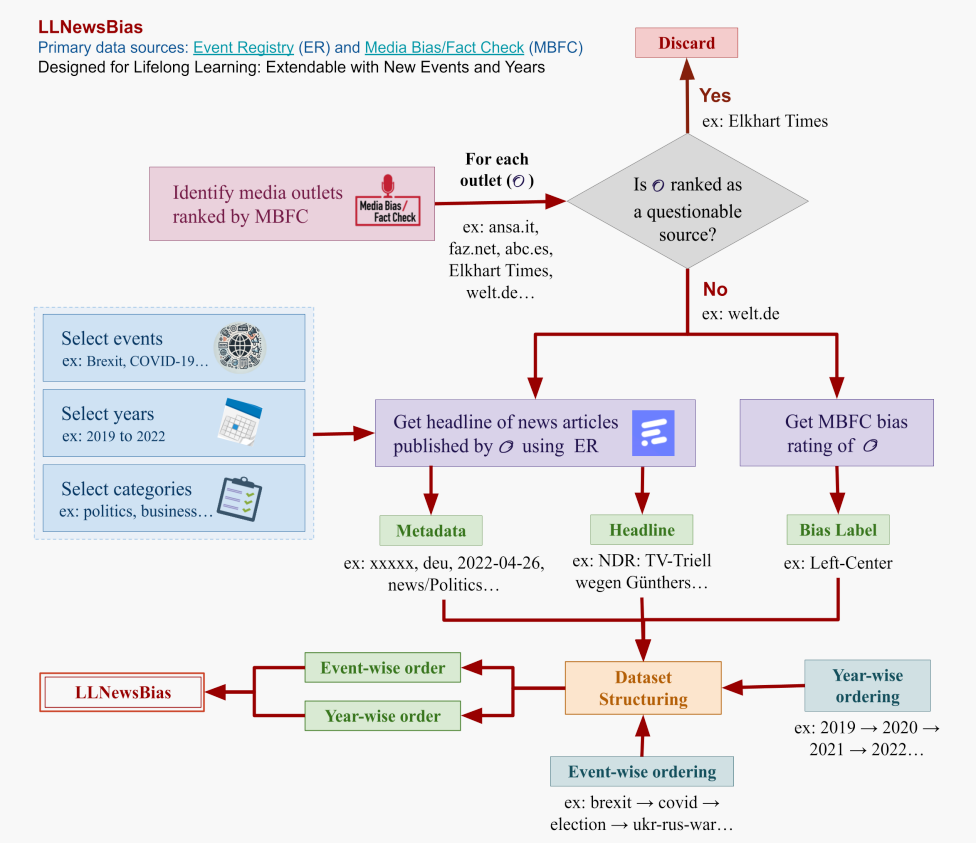


Figure 1: Data Collection Framework. The framework uses MBFC for bias labeling and ER for headline retrieval.

Table 1: Summary of Dataset Statistics.

Language-wise Distribution			
Catalan	882	Romanian	17,038
Croatian	13,929	Russian	10,511
Czech	1,876	Slovak	5,642
Danish	4,330	Spanish	83,940
Dutch	10,905	Swedish	6,441
Finnish	1,512	Ukrainian	10,616
French	85,007	Italian	48,450
Hungarian	105		
Event-wise Distribution			
Brexit	32,286	COVID	309,329
Election	3,829	Ukraine	10,616
Year-wise Distribution			
2019	20,664	2021	4,638
2020	258,871	2022	71,887

3.3.2 ***Directory Structure.*** The dataset is organized in a main 'data' directory with subdirectories categorized by events ('brexit', 'covid', 'election', 'ukr-rus-war') and years (2019-2022). Additional subdirectories consolidate data across all events (ordered_events) and all years (ordered_years). Each subdirectory contains .csv files for training and testing, structured across the following columns.

- **news outlet**: The name of the news outlet.
- **article_ID**: A unique identifier for the raw news article in the Event Registry platform from which the headlines are extracted.
- **language**: The source language of the published news article.
- **date**: The date on which the news was published.
- **headline_text**: The text of the news headline.
- **news_category**: The category assigned by Event Registry.
- **political_bias**: The political bias of the news outlet as provided by the bias rating portal Media Bias/Fact Check.

The dataset is annotated with bias labels: Left Centre (LC), Least Biased (LB), and Right Centre (RC). To ensure model robustness across varying data distributions, we concatenate and shuffle files for each event and year in four distinct random orders. This prevents overfitting to specific sequences and helps evaluate generalization across diverse configurations. While chronological order is ideal for practical use, this randomized approach tests broader performance, with the original event and year splits provided for user flexibility.

Event-wise Ordering:

(1) brexit → covid → election → ukr-rus-war
(2) election → covid → ukr-rus-war → brexit
(3) brexit → ukr-rus-war → election → covid
(4) covid → brexit → ukr-rus-war → election

Year-wise Ordering:

(1) 2019 → 2020 → 2021 → 2022
(2) 2021 → 2020 → 2022 → 2019
(3) 2019 → 2022 → 2021 → 2020

(4) 2020 → 2019 → 2022 → 2021

The dataset captures the distribution of headlines related to various events over the years, reflecting the temporal dynamics of news coverage and the evolving reporting on these events. The differences in coverage levels reveal important patterns in media attention, which are essential for developing datasets that support lifelong learning models.

4 Potential Use-Cases

Our dataset introduced in this study has a wide range of potential use-cases, particularly in the fields of natural language processing and media studies. It is particularly valuable for research and applications that require understanding and predicting news bias in a continual, multilingual environment. Below we list some potential use cases:

- **Lifelong learning for news bias prediction:** Our dataset is ideal for developing and testing lifelong learning models. It allows models to adapt to new events and evolving entities. With its year-wise structure from 2019 to 2022, the dataset addresses the challenges of emerging events and domain shifts (e.g., Brexit, COVID-19, Ukraine-Russia War), providing the data needed to develop and evaluate robust models.

- **Domain Adaptation in Multilingual Contexts:** Our dataset enables researchers to investigate domain adaptation techniques in a multilingual context, featuring headlines in 17 languages. This facilitates the development of models that generalize across languages and adapt to various cultural and political contexts, ensuring accurate bias prediction. It addresses the challenges faced by generic models in the news domain, which often struggle with topic and language diversity.

- **Sparse Experience Replay for Continual Learning:** Our dataset is particularly well-suited for the news domain, supporting efficient experience replay by allowing the selection of specific topics and categories. With its event-wise and year-wise classifications, our dataset enhances memory utilization, improves generalization, reduces catastrophic forgetting, and ensures that models remain accurate and up-to-date in real-time applications.

In a nutshell, our dataset serves as a valuable resource for advancing news bias prediction, particularly in the context of lifelong learning, by providing a flexible framework for integrating new events and years. Unlike many news-based datasets with timestamps, it offers structured annotations and contextual information that enhance the understanding of evolving trends in news coverage, making it particularly suitable for lifelong learning applications. It supports a range of research activities, from model development and evaluation to the exploration of new techniques for handling dynamic and multilingual news environments.

5 Limitations

Several limitations are associated with the dataset presented in this article and should be carefully considered in any further research or analysis:

- **Data Collection Issues:** The dataset was gathered using Media Bias Fact/Check (MBFC) and the paid version of Event Registry (ER). MBFC is publicly accessible, while ER provided comprehensive but limited coverage, potentially missing relevant articles. The use of ER's paid version also restricted the extent of data collection.
- **Sample Size:** The dataset is constrained by its focus on four major events over a span of four years. This limited number of events and time frame may not fully capture the broader spectrum of news and media biases, affecting the diversity of the samples.
- **Biases:** Selection bias is a significant factor, as only news outlets labelled by MediaBiasFactCheck were included. This restriction may limit the number of languages and perspectives represented in the dataset, thereby influencing the overall analysis.
- **Contextual Factors:** The dataset is limited by its temporal scope, covering only four specific events over four years. While it reflects the dynamic nature of news media, it does not account for all future events and years to come.

6 Conclusions

In this study, we present LLNewsBias, a comprehensive dataset designed to tackle the challenges of detecting and analyzing political bias in multilingual news headlines. By spanning four major global events from 2019 to 2022 across 17 languages, this dataset provides a valuable resource for research in natural language processing and media studies. Our framework supports both event-wise and year-wise analysis, emphasizing lifelong learning and enabling models to adapt continuously to new data. The dataset's potential use cases include enhancing bias prediction models, facilitating domain adaptation in multilingual contexts, and improving model robustness. While LLNewsBias offers significant contributions, we also acknowledge limitations such as potential biases in data collection, sample size constraints, and contextual factors. Addressing these challenges in future work will be crucial for maximizing the dataset's impact, ultimately contributing to fairer and more balanced journalism.

7 Acknowledgments

This work was supported by the Slovenian Research Agency and National grants (CRP V2-2272; V5-2264; CRP V2-2146) and by the European Union through enrichMyData EU HORIZON-IA project under grant agreement No 101070284 and ELIAS HORIZON-RIA project under grant agreement No 101120237.

References

[1] Jason Armitage, Endri Kacupaj, Golsa Tahmasebzadeh, Swati, Maria Maleshkova, Ralph Ewerth, and Jens Lehmann. 2020. Mlm: a benchmark dataset for multitask learning with multiple languages and modalities. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2967–2974.

[2] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, 107–110.

[3] Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: global data on events, location, and tone, 1979–2012. In *ISA annual convention*. Vol. 2, 1–49.

[4] Swati Swati, Adrian Mladenić Grobelnik, Dunja Mladenić, and Marko Grobelnik. 2023. A commonsense-infused language-agnostic learning framework for enhancing prediction of political bias in multilingual news headlines. *Knowledge-Based Systems*, 277, 110838.

[5] Swati Swati, Dunja Mladenić, and Tomaž Erjavec. 2021. Eveout: an event-centric news dataset to analyze an outlet's event selection patterns. *Informatica*, 45, 7.

[6] Swati Swati, Dunja Mladenić, and Marko Grobelnik. 2023. An inferential commonsense-driven framework for predicting political bias in news headlines. *IEEE Access*.

Creating Local World Models using LLMs

Mark David Longar
Jožef Stefan Institute
Ljubljana, Slovenia

Erik Novak
Jožef Stefan Institute
Ljubljana, Slovenia

Marko Grobelnik
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

A key limitation of state-of-the-art large language models is their lack of a consistent world model, which hinders their ability to perform unseen multi-hop reasoning tasks. This paper addresses this by extracting local world models from text into a systematic first-order logic framework, enabling structured reasoning. Focusing on the educational domain, we present a multi-step approach using Prolog to represent and reason with these models. Our method involves segmenting educational texts, generating Prolog definitions, and merging them into a comprehensive knowledge graph. We successfully extracted several small models and manually verified their accuracy, demonstrating the potential of this approach. While promising, our results are currently limited to small-scale models.

Keywords

Large language models, local world models, knowledge representation, educational technology, structured reasoning, knowledge graphs

1 Introduction

In recent years, Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP), offering unprecedented capabilities in understanding, reasoning over, and generating human-like text. Despite their impressive performance across various language tasks, a significant limitation persists – the absence of a consistent and coherent world model within these systems [8]. This limitation hampers their ability to perform advanced reasoning tasks that require not only textual understanding but also logical consistency and structured knowledge representation.

While current LLMs are powerful, they are inherently constrained by their reliance on statistical correlations within vast datasets, often resulting in shallow and contextually inconsistent reasoning. To address this limitation, we propose an approach for extracting local world models, i.e., small, context-specific representations of knowledge that capture the relationships and rules governing a particular domain or scenario. The approach is multi-step. First, the input text is segmented into manageable parts. Each segment is analyzed to extract key concepts and their interrelationships, which are then represented as Prolog definitions. Then, the definitions are merged into a comprehensive knowledge graph that reflects the structure and content of the input text.

We focus specifically on the educational domain, where the ability to generate and utilize local world models could significantly enhance the effectiveness of AI-driven educational tools, e.g. by providing LLMs a framework for responding with logically consistent and pedagogically sound explanations. Moreover, by modifying some of the components, the approach can also be applied to other domains, such as industry, finance, and law.

The remainder of the paper is as follows: Section 2 presents the related work on LLMs and creating world models. Next, the proposed approach is described in Section 3. The experiment setting is presented in Section 4, followed by the experiment results in Section 5. We discuss the results in Section 6 and conclude the paper in Section 7.

2 Related Work

The recent surge in large language models, such as GPT-3 [3] and GPT-4 [1], has significantly advanced natural language processing, showing emergent reasoning abilities across various tasks. However, despite their impressive performance, LLMs are often criticized for lacking factual consistency, interpretability, and logical coherence, especially in complex, multi-hop reasoning tasks [8]. To address these shortcomings, efforts have been made to integrate LLMs with structured knowledge frameworks, like knowledge graphs (KGs) and ontologies, to enhance reasoning and knowledge flow between structured data and language models [9].

