Zbornik 23. mednarodne multikonference

INFORMACIJSKA DRUŽBA

Zvezek C

Proceedings of the 23rd International Multiconference

INFORMATION SOCIETY

Volume C

http://is.ijs.si

IS 2020

Odkrivanje znanja in podatkovna skladišča • SiKDD

Data Mining and Data Warehouses • SiKDD

Uredili / Edited by

Dunja Mladenić, Marko Grobelnik

5. oktober 2020 / 5 October 2020

Ljubljana, Slovenia

Zbornik 23. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2020

Zvezek C

Proceedings of the 23rd International Multiconference

INFORMATION SOCIETY – IS 2020

Volume C

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

Uredila / Edited by

Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

5. oktober 2020 / 5 October 2020
Ljubljana, Slovenia

Urednika:

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, Lana Zemljak
Oblikovanje naslovnice: Vesna Lasič

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

Ljubljana, oktober 2020

Informacijska družba
ISSN 2630-371X

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=33077251
ISBN 978-961-264-192-4 (epub)
ISBN 978-961-264-193-1 (pdf)

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2020

Triindvajseta multikonferenca Informacijska družba (http://is.ijs.si) je doživela polovično zmanjšanje zaradi korone. Zahvala za preživetje gre tistim predsednikom konferenc, ki so se kljub prvi pandemiji modernega sveta pogumno odločili, da bodo izpeljali konferenco na svojem področju.

Korona pa skoraj v ničemer ni omejila neverjetne rasti IKTja, informacijske družbe, umetne inteligence in znanosti nasploh, ampak nasprotno – kar naenkrat je bilo večino aktivnosti potrebno opraviti elektronsko in IKT so dokazale, da je elektronsko marsikdaj celo bolje kot fizično. Po drugi strani pa se je pospešil razpad družbenih vrednot, zaupanje v znanost in razvoj. Celo Flynnov učinek – merjenje IQ na svetovni populaciji – kaže, da ljudje ne postajajo čedalje bolj pametni. Nasprotno - čedalje več ljudi verjame, da je Zemlja ploščata, da bo cepivo za korono škodljivo, ali da je korona škodljiva kot navadna gripa (v resnici je desetkrat bolj). Razkorak med rastočim znanjem in vraževerjem se povečuje.

Letos smo v multikonferenco povezali osem odličnih neodvisnih konferenc. Zajema okoli 160 večinoma spletnih predstavitev, povzetkov in referatov v okviru samostojnih konferenc in delavnic in 300 obiskovalcev. Prireditev bodo spremljale okrogle mize in razprave ter posebni dogodki, kot je svečana podelitev nagrad – seveda večinoma preko spleta. Izbrani prispevki bodo izšli tudi v posebni številki revije Informatica (http://www.informatica.si/), ki se ponaša s 44-letno tradicijo odlične znanstvene revije.

Multikonferenco Informacijska družba 2020 sestavljajo naslednje samostojne konference:

- Etika in stroka
- Interakcija človek računalnik v informacijski družbi
- Izkopavanje znanja in podatkovna skladišča
- Kognitivna znanost
- Ljudje in okolje
- Mednarodna konferenca o prenosu tehnologij
- Slovenska konferenca o umetni inteligenci
- Vzgoja in izobraževanje v informacijski družbi

Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi tudi ACM Slovenija, SLAIS, DKZ in druga slovenska nacionalna akademija, Inženirska akademija Slovenije (IAS). V imenu organizatorjev konference se zahvaljujemo združenjem in institucijam, še posebej pa udeležencem za njihove dragocene prispevke in priložnost, da z nami delijo svoje izkušnje o informacijski družbi. Zahvaljujemo se tudi recenzentom za njihovo pomoč pri recenziranju.

V 2020 bomo petnajstič podelili nagrado za življenjske dosežke v čast Donalda Michieja in Alana Turinga. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejela prof. dr. Lidija Zadnik Stirn. Priznanje za dosežek leta pripada Programskemu svetu tekmovanja ACM Bober. Podeljujemo tudi nagradi »informacijska limona« in »informacijska jagoda« za najbolj (ne)uspešne poteze v zvezi z informacijsko družbo. Limono je prejela »Neodzivnost pri razvoju elektronskega zdravstvenega kartona«, jagodo pa Laboratorij za bioinformatiko, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani. Čestitke nagrajencem!

Mojca Ciglarič, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD
INFORMATION SOCIETY 2020

The 23rd Information Society Multiconference (http://is.ijs.si) was halved due to COVID-19. The multiconference survived due to the conference presidents that bravely decided to continue with their conference despite the first pandemics in the modern era.

The COVID-19 pandemics did not decrease the growth of ICT, information society, artificial intelligence and science overall, quite on the contrary – suddenly most of the activities had to be performed by ICT and often it was more efficient than in the old physical way. But COVID-19 did increase downfall of societal norms, trust in science and progress. Even the Flynn effect – measuring IQ all over the world – indicates that an average Earthling is becoming less smart and knowledgeable. Contrary to general belief of scientists, the number of people believing that the Earth is flat is growing. Large number of people are weary of the COVID-19 vaccine and consider the COVID-19 consequences to be similar to that of a common flu dispute empirically observed to be ten times worst.

The Multiconference is running parallel sessions with around 160 presentations of scientific papers at twelve conferences, many round tables, workshops and award ceremonies, and 300 attendees. Selected papers will be published in the Informatica journal with its 44-years tradition of excellent research publishing.

The Information Society 2020 Multiconference consists of the following conferences:

- Cognitive Science
- Data Mining and Data Warehouses
- Education in Information Society
- Human-Computer Interaction in Information Society
- International Technology Transfer Conference
- People and Environment
- Professional Ethics
- Slovenian Conference on Artificial Intelligence

The Multiconference is co-organized and supported by several major research institutions and societies, among them ACM Slovenia, i.e. the Slovenian chapter of the ACM, SLAIS, DKZ and the second national engineering academy, the Slovenian Engineering Academy. In the name of the conference organizers, we thank all the societies and institutions, and particularly all the participants for their valuable contribution and their interest in this event, and the reviewers for their thorough reviews.

For the fifteenth year, the award for life-long outstanding contributions will be presented in memory of Donald Michie and Alan Turing. The Michie-Turing award was given to Prof. Dr. Lidija Zadnik Stirn for her life-long outstanding contribution to the development and promotion of information society in our country. In addition, a recognition for current achievements was awarded to the Program Council of the competition ACM Bober. The information lemon goes to the "Unresponsiveness in the development of the electronic health record", and the information strawberry to the Bioinformatics Laboratory, Faculty of Computer and Information Science, University of Ljubljana. Congratulations!

Mojca Ciglarič, Programme Committee Chair
Matjaž Gams, Organizing Committee Chair

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
prof. Toby Walsh, Australia

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Marjetka Šprah
Mitja Lasič
Blaž Mahnič
Jani Bizjak
Tine Kolenik

Programme Committee

Mojca Ciglarič, chair
Bojan Orel, co-chair
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
Špela Stres
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Tanja Urbančič
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah

KAZALO / TABLE OF CONTENTS

Zbornik 23. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2030

Zvezek C

Proceedings of the 23rd International Multiconference

INFORMATION SOCIETY – IS 2020

Volume C

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

Uredila / Edited by

Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

5. oktober 2020 / 5 October 2020
Ljubljana, Slovenia

PREDGOVOR

Tehnologije, ki se ukvarjajo s podatki so v devetdesetih letih 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, prišlo je do standardizacije procesov, povpraševalnih jezikov itd. 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 – pojavilo se je t.i. skladiščenje podatkov (data warehousing), ki je postalo standarden del informacijskih sistemov v podjetjih. Paradigma OLAP (On-Line-Analytical-Processing) zahteva od uporabnika, da še vedno sam postavlja sistemu vprašanja in dobiva nanje odgovore in na vizualen način preverja in išče izstopajoče situacije. Ker seveda to ni vedno mogoče, se je pojavila potreba po avtomatski analizi podatkov oz. z drugimi besedami to, da sistem sam pove, kaj bi utegnilo biti zanimivo za uporabnika – to prinašajo tehnike odkrivanja znanja v podatkih (data mining), ki iz obstoječih podatkov skušajo pridobiti novo znanje in tako uporabniku nudijo novo razumevanje dogajanj zajetih v podatkih. Slovenska KDD konferenca 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 after mid 90’s. 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. At this point, data warehousing with On-Line-Analytical-Processing entered as a usual part of a company’s information system portfolio, requiring from the user to set well defined questions about the aggregated views to the data. Data Mining is a technology developed after year 2000, offering automatic data analysis trying to obtain new discoveries from the existing data and enabling a user new insights in the data. In this respect, the Slovenian KDD conference (SiKDD) covers a broad area including Statistical Data Analysis, Data, Text and Multimedia Mining, Semantic Technologies, Link Detection and Link Analysis, Social Network Analysis, Data Warehouses.

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Janez Brank, Department of Artificial Intelligence, Jožef Stefan Institute, Ljubljana

Marko Grobelnik, , Department of Artificial Intelligence, Jožef Stefan Institute, Ljubljana

Branko Kavšek, University of Primorska, Koper

Aljaž Košmerlj, Qlector, Ljubljana

Dunja Mladenić, Department of Artificial Intelligence, Jožef Stefan Institute, Ljubljana

Inna Novalija, Department of Artificial Intelligence, Jožef Stefan Institute, Ljubljana

Luka Stopar, Sportradar, Ljubljana

A Dataset for Information Spreading over the News

Abdul Sittar
Jožef Stefan Institute
Ljubljana, Slovenia
abdul.sittar@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

ABSTRACT

Analysing the spread of information related to a specific event in the news has many potential applications. Consequently, various systems have been developed to facilitate the analysis of information spreading, such as detection of disease propagation and identification of the spreading of fake news through social media. The paper proposes a method for tracking information spread over news articles. It works by comparing subsequent articles via cosine similarity and applying a threshold to classify into three classes: "Information-Propagated", "Unsure" and "Information-not-Propagated". There are several open challenges in the process of discerning information propagation, among them the lack of resources for training and evaluation. This paper describes the process of compiling corpus from the Event Registry global media monitoring system. We focus on information spreading in three domains: sports (i.e. the FIFA World Cup), natural disasters (i.e. earthquakes), and climate change (i.e. global warming). This corpus is a valuable addition to currently available dataset to examine the spreading of information about various kind of events.

KEYWORDS

Datasets, Information propagation, News articles

1 INTRODUCTION

Information spreading has received significant attention due to its various market applications such as advertisement. did the information about a specific product reach to the public of a specific region? This could be one of the significant research questions. Research in this area considers influential factors in the process of information spreading such as the economic condition of a specific area related to how textual or visual content is helping to advertise a product. Information spreading analytics can also be used in shaping policies, e.g., in media companies to understand if there is a need to improve the content before publishing it. Health organizations may be interested to know the patterns of spreading of a cure for a certain disease. Environmental scientists are perhaps attentive to see whether spread of news about climate changes inside the country is similar to what is being reported internationally.

Domain-specific gaps in information spreading are ubiquitous, and may exist due to economic conditions, political factors, or linguistic, geographical, time-zone, cultural and other barriers. These factors potentially contribute to obstructing the flow of local as well as international news. We believe that there is a lack of research studies which examine, identify and uncover the reasons for barriers in information spreading. Additionally, there is limited availability of datasets containing news text and metadata including time, place, source and other relevant information.

When a piece of information starts spreading, it implicitly raises questions such as:

(1) How far does the information in the form of news reach out to the public?
(2) Does the content of news remain the same or changes to a certain extent?
(3) Do the cultural values impact the information especially when the same news will get translated in other languages?

This paper presents a corpus that focuses on information spreading over news and that hopes to answer some of the above questions (This corpus is published as an online resource at). We present the use of a news repository to produce a corpus and then analyze information propagation. We present a novel methodology for automatically assembling the corpus for this problem and validate it in three different domains. We focused on a combination of rich- and low resource European languages, in particular English, Portuguese, German, Spanish, and Slovene. Three different types of events are targeted in the data collection procedure to potentially involve different information spreading behaviors in our society. These events are sports (FIFA World Cup, 2,695 articles), natural disasters (earthquakes, 3,194 articles), and climate change (global warming, 1,945 articles). The three types of events were chosen based on their popularity and diversity. A list of sub-events was observed from top websites related to the three events and we selected those which were the most popular in the countries with the selected national languages. For sports, a list of countries with their national sports was fetched and then filtered for national language[1], [2]. Based on popularity, we selected the FIFA world cup. Similarly, for natural disasters, lists of natural disasters were collected by country taking the national language into account, for instance, for Slovenia we looked for this country in the natural disaster category on Wikipedia[3]. Earthquakes[4] and global warming[5] were found to be the most prevalent, thus a dataset for each was collected. Table 1 shows the selected events and other related events ordered by prevalence.

Table 1: List of events

Selected events	Other events (ordered by popularity)
Football	Basketball, Baseball, Boxing, Tennis, Cycling
Earthquake	Floods, Tsunamis, Landslides, Hurricane, Volcanic eruptions
Global warming	CO_2 emissions, Chemical consumption

The paper makes the following contributions to science:

(1) a novel methodology to collect a domain-specific corpus from news repository;
(2) semantic similarity between news articles;

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

[1] http://www.quickgs.com/countries-and-their-national-sports/
[2] https://www.topendsports.com/
[3] https://en.wikipedia.org/wiki/Category:Natural_disasters_in_Slovenia
[4] https://en.wikipedia.org/wiki/List_of_earthquakes_in_2020
[5] https://www.theguardian.com/environment/2011/apr/21/countries-responsible-climate-change, [6]

(3) an annotated dataset encoding the level of information spreading from an article.

The rest of the paper is organized as follows: in Section 2 we discuss prior work about information spreading; in Section 3 we describe the data collection methodology; Section 4 describes semantic similarity and dataset annotation; and Section 5 gives the conclusions.

2 RELATED WORK

Information spreading is prevalent in our society. It plays a vital part in tasks that encompass the spreading of innovations [9], effects in marketing [6], and opinion spreading [4]. News spreading provides information to consumers that can be used for decision making and potentially contribute to shaping national and international policies. There are several types of media involved, such as print media, broadcast, and internet media. Internet is considered as a building block for connecting individuals worldwide, while news reflects current significant events for people [7]. Apart from news, online social media proved to be a remarkable alternative to support information spreading in an emergency [8, 5]. Social connection plays a vital role in news spreading. Especially the structure of network reflecting who is connected to whom, crucially increases the proportion of information spreading. Network structure analysis comes with a hypothesis related to the strength of the connections, namely that information will spread further in a situation where there exist many weak connections rather than clusters of strong [2].

While, in general, there are not many dataset that would help in modelling information spreading, there are some corpora for detecting the spreading of information about diseases [3] and fake news in social media [10]. There is currently no multilingual dataset of news articles for analysis of information propagation composed from a variety of event-centric information such as sports, natural disasters, and climate changes. This provides additional motivation for our work.

3 DATA COLLECTION METHODOLOGY

In order to collect news originating from different sources, in different languages, and targeting diverse events, we used Event Registry, a platform that identifies events by collecting related articles written in different languages from tens of thousands of news sources [9]. Using Event Registry APIs [7], we fetched a list of articles about each event in the following languages: English, Spanish, German, Portuguese, and Slovenian. Figure 1 shows the data collection process.

Each article was parsed from the JSON response and stored in CSV files. Each article was connected with the available list of relevant information such as the language of the article, event type, publisher, title, date, and time. Figure 2 shows the metadata of articles.

The number of collected articles in each domain varies considerably, and also varies across the languages within each domain. Table 2 shows statistics about each dataset.

4 SEMANTIC SIMILARITY BETWEEN NEWS ARTICLES

We have represented the cross-lingual news articles by monolingual (English) Wikipedia concepts using the Wikifier service[8]. This service uses a page-rank based method to identify a coherent set of relevant concepts from Wikipedia [1]. We retrieved a list of Wikipedia concepts for each article. After representing each article with a list of Wikipedia concepts, the tf-idf score was computed using the popular machine learning library Scikit-Learn[9]. Using the same library, cosine similarity was calculated between tf-idf representation of news articles across all five languages. In the process of computing similarity between the articles, for each article we calculated its cosine similarity to all other articles and stored the results in a CSV file. The results were then sorted based on the publishing time of articles and we kept only the calculations of similarity to articles that are published later that the article in hands. Since we are interested in information propagation, we do not need to compare an article to those articles which have been published before it. As a result, we had a multiple similarity score for each article where each score show the similarity with other articles. Cosine similarity varies between zero and one, zero meaning no similarity and one meaning maximum similarity, i.e., a duplicate article.