In the field of ontology and KG development, early initiatives like Cyc [6] laid the groundwork for large-scale structured knowledge representation. More recent efforts [8, 5] have explored using LLMs to assist in ontology generation and KG construction. While LLMs can automate parts of the ontology development process, they struggle with ensuring logical consistency and managing complex domain-specific knowledge [5, 2]. Complementary approaches, like using LLMs for ontology learning [2] and structured knowledge extraction [10], highlight the need for human validation and formal methods to ensure accuracy.

Our work builds on these insights by focusing on using LLMs to extract structured local world models in the form of Prolog-based representations. This approach addresses the limitations of LLMs in handling complex reasoning and provides a more robust, logically consistent framework for educational applications.

3 Methodology

This section introduces the approach for creating local world models by generating and utilizing structured data in Prolog. The methodology is designed to systematically identify and map the concepts and their interrelationships within a given educational document, such as a textbook, facilitating the generation of a knowledge graph.

3.1 Document segmentation

To manage the document's complexity and ensure accurate concept extraction, the source material was divided into several shorter parts, each up to 10 pages long. This segmentation was crucial in allowing us to focus on smaller, more manageable sections of the content, enabling a thorough analysis and avoiding problems that come with long-context LLM outputs. The length of each part was determined based on the natural divisions within

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.22

the text, such as chapters or major sections, to maintain the coherence of concepts within each segment.

3.2 Generating Prolog definitions

For each segmented part, we created a prompt to generate Prolog definitions of the concepts and their relationships. The prompt was carefully crafted to guide the extraction of educational content in a structured format. It consisted of three main components: the context, the predicates and the structured output.

Context. A description of the educational context and a brief narrative to position the content within a learning scenario. This helped to align the LLM-extracted concepts and relationships with our downstream tasks. The following is an example of the prompt used:

> *You are a teacher and an expert in natural language processing (NLP). You wrote a chapter in an NLP textbook and would like to convert the content of the chapter into a classroom lesson. You would like to step into the shoes of a student in order to understand their learning process of this material. You need to understand which concepts are being taught and their relationships.*

Predicates. List of predicates and their descriptions, which were essential for identifying concepts (`isConcept(A)`), prerequisites (`isPrerequisiteOf(A, B)`), and sections (`isSection(S)`). These predicates were used to simulate the learning process, where concepts are linked to sections. A concept may have prerequisite concepts or sections that must be understood before a student can advance to learning the concept.

Structured output. Clear instructions to output the extracted predicates in the form of a Prolog program. The LLM responding in a structured format a crucial part of our approach, as it has been shown that structured responses can improve LLM reasoning and generation quality [13].

In summary, this prompt allowed us to extract detailed summaries of the concepts taught and their relationships, which were then represented in Prolog. Each segment was processed independently to generate a corresponding Prolog program.

3.3 Merging Prolog definitions

After generating the Prolog definitions for each segment, the next step was to merge them into a single cohesive program. To achieve this, we created a prompt, which was nearly identical to the first, but with instructions to combine the disjoint parts into one integrated Prolog program added to the end of the prompt:

> *Now you need to combine the parts into a single Prolog program. Make sure to include all the concepts and relationships, but also properly connect them. Merge concepts from different sections where necessary and make sure to include all the sections and their relationships.*

3.4 Use of the knowledge graph

The generated knowledge graph, represented by the Prolog program, was then used to recommend the next steps in the learning process. Using the structured output, we created a detailed concept map that helped identify key learning paths and prerequisites. Prolog (specifically SWI Prolog [11]) was chosen for this task because it can handle structured data, is widely used (increasing the likelihood that LLMs have encountered it during training), and can be executed and analyzed immediately.

4 Experiment Setting

This section outlines the experiment setting for evaluating our approach to extracting local world models from educational texts and generating structured Prolog representations. We describe the data sources, the large language model used, and the evaluation framework.

4.1 Data sources

We evaluated our approach on two widely used textbooks in deep learning and natural language processing. These texts were chosen because they are relevant to both structured reasoning tasks and the representation of complex, multi-step concepts. The following chapters were selected for analysis:

Deep Learning Preliminaries from the book *Dive into Deep Learning* [12]. This chapter provides foundational knowledge of deep learning, covering key concepts such as linear algebra, calculus, and probability, which are essential for understanding the field. The textbook's teaching approach is highly hands-on, with a significant portion devoted to code. It is open-sourced, and we used the Markdown files provided on their GitHub page[1].

Chapter 2: Regular Expressions, Tokenization, and Edit Distance from *Speech and Language Processing* [4]. This chapter introduces basic NLP techniques, focusing on regular expressions and tokenization, which are pivotal in text preprocessing tasks.

4.2 Used large language model

We employed GPT-4o via the ChatGPT interface to extract concepts and their interrelationships. We leveraged the model's multimodal capabilities, allowing it to process text and PDF documents.

4.3 Evaluation Framework

We developed an evaluation framework to assess the performance of our approach based on three primary aspects: accuracy, completeness, and consistency. To validate the results, we manually reviewed the extracted knowledge graphs and compared them with the source texts. We ensured that the extracted concepts were accurate, complete, and logically consistent.

Assessment Criteria. The following criteria were used to evaluate the effectiveness of our approach:

- *Accuracy.* This aspect examines how accurately the approach extracted the concepts and their relationships from the text. We evaluated the correctness of each Prolog definition against the source material.
- *Completeness.* This evaluates whether the system captured all the key concepts from the educational material. The assessment ensured that no significant concepts or relationships were omitted during extraction.
- *Consistency.* This aspect assesses the extent to which the extracted models maintained logical coherence across different

[1] https://github.com/d2l-ai/d2l-en

segments of the text. This was crucial in determining whether the segmented Prolog definitions could be merged into a cohesive KG.

5 Results

In this section, we review the knowledge graphs of the two tested texts generated by our model.

5.1 *Dive into Deep Learning*

The selected chapter covered six sub-chapters in the following order: Data Manipulation, Data Preprocessing, Linear Algebra, Calculus, Automatic Differentiation, and Probability and Statistics. The results are represented by the graph in Figure 1.

The system accurately identified three major independent branches of the chapter – Linear Algebra, Calculus, and Probability and Statistics – which reflects the structure of the source material. The extracted knowledge graph also logically restructured the content in ways that differed from the original organization but made sense pedagogically. This restructuring highlights the logical flow of how data handling techniques naturally feed into more abstract mathematical concepts despite differing from the original structure.

However, some omissions and reassignments were noted, particularly within the Linear Algebra section. Concepts such as vectors and matrices were omitted, likely due to the high-level nature of the extraction process. Additionally, matrix multiplication, though identified, was separated from Linear Algebra basics and Tensor operations. This disjunction represents a slight deviation from the expected conceptual hierarchy.

Similarly, in the Calculus section, the extracted model restructured the sequence of topics. This restructuring captured the relationship between fundamental calculus concepts and their practical applications in machine learning. Furthermore, the system included concepts like Gradient Descent and Backpropagation which were only briefly mentioned in the source material.

5.2 *Speech and Language Processing*

The Regular Expressions section, seen in Figure 2, was extracted accurately, capturing the core concepts effectively. However, a noticeable limitation was the loss of the original sequencing of the concepts presented in the textbook. While the key ideas were identified, the pedagogical flow, which is essential for gradual learning, was somewhat disrupted in the extraction process.

For the other sections, including Tokenization and Edit Distance, the model extracted only the most prominent concepts, omitting many important details. As a result, these sections are less comprehensive than they need to be for in-depth understanding. Despite this, the overall connections between sections in the knowledge graph were logically structured, showing that the system was still able to create a coherent representation of the material at a high level.

It is important to note that this textbook is significantly more information-dense and longer compared to the *Dive into Deep Learning* book. This added complexity exposed some limitations in the current approach, mainly when dealing with texts that require detailed extraction of concepts and their interrelationships. The model's ability to handle such dense material is limited by its tendency to focus on top-level ideas while losing much of the depth and sequencing provided in the source text. Additionally, there were rare occasions where the output required manual interventions to fix inconsistent formatting of the Prolog variable names.

6 Discussion

Our approach to extracting local world models from educational texts demonstrated strong performance in generating logically coherent knowledge graphs from high-level concepts, but certain limitations were identified. The synthetic data generation effectively captured core concepts from both textbooks, particularly in structuring major branches such as Linear Algebra, Calculus, and Probability from *Dive into Deep Learning*. However, some restructured sections, while logical, differed significantly from the source material's flow.

In the *Speech and Language Processing* textbook, the Regular Expressions subsection was extracted with sufficient accuracy. Other sections, such as Tokenization and Edit Distance, suffered from detail omissions, where only top-level concepts were extracted. This issue was more prominent due to the higher information density of the NLP textbook, exposing limitations in handling detailed, densely packed content.

Regarding the evaluation framework, the model generally performed well on metrics like accuracy and consistency but struggled with completeness in more detailed sections. The model's tendency to restructure content logically, though sometimes deviating from the original, suggests that while it captures core relationships, further refinements are needed to preserve pedagogical flow and details.

6.1 Potential improvements

To address the limitations, improving the prompt engineering could lead to more detailed extractions while maintaining the structure of the source material. Additionally, enhancing the model's ability to handle complex, dense information would mitigate the loss of key concepts. Future iterations may benefit from automated post-processing checks to ensure logical consistency and reduce manual interventions. Overall, while the approach shows promise, refining it to handle finer details and complex sequences more effectively will be essential for broader applications.

7 Conclusion and Future work

In this paper, we proposed a novel approach to extracting local world models from educational texts by generating structured Prolog representations. Our methodology demonstrated the ability to capture core concepts and their interrelationships in a logical and coherent manner, especially in the *Dive into Deep Learning* textbook. However, the results from the more information-dense *Speech and Language Processing* text revealed limitations, particularly in handling detailed content, large knowledge graphs, as well as preserving pedagogical flow.

The use of Prolog proved effective in organizing educational material, allowing for structured reasoning and enabling applications in AI-driven educational tools. Despite these successes, certain challenges remain, such as the omission of detailed concepts and the system's occasional tendency to deviate from the original sequence of topics.

Future work will address these limitations by improving the prompt engineering and enhancing the system's ability to handle complex, information-dense material. Additionally, we plan to explore automating the segmentation process and scaling up the

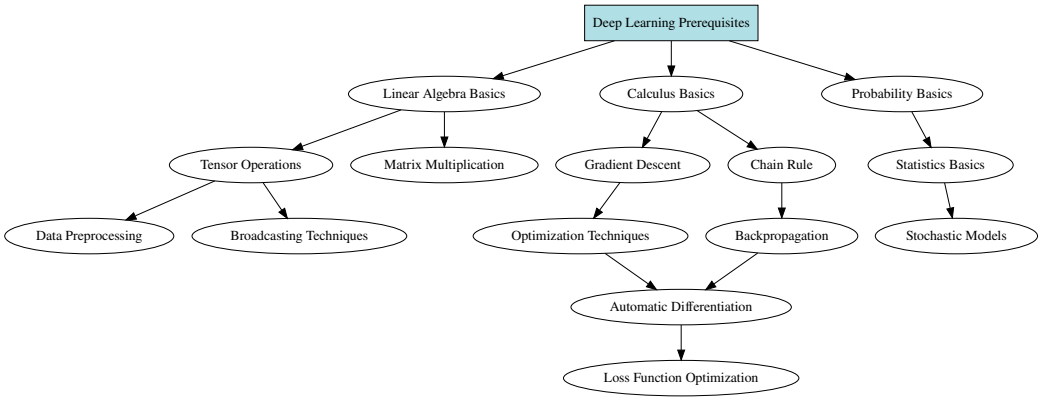


Figure 1: Knowledge graph of the *Preliminaries* section from *Dive into Deep Learning*.

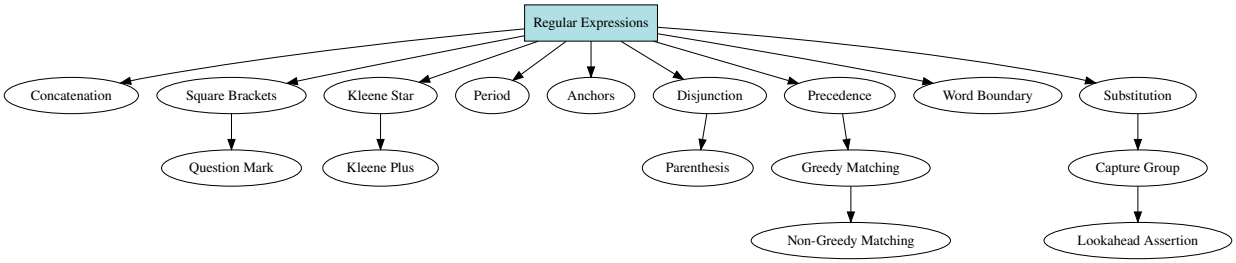


Figure 2: Knowledge graph of the *Regular Expressions* section from *Speech and Language Processing*.

model to generate larger, more intricate knowledge graphs. Other potential directions include integrating retrieval-augmented generation [7] to enrich knowledge extraction and comparing generated world models across different texts to evaluate their pedagogical alignment. Self-evaluation and correction mechanisms could also be introduced to improve accuracy and completeness.

Acknowledgments

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 project Humane AI Net (Grant No. 952026).

References

[1] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[2] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. "LLMs4OL: Large language models for ontology learning". In: *International Semantic Web Conference*. Springer. 2023, pp. 408–427.

[3] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. (Visited on 08/27/2024).

[4] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released August 20, 2024. 2024. URL: https://web.stanford.edu/~jurafsky/slp3/.

[5] Vamsi Krishna Kommineni, Birgitta König-Ries, and Sheeba Samuel. "From human experts to machines: An LLM supported approach to ontology and knowledge graph construction". In: *arXiv preprint arXiv:2403.08345* (2024).

[6] Douglas B Lenat. "CYC: A large-scale investment in knowledge infrastructure". In: *Communications of the ACM* 38.11 (1995), pp. 33–38.

[7] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[8] Fabian Neuhaus. "Ontologies in the era of large language models–a perspective". In: *Applied ontology* 18.4 (2023), pp. 399–407.

[9] Shirui Pan et al. "Unifying large language models and knowledge graphs: A roadmap". In: *IEEE Transactions on Knowledge and Data Engineering* (2024).

[10] Mohammad Javad Saeedizade and Eva Blomqvist. "Navigating Ontology Development with Large Language Models". In: *European Semantic Web Conference*. Springer. 2024, pp. 143–161.

[11] Jan Wielemaker et al. "SWI-Prolog". In: *Theory and Practice of Logic Programming* 12.1-2 (2012), pp. 67–96. ISSN: 1471-0684.

[12] Aston Zhang et al. *Dive into Deep Learning*. https://D2L.ai. Cambridge University Press, 2023.

[13] Pei Zhou et al. "How FaR Are Large Language Models From Agents with Theory-of-Mind?" In: *arXiv preprint arXiv:2310.03051* (2023).

Semantic video content search and recommendation

Mark David Longar*
Jožef Stefan Institute
Ljubljana, Slovenia

Jakob Fir*
University of Ljubljana
Ljubljana, Slovenia

Bor Pangeršič*
University of Ljubljana
Ljubljana, Slovenia

Abstract

The rapid growth of video streaming platforms has intensified the demand for personalized content recommendations. However, current solutions often rely on historical user data, leading to challenges like the cold start problem and overlooking users' immediate preferences. We present a conversational recommendation system that leverages large language models (LLMs) to generate keyword-based content and query descriptions. By integrating Retrieval-Augmented Generation (RAG), our system efficiently retrieves relevant content, independent of prior user interactions, and ensures consistent performance across languages. Preliminary testing shows our system outperforms the RAG baseline by up to 24% in less descriptive queries and demonstrates consistent performance across three languages. While the results are promising, further evaluation focusing on user interaction and satisfaction is necessary. Our approach can potentially be extended to other recommendation systems, offering broader applicability and enhanced content personalization.

Keywords

large language models, recommendation system, search system, retrieval augmented generation

1 Introduction

The surge in video streaming platforms has accelerated the demand for personalized content recommendations. As these platforms expand their libraries and user bases, the challenge of delivering precise, user-specific recommendations intensifies. In this dynamic environment, streaming services must quickly adapt to provide accurate recommendations, which are crucial for maintaining user engagement and ensuring satisfaction.

Existing recommendation systems primarily rely on historical user interaction data, such as viewing history and ratings. This dependence leads to significant challenges, such as the cold start problem, where new users or newly added content lack sufficient data for accurate recommendations. Additionally, these systems often fail to account for users' immediate preferences, which can change dynamically due to various factors such as mood, viewing context (e.g., watching alone or with a group), or recent events in the user's life. This gap highlights the need for more adaptive and responsive recommendation mechanisms.

Recent advancements in Large Language Models (LLMs) present an opportunity to address these limitations. LLMs offer significant potential due to their emergent reasoning abilities, their capacity to extract high-quality representations of textual features, and their ability to leverage the vast external knowledge encoded within them [10], [7]. By harnessing LLMs, it is possible to create a recommendation system that interacts with users to capture their immediate preferences, thereby overcoming the cold start problem and enhancing the relevance of recommendations. Additionally, ensuring consistency in the quality of recommendations across different languages is increasingly important as many streaming services operate globally.

Our approach utilizes LLMs to generate keyword descriptions for both content and user queries. These keywords serve as the basis for recommendations, with a Retrieval-Augmented Generation (RAG) [6] model efficiently retrieving relevant content. By crafting query keywords using LLMs, the system adapts to user preferences in real time, providing relevant and language-consistent recommendations.

This paper makes the following contributions: **(1) Development of a Keyword-Based Recommendation System:** We introduce a novel approach that utilizes LLMs to generate keyword-based descriptions for content and user queries, enabling more personalized and adaptive recommendations. **(2) Exploration of Two User Interaction Models:** We propose and evaluate two distinct interfaces for user interaction—a conversational chat-based model and a structured question-answering model, where the system refines recommendations through a series of targeted yes/no questions generated by the LLM. **(3) Comprehensive Evaluation Strategy:** We outline a detailed plan for evaluating the system's performance in a production environment, focusing on its ability to deliver consistent, high-quality recommendations across different languages and user contexts.

2 Related Work

Recommender systems have progressed from techniques such as collaborative filtering and matrix factorization to more complex models that incorporate deep learning. The advent of large language models (LLMs) has enabled innovative methods for interacting with these systems [11], particularly when combined with retrieval techniques [9]. One of the most promising advancements in this area is the use of Retrieval-Augmented Generation (RAG) models, which integrate the powerful text generation capabilities of LLMs with retrieval-based methods to improve recommendation accuracy and relevance [6].

Recent advancements in conversational recommender systems have focused primarily on integrating LLMs with traditional recommender systems or fine-tuning LLMs using user-item interaction data [9], [10], e.g., [8], [4], and [5]. These approaches, while effective, often rely heavily on historical user data, leading to challenges such as the cold start problem. This reliance underscores the need for novel methods that reduce dependency on past interactions and leverage real-time retrieval mechanisms to enhance content recommendations [2].

To address these challenges, recent work by Di Palma et al. (2023) [2] introduced a Retrieval-Augmented Recommender System, which combines the strengths of LLMs and retrieval-based methods. Their approach employs LLMs both at the conversational layer and the backend retrieval process, thereby improving recommendation relevance, particularly in scenarios with sparse data or new users. Their experimental results demonstrated that

*All authors have contributed equally.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.10

this RAG-based framework performs comparably to state-of-the-art systems, even in zero-shot scenarios, underscoring the potential of such an approach to mitigate cold start and hallucination problems inherent in LLMs.

Our approach builds on the strengths of RAG-based models by introducing a keyword-based recommendation system that operates within a RAG framework. This system ensures consistent performance across multiple languages and adapts to real-time user preferences without relying on historical user data.

3 Data

The data used in this study was provided by our partner United Cloud, who operate a multinational streaming service in the Balkan region, EON TV[1]. The EON platform encompasses a variety of content, such as video-on-demand (VOD) movies and TV shows, as well as live TV channels. We focused exclusively on VOD movie data, although our approach is capable of accommodating multiple content types.

The VOD movies data set comprises nearly 5000 movies in various languages. Each movie is accompanied by a brief description averaging around 460 characters (5-6 sentences) in multiple languages. In cases where multiple translations were available, we opted for the original language of the movie; otherwise, we chose the first available translation.

4 Methodology

4.1 Recommendation Mechanism

The core of our recommendation system is the generation of textual representations of content. Instead of using movie descriptions directly, we employ the LLM to generate a set of English keywords and related movies. This approach prevents the model from overemphasizing less relevant details, such as specific plot points, that may not be central to the user's query. User queries follow a similar approach, where the LLM generates a set of relevant keywords, as well as any possibly relevant movies.

One of the key advantages of this method is its ability to abstract core concepts from user queries using the LLM, aligning better with the keywords generated from movie descriptions. The LLM-generated keywords from both the movie descriptions and user queries are designed to encapsulate the essential topics and themes. By aligning the keywords generated from movie descriptions with those derived from user queries, our system enhances the relevance of the recommendations. This alignment is crucial in ensuring that the retrieved movies resonate with the user's expressed interests, even when these interests are not articulated well. Furthermore, the use of in-context learning allows the system to maintain its performance without extensive fine-tuning [3], making it both efficient and effective.

The rest of the recommendation system follows the Retrieval-Augmented Generation (RAG) [6] pipeline (see Figure 1). The RAG pipeline operates by first generating textual representations of movies, which are then embedded into a vector space. These embeddings are stored in a vector database, allowing for efficient similarity searches. When a user submits a query, the system generates a corresponding representation, embeds it into the same vector space, and retrieves the top k most similar movie embeddings from the database. This process ensures that the recommendations are both contextually relevant and semantically aligned with the user's input.

[1]No EON user data was used.

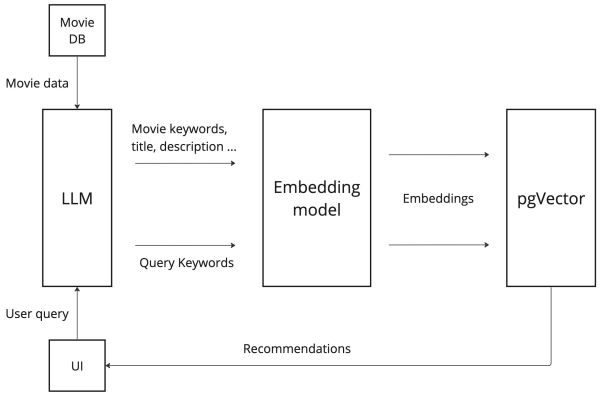


Figure 1: Overview of the Recommendation Pipeline.

4.2 User Interface

Our proposed user interface designs (see Figure 2) offer two main ways for users to interact with our recommendation core. Besides a direct search, where the user submits a query and receives recommendations in a single step, we propose: **(a)** A chatbot, which assists users in narrowing down their options through a conversational interface. The chatbot provides recommendations at each response, allowing for a multi-step interaction that refines the search results progressively. **(b)** An inquisitive method, where an agent asks the user a series of Yes/No questions to narrow down the search. Keywords are generated based on the user's responses, making it particularly useful for users who are uncertain about what they want to watch. This approach shifts the burden of knowing what to query from the user to the system, streamlining the recommendation process.

Each of these designs aims to enhance user engagement and satisfaction by providing tailored interactions that cater to different user preferences and needs.

5 Evaluation

We have developed a twofold approach for addressing the evaluation of our model:

First, to gauge the effectiveness of our keyword-based approach for recommendation, we curated a small multilingual evaluation dataset to test our core recommendation mechanism. This dataset includes queries in various languages along with their expected recommendations. We compared the performance of our mechanism with a baseline RAG system that directly embedded user queries and movie descriptions.

Second, to assess the efficiency and user satisfaction of our system in real-world situations, we have devised an evaluation plan to test our system in production. This strategy utilizes a structured A/B testing framework to conduct precise comparisons between our semantic recommendation system and conventional search, addressing distinct aspects of user experience and system performance.

5.1 Evaluation dataset

To create our evaluation dataset, we carefully selected 25 movies across multiple languages, including both well-known and lesser-known titles. For each movie, we formulated two types of queries to assess the system's retrieval accuracy: *Descriptive* and *General* queries.

The *Descriptive* queries were designed to simulate scenarios where the user knows exactly what they are looking for. For instance, a query for the movie *Messi (2014)* might be, "I am looking for inspirational documentaries about famous athletes, such as Lionel Messi and his rise through football." In contrast, the *General* queries were intended to test situations where the user has only a rough idea of what they want to watch, which is likely more common in real-world environments. An example of a general query for the same movie might be, "soccer movies that will inspire me."

To evaluate the system's performance across different linguistic contexts, we manually translated these queries into English, Serbian, and Slovenian. We then compared the performance of our keyword-based retrieval mechanism against a baseline RAG model that directly used user queries and movie descriptions without generating keywords.

5.2 Experiment Design

We have divided our user base into four distinct groups to facilitate a detailed comparative analysis, aligned with our proposed user interface designs:

Baseline Group: This control group doesn't use our system, but instead finds movies and receives recommendations based on the traditional recommendation methods, a common practice in the industry.

Direct Semantic Search Group: This control group interacts with a straightforward search interface. Users submit a query and receive recommendations in a single step. This approach provides immediate suggestions based on the user's input, mimicking traditional full-text search practices.