[7] https://github.com/EventRegistry/event-registry-python/blob/master/eventregistry/examples/QueryArticlesExamples.py
[8] http://wikifier.org/info.html
[9] https://scikit-learn.org/stable/

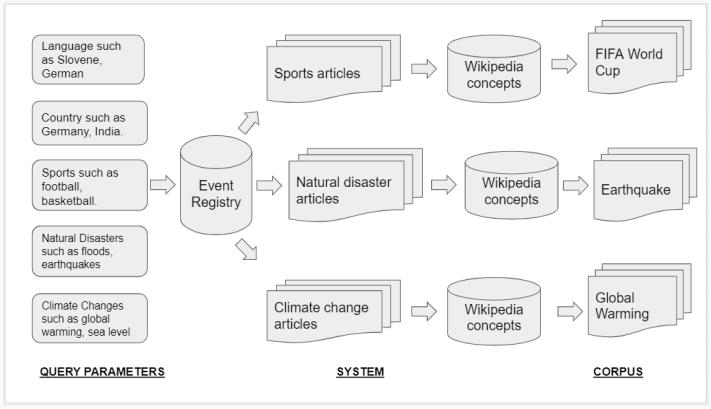

Figure 1: Data collection methodology

A	B	C	D	E	F	G	H	I
Language+I	Event	Weight	Class	Article Title	Publisher	Publishing Time	Website	Article URL
English10	FIFA World Cup	0.991	Information-Propagated	Football: Be	Channel Ne	2020-04-29T16:0	channelne	https://www
English100	FIFA World Cup	0.55	Unsure	Despite beat	OnlineNige	2020-04-28T11:0	news2.on	https://news
English1000	FIFA World Cup	1	Information-Propagated	Norman Hur	Borehamwo	2020-04-10T11:0	borehamv	https://www
English101	FIFA World Cup	0.195	Information-Not-Propagated	Qatar prepa	web4.inside	2020-04-28T10:0	web4.insi	https://www
English102	FIFA World Cup	1	Information-Propagated	Despite beat	Legit.ng	2020-04-28T09:4	legit.ng	https://www
English103	FIFA World Cup	0.199	Information-Not-Propagated	Lungu eulog	Zambia Dail	2020-04-28T09:0	daily-mail.	http://www.
English104	FIFA World Cup	1	Information-Propagated	100 General	My London	2020-04-28T08:5	mylondon	https://www
English105	FIFA World Cup	0.272	Information-Not-Propagated	Nigeria: Oge	allAfrica	2020-04-28T07:4	allafrica.c	https://allafr
English106	FIFA World Cup	0.304	Information-Not-Propagated	What really	Coventry Te	2020-04-28T07:1	coventryt	https://www
English107	FIFA World Cup	0.331	Information-Not-Propagated	From Abdelg	The Nation	2020-04-28T06:4	thenation	https://www
English11	FIFA World Cup	0.906	Information-Propagated	Beckenbaue	Business St	2020-04-29T15:5	business-s	https://www
English110	FIFA World Cup	0.232	Information-Not-Propagated	Analysts' Co	Vancouver	2020-04-27T23:5	whitecaps	https://www
English111	FIFA World Cup	1	Information-Propagated	Indian footb	Scroll	2020-04-27T23:1	scroll.in	https://scrol
English112	FIFA World Cup	0.369	Information-Not-Propagated	Taggart's th	SBS Austral	2020-04-27T22:2	theworldg	https://thew
English113	FIFA World Cup	0.257	Information-Not-Propagated	VAN DIEST:	Toronto Su	2020-04-27T22:2	torontosu	https://toro
English114	FIFA World Cup	0.3	Information-Not-Propagated	Ronaldinho	SBS Austral	2020-04-27T22:1	theworldg	https://thew
English115	FIFA World Cup	0.379	Information-Not-Propagated	Liverpool co	Paisley Gat	2020-04-27T19:1	paisleygat	https://paisl
English116	FIFA World Cup	0.245	Information-Not-Propagated	Manchester	TODAY	2020-04-27T17:2	today.ng	https://www
English117	FIFA World Cup	0.331	Information-Not-Propagated	East Bengal	Indian Expr	2020-04-27T17:2	indianexp	https://india
English118	FIFA World Cup	0.254	Information-Not-Propagated	East Bengal	Firstpost	2020-04-27T17:1	firstpost.c	https://www
English119	FIFA World Cup	0.859	Information-Propagated	General kno	Radio Time	2020-04-27T16:4	radiotime	https://www
English12	FIFA World Cup	1	Information-Propagated	Beckenbaue	timesofmal	2020-04-29T15:4	timesofma	https://time
English120	FIFA World Cup	0.841	Information-Propagated	Argentine st	Legit.ng	2020-04-27T16:3	legit.ng	https://www

Figure 2: Articles with metadata

Table 2: Statistics about dataset

Dataset	Domain	Event type	Articles per Language					Total Articles
			Eng	Spa	Ger	Slv	Por	
1	Sports	FIFA World Cup	983	762	711	10	216	2682
2	Natural Disaster	Earthquake	941	999	937	19	251	3147
3	Climate Changes	Global Warming	996	298	545	8	97	1944

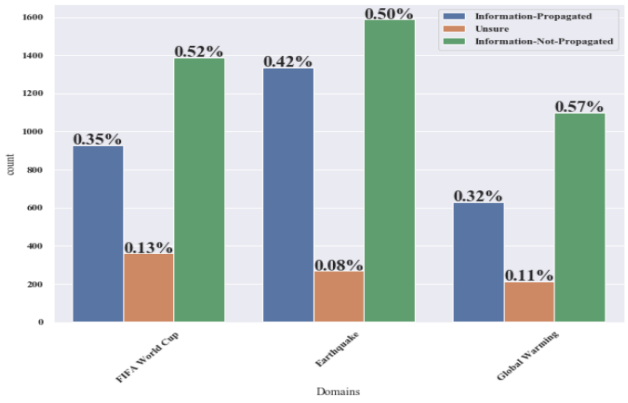

Figure 3: Class distribution for all domains

4.1 Dataset annotations

The results of the semantic similarity calculation were in the form of a table where rows shown the list of articles and columns shown the corresponding similarity score in the range 0..1 with all the other articles. This similarity score was calculated using cosine between TF-IDF representation of news articles (See Section **??**). First, we excluded those articles which had scored 1.0, as they were considered as a copy of the article. We then, for each article, chose an article which had the highest similarity score to it from the list of all articles. After performing this step, we had one similarity score for each article which shows either that the information spread to a certain extent (if >0) or not (if 0). To decide about the class label whether the information is spreading or not, we divided the scores into three intervals. The first is Similarity $\geq$ 0.7, the second is 0.7 > Similarity $\geq$ 0.4, and the third is Similarity < 0.4. Articles that have scores in the first interval were labeled as "Information-Propagated". The second interval was considered as unclear whether the information from the article propagated or not such articles were labeled as "Unsure". The lowest interval was considered as a signal for no propagation and labeled "Information-not-Propagated". For instance, low similarity can be of an article about a sports ground which mentions the population of the city and another article that discusses the population itself. We have manually examined concepts of articles in each class. Figure 3 shows the distribution of class labels in FIFA World Cup, Earthquake, and Global Warming dataset respectively.

4.2 Evaluation of dataset

Each article was annotated with a label based upon the similarity score threshold of each article with other articles (See Section 4.1). For evaluation of the dataset we have checked the content of the corresponding articles which were responsible for a specific class label. We performed the evaluation of labelling by manually inspecting a subset of pairs of articles. If a pair, for instance, were labelled as "Information-Propagated" then two articles should have text discussing more or less the same event, both in mono- and cross-lingual settings.

We have randomly chosen 10 articles with their corresponding articles considering all languages in each class and in each dataset. In this way, we have manually checked 180 articles. Table 3 shows these pairs of articles for evaluation in each dataset. We scanned each article manually for all languages, using Google Translator for Portuguese, German, Slovene and Spanish to translate them into English.

Table 3: Selected articles for evaluation

Domains	Percentage of correctly labelled pairs
Global Warming	100%
Earthquake	93%
FIFA World Cup	100 %

Evaluation results shown that the annotation was significantly related to information spreading. Articles in the "Information-Propagated" class show that most articles were an exact or paraphrased copy of each other, with some articles published within few hours after each other. Articles in the "Unsure" class were typically also relevant to the event but involved extra and different discussions. Lastly, in the third class "Information-Not-Propagated", articles involved only keywords related to event but discussion was about other topics. Moreover, here the gap in the publishing time was quite large.

5 CONCLUSIONS

This paper proposed a methodology and explained the process of data collection from a news repository to provide a corpus for event-centric information propagation between news articles. This corpus covers three domains and each dataset corresponds to one event type (FIFA World Cup, Earthquake, and Global Warming). The corpus is available to others for the evaluation of techniques for information spreading as it allows the analysis of cross-lingual news articles published by different publishers located geographically in different places.

In the future, we plan to add more attributes to each dataset. For instance, for now, we only know the publisher of a news article but in the future, we would like to include the publisher profile and the economic condition of a country from where the information is published. Also, we plan to apply and evaluate different techniques to analysis information propagation barriers.

6 ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency and the project leading to this publication has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. In *Proceedings of Slovenian KDD Conference on Data Mining and Data Warehouses (SiKDD)*.

[2] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science*, 329, 5996, 1194–1197.

[3] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Covid-19: the first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*.

[4] David Liben-Nowell and Jon Kleinberg. 2008. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the national academy of sciences*, 105, 12, 4633–4638.

[5] Kees Nieuwenhuis. 2007. Information systems for crisis response and management. In *International Workshop on Mobile Information Technology for Emergency Response*. Springer, 1–8.

[6] Everett M Rogers. 2010. *Diffusion of innovations*. Simon and Schuster.

[7] Sandeep Suntwal, Susan Brown, and Mark Patton. 2020. How does information spread? an exploratory study of true and fake news. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

[8] Satish V Ukkusuri, Xianyuan Zhan, Arif Mohaimin Sadri, and Qing Ye. 2014. Use of social media data to explore crisis informatics: study of 2013 oklahoma tornado. *Transportation Research Record*, 2459, 1, 110–118.

[9] Duncan J Watts and Peter Sheridan Dodds. 2007. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34, 4, 441–458.

[10] Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. 2020. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9, 1, 7.

Learning to fill the slots from multiple perspectives

Patrik Zajec
patrik.zajec@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Dunja Mladenič
dunja.mladenic@ijs.si
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

We present an approach to train the slot-filling system in a fully automatic, semi-supervised setting on a limited domain of events from Wikipedia using the summaries in different languages. We use the multiple languages and the different topics of the events to provide several alternative views on the data. Our experiments show how such an approach can be used to train the multilingual slot-filling system and increase the performance of a monolingual system.

KEYWORDS

information extraction, slot filling, machine learning, probabilistic soft logic

1 INTRODUCTION

This paper is addressing the slot filling task that aims to extract the structured knowledge from a given set of documents using a model trained for a specific domain and the associated slots. For example, within a news article reporting on an earthquake, the task is to detect the earthquake's magnitude, the number of people injured, the location of the epicentre and other information. We refer to those as a set of *slot keys* or *slots*, to their exact values as a *slot values* and to the named entities from the documents corresponding to those values as *target entities*.

Slot filling is closely related to the task of relation extraction [1] and can be seen as a kind of unary relation extraction. Both tasks can be formulated as classification and are usually approached by first training a classifier with a sentence and tagged entities at the input and the prediction of relation or slot key as the output.

As there is a large number of relations between entities that we might be interested in detecting, there is also a large number of slot keys we seek the slot value for. In order to avoid the resource-intensive process of annotating a large number of examples for each possible slot/relation and to increase the flexibility of training procedures beyond the straight-forward supervised learning, many alternative approaches have been proposed, such as bootstrapping [4], distant supervision [6] and self supervision [5].

As stated both tasks can be performed for different types of documents. We limit our focus to news events on multiple topics (such as natural disasters and terrorist attacks), taking the articles reporting about events as the documents. Since the number of news topics is large, and consequently so is the number of slots, we would like to minimize the need for manual annotations. Furthermore, since the set of topics is not fixed and could expand over time, such a slot filling system should be able to adapt quickly to fill new slots and ideally should not be limited to the English language.

We believe that annotation work can be greatly minimized if we rely on our limited domain to identify and annotate only informative examples and use the additional assumptions to propagate these labels. We also believe that simultaneous training of the system on multiple topics can be advantageous, as we can introduce additional supervision on the common slots and use distinct slots as a source of negative examples.

In this work we use Wikipedia and Wikidata [9] as the source of data. We treat the Wikidata entities that have the point-in-time property specified as events and summary sections of Wikipedia articles about the entity in different languages as news articles. Each entity belongs to a single topic and we adopt the subset of topic-specific properties as slot keys. An automatic exact matching of such values from Wikidata with named entities from Wikipedia articles is rarely successful. We use the successful and unambiguous matches as a set of labeled seed examples.

We formulate the task as a semi-supervised learning problem [8] where the set of base learners is trained iteratively, starting with a small seed set of labeled examples and a larger set of unlabeled examples. In each iteration, the most confident predictions on the examples from unlabeled set are used to increase the training set by assigning pseudo-labels. We introduce an additional component which combines the confidences of multiple base learners for each example.

To the best of our knowledge, we are the first to use the limited domain of news events, which allows the additional assumptions, such as the connection between slots of different topics and the redundancy of reporting in multiple languages, to first train and later boost the performance of a slot-filling system.

The contributions of this paper are the following:

- we combine the data from Wikidata and Wikipedia to setup a learning and evaluation scenario that mimics the learning on news events and articles,
- we demonstrate how simultaneous learning on multiple topics and languages can be used not only to train the multilingual slot-filling system, but to also improve the performance of a monolingual system,
- we show how an inference component can be used to combine predictions from multiple base learners to improve the pseudo-labeling step of the semi-supervised learning process.

2 METHODOLOGY

2.1 Problem Definition

Given a collection of topics $\mathcal{T}$ (such as earthquakes, terrorist attacks, etc.), where each topic t has its own set of slot keys $\mathcal{S}_t$, the goal is to automatically extract values from the relevant texts

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

to fill in the slots. For example, the members of $\mathcal{S}_{earthquakes}$ are *number of injured*, *magnitude* and *location*. For each topic t there is a set of events $\mathcal{E}_t$, each of which took place at some point in time and was reported by several documents in different languages.

The values of all or at least most slot keys (or slots) from $\mathcal{S}_d$ are represented in each of the documents as named entities, which we also refer to as *target entities*. We say most of the slots, since it is possible that an earthquake caused no casualties. It is also possible that some of the documents do not report about the number of casualties as it may be too early to know if there were any. In addition, the documents might contain different values for the same slot key, as for example, the reported number of people injured by an earthquake can increase over time. There may also be several different mentions of the same slot in a particular document, as for example one magnitude might refer to an actual earthquake that the event is about, while the other magnitude might refer to an earthquake that struck the same region years ago.

Our task is actually a two step process. In the first step, the goal is to train a system capable of identifying the target entities for a set of slot keys from the context, which in our case is limited to a single sentence. Such a system is not yet able to recognise the true value for a given slot if there are multiple different candidates, such as selecting the actual magnitude from several reported magnitude values. The goal of the second step is to assign a single correct value to each of the slot keys. We assume that inferring the correctness of a value is a document-level task, since it requires a broader context. Solving the first step is a kind of prerequisite for the second step, so we focus on it in this paper.

2.2 Overview of the proposed method

The system is trained iteratively and starts with a noisy seed set, which grows larger with pseudo-labeled positive and negative examples. Each of the base learners is trained on the set of labeled examples from the topic (or multiple topics) and language assigned to it. The prediction probabilities for each of the unlabeled examples are determined by combining the probabilities of all base learners. This is done either by averaging or by feeding the probabilities as approximations of the true labels into the component, which attempts to derive the true value for each example and the error rates for each learner [7]. The examples with probabilities above or below the specific thresholds are given a pseudo-label and added to the training set.

The seed set is constructed by matching the slot values obtained from Wikidata with named entities found in Wikipedia articles for each event. There are only a handful of unambiguous matches for each slot key, which are labeled as a positive examples, while the negative examples are all other named entities from the articles in which they appeared. Figure 1 shows a high-level overview of the proposed methodology. The entire workflow is repeated in each iteration until no new examples are selected for pseudo-labelling.

2.3 Representing the entities

Each named entity together with its context forms a single example. We annotate each article and extract the named entities with Spacy [1]. To capture the context, we compute the vector representation of each entity by replacing it with a mask token and feeding the entire sentence through a pre-trained version of the XLM Roberta model [3] using the implementation from the Transformers [2] library. Note that the representation of each entity remains fixed throughout the learning process because we have found that the representation is expressive enough for our purposes and it speeds up the training between iterations. Also note that since the entity is masked, it is not directly captured in the representation.

2.4 Selecting the topics

Our assumption is that training the system to detect the slots on multiple topics simultaneously can provide additional benefits. For two topics t and t' there is potentially a set of common slots and a set of topic-specific slots.

For slot s' which appears in both topics the base learner trained on t' can be used to make predictions for examples from t. By combining predictions from learners trained on t and t', we could get a better estimate of the true labels of the examples.

For the slot s, which is specific to the topic t, all examples from the topic t' can be used as negative examples. Selecting reliable negative examples from the same topic is not easy, as we may inadvertently mislabel some of the positive examples.