Chatbot Group: Participants in this treatment group use a conversational interface (interface **a**), where a chatbot assists in narrowing down options. The chatbot provides recommendations at each response, enabling a multi-step interaction that progressively refines the search results. This design enhances engagement by simulating a natural conversation.

Inquisitive Method Group: Users in this group engage with an agent that asks a series of Yes/No questions to narrow down the search (interface **b**). Keywords are generated based on the user's responses.

The evaluation will be conducted continuously, starting with a focused initial phase over the first month post-implementation to address immediate usability and performance issues, followed by ongoing monitoring to capture long-term user engagement and satisfaction.

By implementing this structured evaluation framework, we aim to comprehensively understand the impact and effectiveness of our semantic recommendation system, guiding further refinements and ensuring that the system meets user needs and expectations.

5.2.1 Metrics We would like to measure how users interact with our system in two main ways: First, we would like to know how engaged and satisfied they are with our recommendations, i.e., do users find our system frustrating to navigate, and whether they watch movies recommended by our system. The second set of metrics will aim to capture how different demographics interact with our system, as a major goal is to remove any biases such as language or age.

Engagement and Satisfaction Metrics: These include Click-Through Rate (CTR), which measures the percentage of clicked recommendation links, and Watch Time to gauge the duration users engage with recommended content. Additionally, immediate user reactions are captured through Like/Dislike Ratios, while more detailed user feedback is collected via surveys administered after interactions.

Behavioral Metrics: We analyze User Interaction Patterns, such as search frequency and refinement actions, and System Usage Frequency to determine how different demographics utilize the system and to identify any potential biases in system engagement. We also record the search time and number of queries needed for a decision.

6 Results

The outcomes presented in Table 1 showcase the performance of both models in various query types and languages, as measured by accuracy at the top 5 and top 10 recommendations.

The results reveal that the baseline model surpasses (or matches) the performance of the keyword mechanism in the case of *Descriptive* queries, particularly in terms of Accuracy@5. However, in terms of Accuracy@10, the two models demonstrate relatively similar performance. Conversely, the keyword model shows significant performance enhancements for *General* queries, particularly in Accuracy@10, indicating its capacity to adapt to non-specific content descriptions. Additionally, the keywords model consistently performs well across different languages, whereas the baseline model shows fluctuations of up to 28% across languages.

In summary, the keywords model allows for more general and multilingual queries, while the baseline model excels at retrieving very specific content.

Table 1: Evaluation results on the descriptions and general queries data sets. LLM embeddings were generated using OpenAI's *text-embedding-3-large* model. The Keywords model used *GPT-4o*.

	Accuracy@5		Accuracy@10	
	Keywords	Baseline	Keywords	Baseline
Descriptive Queries				
English	60%	**64%**	**68%**	**68%**
Serbian	56%	**80%**	72%	**84%**
Slovenian	56%	**80%**	72%	**84%**
General Queries				
English	**44%**	28%	**68%**	44%
Serbian	44%	**52%**	**68%**	52%
Slovenian	44%	**56%**	**72%**	56%

6.1 User Interface Implementation

We implemented our proposed interface design using Flutter, which guarantees functionality across a variety of devices, including iOS, Android, Windows, and web browsers. This cross-device compatibility is crucial as it ensures that all users, regardless of their preferred platform, have access to our application. The support for mobile devices is particularly useful in our interrogation design, where users can easily navigate through options by swiping cards left or right.

Additionally, we integrated Tipko [1], a Slovenian transcription service, to facilitate voice-to-text capabilities. This feature

enhances user convenience by enabling voice communication with our chat bot, removing the necessity for typing.

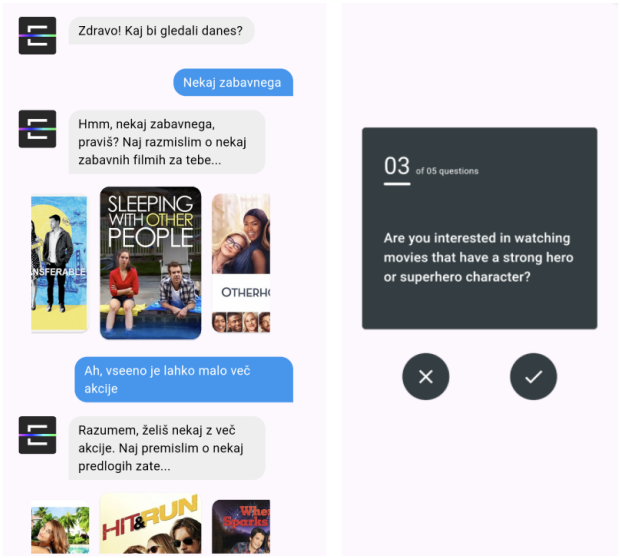


Figure 2: Implementations of our *(a) Chatbot* (left) and *(b) Inquisitive* (right) user interface designs.

7 Discussion

This report introduces a new content recommendation mechanism and three ways to interact with it. Table 1 demonstrates the success of our keyword retrieval model in understanding general user preferences while still performing well when searching for specific content. Moreover, its consistency across languages and its ability to retrieve content using specific descriptions as well as general themes make it well-suited for a diverse user base. Additionally, the keyword model allows seamless integration with both the *Chatbot* and *Inquisitive* methods. Moreover, our system could be extended to dynamically adjust keyword generation based on user-specific factors such as viewing history, local time, weather, and current mood indicators. This personalization ensures that the recommendations are not only relevant to the content but also tailored to the user's immediate context and preferences.

Our approach has some limitations, including the cost per query, which is higher than traditional search, although not exorbitant. Furthermore, our model's performance is commendable given our limited knowledge about the movie content but relies on the assumption that the language model may have more information about a movie than our dataset. It's worth noting that, in the short term, it appears that models are continually improving, becoming faster, more knowledgeable, and more cost-effective.

Lastly, as with any chat application that involves user inputs, security is a crucial consideration. While improvements can be made through better prompting and fine-tuning, ongoing monitoring is essential when the system is in production.

8 Future work

In future work, we plan to further explore methods for improving user experience and personalization. Our initial experiments have involved incorporating the user's time, location, and weather to enhance results. Moving forward, we aim to explore additional integrations, such as the user's calendar. We also intend to expand our user interface by introducing new forms of interaction, such as movie trailers and multiple-choice questions.

To overcome the limitations of our movie information, we are interested in delving deeper into the content by analyzing subtitles using a local language model. Additionally, we aim to broaden our database to include other types of content, such as live channel content and special time-limited events like Eurovision, Eurobasket, and the FIFA World Cup.

Finally, we are interested in the integration of a traditional recommendation models that utilize historical watch data or ratings to re-rank our recommendations.

Acknowledgments

This project was made in collaboration with United.Cloud and In516ht for the 2024 Data Science Competition, organized by The Faculty of Computer and Information Science at the University of Ljubljana. We thank our advisors Slavko Žitnik, Aljaž Košmerlj, Klementina Pirc, and Rebeka Merhar for their contributions.

References

[1] Primož Bratanič. *Transkript app | Samodejna transkripcija slovenskega govora*. May 2024. URL: https://transkript.si/.

[2] Dario Di Palma. "Retrieval-augmented recommender system: Enhancing recommender systems with large language models". In: *Proceedings of the 17th ACM Conference on Recommender Systems*. 2023, pp. 1369–1373.

[3] Elnara Galimzhanova et al. "Rewriting Conversational Utterances with Instructed Large Language Models". In: (Oct. 2023). DOI: 10.1109/wi-iat59888.2023.00014. (Visited on 05/22/2024).

[4] Yunfan Gao et al. "Chat-rec: Towards interactive and explainable llms-augmented recommender system". In: *arXiv preprint arXiv:2303.14524* (2023).

[5] Xu Huang et al. "Recommender ai agent: Integrating large language models for interactive recommendations". In: *arXiv preprint arXiv:2308.16505* (2023).

[6] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[7] Peng Liu, Lemei Zhang, and Jon Atle Gulla. "Pre-train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems". In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 1553–1571.

[8] Zihan Liu et al. "ChatQA: Building GPT-4 Level Conversational QA Models". In: *arXiv preprint arXiv:2401.10225* (2024).

[9] Arpita Vats et al. "Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review". In: *arXiv preprint arXiv:2402.18590* (2024).

[10] Likang Wu et al. "A survey on large language models for recommendation". In: *World Wide Web* 27.5 (2024), p. 60.

[11] Bowen Zheng et al. "Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation". In: *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. 2024, pp. 1435–1448. DOI: 10.1109/ICDE60146.2024.00118.

Continuous Planning of a Fleet of Shuttle Vans as Support for Dynamic Pricing

Filip Stavrov
stavrovf@gmail.com
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Luka Stopar
luka.stopar@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

This paper solves the problem of estimating the number and type of required resources for pickup and delivery of passengers at some time in the future. By combining optimization and sampling methods, as well as making plans based on several statistical samples, we estimate the real values for the required resources and show how the sample values converge towards the real values. Our approach combines machine-learning based demand predictions, for the number of passengers, and a route optimization engine that assigns the passengers into shared shuttle vehicles. In order to validate our method we create a baseline data that is representative of the real values. We test our approach using this baseline data, and we obtain statistically significant results.

KEYWORDS

statistical samples, demand predictions, route optimization engine, sampling techniques, optimization technique

1 INTRODUCTION

The effective allocation of resources is a critical topic in the mobility industry. Anticipating the number and type of resources required can significantly enhance a company's ability to plan accurately for the future. Our work addresses this challenge by focusing on how to estimate the number and type of vehicles needed for passenger pickup and delivery at a future time. The input to our problem consists of machine learning-based demand predictions, which provide estimates of the number of passengers across various routes offered by the company. These predictions are provided daily and further broken down into hourly estimates for each day.

*Both authors contributed equally to this research.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.27

Once we receive these predictions, our goal is to simulate reservations based on this data. For instance, if the predictions indicate that 12 passengers will travel from Ljubljana to Koper on October 20, 2024, we would simulate reservations using sampling techniques. One particular example is creating four separate bookings—one for five passengers, one for three, and two for two passengers each. We will introduce the sampling techniques used in this process in greater detail later on.

After generating these reservations, the next step is to input them into the Route Optimization Engine to generate a plan for that day. This plan will specify the number of vehicles required and the specific reservations each vehicle will serve.

The main hypotheses that our approach explores and experimentally tests are the following:

- H1: We can accurately estimate the number of required resources using optimization methods based on predicted passenger numbers.
- H2: Monte Carlo sampling of historical distributions can effectively model uncertainty in demand predictions, leading to stable resource estimations.
- H3: Creating plans based on several sample values will converge towards the actual number of required resources.

On the other hand, the key assumptions and limitations that underline our research are:

- Prediction Accuracy: We assume that the predictions effectively estimate the number of future passengers.
- Passenger Distribution: We assume that the number of passengers follows a Poisson distribution and that the distributions on different routes are independent.
- Independence: We assume that the passenger distribution and the window type distributions are independent to each other.
- Concept Drift: We assume there is no concept drift in the data, meaning the underlying data patterns do not change over time.

2 RELATED WORK

The problem of resource allocation in the mobility industry, particularly in the context of vehicle routing and passenger demand prediction, has been extensively studied. Traditional methods for vehicle routing often rely on static models that assume known and deterministic demand. However, recent advances in machine learning and optimization have enabled more dynamic approaches that can account for uncertainty and variability in demand. [3][4] For instance, predictive analytics has been employed to forecast passenger demand using historical data, which can then be fed into optimization algorithms to determine the optimal allocation of vehicles. Monte Carlo simulation is another technique commonly used to model uncertainty in demand predictions, providing a probabilistic framework for decision-making under uncertainty. [2] Moreover, dynamic vehicle routing approaches, have demonstrated the benefits of real-time adjustments to routing plans based on updated demand information. [1] The integration of these methodologies into a continuous planning framework is relatively novel and addresses the limitations of static planning approaches, particularly in highly variable and uncertain environments. [1][5]

3 METHODOLOGY

Our methodology begins with demand predictions for the number of passengers, and the ultimate goal is to determine the number and type of vehicles required, as well as the reservations each vehicle will serve. The figure below provides a detailed overview of this process.

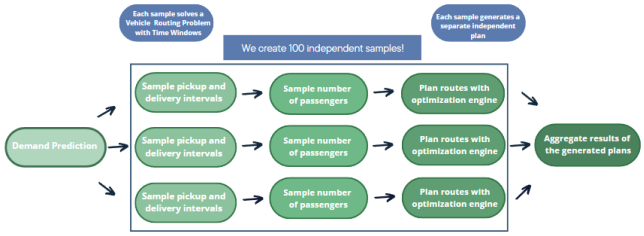


Figure 1. Methodology

Starting with the demand predictions, we apply sampling techniques to simulate reservation data. Specifically, we take the predicted number of passengers for different routes at various times and generate reservations through sampling. This reservation data follows a specific format, including fields such as ID, start location, end location, pickup time, and more. Key attributes include the number of passengers per reservation and the window type, which reflects travel preferences. For instance, some passengers may prefer a private vehicle (VIP), while others are open to sharing the ride. Additionally, the window interval is crucial—it can be a specific time or a more flexible period, affecting both the service pricing and overall experience. These factors will be incorporated into the dynamic pricing model later on.

The process begins with demand predictions and culminates in the generation of reservation data. Critical steps include sampling the number of passengers per reservation, the window type, and the window length. Sampling is done from probabilistic distributions derived from historical data, with the distributions illustrated below.

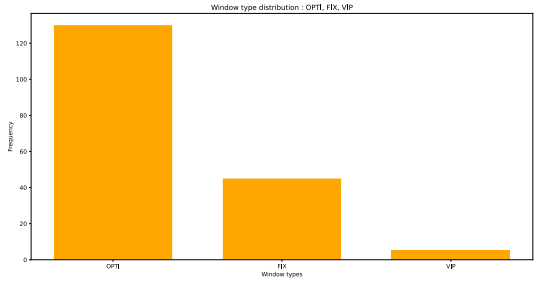


Figure 2. Window type distribution

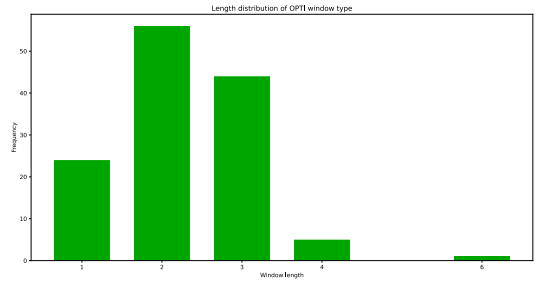


Figure 3. Window length distribution

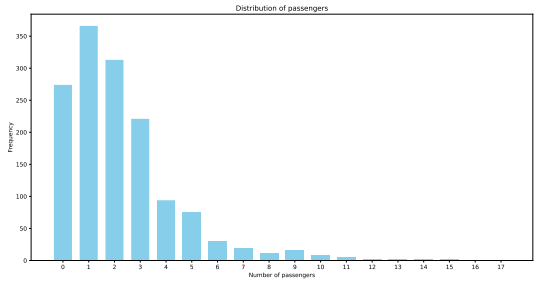


Figure 4. Number of passengers distribution

Please note that from a single demand prediction input file, we generate 100 independent samples of reservation data. This approach introduces uncertainty through probabilistic sampling. Each independent sample is then submitted as a separate job to the Route Optimization Engine, where it solves a vehicle routing problem with time constraints. The output for each job is a plan corresponding to the reservation data. Our final objective is to aggregate these results and analyze the insights they provide.

4 RESULTS

After solving all 100 jobs, we obtained 100 independent plans and began analyzing the results. As shown in the figure below, the distribution of the number of passengers yielded a mean value of 325.87 with a standard deviation of 16.85. For the number of vehicles, the mean was 38.01 with a standard deviation of 3.06. It's notable that the passenger data exhibits significantly more variance compared to the vehicle data. This is expected, as passengers are grouped into visits, and visits are then allocated to vehicles, resulting in less variation in the vehicle count.

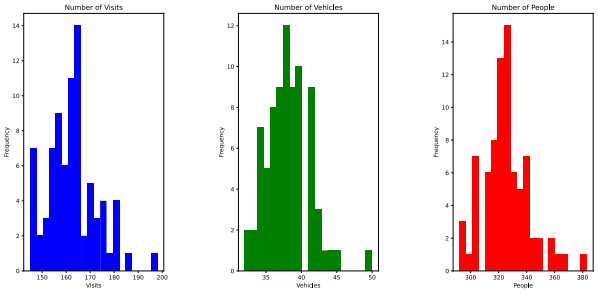


Figure 5. Sampled data: visits, vehicles and passengers distributions

To further validate our approach, we created a baseline using the same data from which the demand predictions were generated. We generated 100 samples from this baseline and submitted them as independent jobs. Upon completion, we compared the baseline results with those of our sampled data. The mean number of vehicles from the baseline was 37.81 with a standard deviation of 3.01, which closely aligns with the values from our sampled data. You can observe the comparison on the figure below.

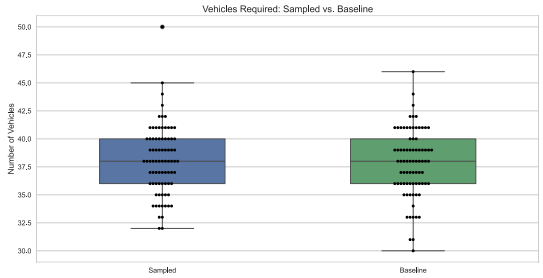


Figure 6. Comparison of required vehicles between sampled and baseline data

We also analyzed the error distribution for the number of vehicles between the baseline and sampled data, finding a mean absolute error of 3.16. This suggests that the difference between the two sets is minor, considering the sampling of data, and it is indicating a good alignment. Additionally, the average number of vehicles in both the sampled and baseline data is quite similar. While the mean absolute error reflects some variability in the sampled values, this is acceptable given the overall similarity to the global mean, and the sampling of values. Thus, despite the variance, the sampled values converge towards the actual values. This error distribution is displayed on the figure below.

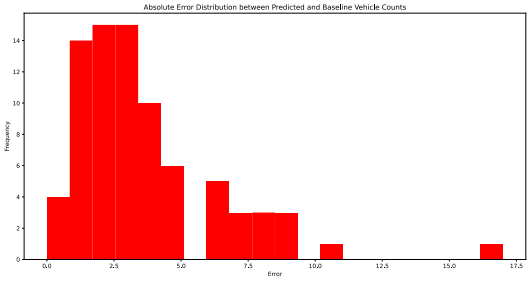


Figure 7. Required vehicles - error distribution

To statistically test whether the sampled and baseline data have the same mean number of vehicles, we conducted a Welch's t-test. The results showed a test statistic of 0.59, a p-value of 0.55, and a 95% confidence interval ranging from -0.64 to 1.23. Given the p-value, we fail to reject the null hypothesis, meaning there is no statistically significant difference between the sampled and baseline vehicle counts. Additionally, the range of the mean difference of vehicles between the sampled and the baseline data, which is from - 0.64 to 1.23, falls within our practical significance threshold of up to 2 vehicles, further supporting the similarity between the two datasets. This indicates that we can effectively estimate the number of required resources by applying optimization techniques on top of the demand prediction values.

We also analyzed the mean number of vehicles and observed that this value converges toward the actual values as the number of samples increases. This is shown on the figure below.

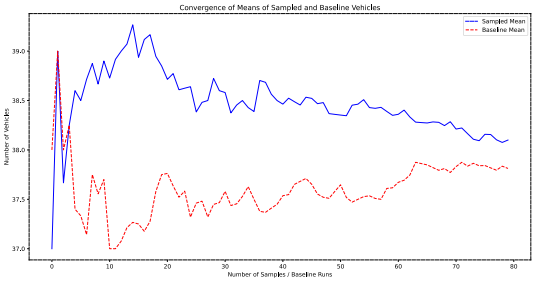


Figure 8. Convergence of means of sampled vehicles

Finally, after obtaining both the number of passengers and the number of vehicles, we decided to fit a linear regression to explore whether we could simplify the process and avoid the detailed approach previously described. As illustrated in the figure below, the regression line serves as a reasonable estimator for the number of vehicles based on the number of passengers. However, this model struggles to capture the non-linear relationships influenced

by various optimization types, window lengths, and travel modes, resulting in considerable variance around the regression line. While it is generally true that a higher number of passengers correlates with an increased number of vehicles, this relationship can be misleading. Different travel types can accommodate more passengers per vehicle, which can disrupt the linear relationship, especially in cases where these travel types dominate. Consequently, although the linear regression provides a solid approximation, it overlooks essential non-linear factors that are critical to our analysis. Our approach, which integrates these factors, demonstrates greater robustness and effectiveness. The linear regression line and the data correlation are presented in the figure below.

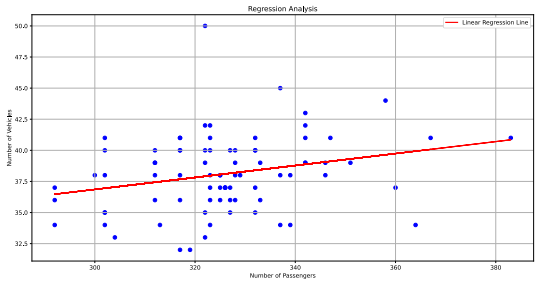


Figure 9. Regression Analysis

5 CONCLUSION

In conclusion, our findings demonstrate that we can effectively estimate the number of required resources by employing optimization methods based on predicted passenger numbers. As the number of samples increases, the sampled values consistently converge toward the actual resource requirements, reinforcing the reliability of our approach. Alternative methods, such as linear regression, fail to adequately address the non-linear complexities inherent in resource allocation, such as varying optimization types and window lengths. Our method, which incorporates these factors, proves to be a far more accurate and effective solution for resource estimation in the mobility industry.

ACKNOWLEDGMENTS

Our research is part of a broader, multi-partner initiative called CONDUCTOR. The primary objective of this project is to design, integrate, and demonstrate advanced, high-level traffic and fleet management systems. These systems aim to optimize the transport of passengers and goods efficiently on a global scale, ensuring seamless multimodality and interoperability. The CONDUCTOR project is co-funded by the European Union's Horizon Europe research and innovation programme under the Grant Agreement No 101077049.

REFERENCES

[1] Berbeglia, G., Cordeau, J. F., & Laporte, G. (2010). Dynamic pickup and delivery problems. Transportation Research Part B: Methodological, 44(5), 667-684. https://doi.org/10.1016/j.trb.2009.10.004

[2] Ulmer, M. W., Thomas, B. W., & Mattfeld, D. C. (2018). Preemptive depot returns for same-day delivery under uncertain customer availability. European Journal of Operational Research, 269(2), 356-371. https://doi.org/10.1016/j.ejor.2017.08.008

[3] Bertsimas, D., & Sim, M. (2004). The Price of Robustness. Operations Research, 52(1), 35-53. https://doi.org/10.1287/opre.1030.0065

[4] Ghiani, G., Guerriero, F., Laporte, G., & Musmanno, R. (2003). Real-time vehicle routing: Solution concepts, algorithms and parallel computing strategies. European Journal of Operational Research, 151(1), 1-11. https://www.sciencedirect.com/science/article/abs/pii/S0377221702009153

[5] Psaraftis, H. N., Wen, M., & Kontovas, C. A. (2016). Dynamic vehicle routing problems: Three decades and counting. Networks, 67(1), 3-31. https://doi.org/10.1002/net.21628

Knowledge graph Extraction from Textual data using LLM

Khasa Gillani
khasagillani22@gmail.com
Jožef Stefan Postgraduate School
Ljubljana, Slovenia

Erik Novak
erik.novak@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Klemen Kenda
klemen.kenda@ijs.si
Jožef Stefan Institute and Qlector
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute and
Jožef Stefan Postgraduate School
Ljubljana, Slovenia

ABSTRACT

The advent of Large Language Models (LLMs), such as ChatGPT and GPT-4, has revolutionized natural language processing, opening avenues for advanced textual understanding. This study explores the application of LLMs in developing Knowledge graphs from textual data. Knowledge graphs offer a structured representation of information, facilitating enhanced comprehension and utilization of unstructured text. We intend to construct Knowledge graphs that capture relationships and entities within diverse textual datasets by harnessing LLMs' contextual understanding and language generation capabilities. The primary goal is to explore and understand how well LLMs can identify and extract relevant entities and relationships from textual data using prompt engineering while contributing to structured knowledge representation.

KEYWORDS

Knowledge graph, Large Language Models, prompt engineering, information extraction, textual data

1 INTRODUCTION

In an era where data is ubiquitous, efficient organization, retrieval, and interpretation of textual information are crucial. Knowledge graphs, representing facts and relationships in structured forms, play a pivotal role in various AI applications, from enhancing search engines to powering recommendation systems. However, the construction of these graphs is often hindered by the complexity and variability of human language. This paper explores the potential of Large Language Models, like GPT-4, to revolutionize this process. By leveraging their advanced natural language understanding capabilities, we aim to automate and refine the extraction of knowledge from textual datasets. The fundamental purpose of this research is to understand the extent to which LLMs can identify and extract relevant entities and relationships from textual data and then build a Knowledge graph using the extracted information.

The motivation behind this study stems from the growing need to effectively manage and utilize the vast amounts of textual data generated daily. Knowledge graphs offer a structured and intuitive way to represent information, but their construction is often labor-intensive and requires expert knowledge. However, constructing Knowledge graphs from unstructured text is intricate and depends on sophisticated natural language processing (NLP) methods, including named entity recognition (NER) and relation extraction. The advancement of LLMs like GPT-4 presents an opportunity to automate and improve this process as illustrated in Figure 1. Utilizing LLMs can lead to more efficient, scalable, and accurate Knowledge graph construction, thereby unlocking new possibilities in information management and AI applications.