2.5 Using multiple languages

Articles from different languages offer in some ways different views on the same event. The slot values we are trying to detect should appear in all the articles, as they are highly relevant to the event.

The values for slots such as location and time should be consistent across all articles, whereas this does not necessarily apply to other slots such as the number of injured or the number of casualties. Matching such values across the articles is therefore not a trivial task, and although a variant of soft matching can be performed, we leave it for the future work and limit our focus only on the values that can be matched unambiguously.

We can combine the predictions of several language-specific base learners into a single pseudo-label for entities that can be matched across the articles.

2.6 Assigning pseudo labels

Each iteration starts with a set of labeled examples X_l, a set of unlabeled examples X_u and a set of base learners trained on X_l. Base learners are simple logistic regression classifiers that use vector representations of entities as features and classify each example x as a target entity for the slot key s or not.

Each base learner $\bar{f}^s_{t,l}$ is a binary classifier trained on the labeled data for the slot key s from the topic t and the language l. Such base learners are *topic-specific* as they are trained on a single topic t. Base learners $\bar{f}^s_l$ are trained on the labeled data for the slot key s from the language l and all the topics with the slot key s. Such base learners are *shared* across topics, as they consider the examples from all the topics as a single training set. We use the classification probability of the positive class instead of hard labels, $\bar{f}^s_{t,l}(x), \bar{f}^s_l(x) \in [0, 1]$.

For each entity x from a news article with the language l reporting on the event e from the topic t we obtain the following predictions:

- $\bar{f}^s_{t',l}(x)$ for each $s \in \mathcal{S}_t$ and all such t' that $s \in \mathcal{S}_{t'}$, that is the probability that x is a target entity for the slot key

[1] https://spacy.io/

[2] https://huggingface.co/transformers/

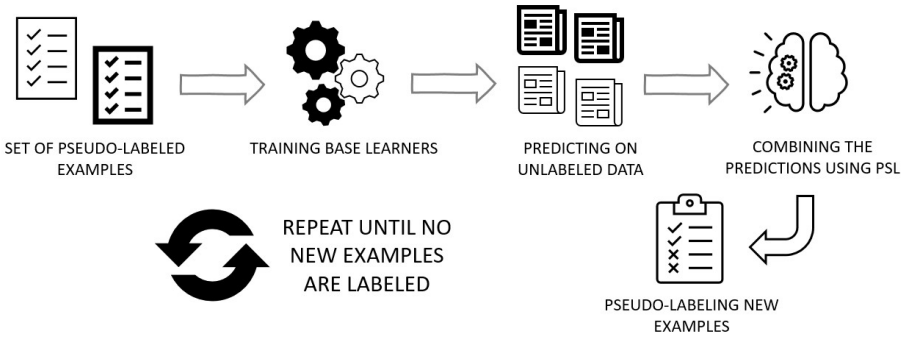

Figure 1: High-level overview of the proposed methodology.

s, where s is a slot key from the topic t, using the *topic-specific* base learner trained on examples from the same language on the topic t' that also has the slot key s,

- $\bar{f}^s_{t,l'}(x)$ which equals $\bar{f}^s_{t,l'}(y)$ for each $s \in \mathcal{S}_t$ and for each language l' such that there is an article reporting about the same event e in that language and contains an entity y which is matched to x,
- $\bar{f}^s_l(x)$ for each $s \in \mathcal{S}_t$, using the *shared* base learner, which is on examples from all topics t' that have the slot key s.

Predictions from multiple base learners for each x and s are combined as a weighted average to obtain a single prediction $f^s(x)$. The weight of each base learner $\bar{f}$ is determined by its error rate $e(\bar{f})$ which is estimated using an approach from [7] using both unlabeled and labeled examples. This is done by introducing the following logical rules (referred to as *ensemble rules* in [7]) for each of the base learners $\bar{f}^s$ predicting for x:

$$\bar{f}^s(x) \wedge \neg e(\bar{f}^s) \rightarrow f^s(x),\ and\ , \bar{f}^s(x) \wedge e(\bar{f}^s) \rightarrow \neg f^s(x),$$
$$\neg\bar{f}^s(x) \wedge \neg e(\bar{f}^s) \rightarrow \neg f^s(x),\ and\ , \neg\bar{f}^s(x) \wedge e(\bar{f}^s) \rightarrow f^s(x).$$

The truth values are not limited to Boolean values, but instead represent the probability that the corresponding ground predicate or rule is true. For a detailed explanation of the method we refer the reader to [7]. We introduce a prior belief that the predictions of base learners are correct via the following two rules:

$$\bar{f}^s(x) \rightarrow f^s(x),\ and\ , \neg\bar{f}^s(x) \rightarrow \neg f^s(x).$$

Since each x can be target entity for at most one slot key, we introduce a *mutual exclusion* rule:

$$\bar{f}^s(x) \wedge f^{s'}(x) \rightarrow e(\bar{f}^s).$$

The rules are written in the syntax of a Probabilistic soft logic [2] program, where each rule is assigned a weight. We assign a weight of 1 to all *ensemble rules*, a weight of 0.1 to all *prior belief* rules and a weight of 1 to all *mutual exclusion* rules. The inference is performed using the PSL framework [3]. As we obtain the approximations for all $x \in X_u$, we extend the set of positive examples for each slot s with all x such that $f^s(x) >= T_p$ and the set of negative examples with all x such that $f^s(x) <= T_n$, for predefined thresholds T_p and T_n.

3 EXPERIMENTS

3.1 Dataset

To evaluate the proposed methodology, we have conducted experiments on two topics: *earthquakes* and *terrorist attacks*.

[3]https://psl.linqs.org/

We have collected the Wikipedia articles and Wikidata information of 913 earthquakes from 2000 to 2020 in 6 different languages, namely English, Spanish, German, French, Italian and Dutch. We have manually annotated the entities of 85 English articles using the slot keys *number of deaths*, (*number of injured* and *magnitude*, which serve as a labeled test set and are not included in the training process. In addition, we have collected the data of 315 terrorist attacks from 2000 to 2020 with the articles from the same 6 languages.

3.2 Evaluation Settings

The evaluation for each approach is performed on the labeled English dataset, where 76 entities are labeled as number of deaths, 45 as number of injured and 125 as magnitude. The threshold values for the pseudo-labeling are set to $T_p = 0.6$ and $T_n = 0.05$. The approaches differ by the subset of base learners used to form the combined prediction and by the weighting of the predictions.

Single or multiple languages. In single language setting, only English articles are used to extract the entities and train the base learners. In the multi-language setting, all available articles are used and the entities are matched across the articles from the same event.

Single or multiple topics. In the single topic setting only the examples from the *earthquake* topic are used. In the multi-topic setting, the examples from *terrorist attacks* are used as negative examples for the slot key *magnitude*, the base learners for the slot keys *number of deaths* and *number of injured* are combined as described in the section 2.6.

Uniform or estimated weights. In the uniform setting all predictions of the base learners contribute equally, while in the estimated setting the weights of the base learners are estimated using the approach described in the section 2.6.

3.3 Results and discussion

The results of all experiments are summarized in the table 1. Since the test set is limited to the topic *earthquake* and English, only a subset of base learners was used to make the final predictions. We report the average value of precision, recall and F1 across all slot keys. The threshold of 0.5 was used to round the classification probabilities.

Single iteration. Approaches in which base learners are trained on the initial seed set for a single iteration achieve higher precision with the cost of a lower recall. We observe that they distinguish almost perfectly between the slots from the seed set and

Table 1: Results of all experiments. The column *Single iteration* reports the results of approaches where base learners were trained on the seed set only. Results where base learners were trained in the semi-supervised setting with different weightings of the predictions are reported in the columns *Uniform weights* and *Estimated weights*. The values of precision, recall and F1 are averaged over all slot keys.

	Single iteration			Uniform weights			Estimated weights		
Model	P	R	F1	P	R	F1	P	R	F1
Single language, single topic	0.94	0.64	0.76	0.83	0.75	0.77	0.84	0.76	0.79
Multiple languages, single topic	0.94	0.64	0.76	0.82	0.74	0.76	0.83	0.75	0.77
Single language, multiple topics	0.91	0.76	0.83	0.83	0.83	0.83	0.86	0.83	0.84
Multiple languages, multiple topics	0.93	0.76	0.83	0.82	0.83	0.82	0.84	0.84	0.84

produce almost no false positives. Using one or more languages has almost no effect on the averaged scores when the number of topics is fixed. When using multiple topics, a higher recall is achieved without a significant decrease in precision. All incorrect classifications of the slot *number on injured* are actually examples of the *number of missing* slot that is not included in our set and likewise almost all incorrect classifications for the slot *magnitude* are examples of the slot *intensity on the Mercalli scale*. This could easily be solved by expanding the set of slot keys and shows how important it is to learn to classify multiple slots simultaneously.

Semi-supervised. Approaches in which base learners are trained iteratively trade precision in order to significantly improve recall. Most of the loss of precision is due to misclassification between slots *number of deaths* and *number of injured*, similar as the example *"370 people were killed by the earthquake and related building collapses, including **228** in Mexico City, and more than 6,000 were injured."* where 228 was incorrectly classified as number of injured and not the number of deaths. The use of multiple topics reduces misclassification between these slots and further improves the recall as new contexts are discovered by the base learners trained on *terrorist attacks*.

Uniform and estimated weights. Using the estimated error rates as weights for the predictions of base learners shows a slight improvement in performance. It may be advantageous to estimate multiple error rates for *topic-specific* base learners, as they tend to be more reliable in predicting examples from the same topic. We believe that more data and experimentation is needed to properly evaluate this component. A major advantage is its flexibility, since we can easily incorporate prior knowledge of the slots or additional constraints on the predictions in the form of logical rules.

4 CONCLUSION AND FUTURE WORK

We presented an approach for training the slot-filling system which can benefit from large amounts of data from Wikipedia. The experiments were performed on a relatively small dataset and show that the proposed direction seems promising. However, the right test of our approach would be to apply it to a much larger number of topics and events, which will be done in the immediate next step. Furthermore, the current approach needs to be evaluated in more detail.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and NAIADES European Unions project under grant agreement H2020-SC5-820985.

REFERENCES

[1] Nguyen Bach and Sameer Badaskar. 2007. A Survey on Relation Extraction. Technical report. Language Technologies Institute, Carnegie Mellon University.

[2] Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss markov random fields and probabilistic soft logic. *The Journal of Machine Learning Research*, 18, 1, 3846–3912.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[4] Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning. In *Proceedings of AAAI*.

[5] Xu ming Hu, Lijie Wen, Y. Xu, Chenwei Zhang, and Philip S. Yu. 2020. Selfore: self-supervised relational feature learning for open relation extraction. *ArXiv*, abs/2004.02438.

[6] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 1003–1011.

[7] Emmanouil Platanios, Hoifung Poon, Tom M Mitchell, and Eric J Horvitz. 2017. Estimating accuracy from unlabeled data: a probabilistic logic approach. In *Advances in Neural Information Processing Systems*, 4361–4370.

[8] Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109, 2, 373–440.

[9] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57, 10, 78–85.

Knowledge graph aware text classification

Nela Petrželková*
Jožef Stefan Institute
Ljubljana, Slovenia
nela.petrzelkova@seznam.cz

Blaž Škrlj
Jožef Stefan Institute and
Jožef Stefan Int. Postgraduate School
Ljubljana, Slovenia
blaz.skrlj@ijs.si

Nada Lavrač
Jožef Stefan Institute
Ljubljana, Slovenia
nada.lavrac@ijs.si

ABSTRACT

Knowledge graphs are becoming ubiquitous in many scientific and industrial domains, ranging from biology, industrial engineering to natural language processing. In this work we explore how one of the largest currently available knowledge graphs, the Microsoft Concept Graph, can be used to construct interpretable features that are of potential use for the task of text classification. By exploiting graph-theoretic feature ranking, introduced as part of the existing tax2vec algorithm, we show that massive, real-life knowledge graphs can be used for the construction of features, derived from the relational structure of the knowledge graph itself. To our knowledge, this is one of the first approaches that explores how interpretable features can be constructed from the Microsoft Concept graph with more than five million concepts and more than 80 million IsA relations for the task of text classification. The proposed solution was evaluated on eight real-life text classification data sets.

KEYWORDS

knowledge graphs, text classification, feature construction, semantic enrichment

1 INTRODUCTION

Text classification is the process of assigning labels to text according to its content. It is one of the fundamental tasks in Natural Language Processing (NLP) with various applications such as spam detection, topic labeling, sentiment analysis, news categorization and many more [1]. In recent years, *knowledge graphs*—real-life graph-structured sources of knowledge—are becoming an interesting source of background knowledge, potentially useful in contemporary machine learning [2]. Knowledge graphs, such as DBPedia[1] or the Microsoft Concept Graph[2] span tens of millions of triplets of the form subject-predicate-object, and include many potentially interesting relations, from which a given machine learning algorithm can potentially benefit.

In this work we propose an approach to scalable *feature construction* from one of the largest freely available knowledge graphs, and demonstrate its utility on multiple real life data sets. The main contributions of this work are as follows:

(1) We propose an extension to the tax2vec [3] algorithm for semantic feature construction, adapting it to operate with real-life knowledge graphs comprised of tens of millions of triplets.
(2) The proposed method is extensively empirically evaluated, indicating that the proposed semantic feature construction aids the classification performance on many real-life datasets.
(3) The implemented method is freely available[3] with a simple-to-use, scikit-learn API.

The paper is structured as follows. Section 2 presents the background and related work. Section 3 presents the proposed approach to semantic feature construction using the information from a given knowledge graph. Section 4 describes the experimental setting and the results, followed by a summary and further work in Section 5.

2 BACKGROUND AND RELATED WORK

In text classification tasks, characterized by short documents or small amounts of documents, deep learning methods are frequently outperformed by more standard approaches, including SVMs [4]. In such settings, it was shown that approaches capable of using semantic context may outperform the naïve learning approaches, the examples are among other based on Latent Dirichlet Allocation [5], Latent Semantic Analysis [6] or word embeddings [7], which is referred to as first-level context.

Second-level context can be introduced by adding *background knowledge* into a learning process, which may help to increase performance and improve interpretability. Usage of knowledge graphs also helped in classification with extending neural network based lexical word embedding objective function [8]. Elhadad et al. [9] present an ontology-based web document, while Kaur et al. [10] propose a clustering-based algorithm for document classification that also benefits from knowledge stored in the underlying ontologies. Use of hypernym-based features was performed already in e.g., the Ripper rule learning algorithm [11]. Wang and Domeniconi [12] used the derived background knowledge from Wikipedia for text enriching. In short document classification, it was shown that the tax2vec algorithm (described below) can help those classifiers gain better results by adding *extra semantic knowledge* to the feature vectors.

The tax2vec [3] is an algorithm for *semantic feature construction* that can be used to enrich the feature vectors constructed by the established text processing methods such as the tf-idf. It takes as input a labeled or unlabeled corpus of documents and a word taxonomy, i.e. a directed graph to which parts of a given document map to. It outputs a matrix of *semantic feature vectors* where each row represents a semantics-based vector representation of one input document. It makes it by mapping the words from the document to a given taxonomy, WordNet or in this work Microsoft Concept Graph, by which it creates the collection of terms for each document and from it, a *corpus taxonomy*—a relational structure specific to the considered document space. The terms presented in the corpus taxonomy represent the potential features.

[1] https://wiki.dbpedia.org/
[2] https://concept.research.microsoft.com/Home/Introduction
[3] https://github.com/SkBlaz/tax2vec

Information society '20, October 5–9, 2020, Ljubljana, Slovenia

3 KNOWLEDGE GRAPH-BASED SEMANTIC FEATURE CONSTRUCTION

Semantic features are constructed as follows. With the help of spaCy library [13], we first find *nouns* in each document in the corpus and for every noun we find all *hypernyms* in the associated knowledge graph. Next, we add the most frequent n such hypernyms to the document-based taxonomy (the number in the third column in Table 1). We identified this step as critical, as the crawl-based knowledge graphs are commonly noisy, and prunning out *uncertain relations* is of high relevance. After performing this for all documents in the corpus, document-based taxonomies are concatenated into corpus-based taxonomy. Next, we perform feature selection, discussed next.

3.1 Feature selection

During feature selection we choose a predefined number of features within the set of features with the goal to *select* the most useful or important features. Hence, from the set of hypernyms which we constructed from the knowledge graph, we choose only top d features (= dimension of the space) based on one of the heuristics described below. **Closeness centrality** of a node is a measure of centrality in a network, calculated as $C(x) = \frac{1}{\sum_y d(y,x)}$, where $d(y,x)$ is the distance (path length) between vertices x and y. The bigger the closeness centrality value a given node has, the closer it is to all other nodes. The **rarest terms** are the most document-specific and are more likely to provide more information than the ones frequently occurring. Hence this heuristic simply takes overall counts of all the hypernyms, sorts them in ascending order by their frequency of occurrence and takes the top d. The **mutual information** between two random discrete variables represented as vectors X_i (the i-th hypernym feature) and Y (the target binary class) is defined as follows:

$$MI(X_i, Y) = \sum_{x,y \in \{0,1\}} p(X_i = x, Y = y) \log_2 \frac{p(X_i = x, Y = y)}{p(X_i = x)p(Y = y)}$$

where $p(X_i = x)$ and $p(Y = y)$ correspond to marginal distributions of the joint probability distribution of X_i and Y. Tax2vec computes the mutual information (MI) between all hypernym features and a given class. So for each target class a vector of mutual information scores is obtained, corresponding to MI between individual hypernym features and a given target class. Then the MI scores for each target class are summed up and the final vector is obtained. The features are sorted by MI scores in descending order and the first d features are chosen as the final semantic space. The **personalized PageRank** algorithm takes as an input a network and a set of starting nodes in the network and returns a vector assigning a score to each node. The scores are calculated as the stationary distribution of the positions of a random walker that starts its walk on one of the starting nodes and, in each step, either randomly jumps from a node to one of its neighbors (with probability p) or jumps back to one of the starting nodes (with probability 1-p). In our experiments probability p was set to 0.85. The tax2vec exploits the idea initially introduced in [14], where personalized PageRank scores are computed w.r.t. the terms, present throughout the document space. This way, a graph-based, completely unsupervised ranking is obtained, and is used in similar manner to other feature selection heuristics discussed in the previous paragraphs. In this section we introduce how the knowledge graph is used for semantic feature construction, how the text is being processed prior to that and how are semantic features used after that.

Table 1: Part of the Microsoft Concept Graph. The row is in form of hypernym - hyponym - frequency of relation

social network	facebook	4987
symptom	fever	4966
sport	tennis	4964
fruit	strawberry	4824
activity	fishing	4789

3.2 Microsoft Concept Graph

We are using Microsoft Concept Graph[4] [15] [16] for obtaining the extra semantic information. This large relational graph consists of more than 5.4 million concepts that are a part of more than **80 million triplets**. It was created by harnessing billions of web pages, so it is very general and various, offering a lot knowledge to add to our text we want to classify. It contains mostly IsA relations, which was the part we use to obtain hypernyms for nouns in the input text and enrich the feature vectors by some of them. A part of the downloaded knowledge graph is shown in Table 1. The number in the third column is the count of times this relation was found when creating the knowledge graph, so a frequency of the relation's occurrence. We removed relations that had frequency of one, which immediately reduced the graph approximately to half the size and removed mostly noisy relations. Later we used the NetworkX library [17] to transform the Microsoft Knowledge Graph from bare text to a directed graph. This step makes the subsequent exploitation of the knowledge graph easier.

3.3 Proposed approach extending tax2vec

Firstly, we tokenize each document and assign part-of-speech tags to the tokens with the help of the spaCy library [13]. Then for each noun in the text, we find its hypernyms in the knowledge graph. The number of hypernyms for each noun is a parameter chosen by the user, we choose those hypernyms based on the highest frequencies of relation between the current noun and the hypernyms. As shown later in the paper, bigger number of hypernyms does not help a lot, but increases execution time significantly, so it is more sensible to choose a smaller number. Then we create a document-based taxonomy, which is a directed graph where edges are created as hypernym-noun for each hypernym and each noun. We merge the document-based taxonomies into one corpus-based taxonomy (maintaining unique nodes, mergeGraph method in the pseudocode) and on it we perform one of the above mentioned heuristics to choose the best d hypernyms. Those steps are outlined in Algorithm 1.

4 EXPERIMENTS AND RESULTS

This section presents the setting of the experiments and the data sets on which the experiments were conducted. We also describe the metrics used to estimate classification performance.

4.1 Data sets

We conducted the experiments on eight different data sets, which are described below. They were chosen intentionally from different domains and the basic information about them can be seen in Table 2.

[4] https://concept.research.microsoft.com/

Data: corpus, knowledgeGraph, maxHypernyms
corpusTaxonomy = [ ];
foreach *doc* ∈ *corpus* **do**
  documentTaxonomy = [ ];
  tokens = tokenize(*doc*);
  foreach *token* ∈ *tokens* **do**
    if *token is noun* **then**
      edges = knowledgeGraph.edgesFrom(*token*);
      foreach *edge* ∈ *edges* **do**
        if *len*(*documentTaxonomy*) >= *maxHypernyms* **then**
          break;
        documentTaxonomy.add(*edge* ∈ *edges*)
  corpusTaxonomy.mergeGraph(documentTaxonomy)
featureSelection(corpusTaxonomy)
Result: Selected semantic features

Algorithm 1: Semantic feature construction.

Table 2: Data sets used for evaluation of knowledge graph's extra features impact on learning.

Data set	Classes	Words	Unique w.	Documents
PAN 2017 Gender	2	5169966	607474	3600
PAN 2017 Age	5	992742	185713	402
SMSSpam	2	86910	15691	5571
CNN-news	7	1685642	159463	2107
MedicalRelation	18	1136326	66235	22176
Articles	20	5524333	178443	19990
SemEval2019	2	295354	39319	13240
Yelp	5	1298353	88539	10000

PAN 2017 (Gender) Given a set of tweets per user, the task is to predict the user's gender [18].

PAN 2017 (Age) Given a set of tweets per user, the task is to predict the user's age group [19].

CNN News Given a news article (composed of a number of paragraphs), the task is to assign to it a topic from a list of topic categories. [20].

SMS Spam Given a SMS message, the task is to predict whether it is a spam or not. [21].

Medical Relations Given an article with biomedical topic, the task is to predict the relationship between the medical terms annotated. [22].

SemEval 2019 Given a tweet, the task is to predict whether it contains offensive content [23].

Articles Given an web article, the goal is to assign to it a topic. [24].

Yelp Given an review of a restaurant, the goal is to predict the ranking from one to five stars.

Settings. In all the datasets the stop words were removed. Stop words are for example "the", "is", "are" etc. There is no universal list of stop words in NLP research, however we used NLTK (Natural Language Toolkit) [25] for filtering stop words. The documents were tokenized with the help of spaCy's NLP tool. The data sets were divided into 90% training data and 10% test data by using random splits. Number of hypernyms for each noun was 10. We used linear SVM classifier for classification and F_1 measure for performance.

4.2 Results

Figure 1 shows that on some datasets (namely Yelp, PAN 2017 Age, PAN 2017 Gender and on SemEval 2019 and Articles) the extra semantic features constructed from the knowledge graph help in some cases. We compare those results to the classification without any semantic features which is plotted as a grey horizontal line. On the other hand, on the datasets CNN News, Medical Relation and SMS Spam we didn't see any improvement with the addition of semantic features. Figure 2 shows the relation between feature space size and the execution times.

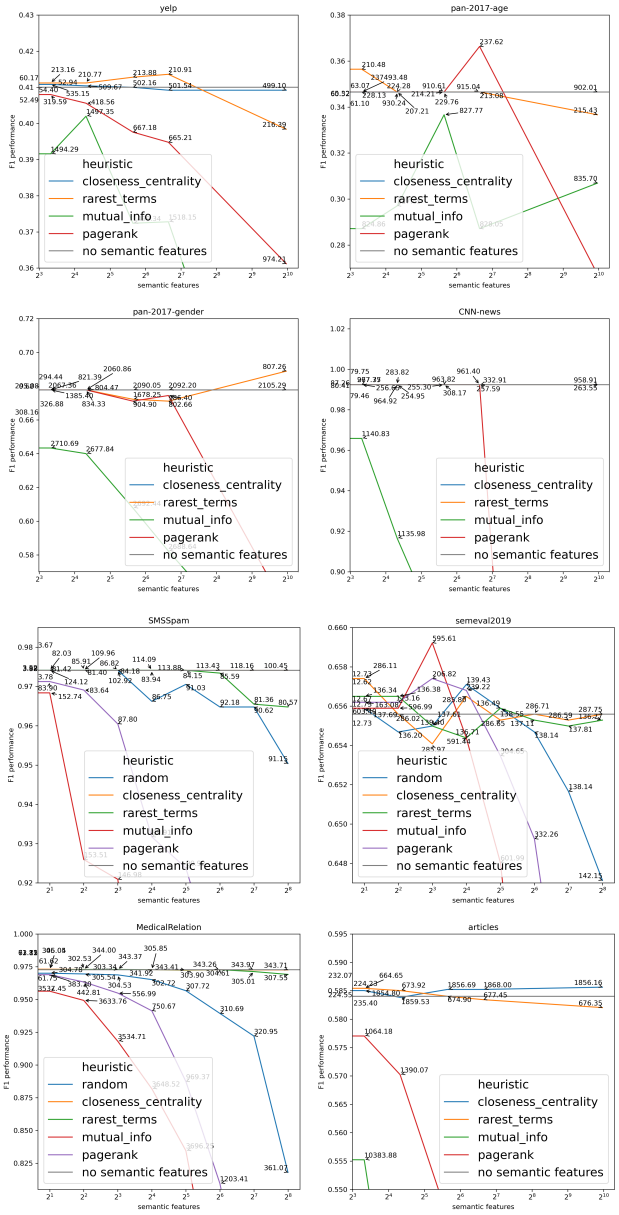

Figure 1: Results of text classification on data sets Yelp, pan-2017-age, pan-2017-gender, CNN News, SMSSpam, SemEval 2019, Medical Relation and Articles with execution times as the numbers in the plot.

5 CONCLUSION

We showed that information from a large, real-life knowledge graph can improve text classification. Our approach aims at short texts like tweets, shorter articles, messages and similar. We firstly process the document with spaCy, find nouns with their corresponding hypernyms, from which we create a taxonomy and from that we later choose the most helpful features with one

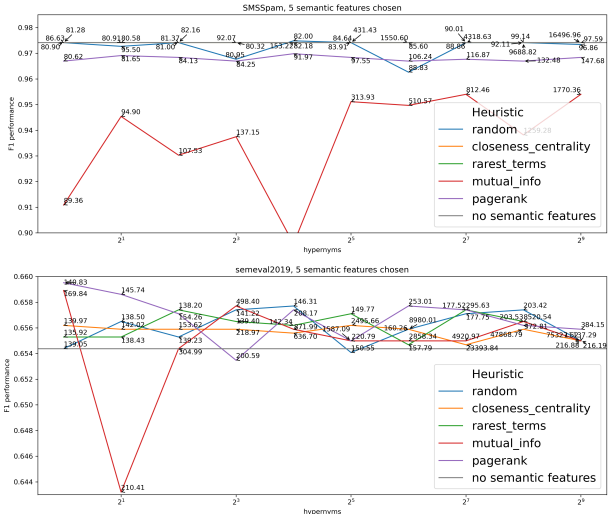

Figure 2: Results of text classification on data sets SMSSpam and SemEval 2019 with execution times as the numbers in the plot.

of the heuristics. The result remains *interpretable*, which is an advantage of this approach. This approach could be potentially improved by performing some type of word sense disambiguation and by finding objects in texts, which consists of more than one word. Further, other knowledge graphs can be used for the hypernym search. Also, because the hypernym search in each document is independent, the documents can be processed in parallel; however, such processing can be memory-intensive, which is to be addressed.

ACKNOWLEDGMENTS

The work of BŠ was financed via a junior research grant (ARRS). This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The authors acknowledge also the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103), the project TermFrame - Terminology and Knowledge Frames across Languages (No. J6-9372) and the ARRS ERC complementary grant SDM-Open.

REFERENCES

[1] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. E. Barnes, and D. E. Brown. 2019. Text classification algorithms: A survey. *CoRR*, abs/1904.08067.

[2] Q. Wang, Z. Mao, B. Wang, and L. Guo. 2017. Knowledge graph embedding: a survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*.

[3] 2020. Tax2vec: constructing interpretable features from taxonomies for short text classification. *Computer Speech & Language*.

[4] F. Rangel, P. Rosso, M. Potthast, and B. Stein. 2017. Overview of the 5th author profiling task at pan 2017: gender and language variety identification in twitter. *Working Notes Papers of the CLEF*.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation.

[6] T. K. Landauer. 2006. Latent semantic analysis.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. [n. d.] Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*.

[8] A. Celikyilmaz, D. Hakkani-Tür, P. Pasupat, and R. Sarikaya. 2015. Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems. In.

[9] M. K. Elhadad, K. M. Badran, and G. I. Salama. 2018. A novel approach for ontology-based feature vector generation for web text document classification.

[10] R. Kaur and M. Kumar. 2018. Domain Ontology Graph Approach Using Markov Clustering Algorithm for Text Classification. *Advances in Intelligent Systems and Computing*, 632.

[11] S. Scott and S. Matwin. 1998. Text classification using WordNet hypernyms. In *Usage of WordNet in Natural Language Processing Systems*.

[12] P. Wang and C. Domeniconi. 2008. Building semantic kernels for text classification using wikipedia. In (August 2008).

[13] M. Honnibal and I. Montani. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, (2017).

[14] J. Kralj, M. Robnik-Sikonja, and N. Lavrac. 2019. Netsdm: semantic data mining with network analysis. *Journal of Machine Learning Research*, 20, 32, 1–50.

[15] J. Cheng, Z. Wang, J.-R. Wen, J. Yan, and Z. Chen. 2015. Contextual text understanding in distributional semantic space. In *ACM International Conference on Information and Knowledge Management (CIKM)*.

[16] W. Wu, H. Li, H. Wang, and K. Q. Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *ACM International Conference on Management of Data (SIGMOD)*.

[17] A. A. Hagberg, D. A. Schult, and P. J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, 11 –15.

[18] F. Rangel, P. Rosso, M. Potthast, and B. Stein. [n. d.] Overview of the 5th author profiling task at pan 2017: gender and language variety identification in twitter.

[19] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations.

[20] M. Qian and C. Zhai. 2014. Unsupervised feature selection for multi-view clustering on text-image web news data, 1963–1966.

[21] T. A. Almeida and J. M. G. Hidalgo. 2011. Sms spam collection v. 1. http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/. (2011).

[22] 2015. Medical information extraction. https://appen.com/datasets/medical-sentence-summary-and-relation-extraction/. (2015).

[23] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

[24] 2019. Text classification 20. https://www.kaggle.com/guiyihan/text-classification-20. (2019).

[25] S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

EveOut: Reproducible Event Dataset for Studying and Analyzing the Complex Event-Outlet Relationship

Swati
swati@ijs.si
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

Tomaž Erjavec
tomaz.erjavec@ijs.si
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

ABSTRACT

We present a dataset consisting of 77,545 news events collected between January 2019 and May 2020. We selected the top five news outlets based on Alexa Global Rankings and retrieved all the events reported in English by these outlets using the *Event Registry API*. Our dataset can be used as a resource to analyze and learn the relationship between events and their selection by the outlets. It is primarily intended to be used by researchers studying bias in event selection. However, it may also be used to study the geographical, temporal, categorical and several other aspects of the events. We demonstrate the value of the resource in developing novel applications in the digital humanities with motivating use cases. Website with additional details is available at *http://cleopatra.ijs.si/EveOut/*.

KEYWORDS

Dataset, News Event Analysis, Event selection bias, News coverage

1 INTRODUCTION

News outlets are constantly faced with the task of selecting events they will report on, dependent on the perceived interest of the event to their readership. This can be driven by various factors, such as the geographical origin of the event, involvement of well-known persons, etc. Such selection requires monitoring of current affairs to determine their news value for the outlet.

Machine learning tools may help outlets to deal with the large numbers of events, help them explore strategies for selecting publishable events, and build dedicated decision support systems for this task. The effectiveness of these systems depends on the availability of news event collections complemented by relevant event details such as date, category, country of occurrence, brief description, etc.

In this paper we introduce EveOut, the first large publicly available data set of 77,545 English news events with a variety of features collected between January 2019 and May 2020. It includes events in eight different categories of news, i.e. business, politics, technology, environment, health, science, sports, and arts-and-entertainment. We hope that EveOut will encourage publishers and others involved in the news production process to develop tools to enhance digital journalism. The data set would also allow researchers from digital humanities to study and analyze the relationship and impact of different features on the selection of events by the outlets.

1.1 Contributions

The paper makes the following three contributions to science:

- The dataset generation scripts, which provide a structured reproducible approach to building a publicly available dataset of news events with varied features. This will not only speed up the development of future versions of EveOut, but will also help to create custom datasets with the desired outlets and features.
- The compilation of EveOut, a novel dataset with a rich range of event features and spanning multiple news categories.
- Identification of possible use cases intended to facilitate the creation of tools to improve digital journalism and to help researchers study the complex relationship between events and news outlets.

2 DATASET

Several news outlets may cover a single world event as a story in a variety of different ways. A collection of one or more stories, all of which describe the same world event, is referred to as an *'event'* in the entire paper. In the following subsections, we define our data generation process and provide statistics on the resulting dataset.

2.1 Data Source

We use **Event Registry**[1] [4] as the data source which monitors, collects, and provides news articles from news outlets around the world in over 30 languages. It also identifies the major incidents reported in the articles and aggregates them into clusters known as events. For example, *"missiles launched by Iran at US forces in Iraq"* is an event reported across the globe in over 3,200 news articles.