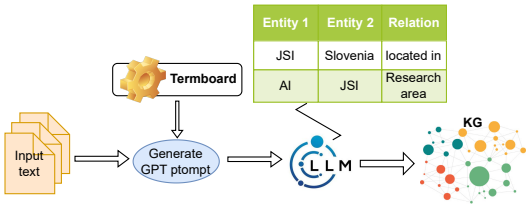


Figure 1: Overview of proposed approach where input text is processed through a Termboard to generate a structured prompt for LLM, creating an entity-relation table to build a Knowledge graph (KG).

2 BACKGROUND

An overview of recent research in Large Language Models and Knowledge graphs is provided in this section, which also emphasizes the potential for their integration.

2.1 Large Language Model (LLM)

Large Language Models are advanced AI systems pre-trained on extensive data, enabling them to comprehend and produce human language. Their recent surge in popularity is due to their proficiency in various language-processing tasks, including text completion, translation, summarization, and answering questions. These models, primarily based on transformer architecture, utilize self-attention mechanisms through encoder-decoder modules. Encoders transform input text into numerical embeddings that reflect the context and meaning, while decoders use these embeddings to generate coherent and pertinent textual output. The large language models feature a decoder-only architecture and, thus, make a prediction of the target output text using only the decoder module. The training paradigm for these models is to predict the next word in the sentence. Generally, large-scale decoder-only LLMs such as ChatGPT [7] and GPT-4 [2], focus on human-like language output, predicting subsequent words based on the preceding text for tasks like text generation.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.15

Table 1: Simplified comparison between Large Language Models (LLMs) and Knowledge graphs (KGs)

Feature	LLM	KG
Knowledge type	Broad, general knowledge	Structured, domain-specific knowledge
Data handling	Flexible, can process varied inputs	Requires structured data
Accuracy	May lack precision in understanding	Highly accurate with structured data
Understanding	Can interpret and generate language	Designed for specific queries and relationships
Adaptability	Adapts to new information by retraining	Adaptable when updated with new data
Transparency	Often seen as "black boxes" with unclear reasoning	Clear decision-making pathways
Error rate	Can make mistakes due to broad generalizations	Can be prone to errors if data is incorrect or missing
Complexity	Handles complex language tasks	Manages complex relationships and attributes
Usage	Broad applications in text generation, translation, etc.	Used for specific tasks like recommendations, search optimization
Scalability	Scales with computational power	Scales with the amount of structured data available

2.2 Knowledge graph (KG)

Knowledge graphs are structured representations of information that depict the relationships between entities in a specific domain. They are used extensively in various applications, such as search engines, recommendation systems, and question-answering systems. These graphs use detailed connections between data to help with smart thinking, finding specific information easily, and running applications that use knowledge. Hence, allows us to better understand and use information across multiple fields.

Knowledge graphs provide a structured way of representing interconnected knowledge. They are precise and consistent, aiding in decisive and informed decision-making. KGs are particularly valuable for their interpretability and explainability due to the explicit representation of entities and relationships. They can capture domain-specific information accurately and evolve to incorporate new data. However, KGs may suffer from incompleteness and may not always reflect the most recent or unseen facts. They also typically cannot understand natural language in an unstructured format [3][6]. Moreover, KGs are preferred in scenarios where explainability and interpretability are crucial, as they provide structured knowledge representation.

2.3 Combining LLM and KG

The comparison between Large Language Models and Knowledge graphs (Table 1) can be supported by various references that highlight their respective strengths and weaknesses [4]. Large Language Models like ChatGPT [7] are celebrated for their generalizability and ability to process diverse text data, allowing them to perform various language-related tasks without extensive task-specific training. They can act as reservoirs of general knowledge, aiding in information synthesis and research. Their proficiency in language processing is useful in tasks like natural language understanding and sentiment analysis. However, they can suffer from hallucinations, where they generate plausible but factually incorrect information. Their "black-box" nature makes it difficult to understand the internal decision-making processes, and they can be indecisive, producing uncertain responses to ambiguous inputs. Additionally, while they have vast general knowledge, they may not be up-to-date with domain-specific or the latest information. Critics of LLMs argue that these models lack transparency and interoperability.

Recent research [3] [4]efforts are, however, improving LLM's interpretability through techniques like attention mechanisms and model introspection. KGs also present advantages over LLMs by providing knowledge about long-tail entities, thus improving recall for knowledge computing tasks. However, both LLMs and KGs can perpetuate biases present in their training data or construction methodologies. In conclusion, both LLMs and KGs have their unique strengths and challenges. While LLMs excel in general language processing and knowledge extraction from vast corpora, KGs provide a structured and interpretable way to organize explicit knowledge. These differences underscore the potential benefits of integrating LLMs and KGs to create more robust AI systems that leverage the strengths of both approaches.

3 PROOF OF CONCEPT: ANALYSIS AND KNOWLEDGE GRAPH GENERATION

This section demonstrates how to process and analyze textual data to build a Knowledge graph using LLM. It is important to mention that prompt engineering [5] is of great importance when it comes to the results generated from ChatGPT. Since it is a generative model, small variations in the input sequence can create large differences in the produced output as demonstrated below. We use two different textual files containing contextual data: (i) APRIORI proposal (containing project details, job description, potential candidate skills, hosting organizations, etc.) and (ii) ADRIA Motorhome instruction manual (containing textual as well as tabular data). Moreover, building KG out of the ADRIA instruction manual has potential applications for the manufacturing industry.

3.1 Using ChatGPT Prompts:

We compare ChatGPT-3.5 and GPT-4 extracted entities and relations using the same prompts. We use Termboard[1] which offers customized ChatGPT prompts to create terms, entities, and relations to visualize larger graphs from the provided text.

Prompt: Extract an ontology and create a table of relations with 3 columns in this order: source, target, and relation name. Also Create a table with 2 columns: put in the first column the name of the term and in the second column an elaborate definition of the term. Use this text as a basis: ¨APRIORI¨- (contains textual data about the job description, candidate skills, project description, hosting organization, etc).

Observing the Knowledge graphs generated by ChatGPT-3.5 (Figure 2) and GPT-4 (Figure 3); we notice, that it didn't extract all entities and relations and missing terms/concepts. For this reason, we ran the second prompt, where we redefined a more detailed prompt to ask GPT-4 to explicitly generate a comprehensive ontology including all entities and relations from the provided text, categorize entities into types like Persons, Organizations,

[1]https://termboard.com/

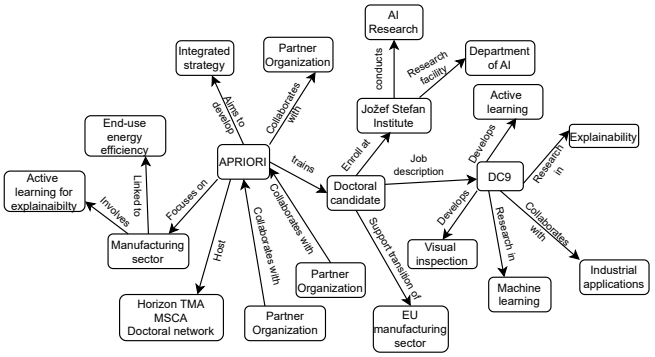


Figure 2: The KG generated using ChatGPT-3.5 contains 20 entities. It was able to extract entities and link them to relations, but it failed in abstracting concepts and specifying entities (i.e. partner organizations, location, etc.).

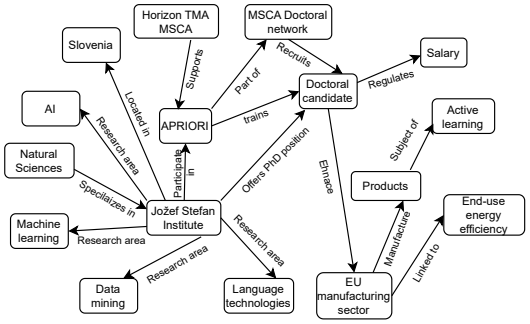


Figure 3: The KG generated by GPT-4 contains 16 entities. It was able to identify abstract concepts, and geographic entities that ChatGPT-3.5 doesn't. Extracted more elaborated entities with relations.

and concepts, and Geographic Locations, and then identify the relations between these entities. Providing additional information to GPT-4 resulted in an improved Knowledge graph (Figure 4). However, ChatGPT-3.5 didn't produce a quality graph (Figure 5) compared to Figure 2.

3.2 Python Implementation

We use a free, open-source library called spaCY [2] for advanced NLP in Python. We employ the named entity recognition technique to identify named entities from a given text using the spaCY model (en-core-web-sm). We used a chunk of textual data from the ADRIA Motorhome manual for experiment purposes. Table 2 compares entities, relations, and triplets extracted from the raw texts. The table shows that the number of triplets extracted by algorithms is similar–(Figure 6 and Figure 7). However, the number of entities that spaCY extracts are larger but not every pair of entities is connected by meaningful relation, leading to fewer triplets. Thus defeating the purpose of creating a Knowledge Base. When using spaCy for entity extraction, the entities are typically recognized based on the named entities present in the text. Named entities are often specific nouns, such as names of people, organizations, locations, dates, or product names. spaCy might not identify it as a specific entity by default. So to extract specific entities, it might need to customize spaCy's NER model

[2]https://spacy.io/models

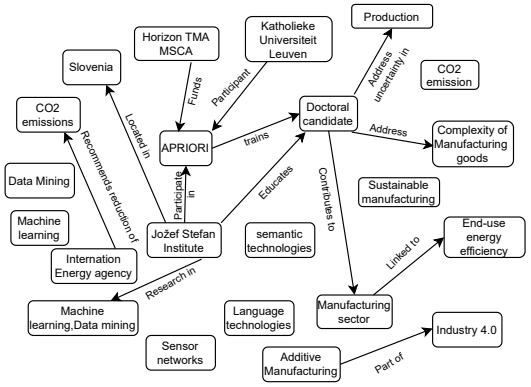


Figure 4: The KG generated by GPT-4 contains 22 entities. It Identified more key entities and relevant concepts and identified suitable relations to connect them (i.e. participant-Katholieke Universiteit Leuven). However, it didn't cover all relations and classes (i.e. skills). We also notice a few duplicated entities(i.e. data mining, CO2 emission, etc.) and some independent entities (i.e. sustainable manufacturing).

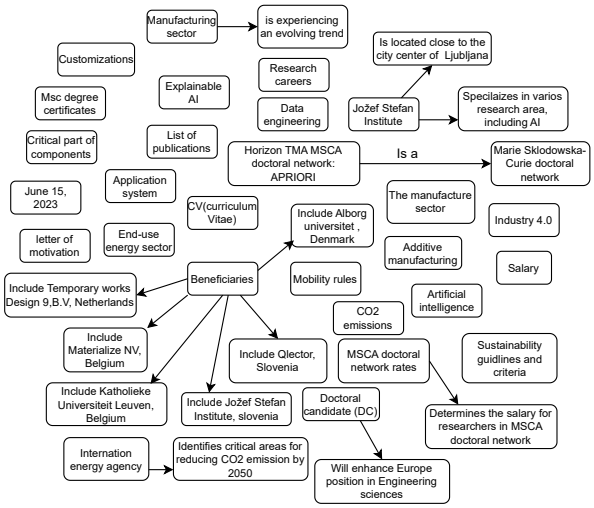


Figure 5: ChatGPT-3.5 was able to extract a larger number of entities but it was not successful at abstracting concepts and missing relations. Entities and relations found frequently represented complete sentences rather than concepts. This occurs because ChatGPT is a conversational model trained on a task to create responses to a given prompt and is not particularly trained to recognize entities and relations

or provide additional context for better recognition. Hence results can be improved by pre-processing data into a structured format.

4 EVALUATION

When there is no ground truth data available, creating an automated evaluation metric for a Knowledge graph becomes challenging. In such cases, the evaluation relies on qualitative principles to assess the results. Based on the practical framework defined in the study [1], the following principles were identified:

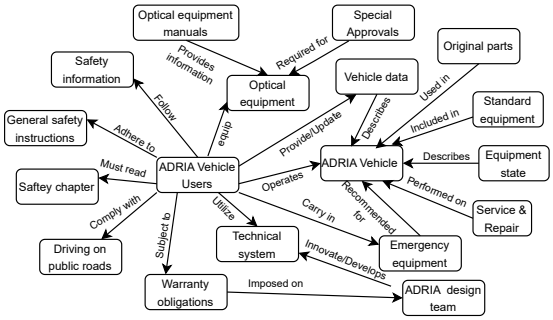


Figure 6: The KG generated by GPT-4 contains 18 entities using the ADRIA motorhome instruction manual. It extracted concepts relevant to ADRIA users and vehicle instructions, their functions, and how they are connected.