To construct an event, Event Registry follows a series of steps. News aggregation is the first step in which RSS feeds are constantly monitored for new articles. The next major step is the semantic event information extraction, which retrieves information from the articles in a structured way to be used in subsequent steps. Clustering algorithms are then used to group articles that describe the same event. In the last step, the article clusters are marked as events and are annotated with rich metadata such as a unique id to track the event coverage, categories to which it may belong, geographical location, sentiment, etc. As a result, its extensive temporal coverage can be used effectively to study the complex correlation between events and news outlets.

[1] https://eventregistry.org

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

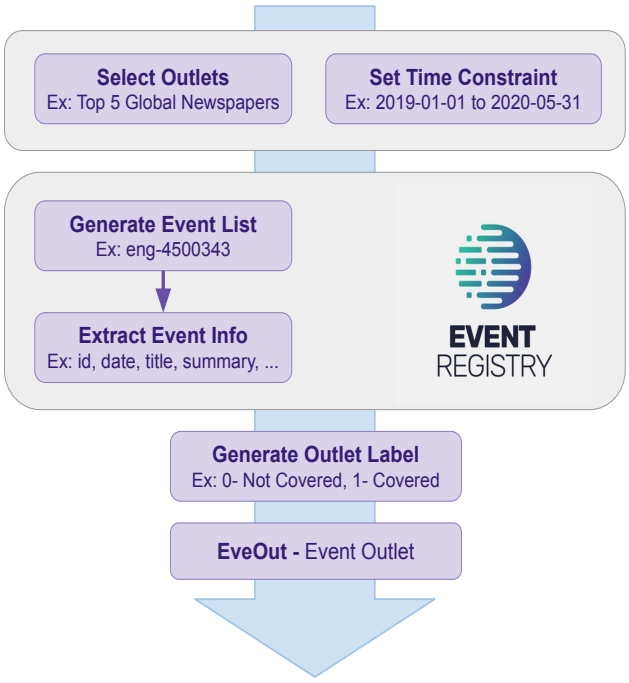

Figure 1: EveOut dataset generation process.

Table 1: Description of the dataset attributes.

Attribute	Description
uri	a unique event identifier
title	title of the event in English
event_date	date in yyyy-mm-dd format
sentiment	event sentiment
categories	event categories
loc_country	country where the event occurred
loc_continent	continent where the event occurred
total_article_count	total number of articles published
article_count	total number of articles published in English
summary	summary of the event
outlet_list	list of outlets that reported the event

2.2 Data Generation Process

To generate the dataset we adopted an automated approach which is depicted in Figure 1. We use Event Registry API to collect event related information mentioned in Table 1. The script is designed to simplify the release of future versions and to be able to replicate the process of generating custom datasets. The outlined process is the result of the resource's core requirement to best address the potential use-cases referred to in Section 4.

For data generation, we first selected the top five news outlets based on Alexa Global Rankings [2]. We then used an explicit temporal query (Q_t) to retrieve all events in all news categories from the Event Registry API. $Q_t = \{Q_{text}, Q_{time}\}$ consists of the text component Q_{text} and the time component Q_{time}. Next, we set the time limit $Q_{time} = [Q_{sd}, Q_{ed}]$ for extracting events that occurred within the specified time where, Q_{sd} = '2019-01-01' and Q_{ed} = '2020-05-31' signify the event's start date and end date. Since the outlet's event selection policy may change over time, we selected this time frame as recent data tends to be more reliable in predicting event coverage patterns. We then set $Q_{text} = \{Q_{out}, Q_{lang}, Q_{cat}\}$ where, Q_{out} = {'*nytimes*', '*indiatimes*', '*washingtonpost*', '*usatoday*','*chinadaily*'}, Q_{lang} = {'*eng*'}, and Q_{cat} = {'*politics*', '*business*', '*sports*', '*arts and entertainment*', '*science*', '*technology*', '*health*', '*environment*'} represent the outlets, languages and news categories respectively.

From the extracted event list, we first excluded events that were not covered by any of the selected outlets. We then extracted individual outlets from the event's outlet list and created a column in the dataset to represent each of them. We use a binary scalar value to indicate whether the outlets covered the event or not. The event coverage by the outlets is not uniform, which can be visualized in Figure 2.

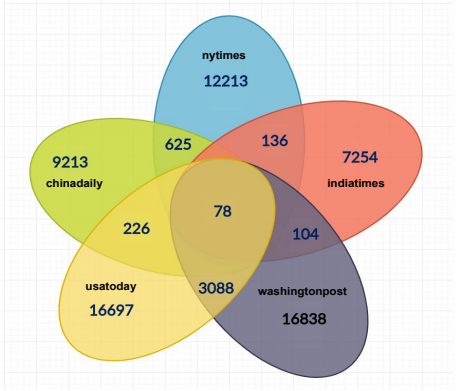

Figure 2: Distribution of event coverage by the outlets.

3 AVAILABILITY

The GitHub repository containing the scripts is available at ***https://github.com/Swati17293/EveOut***. To facilitate discoverability and preservation, the full data set is archived as an online resource at ***https://doi.org/10.5281/zenodo.3953878***. EveOut is available in three common formats (JSON, XML, and CSV) for direct download and use. The documentation meets the requirements of the *FAIR Data principles*[3] with all necessary metadata defined. Under the *Creative Commons Attribution 4.0 International license*, it is freely available to make it reusable for almost any purpose. A separate web page with detailed statistics and illustrations can be found at ***http://cleopatra.ijs.si/EveOut/*** for in-depth analysis.

3.1 Reusability

The resource is currently being used for individual projects and as a contribution to the project's deliverables of the Marie Skłodowska-Curie CLEOPATRA Innovative Training Network[4]. A major part of this project aims to provide a temporal, cross-lingual analysis of concepts around different events, exploring how language impacts the mediatic narratives built by the media. It also aims to analyse news reporting bias and multiple media

[2]https://www.alexa.com/topsites/category/Top/News/Newspapers

[3]http://www.nature.com/articles/sdata201618/

[4]http://cleopatra-project.eu/

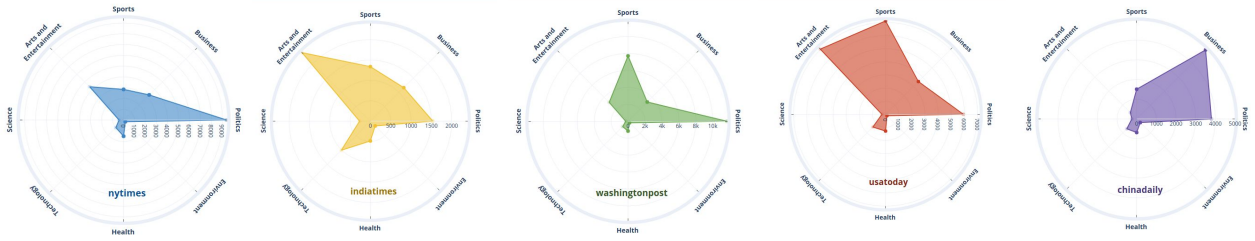

Figure 3: Overview of the category-wise event coverage by the outlets.

narratives which would enable to filter out appropriate information which then will be used to build information representation tools. Since EveOut serves as the basis for the study and analysis of events and their attributes, it is ideally suited to the project needs.

4 POTENTIAL USE CASES

4.1 Examine Event-Selection Bias

It is important for a journalist to know which event is worthy enough to be published. Even readers would be interested to know the factors that affect this selection. An automated solution can be devised using EveOut to provide an overview of the event and to visualize differences in coverage.

4.2 Outlet Prediction

EveOut is designed to predict the likelihood of an event being covered by the outlet. It would enable the publishers of the outlets to assess the significance of the event. In addition, it may also be used by independent editors who prefer to report on events covered by mainstream outlets.

5 STATISTICS AND ANALYSIS

In this section we provide further information about the data contained in EveOut, focusing explicitly on the distribution of events between the outlets.

With regard to the distribution of event categories covered by the outlets, as shown in Figure 3, *'politics'* is the most common category, while *'environment'* is the least common category. It is also worth noting that each outlet focuses on the different categories of events aside from *'politics'*. For instance, *'indiatimes'* focuses more on events related to *'arts and entertainment'*, whereas *'chinadaily'* tends to cover more *'business'* related events.

As far as the coverage of the event over time is concerned, it is also inconsistent as depicted in Figure 6. Furthermore, the event-coverage of *'usatoday'* and *'washingtonpost'* is slightly inconsistent. It is also interesting to note the sharp decline in coverage by *'usatoday'* in *'Aug* 2019' and by *'washingtonpost'* in *'May* 2020'.

The drop in the graph for washingtonpost in *'May* 2020 is due to its event preference. It is evident from washingtonpost's radial graph in Figure 3 that its coverage is biased towards politics and sports. These two categories alone represent around 50% of events in the dataset. However, this percentage dropped to 40% in *'May* 2020 and, as a result, the coverage of washingtonpost dropped significantly. Increase of event coverage in *'Mar* 2019 is also attributed to the fact that about 56% of events were from these two categories. In nutshell, if the outlet favors a certain category of events and, in a specific time frame, and events of that category are high/low than usual, it will be reflected in the outlet's coverage pattern.

Figure 4 reveals that instead of favoring events with neutral sentiment, outlets tend to favor events with positive sentiment. In addition, event coverage by *'usatoday'* and *'washingtonpost'* is quite diverse with respect to sentiments.

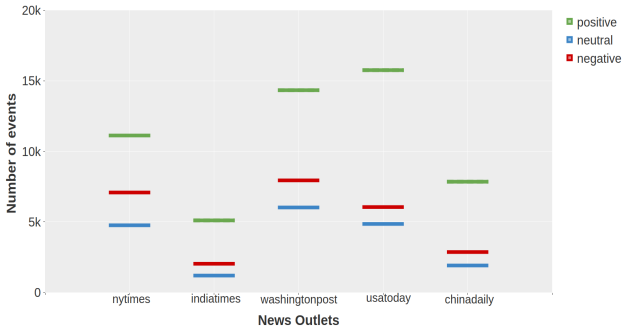

Figure 4: Distribution of event coverage by the outlets with respect to sentiments.

In terms of the sentiments used in each category as plotted in Figure 5, it is worth noting that *'technology'* and *'sports'* events are mostly positive.

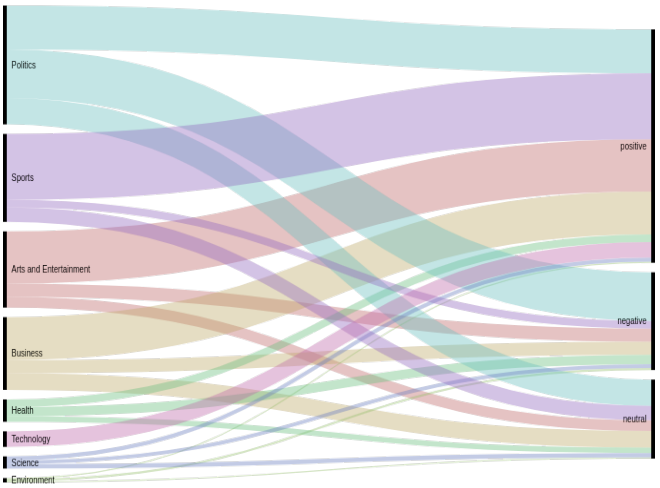

Figure 5: Distribution of category over sentiments.

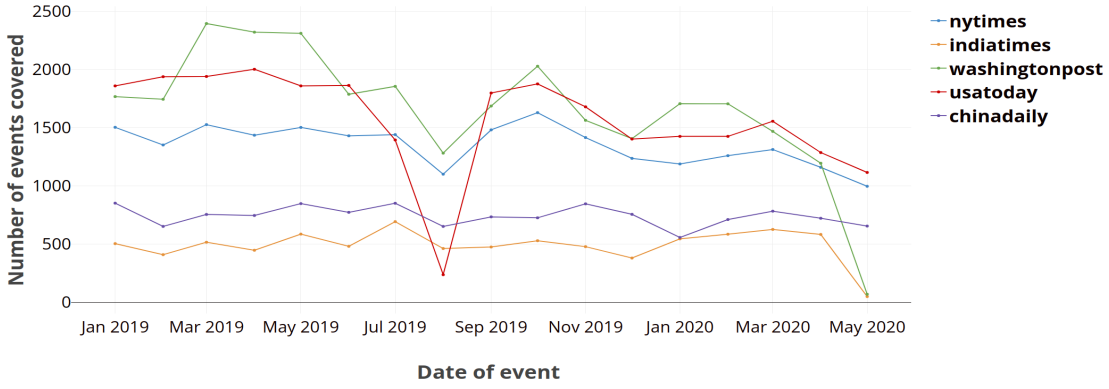

Figure 6: Distribution of the event coverage by the outlets over time.

6 RELATED WORK

There are a number of datasets that focus on news articles [7]. As far as the availability of event-centric datasets is concerned, there is a scarcity of publicly available datasets. There are few related research on the event data [3, 1], but the extracted/generated datasets for the experiments is also not publicly accessible.

GDELT [5] is the most popular, very large and publicly available event-oriented news dataset. It contains data in multiple languages from a wide range of online publications. It's collection of world events is centered on location, network and temporal attributes. There is no attribute defining the outlet list for the event in the dataset. As a result, there is a lack of knowledge essential to the analysis of the event-outlet relationship that is the foundation of our dataset.

In addition, the existing event datasets [6, 2] are category-dependent (*politics/healthcare/disaster etc.*) which renders them useful for specific research purposes only. Therefore, by providing a generalized event-centric news dataset, EveOut addresses the stated dataset bottleneck.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced the EveOut dataset, which covers events reported by the top five global news outlets for over 17 months. We have ensured that the dataset complies with the *FAIR principles*. In conjunction with the data set, we provide the source code for reproducing the dataset with varied features. For instance, it is possible to generate a reduced version of EveOut, focused on just one category, say *'politics'*. Specific outlets, dates, and languages can also be specified in accordance with the requirements. We illustrate potential use cases to show how the dataset could be used to study the pattern of event coverage of an individual outlet and to predict whether or not the outlet will cover a specific event. Researchers from digital humanities can also use it for an in-depth analysis of complex event-outlet relationships. In the future , we intend to extend the dataset to include events described in different languages.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. 2018. Selection bias in news coverage: learning it, fighting it. In *Companion Proceedings of the The Web Conference 2018*, 535–543.

[2] Cindy Cheng, Joan Barceló, Allison Spencer Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. Covid-19 government response event dataset (coronanet v. 1.0). *Nature Human Behaviour*, 1–13.

[3] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2018. Bias-aware news analysis using matrix-based news aggregation. *International Journal on Digital Libraries*, 1–19.

[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] Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: global data on events, location, and tone, 1979–2012. In *ISA annual convention*. Volume 2, 1–49.

[6] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47, 651–660.

[7] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, X. Xie, Jianfeng Gao, Winnie Wu, and M. Zhou. 2020. Mind: a large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3597–3606. DOI: 10.18653/v1/2020.acl-main.331. https://www.aclweb.org/anthology/2020.acl-main.331.

Ontology alignment using Named-Entity Recognition methods in the domain of food

Gorjan Popovski[1,2*], **Tome Eftimov**[1], **Dunja Mladenić**[1,2] and **Barbara Koroušić Seljak**[1,2]

[1]Jožef Stefan Institute, 1000 Ljubljana, Slovenia

[2]Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia

{gorjan.popovski, tome.eftimov, dunja.mladenic, barbara.korousic}@ijs.si

Abstract

In recent years, a great amount of research has been done in predictive modeling in the domain of healthcare. Such research is facilitated by the existence of various biomedical vocabularies and standards which play a crucial role in understanding healthcare information. In addition, the Unified Medical Language System (UMLS) links together biomedical vocabularies to enable interoperability. However, in the food domain such resources are scarce. To address this issue, this paper explores a methodology for ontology alignment in the domain of food by leveraging Named-Entity-Recognition (NER) methods based on different semantic resources. It is based on a recently published rule-based NER method named FoodIE, whose semantic annotations are based on the Hansard corpus, as well as a NER tool called Wikifier, from which DBpedia URIs are extracted. To perform the alignment we use the FoodBase corpus, which consists of recipes annotated with food entities and includes a ground truth version which is additionally used for evaluation.

1 Introduction

Information Extraction (IE) is the task of automatically extracting information from unstructured data and, in most cases, is concerned with the processing of human language text by means of natural language processing (NLP) [Aggarwal and Zhai, 2012]. The main idea behind IE is to provide a structure to the information extracted from the unstructured data.

One of the core IE tasks is named-entity recognition (NER), which addresses the problem of identification and classification of predefined concepts [Nadeau and Sekine, 2007]. It aims to determine and identify words or phrases in text into predefined labels (classes) that describe concepts of interest in a given domain. Various NER methods exist: *terminology-driven*, *rule-based*, *corpus-based*, *methods based on active learning (AL)*, and *methods based on deep neural networks (DNNs)*.