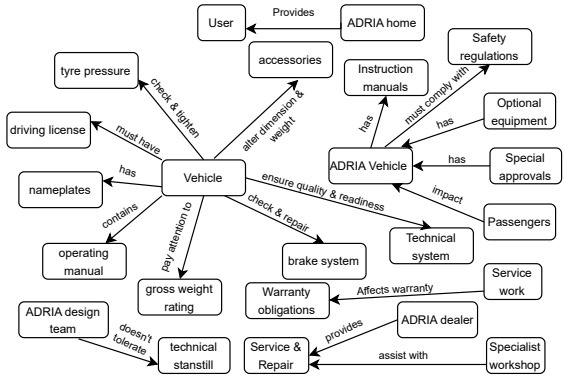


Figure 7: The KG generated by ChatGPT-3.5 contains 24 entities. Extracted more entities relevant to ADRIA vehicles but relations between entities are more generic and entities are duplicated.

- Triplets should be concise.
- Contextual information of entities should be captured.
- The Knowledge graph does not contain redundant triples.
- Entities should be densely connected.
- Relations among different types of entities should be included.
- Knowledge graphs should be organized in structured triples for easy processing by machine.
- For tasks specific to a particular domain, it's essential that the Knowledge graph is tailored and relevant to that specific field

According to these principles, in our use case, we manually inspected the Knowledge graphs generated above, and we can conclude that the ChatGPT-3.5 approach provides a more detailed Knowledge graph without abstract concepts compared to the GPT-4. However, to create these Knowledge graphs, a few steps of refining the answers from ChatGPT are needed. Sometimes the produced output is incorrect and needs to be corrected before proceeding. When we redefined the prompt, GPT-4 identified more specific entities, and concepts compared to ChatGPT-3.5. Even though ChatGPT extracted a larger number of entities, it failed to provide abstract concepts and entity-relation.

In the second part of the experiment, we employed the NER method to extract relations and entities from the given text (i.e. ADRIA). We analyzed that extracted entities are duplicated and relations have some noise and incomplete information. If you have specific patterns or structures in mind that you want to extract entities and relations based on, you may need to customize the relation extraction logic. Alternatively, more advanced natural language processing techniques or pre-trained models designed for relation extraction tasks might provide better results. Also, we analyzed half of the relations-entities extracted by spaCY and ChatGPT are overlapped.

Table 2: Knowledge extraction comparison. (ADRIA motorhome manual dataset)

Algorithm	Entities	Relations	Triplets
GPT-4	18	20	20
ChatGPT-3.5	24	18	18
spaCY	22	14	17

5 CONCLUSION

The proposed exploration of using LLMs for Knowledge graph extraction holds promise for advancing our understanding of how advanced language models can contribute to structured knowledge representation. This paper explores using LLMs to generate Knowledge graphs out of source documents. We utilized ChatGPT-3.5 and GPT-4 models to generate the Knowledge Graphs for two different textual data and compared the structure of the KGs. GPT-4 performed better as it successfully identified more abstract concepts and key entities compared to ChatGPT-3.5. Therefore, it provides insights into the practical application of LLMs in developing structured knowledge from unstructured textual data, with potential applications in knowledge-based AI applications, paving the way for more effective information processing and utilization. In future studies, we intend to use a more formal framework to evaluate the quality of created Knowledge graphs. Such a framework will allow us to efficiently analyze the quality of KG and provide a standardized method to forecast missing linkages between concepts and relationships within a given domain.

ACKNOWLEDGEMENTS

This research is supported by EU funding HE MSCA Project Apriori (GA: 101073551). The author acknowledges the usage of ChatGPT and Grammarly for content paraphrasing, grammar, and error checking.

REFERENCES

[1] Haihua Chen, Gaohui Cao, Jiangping Chen, and Junhua Ding. 2019. A practical framework for evaluating the quality of knowledge graph. In *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding: 4th China Conference, CCKS 2019, Hangzhou, China, August 24–27, 2019, Revised Selected Papers 4*. Springer, 111–122.

[2] R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2, 13.

[3] Jeff Z Pan et al. 2023. Large language models and knowledge graphs: opportunities and challenges. *arXiv preprint arXiv:2308.06374*.

[4] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: a roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

[5] Elvis Saravia. 2022. Prompt engineering guide. (2022).

[6] Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. 2023. Enhancing knowledge graph construction using large language models. *arXiv preprint arXiv:2305.04676*.

[7] Ce Zhou et al. 2023. A comprehensive survey on pretrained foundation models: a history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Solving hard optimization problems of packing, covering, and tiling via clique search

Sándor Szabó
sszabo7@hotmail.com
University of Pécs
Pecs, Hungary

Bogdán Zaválnij
bogdan@renyi.hu
HUN-REN Alfred Renyi Institute of Mathematics
Budapest, Hungary

Abstract

In the paper we propose to convert NP-hard combinatorial optimization problems of packing, covering, and tiling types into maximum or k-clique problems. The key step is to come up with a tactically constructed auxiliary graph whose maximum or k-cliques correspond to the sought combinatorial structure. As an example, we will consider the problem of packing a given cube by copies of a brick. The aim of the paper is two fold to illustrate (i) the modeling power and (ii) the feasibility of the clique approach. Since theoretical tools are not readily available to study the effectiveness of the solution of the resulting clique problems we will carry out carefully conducted numerical experiments.

Keywords

mathematical programming, k-clique problems, combinatorial optimization

1 Introduction

One can see graphs as a mathematical models that can describe various fields of interest. Like numbers, functions, or Linear Programming graph based approach can model interesting problems and aid us in solving them. Some of these approaches are quite straightforward like cliques of people in a social interaction graphs or shortest path problem in a road map. Other approaches are less obvious but still easily constructed, like conflict graphs in a set of codewords where a maximum independent set represents a maximum set of suitable error correcting codes [9].

But the approach of modeling and solving various problems by graphs are more versatile. Namely, we can see graphs as a language for mathematical programming – if certain combinatorial problems can be solved by constructing a suitable auxiliary graph and finding a maximum or k-clique of this graph gives the solution. The authors have already used this approach in connection with mathematical conjectures [1], hyper graph coloring [11], subgraph isomorphism [2], scheduling problems [12], graph coloring problems [13] and protein docking problems in chemistry [8].

Here we would like to give an example, where a hard combinatorial optimization problem can be solved by this approach. For this we chose a simple to understand but numerically hard to solve problem of brick packing popularized by M. Gardner. We will focus on different approaches of how to construct an auxiliary graph in order that to translate this problems into a clique search problem. We will try to investigate how these different approaches – based on packing, covering and tiling– affect the solving time and if they have other consequences as well. First, we describe the basic problem, then we present theoretical discussion of different reformulations, and finally we describe the results of numerical experiments. The emphasis is on the modeling aspect of the computation and not on reaching new records, as the proposed problem was solved in theoretical manner within months of its formulation. Here we use it as a prototype of similar problems, and our aim to show the versatility of our approach, that is model a problem by a graph.

Graphs in this paper will be finite simple graphs. Further all graphs we use will not have loops or double edges. A finite simple graph G can be described with its set of nodes V and a subset E of the Cartesian product $V \times V$. The subset E can be identified by the set of edges of G.

Let $G = (V, E)$ be a finite simple graph. A non-empty subset C of V is called a k-clique if each two distinct nodes of C are adjacent in G and in addition C has exactly k elements. If C has only one element, then we consider it a 1-clique. The 2-cliques of G are the edges of G. A k-clique C of G is called a maximum clique if G does not have any $(k+1)$-clique. For each finite simple graph G there is an integer k such that G contains a k-clique but G does not contain any $(k+1)$-clique. This well defined integer k is called the clique number of G. We state two clique problems formally.

Problem 1. *Given a finite simple graph G and an integer k. Decide if G has a k-clique.*

Problem 2. *Compute the clique number of a given finite simple graph.*

Problem 1 is a decision problem, it is referred as the k-clique problem, and it is an NP-complete problem included in the original list of 21 NP-complete problems by Karp [7]. Problem 2 is an optimization problem and referred as the maximum clique problem, and as the decision problem belongs to the NP-complete class it follows that it belongs to the NP-hard class.

We color the nodes of a finite simple graph G with the colors $1, 2, \ldots, k$ such that each node receives exactly one color and adjacent nodes never receive the same color. Such a coloring of the nodes of G is called a well coloring, a proper coloring, or a legal coloring (the terminology is not unified). The set of nodes of G receiving the color i is called the i-th color class. Clearly, a color class is an independent

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.9

set of G, that is, two nodes from a fixed color class are never adjacent.

If the nodes of a finite simple graph can be legally colored using k colors, then we say that G is a k-partite graph. The reason is that in this situation the nodes of G form a union of k independent sets and these sets are pair-wise disjoint.

In this paper we will focus on the following clique problem.

Problem 3. *Given a finite simple graph G whose nodes are legally colored using k colors. Decide if G has a k-clique.*

Problem 3 is a k-clique problem particularized to case of k-partite graphs. This problem is still an NP-complete problem, as the graph coloring problem can be reduced to such question as shown in [13], and should not be confused with the problem of complete graphs.

The problem class we will be focusing on in the present paper consists of packing, covering, or tiling problems. Obviously many real world and mathematical problems fall into this class, and here we would show some ideas how such problems can be modeled by a suitably constructed auxiliary graph where a k-clique search would solve the original problem.

2 Packing, covering, and tiling

First, we describe the problem class in question. Second, we draw up some basic concepts how these problems can be modeled by graphs.

Let U be a finite ground set and let

$$A_1, \ldots, A_m \tag{1}$$

be subsets of U. A family of subsets

$$B_1, \ldots, B_n \tag{2}$$

with $\{B_1, \ldots, B_n\} \subseteq \{A_1, \ldots, A_m\}$ is called a packing of U if the members of the family (2) are pair-wise disjoint. A family of subsets (2) is called a covering of U if the union of (2) is equal to U. Phrasing it differently, a family of subsets (2) is a covering of U if each element of U belongs to at least one member of the family (2). If a family of subsets (2) is a packing and a covering of U in the same time, then it is called a tiling of U. A tiling of U some times referred as exact covering of U.

A packing of U is called a k-packing if it consists of k subsets of U. Similarly, a covering of U is called a k-covering if it consists of k subsets of U. Finally, a tiling of U is called a k-tiling if it consist of k subsets of U. For a given ground set U and for its given subsets (1) there is an integer k such that U has a k-packing using subsets of the family (1) but there is no any $(k+1)$-packing of U using members of the family (1). This well defined integer k is the packing number of U with respect to the family (1). If the packing number of U is equal to k, then each k-packing of U is called maximum packing of U.

For a given ground set U and for its given subsets (1) there is an integer k such that U has a k-covering using subsets of the family (1) but there is no any $(k-1)$-covering of U using members of the family (1). This well defined integer k is the covering number of U with respect to the family (1). If the covering number of U is equal to k, then each k-covering of U is called minimum covering of U.

We state five problems related to packings, coverings, and tilings in a formal manner. Given a finite set U and its subsets (1).

Problem 4. *Decide if U has a k-packing using the members of the family (1).*

Problem 5. *Decide if U has a k-covering using the members of the family (1).*

Problem 6. *Decide if U has a k-tiling using the members of the family (1).*

Problem 7. *Compute the packing number of U with respect to the family (1).*

Problem 8. *Compute the covering number of U with respect to the family (1).*

Problem 4 can be reduced to Problem 1. We construct a finite simple graph G. The nodes of G are the members of the family (1). Two distinct nodes A_i and A_j are adjacent in G whenever A_i and A_j are disjoint. A k-clique in G corresponds to a k-packing of U.

Problem 5 can be reduced to Problem 3. We sketch the main points of this reduction. We construct a finite simple graph G. The first type of nodes of G are ordered pairs (B, x), where $B \in \{A_1, \ldots, A_m\}$, $1 \leq x \leq k$. The intuitive meaning of the pair (B, x) that the subset B is the x-th member of a k element family of (1). To the node (B, x) we assign the color x. Two nodes receiving the same color will be non-adjacent in G. Therefore the first type nodes of G are legally colored with k colors.

We are adding second type nodes to G. Namely, we are adding the ordered pairs (A, u), where $A \in \{A_1, \ldots, A_m\}$, $u \in U$ and in addition $u \in A$ holds. The intuitive meaning of the pair (A, u) is that the element u is covered by set A. To the node (A, u) we assign u as a color. Two nodes receiving the same color will not be adjacent in G. Thus the second type nodes of G are legally colored using $t = |U|$ colors. Now if we are locating a $(k+t)$-clique in G, then we select exactly k subsets from (1) and each element of U will belong to at least one of these subsets. The missing part of the construction, what we left for the reader, is how the first and second types of nodes are connected by edges.

Problem 6 can be reduced to Problem 3. As a tiling is a packing and covering at the same time, we can add the packing restrictions, namely not connecting two sets if they intersect, to the second type of nodes. On the other hand – in case of equal size sets –, we do not need to count the used sets, so we won't need the first type of nodes, they can be omitted.

The computational difficulties of the k-packing, k-covering, and k-tiling problems are different. It seems that the covering problems are the computationally most demanding and the tiling problems are the most manageable.

3 Gardner's bricks problem

We picked Gardner's problem because it is intuitive and easy to comprehend among such problems that can be reduced to Problem 3 and so it serves as a good illustration of the kind of clique modeling we are dealing with. We do not claim any originality in connection with the problem. We do not prove any new results. Each of the facts we use are known from the folklore and we present them only

for the reader convenience. The problem was raised by Foregger in March 1975 [10], popularized by Gardner in February 1976 [5], and solved by Foregger and Mather in November 1976 [3].

Let us consider a brick B of dimensions $1 \times 2 \times 4$. The brick B is a union 8 unit cubes whose edges are parallel to the coordinate axis. From some reason unknown for us the brick B is referred as canonical brick. Suppose we have a large supply of congruent copies of B and we want to pack as many as possible into a $7 \times 7 \times 7$ cube C. The cube C is a union of 343 unit cubes. Let us divide 343 by 8 with remainder. As $343 = (42)(8) + (7)$, 43 copies of B cannot be packed into C. M. Gardener advanced the question if 42 copies of B can be placed into C. One can place a copy of B into C in any possible rotated position as long the edges of B are parallel to the coordinate axis. (The answer to this question is actually: No, one cannot place 42 bricks into a cube of size $7 \times 7 \times 7$.)

Gardner's problem can be expressed in terms of computing the clique number of a suitable constructed graph G. In other words, Gardner's problem can be reduced to an instance of the maximum clique problem. Let us denote the set of the 343 unit cubes forming C by U. An 8 elements subset v of U is a vertex of G if the union of the elements of v is a congruent copy of B. As it turns out G has 1008 nodes. Two distinct nodes v and v' of G are adjacent in G if v and v' are disjoint. If G contains a (42)-clique, then 42 congruent copies of B can be packed into C. During our numerical experiments a greedy coloring procedure provided a legal coloring of the nodes of G using 42 colors. Note that this is just a coincidence, it could've happened otherwise. Thus we are facing with a particular case of the k-clique problem stated in Problem 3. The nodes of G are legally colored with 42 colors and we are looking for a (42)-clique in G. Phrasing it differently, we are looking for a k-clique in a k-partite graph, where $k = 42$.

We introduce a coordinate system whose origin coincides with a corner of the cube C.

Observation 1. *If* 42 *congruent copies of the brick* B *can be packed into* C*, then there is such a packing which contains the congruent copy of* B *whose one corner is the origin. Further the edges of lengths* 1*,* 2*,* 4 *are parallel to the first, second and third coordinate axis, respectively.*

Proof. As $343 = (42)(8) + (7)$ holds, 7 unit cubes of C are not contained by any bricks of the packing. The cube C has 8 corners and so at least one of the corners must be contained by a brick. At this point we introduce a coordinate system whose origin is this corner of C. Then we introduce the first, second, and third coordinate axis to satisfy our requirement. □

The cube C can be sliced into 7 slabs using planes perpendicular to the first coordinate axes. Each slab is a $1x7x7$ slice of the big cube, that is a union of 49 unit cubes. The centers of these cubes are in a plane perpendicular to the first coordinate axis. The 7 unit cubes of C, that are not contained by any brick of the packing, are referred as unpacked unit cubes.

Observation 2. *Two distinct uncovered unit cubes of* C *cannot be in the same slab.*

Proof. Note that a fixed slab can contain only 0, 2 or 4 unit cubes from any brick of the packing. The point is that the numbers 0, 2, 4 are all even. Each slab consists of an odd number of unit cubes. Therefore, each slab must contain an odd number of unpacked unit cubes. The number of slabs is 7 and so each slabs must contain exactly one unpacked unit cube. □

We can also form slabs by slicing C with planes perpendicular to the second coordinate axes. Each of these 7 slabs contains exactly one unpacked unit cube. Finally, slicing C by planes perpendicular to the third axes we get that each of these slabs contains exactly one unpacked unit cubes. These constraints on the uncovered unit cubes are independent, but can also be checked independently during an extended search, and as such can reduce the search space well.

4 Numerical experiments

Gardner's brick packing problem can be turned into various clique search problems and we carried out numerical experiments with them. We will observe that the same geometric problem will lead to very different clique search problems. When we try to pack 42 congruent copies of the canonical brick B into the the big cube C, we get a k-clique problem. When we notice that the nodes of the auxiliary graph can be legally colored using 42 colors we get a k-clique problem in a k-partite graph which is a more tractable search problem. When we try to pack 42 congruent copies of the brick into the cube C together with 7 unit cubes we get tiling problem. When we try to pack 42 congruent copies of the brick into the cube C together with 7 unit cubes and in addition we distinguish the unit cubes among each other we get yet another version of the tiling problem.

In the first approach the auxiliary graph G_1 had 1008 vertices. The nodes of G_1 were legally colored using 42 colors and we tried to locate a (42)-clique in G. Note, that although this graph can be colored with 42 colors it was just a coincidence. There is no theoretical background to this fact. Of course the expectation was that G_1 do not have any (42)-clique.

Let us assume that it is possible to pack 42 congruent copies of the $1 \times 2 \times 4$ canonical brick B into the $7 \times 7 \times 7$ cube C. By Observation 1, we may assume that a brick appear in the packing such that one of the corners of the brick coincides with the origin of the coordinate system and the edges of lengths 1, 2, 4 are along the 1-st, 2-nd, 3-rd coordinate axis. This information can be interpreted such that there a (42)-clique C_2 in G_1 which has a specific node. Namely, the vertex v_1 of G_1 that corresponds to the special corner brick is a node of the of C_2. This suggests to restrict the graph G_1 to the neighbors of the vertex v_1 to get a new graph G_2. Then we are looking for a (41)-clique in G_2. Plainly, the nodes of G_2 are legally colored using 41 colors. This coloring is inherited from the coloring of the nodes of G_1. Since the graph G_2 has fewer vertices than G_1 (actually 960) and we are looking for a smaller clique in G_2 than in G_1. The new clique problem probably requires less computational effort because the graph is smaller, and because we introduced a symmetry breaking to it.

The problem of packing 42 bricks into a bigger cube can be viewed as a tiling problem. Namely, we try to tile the

$7 \times 7 \times 7$ cube C by 42 copies of the canonical brick and 7 additional copies of a unit cube. Thus we are facing to a tiling problem using two different types of tiles and the number of the tiles is given. To ensure that we use 42 bricks we numerate the small cubes as $\{1, \ldots, 7\}$ and ensure in the graph that each small unit cube is used once, that is we do not connect nodes where the unit square is covered by the same small cube. This tiling problem can also be reduced to a clique search problem. We denote the corresponding graph G_3. Tiling problems are more manageable compared with packing problems as during the search back-tracking can be anticipated earlier. However, the graph associated with the tiling in our case has more vertices than the graph associated with the packing, namely it has 10 465 nodes. Therefore only computations can reveal which approach is preferable.

Obviously, in this case we can also fix a brick in the corner. This version will be the G_4 graph.

In the last clique search equivalent of Gardner's problem we construct a graph G_5. In this construction we handle a mixed tiling problem but we utilize the extra information that no two distinct unit cube can appear in the same slab. By Observation 2, this may be assumed. This is done by not connecting two nodes associated with unit cubes if those unit cubes lay in the same slab. This graph is the same size as G_3, as we only delete some edged from it. Also, we can fix a brick in the corner in this case as well, that shall be the G_6 graph.

Once again only numerical experiments can guide us in judging the merits of the possible clique search equivalents of the problems. Further, the preconditioning methods perform differently on the graphs G_1, G_2, G_3, G_4, G_5, G_6 and this adds an extra layer of difficulty to the numerical work. We used a computer with AMD EPYC 7643 processors, `C++`, and gcc v12.1 with settings `-O3 -arch=znver3`.

We made all six graphs and performed k-clique search on them after preconditioning as described in [12, 13]. The preconditioning run for 1-2 hours for the bigger graph, and reduced it by half, namely to around 6 000 nodes for G_3, G_4; and to around 4 000 for G_5, G_6, that is the graphs where we allow only one small cube in a slab. For the smaller graphs (G_1, G_2) the preconditioner runs for a couple seconds but cannot significally reduce the graph. Three of the six graph could be solved after preconditioning: G_2, G_5, and G_6.

The solution time of G_2 (the original graph with fixed brick in the corner) was 50 days. The solution of G_5 was a bit faster, 29 hours. Finally, the graph G_6 could be solved more effectively. The running time was 123 484 seconds, that is 34 hours. This clearly show us the importance of the extra information of slabs.

5 Conclusions

We detailed several k-clique search reformulations of a certain combinatorial problem in terms of constructing suitable auxiliary graphs. We do not claim, that these methods result more efficient practical computations than other approaches. The point we are trying to make is that the clique reformulations open up a possibility to use well tuned clique solvers, including preconditioning, to handle different combinatorial problems in a unified manner as a general solver.

The results presented here have interesting consequences and suggest further research problems. First, and as anticipated, different auxiliary graphs lead to very different search space sizes. And although the usual concept in our research is that bigger graphs usually tend to be harder, that is not always the case. Remarkably, numerical results indicate that the size of the auxiliary graph alone is not as important as the type of the reformulation. Namely, the tiling type auxiliary graphs required less computational effort for clique search even if they were not the smallest graphs. Second, there are additional constraints that can be added to some reformulations while they seemingly cannot be incorporated into others. An example to such a constraint is the fact described after the proof of Observation 1. Namely, that no two distinct uncovered unit cubes can appear in the same slab in Gardner's brick packing problem. That kind of restriction could be incorporated into the tiling version of reformulation, and possibly not applicable to the packing reformulation. Taking advantage of the extra constraint made possible to solve the brick packing problem in reasonable time.

There are other problems that can be solved using similar approaches as detailed in the paper. Authors could solve smaller instances of the Golomb ruler problem or the Salem-Spencer set problem. The results, that lay outside the scope of the present paper, obtained with those instances open up even more interesting considerations.

Acknowledgements

The present research was funded by National Research, Development and Innovation Office – NKFIH Fund No. SNN-135643.

References

[1] K. Corrádi and S. Szabó, A combinatorial approach for Keller's conjecture. *Period. Math. Hungar.* Vol. 21, 91–100, 1990.
[2] M. Depolli, S. Szabó and B. Zaválnij, An Improved Maximum Common Induced Subgraph Solver. *MATCH Commun. Math. Comput. Chem.* 84 pp. 7–28. 2020.
[3] T. H. Foregger, and M. Mather, M. E2524. *The American Mathematical Monthly.* Vol. 83, No. 9 (Nov., 1976), pp. 741–742
[4] D. Hespe, Ch. Schulz, D. Strash. Scalable Kernelization for Maximum Independent Sets. *ACM Journal of Experimental Algorithmics.* Volume 24, Article No.: 1.16, pp 1–22. 2019.
[5] M. Gardner, MATHEMATICAL GAMES – Some elegant brick-packing problems, and a new order-7 perfect magic cube. *Scientific American.* Vol. 234, No. 2 (February 1976), pp. 122–127.
[6] M. R. Garey, and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*, Freeman, New York, 2003.
[7] R.M. Karp. "Reducibility Among Combinatorial Problems." In: Complexity of Computer Computations. New York: Plenum. pp. 85–103. 1972.
[8] K. Rozman, A. Ghysels, B. Zavalnij, T. Kunej, U. Bren, D. Janežič, and J. Konc, Enhanced Molecular Docking: Novel Algorithm for Identifying Highest Weight k-Cliques in Weighted General and Protein-Ligand Graphs. *JOURNAL OF MOLECULAR STRUCTURE.* 1304 p. 137639 Paper: 137639. 2024.
[9] N. J. A. Sloane. *Challenge Problems: Independent Sets in Graphs.* https://oeis.org/A265032/a265032.html
[10] T. H. Foregger. Elementary Problem E2524. *The American Mathematical Monthly.* Vol. 82, No. 3 (Mar., 1975), p. 300.
[11] S. Szabó and B. Zaválnij, Reducing hyper graph coloring to clique search. *Discrete Applied Mathematics.* 264. pp. 196–207. 2019.
[12] S. Szabó, and B. Zaválnij, Clique search in graphs of special class and job shop scheduling. *Mathematics.* 10(5), 697. 2022.
[13] S. Szabó, and B. Zaválnij, Graph Coloring via Clique Search with Symmetry Breaking. *SYMMETRY (BASEL).* 14 : 8 Paper: 1574, 16 p. 2022.

Indeks avtorjev / Author index