Terminology-driven NER methods, also called dictionary-based NER methods [Zhou *et al.*, 2006], match text phrases against concept synonyms that exist in the terminological resources (dictionaries). The main disadvantage of these methods is that only the entity mentions that exist in the resources will be recognized, but the benefit of using them is related to the frequent updates of the terminological resources with new concepts and synonyms.

Rule-based NER methods [Hanisch *et al.*, 2005] use regular expressions that combine information from terminological resources and characteristics of the entities of interest. The main disadvantage of these methods is the manual construction of the rules, which is a time-consuming task and depends on the domain.

Corpus-based NER methods [Alnazzawi *et al.*, 2015; Leaman *et al.*, 2015] are based on an annotated corpus provided by subject-matter experts as well as the use of ML techniques to predict the entities' labels. These methods are less affected by terminological resources and manually created rules. However, their limitation is their dependence on an existence of an annotated corpus for the domain of interest. The construction of the annotated corpus for a new domain is a time consuming task and requires effort by the subject-matter experts to produce it.

To exploit unlabelled data in constructing NER methods, *AL* can be used [Settles, 2010; Tran *et al.*, 2017]. This represents semi-supervised learning in which an algorithm is able to interactively query the user to obtain the desired labels/outputs at new data points. Which examples are sent to the user for labelling is chosen by the algorithm and their number is often much lower than the number of examples required for supervised learning. It usually consists of three components: (1) the annotation interface, (2) the corpus-based NER, and (3) component for querying samples.

2 Related work

2.1 Hansard corpus

The Hansard corpus is a collection of text and concepts created as a part of the SAMUELS project [Alexander and Anderson, 2012; Rayson *et al.*, 2004]. It contains 37 higher level semantic groups, one of which is our topic of interest — *Food and Drink*.

*Contact Author

2.2 FoodIE

FoodIE is a rule-based food Named-Entity Recognition method [Popovski *et al.*, 2019a]. As it is rule-based, it consists of a rule-engine in which the rules are based on computational linguistics and semantic information that describe the food entities.

2.3 Wikifier

Wikifier is a tool that uses an efficient approach for annotating documents with relevant concepts from Wikipedia [Brank *et al.*, 2017]. It is based on a pagerank method to identify a set of relevant concepts. As it provides the location in the document where the annotation occurs, it is effectively a Named-Entity Recognition method. It provides Wikipedia concepts as annotations, additionally assigning DBpedia concepts if they exist.

3 Data

A recent publication provides one of the first annotated corpora, named FoodBase [Popovski *et al.*, 2019b], containing food entities. It consists of two version, a ground truth set referred to as "curated" (containing 1,000 annotated recipes), as well an "un-curated" version, consisting of around 22,000 recipes. The recipe categories that are included are: *Appetizers and snacks*, *Breakfast and Lunch*, *Dessert*, *Dinner*, and *Drinks*. In this paper, we use the *curated* version to perform the ontology alignment as well as evaluate the methodology. This version was manually checked by subject-matter experts, so the false positive food entities were removed, while the false negative entities were manually added in the corpus. An example of a recipe can be found on Figure 1.

4 Ontology alignment

Using FoodIE and the Wikifier tool, we obtain annotations for all 1,000 recipes from the FoodBase.

FoodIE extracts and annotates each recipe with semantic tags from the Hansard corpus. Each annotation contains the location of the extracted entity, i.e. where in the raw text the surface form representing the concept occurs, and its corresponding semantic tags from the Hansard corpus.

The Wikifier tool is used to annotate the recipes with DBpedia URIs. As these are general DBpedia concepts, additional information to filter out food concepts from non-food concepts is required. Webscraping the pages for the URIs provides useful information that can be used to distinguish food from non-food concepts, such as the broader concept/class to which the concept of interest belongs. The post-processing of the DBpedia URIs checks the entity type of the concept and checks if it is one of: "FOOD", "FOODS", "DISH", "INGREDIENT", "FOOD AND DRINK", "BEVERAGE", "PLANT", "ANIMAL", or "FUNGUS". If it does not belong to one of the above entity types, the page is checked for mentions of other URIs which are semantically related to food: "FOOD", "PLANT", "ANIMAL", or "FUNGUS". These URI mentions can occur anywhere in the page and if one of these matches is satisfied, the entity is assumed to be a food entity.

A post-processed example of such an annotation can be found on Figure 2.

Having annotated the recipes with both methods, we can perform the ontology alignment by using the location information for each annotation in each recipe. Each unique concept from both methods (semantic resources) is assigned its unique ID, and then a table is constructed for each concept mapping containing the IDs.

5 Evaluation and experimental setup

5.1 Match types

- True Positives (TP) — these are matches where the whole food concept is correctly annotated;
- False Positives (FP) — these are matches where a non-food concept is annotated as a food concept;
- False Negatives (FN) — these are matches where a food entity is not properly annotated;
- Partial match — these are matches where only some tokens from a food concepts are properly annotated.

5.2 Evaluation metrics

Using the concept of True Positives, False Positives and False Negatives, we compute the widely used evaluation metrics: Precision (P), Recall (R) and F1 Score (F1). They are defined as:

- $P = \frac{TP}{TP+FP}$
- $R = \frac{TP}{TP+FN}$
- $F1 = 2\frac{P \cdot R}{P+R}$

6 Results and discussion

After running the evaluation, we obtain the following results. The matches for both methods are presented in Table 1, while the evaluation metrics are presented in Table 2.

Table 1: Match types.

	FoodIE	Wikifier
TPs	11461	6380
FNs	684	4121
FPs	258	5861
Partial	359	3297

Table 2: Evaluation metrics.

	FoodIE	Wikifier
F_1 Score	0.9605	0.5611
Precision	0.9780	0.5212
Recall	0.9437	0.6076

From the results in the tables it is evident that FoodIE provides more promising results. However, this was expected as this NER method was specifically constructed to only cater to the domain of food. Of especial interest are the matches of type *partial*, since they represent a match where only a subset of the tokens in a food entity are correctly recognized. For example, looking at Figure 1, the first extracted food entity

```
<document>
    <id>0recipe1090</id>
    <infon key="category">Appetizers and snacks</infon>
    <infon key="full_text">
    Mix the dry ranch salad dressing mix, mayonnaise, and milk in a bowl.
    Beat in the cream cheese with an electric mixer until smooth. Mix in Cheddar cheese.
    Cover bowl with plastic wrap, and freeze 30 minutes. Divide mixture in half, and shape into balls.
    Roll each ball in almonds to coat. Cover and refrigerate balls until ready to serve.
    </infon>
    <annotation id="1">
        <location offset="3" length="28"/>
        <text>dry ranch salad dressing mix</text>
        <infon key="semantic_tags"> AG.01.h.02 [Vegetables];AG.01.m [Substances for food preparation];
            AG.01.n.09 [Prepared vegetables and dishes];</infon>
    </annotation>
    <annotation id="2">
        <location offset="9" length="10"/>
        <text>mayonnaise</text>
        <infon key="semantic_tags"> AG.01.l.04 [Sauce/dressing];
            AG.01.n.01 [Food by way of preparation];</infon>
    </annotation>
    <annotation id="3">
        <location offset="12" length="4"/>
        <text>milk</text>
        <infon key="semantic_tags"> AG.01.e [Dairy produce];</infon>
    </annotation>
    <annotation id="4">
        <location offset="20" length="12"/>
        <text>cream cheese</text>
        <infon key="semantic_tags"> AG.01.e [Dairy produce];AG.01.e.02 [Cheese];
            AG.01.n [Dishes and prepared food];AG.01.n.18 [Preserve];</infon>
    </annotation>
    <annotation id="5">
        <location offset="31" length="14"/>
        <text>Cheddar cheese</text>
        <infon key="semantic_tags"> AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];</infon>
    </annotation>
    <annotation id="6">
        <location offset="59" length="7"/>
        <text>almonds</text>
        <infon key="semantic_tags"> AG.01.h.01.f [Nut];</infon>
    </annotation>
</document>
```

Figure 1: Example recipe from the “curated” part of FoodBase.

	urls	text	from	to	matchType
0	http://dbpedia.org/resource/Salad	salad	19	23	PREF
1	http://dbpedia.org/resource/Mayonnaise	mayonnaise	39	48	PREF
2	http://dbpedia.org/resource/Milk	milk	55	58	PREF
3	http://dbpedia.org/resource/Bowl	bowl	65	68	PREF
4	http://dbpedia.org/resource/Cream	cream	83	87	PREF
5	http://dbpedia.org/resource/Cream_cheese	cream cheese	83	94	PREF
6	http://dbpedia.org/resource/Cheese	cheese	89	94	PREF
7	http://dbpedia.org/resource/Cheddar_cheese	Cheddar	140	146	PREF
8	http://dbpedia.org/resource/Plastic_wrap	plastic wrap	172	183	PREF
9	http://dbpedia.org/resource/Mixture	mixture	216	222	PREF
10	http://dbpedia.org/resource/Virus	shape	237	241	PREF
11	http://dbpedia.org/resource/Almond	almonds	273	279	PREF
12	http://dbpedia.org/resource/Refrigeration	refrigerate	300	310	PREF

Figure 2: Wikifier annotation example on a single recipe

should be "dry ranch salad dressing", which is correctly extracted by FoodIE. Looking at Figure 2, the same food entity is only extracted as "salad". Such match types do not factor in the calculation of the evaluation metrics, as it is debatable whether to count them as TPs or FNs. Nevertheless, they are interesting to compare, since even partial matches convey at least some semantic meaning regarding the food entity. Moreover, FP annotations on the same figure are "bowl" and "shape" which are not food entities. Additionally, a recent comparison of existing food NER methods can be found in [Popovski *et al.*, 2020], where the authors compare the performance of FoodIE with NER methods using other food ontologies available in the BioPortal.

Regarding the mapping of the concepts, a total of 348 explicit concept mappings were discovered by the methodology. An example mapping for the concept "garlic" would be:

- A000016: 'garlic', AG.01.h.02.e [Onion/leek/garlic].
- E000029: 'garlic', http://dbpedia.org/resource/Garlic

7 Conclusion and future work

In this work we propose a methodology for ontology alignment by using Named-Entity Recognition methods in the domain of food. It utilizes the newly proposed FoodIE NER method and the Wikifier text annotation tool. Our experimental results show that FoodIE provides more promising results than Wikifier, achieving an $F1$ score of 0.9605, compared to 0.5611. This is expected since FoodIE is specifically designed for the food domain, while Wikifier uses general vocabulary and annotates text with Wikipedia concepts.

For future work, recursive webscraping can be performed to more accurately distinguish between food and non-food annotated concepts from the Wikifier tool. Specifically, this would mean repeating the steps to check if the entity is a food entity or not on the parent nodes in DBpedia. Additionally, more food semantic resources can be included to provide mapping between multiple ontologies. Doing this is dependent on the existence of a NER method that works with concepts from the desired food semantic resource.

Acknowledgements

This research was supported by the Slovenian Research Agency (research core grant number P2-0098), and the European Union's Horizon 2020 research and innovation programme (FNS-Cloud, Food Nutrition Security) (grant agreement 863059). The information and the views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use that may be made of the information contained herein.

References

[Aggarwal and Zhai, 2012] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[Alexander and Anderson, 2012] Marc Alexander and J Anderson. The hansard corpus, 1803-2003. 2012.

[Alnazzawi *et al.*, 2015] Noha Alnazzawi, Paul Thompson, Riza Batista-Navarro, and Sophia Ananiadou. Using text mining techniques to extract phenotypic information from the phenochf corpus. *BMC medical informatics and decision making*, 15(2):1, 2015.

[Brank *et al.*, 2017] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD*, 2017.

[Hanisch *et al.*, 2005] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):S14, 2005.

[Leaman *et al.*, 2015] Robert Leaman, Chih-Hsuan Wei, Cherry Zou, and Zhiyong Lu. Mining patents with tmchem, gnormplus and an ensemble of open systems. In *Proce. The fifth BioCreative challenge evaluation workshop*, pages 140–146, 2015.

[Nadeau and Sekine, 2007] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[Popovski *et al.*, 2019a] Gorjan Popovski, Stefan Kochev, Barbara Koroušić Seljak, and Tome Eftimov. Foodie: A rule-based named-entity recognition method for food information extraction. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, (ICPRAM 2019)*, pages 915–922, 2019.

[Popovski *et al.*, 2019b] Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. FoodBase corpus: a new resource of annotated food entities. *Database*, 2019, 11 2019. baz121.

[Popovski *et al.*, 2020] G. Popovski, B. K. Seljak, and T. Eftimov. A survey of named-entity recognition methods for food information extraction. *IEEE Access*, 8:31586–31594, 2020.

[Rayson *et al.*, 2004] Paul Rayson, Dawn Archer, Scott Piao, and AM McEnery. The ucrel semantic analysis system. 2004.

[Settles, 2010] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[Tran *et al.*, 2017] Van Cuong Tran, Ngoc Thanh Nguyen, Hamido Fujita, Dinh Tuyen Hoang, and Dosam Hwang. A combination of active learning and self-learning for named entity recognition on twitter using conditional random fields. *Knowledge-Based Systems*, 132:179–187, 2017.

[Zhou *et al.*, 2006] Xiaohua Zhou, Xiaodan Zhang, and Xiaohua Hu. Maxmatcher: Biological concept extraction using approximate dictionary lookup. In *Pacific Rim International Conference on Artificial Intelligence*, pages 1145–1149. Springer, 2006.

Extracting structured metadata from multilingual textual descriptions in the domain of silk heritage

M.Besher Massri
Jožef Stefan Institute, Slovenia
besher.massri@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

In this paper, we present a methodology for extracting structured metadata from museum artifacts in the field of silk heritage. The main challenge was to train on a relatively small and noisy data corpus with highly imbalanced class distribution by utilizing a variety of machine learning techniques. We have evaluated the proposed approach on real-world data from five museums, two English, two Spanish, and one French. The experimental results show that in our setting using traditional machine learning algorithms such as Support Vector Machines gives comparable and in some cases better results than multilingual deep learning algorithms. The study presents an effective approach for categorization of text described artifacts in a niche domain with scarce data resources.

KEYWORDS

Information extraction, Text classification, Silk heritage, Transformers, Support Vector Machines.

1 INTRODUCTION

When looking to improve the understanding of silk heritage we find that the data available in the museums often lack semantic information on the artifacts or have them to some extent included in textual descriptions. To facilitate automatic analysis of silk heritage data and support digital modeling of the weaving techniques, we propose multilingual metadata extraction from textual descriptions provided by the museums.

We propose the usage of machine learning techniques to model the target variables, referred here as slots to align with the terminology of information extraction. Using machine learning methods we build a model for each of the target variables in order to annotate the text. This enabled us to add metadata to the silk heritage artifacts of the museums. The domain experts collaborating on Silknow project [9] have identified four kinds of metadata information that would be useful and are contained in texts of at least some of the targeted museums. We treat these as four slots for information extraction, where the list of possible slot values for each of the four was defined by the domain experts. Based on that we formed a multi-class dataset for each slot.

The corpora of text included were in three different languages (English, Spanish, and French) from five different museums, with a total of 500 museum records used in the study. After the data processing and annotation, we generated 24 binary datasets and 19 multi-class datasets (four for English, two for Spanish, and one for French). Using machine learning techniques we trained classifiers on the labeled data examples to predict the labels (slot values) based on the textual descriptions. Despite relatively small and unbalanced data corpora, using sampling techniques and weighted loss function helped mitigate the issue. In an experimental evaluation, we observed that on our data using traditional methods might be as good as using deep learning models when the data is scarce. However, using deep learning allows for building multilingual models that scale across different languages.

The main contribution of this paper is in proposing an approach to adding metadata to historical artifacts based on applying machine learning on multilingual textual descriptions of the artifacts. Moreover, we have defined the learning problem in collaboration with domain experts and performed evaluations on real-world data in English, Spanish, and French. The rest of this paper is structured as follows. Section 2 provides a description of the data, Section 3 describes the proposed methodology, Section 4 gives the results of the evaluation and Section 5 concludes the paper summarizing the approach and the findings.

2 DESCRIPTION OF DATA

We used the SilkNow knowledge graph [8] as our source of data. The source consists of records of different museums in different languages as shown in Table 1. The largest are MET with8364 artifacts in English, VAM with 7231 artifacts in English, and Imatex with 6799 artifacts in Spanish. We have used a subset of the data that contain artifacts with provided metadata and textual descriptions in related fields that were pointed out as relevant by the domain experts. Each record consists of the basic information about the object, such as the title and the museum it belongs to, along with two other sets of attributes, textual attributes, and categorical attributes. Textual attributes hold a textual description of the object in several fields, such as physical description and a technical description. The categorical description holds metadata information, such as technique or materials used. However, the data quality varies across the museums and records. Some museums are rich in both textual and categorical attributes, like the VAM museum, and others have short/low-quality textual attributes like Imatex. Also, some records have a text description in their categorical attributes instead of a single category value.

The metadata fields that we have considered are weaving technique, weave, motifs, and style. The list of labels or slot values for each of the metadata field (i.e. slot for information extraction) were compiled by the domain experts. These values describe the silk artifacts' nature and structure. Each of those slot values is represented by a term and a list of alternatives, up to four alternatives per term. Examples of slot values are satin, twill, and tabby, representing possible values of the weave slot.

Information society '20, October 5–9, 2020, Ljubljana, Slovenia

Museum	Language	Count
CER	Spanish	1296
Garin	Spanish	3101
Imatex	Spanish	6799
Joconde	French	376
MAD	French	763
MET	English	8364
MFA	English	3297
MTMAD	French	663
RISD	English	3338
VAM	English	7231

Table 1: Museums from the Silknow knowledge graph showing the language of the artifacts and the number of artifacts included in the knowledge graph.

3 METHODOLOGY

3.1 Annotating datasets with slot values

Based on the data and target variables, two types of datasets were formed for two types of text classification tasks. The first type is binary classification dataset, in which the target class is one of the slot values. The other is multi-class classification dataset, in which a dataset is formed for each of the four slots in each museum, where the target classes are the slot values that fall under the selected slot in addition to extra "other" class indicating that the example doesn't fall under any of them.

For forming the binary classification dataset we used a simple string matching approach. For each target class in each museum, examples were formed out of textual attributes of the museum records that contain a mention of either one of the possible value terms or its alternatives. Categorical attributes of the same record were used to determine the label of the example. The task is to classify whether the example has the slot value against the other slot values of the same slot. Each item is classified as True if the categorical attributes contain only the target value or one of its alternatives but not any of the other slot values' terms or their alternatives. If there is no mention of the slot value term or alternatives, then it's classified as false. If it contains this slot value' term along with other slot values' terms then it's considered as indeterminate and the example is removed.

To form the multi-class datasets, we merged the datasets of the same museum with target classes representing slot values that fall under the same slot. The true items of each slot value dataset formed the set of the examples with that slot value as the labels. The items that are false in each slot value dataset formed the "Other" class in the multi-class dataset.

3.2 Binary Classification Tasks

For binary classification, we used TFIDF word-vector representation for generating the feature vectors and trained a Linear Support Vector Machines (SVM) as the classifier using scikit-learn library [5]. All dataset were split into train and test using 80-20 stratified split. We performed a grid search with 5-fold cross validation on the training part using the following options:

- stemming, lemmatisation, or none
- max document frequency: [0.95.1.0]
- min document frequency: [0,0.05]
- SVM tolerance: [1e-4,1e-5]

The features were generated from sequences of words, referred to as n-grams, of length 1, 2, and 3. The remaining parameters were left unchanged from their default values. We used nltk [1] library for tokenization, SpaCy [4] for lemmatization, and Snow Ball Stemmer [6] for stemming.

Due to the methodology of data labeling, we sometimes ended up with a highly imbalanced datasets having a lot more negatives than positives. Therefore, in the binary dataset, we took a random subset from the negative examples to match the positive count. In addition, some examples were generated from the same records, by having more than one textual record with mentions of the same class's term/alternatives, therefore, corrections have been applied to the dataset by putting all examples of the same record in either train or test but not in both. This process was done to ensure no leakage occurs by potentially having highly similar textual text in train and text.

3.3 Multi-class Classification Tasks

For multi-class classification, we used a deep learning approach. The architecture consists of a pre-trained transformer, an LSTM layer, a dropout layer, a dense (linear) layer, and finally a soft-max activation layer. For the transformer we used BERT [3], multi-lingual BERT, and XLM-ROBERTA [2]. The loss function used was a cross-entropy loss with Adam as the optimizer. We used PyTorch framework [7] and hugging-face transformers library [10].

Considering that some of the datasets have a large class imbalance, which can be a couple of thousand examples of the majority class and only a few examples of the minority classes, we experimented with several class-weighting schemas. First, we tried assigning weights to the classes in the loss function is inversely proportional to the number of examples of each class. In addition, when we used weighted sampling with return for loading the examples into batches. This had the effect of over-sampling the minority classes and under-sampling the majority classes to achieve as balanced batch representation as possible. Finally, we tried a derivable version of F1 Macro as a loss function where the prediction matrix is taken as a probability rather than a binary value.

4 RESULTS

4.1 Experimental Datasets

The dataset collection methodology was applied to 10 museums and 4 categories holding more than 150 class values overall. However, most of the datasets have no positive items. In this research, we have selected datasets with at least 10 positive examples for binary classification tasks and at least 10 non-other in multi-class tasks. This final list consists of 24 binary datasets and 19 multi-class datasets. These datasets are used for training machine learning classifiers.

4.2 Binary Classification Tasks

For binary Classification, we applied the described methodology on all the datasets with at least 10 positive examples. The results of binary classification are consolidated in Table 2.

The graph in figure 1 displaying the correlation between the number of examples and the F1 score reveals a weak correlation of 0.19. We can see that when having more than 600 examples, we achieve F1 over 0.8. Upon closer inspection on the museum level, we found that the best results are achieved in the MFA museum on motifs and weaving technique and Joconde museums on weave.

Museum	Slot value	Slot	Language	#Exs	Accuracy	Precision	Recall	F1
cer	bordado	weaving technique	Spanish	278	0.89	0.87	0.93	0.9
cer	motivo vegetal	motifs	Spanish	146	0.57	0.56	0.6	0.58
cer	tafetÃ¡n	weave	Spanish	581	0.77	0.9	0.6	0.72
cer	terciopelo	weaving technique	Spanish	118	0.67	0.67	0.67	0.67
garin	brocatel	weaving technique	Spanish	932	0.88	0.85	0.92	0.89
garin	damasco	weaving technique	Spanish	1748	0.9	0.92	0.87	0.89
garin	espolÃn	weaving technique	Spanish	972	0.88	0.89	0.88	0.88
joconde	Satin	weave	French	159	0.91	0.9	0.95	**0.93**
joconde	Taffetas	weave	French	110	0.95	0.92	1	**0.96**
mfa	Lace	motifs	English	190	0.92	0.9	0.95	**0.92**
mfa	plain	weaving technique	English	130	1.00	1.00	1.00	**1.00**
vam	brocade	weaving technique	English	634	0.87	0.87	0.87	0.87
vam	damask	weaving technique	English	480	0.84	0.85	0.83	0.84
vam	Ear	motifs	English	262	0.83	0.84	0.81	0.82
vam	Edge	motifs	English	178	0.81	0.87	0.72	0.79
vam	embroidery	weaving technique	English	1614	0.85	0.86	0.83	0.84

Table 2: Results for the binary classification task.

Overall the best results are achieved by MFA and Joconde with an average F1 of .96 and .95 respectively followed by Garin, VAM, and CER with the average F1 of .89, .81, and .72 respectively.

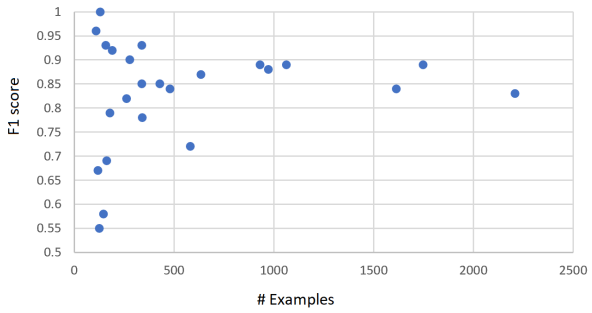

Figure 1: F1 score vs #Examples showing good performance on the largest datasets, when the number of examples is at least 600.

4.3 Multi Class Classification Class

4.3.1 Use Case: Detecting Weave Slot from VAM museum. We selected the VAM Weave slot as a use case dataset to perform hyperparameter tuning and select the best configurations for weighting. The dataset contains 2760 items with a baseline of 52.9% distributed across 4 classes: Satin, Tabby, Twill, and Other. The dataset was split into train, test, and validation in the form of 60-20-20 split. The results in Table 3 show that using class weighting in both loss function and sampling provides the best results w.r.t both classification accuracy and F1. Using F1 as a loss function sometimes provided good results but was discarded as it was not stable across different datasets. In addition, decreasing the learning rate improved results and stabilized the training curve. Finally, using the XLM-ROBERTA transformer showed an improvement in accuracy. The number of epochs was determined based on the accuracy performance of the validation dataset. The training would stop when the accuracy did not improve for the last 15 epochs. The accuracy (F1 micro) was chosen over F1 macro because of the large fluctuation in F1 macro value across training epochs caused by having minority classes with few examples.

Model configuration	Accuracy	F1
Base model	84.6	43.1
Weighted loss	82.1	47.2
Weighted sampling	82.6	52.2
F1 loss function	77.5	59.1
weighted sampling and f1 loss	52	22.8
Weighted loss and weighted sampling	84.8	54.7
+ Learning rate 1e-4 $\longrightarrow$ $5e-6$	86.1	57.9
Multi-Lingual BERT	85.3	55.2
XLM-ROBERTA	87.5	53.6

Table 3: Comparison between different model configuration on the Weave Slot detection in VAM Dataset

Comparing the learning curves of BERT and multi-lingual BERT in figure 2 reveals that despite the comparable results, the multi-lingual BERT took double the number of epochs to stabilize and finish training compared to its BERT counterpart. This can be due to the fact that Multi-lingual BERT is trained in many languages and it needs more fine-tuning to adapt to any certain language, whereas the BERT transformer was trained in English-only documents.

4.3.2 Generalizing towards all datasets. After we experimented with different parameter settings, we decided to use the following parameters on all the datasets: Weighted Loss function and Weighted Sampling for batches; learning rate of $5 * 10^{-6}$; batch size of 16 for BERT and 12 for multi-lingual BERT and XLM-ROBERTA, due to memory limits; 1024 Units for LSTM Layer; dropout layer of 0.5.

Moreover, the datasets were tested against three types of transformer: Language-Specific BERT, Multilingual BERT, and XLM-ROBERTA, as well as the SVM classifier. The accuracy results in Table 4 show that on most of the datasets SVM performs better or comparable to the deep learning models.

Museum	Lang	Slot	Baseline	# Cls	# Exs	SVM	BERT	Multi BERT	XLM-ROBERTA
VAM	English	Weave	52.9	4	2760	82.8	86	85.3	**87.5**
VAM	English	Weaving Technique	35.9	14	3525	77.6	**80.1**	78	78
VAM	English	Motifs	84.8	9	5500	**91**	**90.6**	87.4	87
CER	Spanish	Weave	59.3	5	945	**75.1**	**75.1**	64	72
CER	Spanish	Weaving Technique	61.1	11	720	**74.3**	**74.1**	71.5	66
Joconde	French	Weave	55.6	4	180	66.7	30.6	86.1	**91.7**
Joconde	French	Weaving Technique	60	5	150	**97.2**	70	76.7	63.3

Table 4: Results for the multi-class classification task.

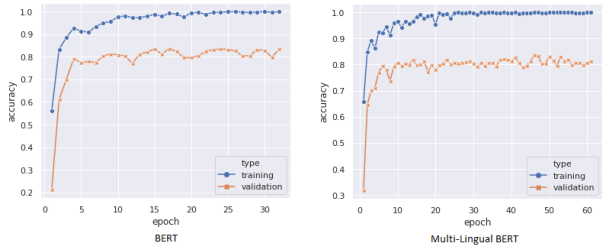

Figure 2: Comparison of a learning curve between BERT and Multi-Lingual BERT as a transformer in the deep learning model trained on the VAM museum Weave Slot dataset.

5 CONCLUSION AND FUTURE WORK

We propose an approach to extracting metadata from a multilingual text description of silk heritage domain museum artifacts. The datasets had several specifics that made the model development a non-trivial task. First, the size of the dataset sometimes was too small to train a model. Second, some class values have considerably more examples than others, which caused the datasets to be imbalanced. Finally, in the preparation phase, the datasets were labeled to accommodate the described issues, which in itself is an approximation and carries an inherent error rate. We have improved the performance of the model by over-sampling minority classes, under-sampling majority classes, and using a class-weighted loss function. In addition, by performing cross-validation in the binary classification case or adding a dropout layer and validating based on a validation dataset, we managed to mitigate some of the over-fitting behavior caused by having a little amount of data. We believe that the over-fitting could be mitigated further by using regularization on the LSTM layer, as well as using weight-decaying in the optimizer.

The experimental results show that with low data quality and having not enough data, traditional methods such as SVM in some cases outperform deep neural network models. We expect that the results could be improved by having an assembly of those models instead of using one of them only, which is a part of the future work. Furthermore, one can fine-tune each model independently to achieve better performance.

In future work, we plan to test cross-museum learning by training on one museum and predicting other museums both in the same language and in different languages using multi-lingual transformers. This has practical value for labeling the data in the museums that do not contain metadata information but do have suitable textual descriptions of the artifacts.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and SilkNow European Unions Horizon 2020 project under grant agreement No 769504.

REFERENCES

[1] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, (July 2020), 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. https://www.aclweb.org/anthology/2020.acl-main.747.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[4] Matthew Honnibal and Ines Montani. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, (2017).

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[6] Martin F. Porter. 2001. Snowball: a language for stemming algorithms. Published online. Accessed 11.03.2008, 15.00h. (2001). http://snowball.tartarus.org/texts/introduction.html.

[7] [n. d.] Pytorch: an imperative style, high-performance deep learning library. In.

[8] 2020. Silknow knowledge graph data. https://github.com/silknow/converter/tree/master/output. (2020).

[9] 2020. SilkNow project. https://silknow.eu/. (2020).

[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. [n. d.] Huggingface's transformers: state-of-the-art natural language processing.

Hierarchical classification of educational resources

Gregor Žunič
Jožef Stefan Institute
Ljubljana, Slovenia
gregor.zunic@ijs.si

Erik Novak
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
erik.novak@ijs.si

ABSTRACT

This paper describes an approach to automate the process of labelling hierarchically structured data. We propose a top-down level-based approach with SVMs to classify the data with scientific domain labels. The model was trained on labeled open education lectures and returns high accuracy predictions for lectures in the English language. We found that our model performs better with the traditional text extraction method TF-IDF than with pre-trained language model XLM-RoBERTa.

KEYWORDS

hierarchical classification, support vector machine, multi-class classification, machine learning, open educational resources

ACM Reference Format:
Gregor Žunič and Erik Novak. 2020. Hierarchical classification of educational resources. In *Proceedings of Slovenian KDD Conference (SiKDD'20).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Manually labeling data can be tedious work; one must have sufficient background knowledge about the data and have clear instructions in the labeling process. This becomes even more difficult when the data needs to be annotated with hierarchically structured labels.

In this paper we present a top-down level-based approach using support vector machines (SVMs) for labeling open education resources (OERs). The labels are in a hierarchical structure and represent different scientific domains. We compare different lecture representations using TF-IDF and XLM-RoBERTa and find that the TF-IDF representations yield better results. Even though the paper focuses on OERs the method can be generalized to any textual data set.

The remainder of the paper is structured as follows. Section 2 describes the related work done on the topic of hierarchical classification. Next, we present the data used in the evaluation in Section 3. The methodology is described in Section 4. The evaluation setting and its results are described in Section 5 followed by a discussion in Section 6. We present the future work in Section 7 and conclude the paper in Section 8.

SiKDD'20, October 2020, Ljubljana, Slovenia

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

2 RELATED WORK

There are two approaches to hierarchically classify the data: (1) the Big-bang, and (2) the Top-down level-based approach [4, 8, 9].

The big-bang approach works by training (complex) global classifiers which consider the entire class hierarchy as a whole. Each global classifier is binary and decides if the material fits the entire hierarchy (entire hierarchy is for example "Computer Science/Machine Learning/Support Vector Machine"). The advantage of this approach is that it avoids class-prediction inconsistencies across multiple levels. The major drawback of this approach is the high complexity due to the enforcing the model to correctly predict the whole hierarchy branch, which can be difficult to achieve.

The top-down level-based approach works by training local classifiers at each level to distinguish between its child nodes. An example will first, at the root level, be classified into a second-level category. It will then be further classified at the lower level category until it reaches one or more final categories where it can not be classified any further. The main advantage of this model is its simplicity. The disadvantage is the difficulty to detect an error in the parent category which could lead to false classification.

The most common implementation of a local classifier [3] is the support vector machine [7, 11]. In the later papers they propose to train separate SVMs for every level of a branch in the hierarchy.

3 DATA SET

The data set used in the experiment consists of 28,769 OER lectures available at Videolectures.NET [10], an award winning video OER repository. For each lecture we collected the following metadata: title, description, labels, language, authors, date published and the length of the lecture. The description is present in 58% of the lectures. The data set contains 24532 lectures in English, 3930 in Slovene and 307 lectures in other 16 languages.

Preprocessing. For our methodology we used only the lecture's title, description, language and categories. Each lecture is labeled with one or more scientific (sub-)domains most relevant for the lecture (e.g. "Computer Science", "Computer Science/Crowd Sourcing"). Figure 1 shows the distribution of lectures per number of labels.

Almost half of the lectures have more than one label. Lectures with no labels are placed under the "No Labels" category. These lectures are mostly introductory speakers' presentations in conferences. We focus on predicting a single label with high accuracy. We prescribed to only have one label per lecture. We achieve this by duplicating a lecture n times, where n is the number of labels of the lecture and assign a distinct label to each duplicate. Although the duplicates may reduce the performance of the models we do not reduce the already small number of lectures used during the

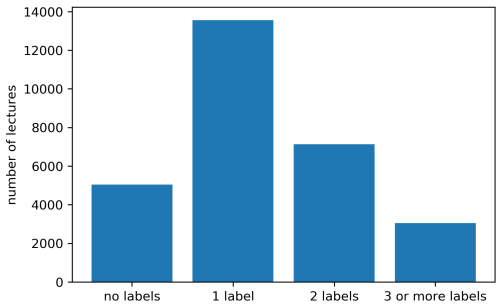

Figure 1: Distribution of lectures per number of corresponding labels. Most of the lectures have only one label.

training process. Figure 2 shows the top scientific domain labels in the data set.

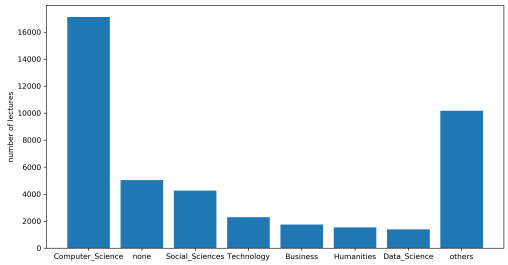

Figure 2: Top scientific domain labels in the data set. The most frequent label is Computer_Science.

The most frequent label is "Computer Science". In addition, a large number of lectures are not labeled; this is because a lot of lectures are presentations that do not correspond to any of the scientific domains. The data set is unbalanced on both domain and sub-domain levels.

4 METHODOLOGIES

In this section we describe the methods used to perform the feature extraction of the text, the implementation of multi class classifier model and the lectures' weights.

The input to the classifier is a raw string created by concatenating the title and the description if the description is available. It is then converted to a vector. In this paper we experimented with two approaches: TF-IDF and XLM-RoBERTa.

4.1 Feature Extraction

TF-IDF. Each lecture is represented with a vector of its TF-IDF values [6]. TF measures how frequently a term occurs in a lecture's text. The IDF is a measure of how much information the word provides. If it is common across all lectures its value is close to 0. The terms with the highest TF-IDF scores are usually the ones that characterize the topic of the lecture best.

The size of the lecture's vector representation is exactly the same as the total number of unique words. Since most of the features are zero the lecture vectors are sparse.

XLM-RoBERTa. The model is based on the RoBERTa model released in 2019. It is a large language model trained on 2.5 TB of CommonCrawl data [2]. The model achieves state-of-the-art performance on cross-lingual classification, sequence labeling and question answering. The most useful feature of the model is that it does not require the sentence language as an input. In theory, it extracts the same vectors for similar words in 100 languages.

The length of the vector that the model outputs is 768. To extract the features a CUDA-enabled GPU is required and the model training is very slow.

4.2 Multi-class SVM Classifier

We chose the top-down level-based approach for our classifier. The raw text input is firstly vectorized following one of the two feature extraction approaches described in Section 4.1. The vector is then input to the main SVM which determines the first category. Then the input is handled by the second SVM, trained specifically for sub-labels of first classified category. If a sub-label tops the threshold of 0, this step is repeated, otherwise the model outputs the lowest level parent category.

For example "Computer Science" is the first determined category. Then the input is handled by the SVM trained on sub-labels of "Computer Science", which determines that the input does not match with any of the sub-labels. The model puts the lecture in the "Computer Science" category. This is visually explained in figure 3.

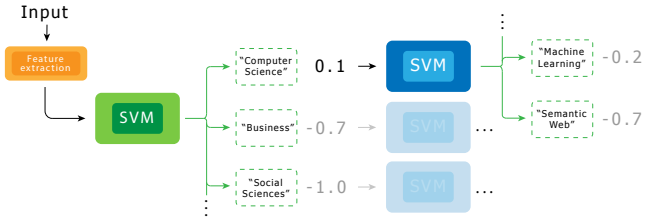

Figure 3: Visual representation of hierarchical SVM classifier. The example shows a lecture classified as belonging to the "Computer Science" category

Each SVM is an implementation of a multi-class classifier using the one-vs-rest approach. Predicted class should always be dominant otherwise the recommendation is not relevant.

4.3 Lecture Weights

Each lecture is assigned a weight of $\frac{1}{n^x}$, $x = 4$, where n is the number of total labels in the original lecture and x is a parameter. If $x < 4$ the accuracy is greatly reduced, if $x > 4$ the accuracy is increased by a small margin. It converges when $x \to \infty$. When increasing the parameter x the weight comes closer to 0 which means that the model accounts for data less during training. This means that the 4^{th} power is a sufficient balance between excluding some data and reducing the accuracy.

The other approach could be to ignore multi-label lectures during testing phase ($\frac{1}{n^\infty}$).

Because some labels are so scarce, we limit ourselves to labels with at least 20 lectures. This reduces the total number of labels in the data set from 502 to 244.

5 EVALUATION

5.1 Parameters and Specifications

SVM. The SVM implementation used in the evaluation is the LinearSVC [1] with the default parameters.

XLM-RoBERTa. The model used for representation generation is the hugging face's pretrained model [5] which was trained on default parameters found in the paper [2]. The training was executed on the Google Colab (online hosted Jupyter notebook) free tier machine (12GB RAM, dual core CPU, NVIDIA K80).

5.2 Results

Table 1 shows the performance of the different models with linear kernel. We have also evaluated other kernels (polynomial, RBF, sigmoid), but the performance was worse than using linear kernel. That is why we omitted them from the performance table.

TF-IDF with linear kernel SVM. Using the TF-IDF method for feature extraction we found that the SVMs performed the best with linear kernel. One explanation for such results is that the dimension of the features is large (more than 60k), which means that other more advance kernels might lead to over-fitting.

XLM-RoBERTa with linear kernel SVM. The model's performance was worse than using TF-IDF. The accuracy of the main classifier was 19% compared to 70% when using TF-IDF. The other SVM kernels (polynomial, RBF, sigmoid) performed worse compared to linear kernel. Table 1 shows the performance of the model.

SVM. The problem with current SVM implementation is that it can only put the lecture in one category. One way to solve the issue of only one label would be to firstly predict one label. Then, if the user (editor) wants another prediction, the model can output the prediction with second highest certainty.

TF-IDF vs XLM-RoBERTa. The advantage of choosing XLM-RoBERTa over of TF-IDF is that it works with 100 languages. The vector outputs are similar [2] for all languages. This was proven by translating the same text input into multiple languages (using Google Translate) and the predicted category did not change. When using TF-IDF you have to split the original data set into subsets containing a single language and train the model from scratch. That would be possible with enough data. For some languages (German, French) the the data set contains less than 30 lectures, which means that you can not train an SVM sufficiently.

6 DISCUSSION

Unbalanced Data Set. We found the SVM trained on an oversampled data set to be working better than the SVM trained on the raw data set. Due to the unbalanced data if the data set is not resampled the bias towards the strongest category (*Computer Science*) is strongly presented. For example neutral words such as " ", "the" etc. are classified as belonging in *Computer Science* category.

Comparing Word Embedding Techniques. The TF-IDF approach performs much better than XLM-RoBERTa which is surprising. Pre-trained models usually perform better than legacy feature extractors. The reason could be that the hyper parameters of the model were not set correctly, but we did not find the right balance for the model to perform any better. The production versions could include both models. For languages with a lot of data in the data set, the model would opt for SVMs trained on features extracted using TF-IDF, because of the better performance. All other languages would be handled by SVMs trained by XLM-RoBERTa, because the classifier performs much better than random.

The TD-IDF method could also be used to classify lectures that are in the non-english languages by firstly translating the text to English before using them during training. With this approach the model could work in all languages and retain the simplicity of TF-IDF. Note that that this approach would be strongly dependant on the quality of the translations.

Weighting the errors during the training process. We did not use the hierarchy structure for calculating the error between the predicted and the actual labels hence all the errors types during training were the same. This is not ideal because the error should be more significant when the classifier incorrectly predicts the main branch versus when it incorrectly predicts a lower level label. For example, if we take a lecture that is labeled as "Computer Science/Machine Learning" then the error should be bigger if our classifier predicts the "Biology" label rather than the "Computer Science/Semantic Web" label.

7 FUTURE WORK

We intend to improve the performance of the XLM-RoBERTa and to experiment with other language models and try to achieve better performance.

One additional direction for future work might be training a multiclass classifier to predict more than one label to a given lecture. We tried implementing the multi label output classifier using the MultiOutputClassifier wrapper on SVM but the precision of the model was noticeably lower.

The model is ready to be used in production in Videolectures.NET as a recommender engine to help the editors. The service could either be wrapped in a *Flask* microservice or directly into Videolectures.NET's backend.

8 CONCLUSION

In this paper we explore a top-down level-based approach for classifying OER lectures with scientific domain labels. We used oversampling to handle label unbalance and experimented with two text representation approaches, TF-IDF and XLM-RoBERTa. We found that the model using the TF-IDF representations gives better results.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and X5GON European Unions Horizon 2020 project under grant agreement No 761758.

REFERENCES

[1] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116* (2019).

parent category	TF-IDF acc.	TF-IDF recc.	TF-IDF F	TF-IDF prec.	XLM-RoBERTa acc.	XLM-RoBERTa recc.	XLM-RoBERTa F	XLM-RoBERTa prec.	materials
Root	70%	69%	72%	75%	19%	11%	19%	68%	27009
Computer Science	59%	59%	60%	61%	9%	4%	8%	50%	12935
Machine Learning	60%	55%	59%	64%	11%	5%	9%	26%	3260
Semantic Web	75%	71%	75%	79%	23%	20%	31%	68%	454
Computer Vision	82%	79%	81%	83%	57%	55%	59%	63%	140
Social Sciences	73%	72%	73%	74%	35%	24%	34%	60%	2928
Society	74%	72%	72%	72%	36%	28%	38%	60%	890
Politics	76%	66%	75%	86%	59%	43%	54%	73%	83
Law	96%	96%	96%	96%	57%	41%	51%	67%	112
Journalism	100%	100%	100%	100%	91%	88%	90%	92%	53
Technology	84%	82%	82%	82%	50%	43%	50%	60%	970
Nanotechnology	69%	59%	69%	83%	46%	37%	46%	62%	78
Business	74%	72%	73%	74%	43%	36%	43%	54%	1009
Transportation	63%	53%	61%	71%	33%	22%	32%	56%	267
Humanities	85%	83%	84%	85%	55%	48%	55%	65%	873
Biology	71%	66%	67%	68%	23%	17%	22%	31%	430
Science	78%	77%	78%	79%	53%	51%	52%	53%	656
Medicine	89%	88%	89%	90%	39%	34%	48%	83%	326
Computers	83%	83%	83%	83%	55%	48%	53%	59%	731
Mathematics	89%	87%	89%	91%	41%	36%	38%	40%	421
Physics	86%	81%	85%	89%	36%	32%	38%	46%	227
Arts	88%	87%	85%	83%	45%	40%	49%	63%	338
Visual Arts	100%	100%	100%	100%	62%	56%	70%	92%	159
Design	52%	46%	55%	68%	23%	9%	14%	30%	104
Chemistry	100%	100%	100%	100%	85%	83%	91%	100%	161
Environment	94%	94%	93%	92%	71%	66%	73%	81%	161
Earth Sciences	73%	67%	74%	82%	50%	51%	50%	49%	27

Table 1: Comparison of model performance using the linear kernel. The performance of the TF-IDF approach is better than that of XLM-RoBERTa.

[3] Susan Dumais and Hao Chen. 2000. Hierarchical Classification of Web Content. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 256–263. https://doi.org/10.1145/345508.345593

[4] A. D. Gordon. 1987. A Review of Hierarchical Classification. *Journal of the Royal Statistical Society: Series A (General)* 150, 2 (1987), 119–137. https://doi.org/10.2307/2981629 arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2981629

[5] huggingface. 2020. huggingface.co - pretrained models. https://huggingface.co/transformers/pretrained_models.html.

[6] J.D. Rajaraman, A.; Ullman. 2011. Mining of Massive Datasets. pp. 1–17. http://i.stanford.edu/~ullman/mmds/ch1.pdf.

[7] Ahmad Shalbaf, Reza Shalbaf, Mohsen Saffar, and Jamie Sleigh. 2020. Monitoring the level of hypnosis using a hierarchical SVM system. *Journal of Clinical Monitoring and Computing* 34, 2 (2020), 331–338. https://doi.org/10.1007/s10877-019-00311-1

[8] Carlos N. Silla and Alex A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22, 1 (2011), 31–72. https://doi.org/10.1007/s10618-010-0175-9

[9] Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. 2003. Performance measurement framework for hierarchical text classification. *Journal of the American Society for Information Science and Technology* 54, 11 (2003), 1014–1028. https://doi.org/10.1002/asi.10298 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.10298

[10] VideoLectures.Net. 2020. VideoLectures.NET - VideoLectures.NET. https://videolectures.net/. Accessed: 2020-08-20.

[11] S. V. M. Vishwanathan and M. Narasimha Murty. 2002. SSVM: a simple SVM algorithm. 3 (2002), 2393–2398 vol.3.

Are You Following the Right News-Outlet? A Machine Learning based approach to outlet prediction

Swati
swati@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

ABSTRACT

In this work, we propose a benchmark task of outlet prediction and present a dataset of English news events tailored to the proposed task. Addressing this problem would not only allow readers to choose and respond to relevant and broader facets of events but also enable the outlets to examine and report on their work. We also propose a neural network based approach to recommend a list of probable outlets covering an event of interest. Evaluation results reveal that even in its simplest form, our model is capable of predicting the outlet significantly better than the existing rule based approaches. The proposed model will also serve as a baseline for evaluating approaches intended to address the task. Implementation scripts can be found at *https://github.com/Swati17293/outlet-prediction.*

KEYWORDS

News bias, Event Selection bias, News coverage, News Event Analysis, Recommendation System

1 INTRODUCTION

The advancement in the field of Natural Language Processing [9, 10, 5, 4] over the last decade, has made solutions to complex machine learning problems more convenient. The problems such as machine translation, text summarization, and segmentation are being solved much more efficiently than ever before. Consequently, it offered the researchers the opportunity to use these advanced techniques to solve problems in a variety of contexts such as news bias analysis. This analysis task is poised as the identification of the inherent bias present in the news production and its coverage process. It occurs when a news outlet publishes a news story selectively or incorrectly.

If the news is biased, then it can bias the thought process and decision making of the person listening, watching, and/or reading it [12]. It can have several direct or indirect implications whether political or social. For example, if the news shows only the positive or negative side of a political party; it has been observed to influence the public vote [2]. Not only politics but also the news about the disaster or spread of viral disease affects the belief system of the general public.

There are numerous events that happen continuously, and any form of bias can arise in numerous possible ways. It is not possible for any single outlet to capture every event. Thus, an outlet is forced to select a set of reporting events. Several factors, such as the geographical origin of the event, the involvement of an elite person or country, etc. influences such selection. Also the procedure requires rigorous monitoring of current affairs to determine the news value, and may result in event selection bias also known as gatekeeping bias.

However, no well-established automated method reveals to users the outlets that will cover the event of their interest. This drives the motivation of this study. The aim is to predict a list of outlets reporting on a given event. Addressing this problem would not only allow readers to choose and respond to relevant and broader facets of events but also enable the outlets to examine and report on their work. For instance, some outlets tend to publish events covered by well-established outlets. Instead of waiting for the news to be published, the proposed system will help them to get an insight into the degree of predictability of event selection by the major outlets.

1.1 contributions

We make the following contributions in this context:

- We propose a benchmark task of outlet prediction and present a dataset of English news events tailored to the proposed task.
- We provide a neural network model that can serve as a baseline for evaluating approaches intended to address the task.

The GitHub repository containing our code is available at ***https://github.com/Swati17293/outlet-prediction***.

1.2 Problem Statement

The problem is addressed as an outlet prediction task in which the bias is examined by comparing the learning ability of a classifier trained to predict the probability of event coverage by an outlet.

2 LITERATURE REVIEW

During the different stages of news production, various forms of news bias arise as described by Baker et al. [1]. The first stage begins with the selection of events also called gatekeeping, where an outlet selects or rejects an event for reporting. The selection process is driven by a number of factors, such as the geographical origin of the event, the involvement of an elite person or country, etc., and requires rigorous monitoring of current affairs to determine the news value. To our knowledge, only a few methods have been suggested that explicitly attempt to examine this bias.

Saez-Trumper et al. [11] attempted to identify bias in online news sources and social media groups surrounding them. They studied the disparity in the selection of events based on the quantity and exclusivity of stories published by 80 mainstream news

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

outlets across the globe over a span of two weeks. From the review, it is found that there is a weak correlation between the quantity and exclusivity of news articles published by the outlets. It is also discovered that both the news and social media follow the same pattern of selection of events in similar geographical areas. However, media in the same region often choose the same events and publish similar-length posts.

Bourgeois et al. [3] used a matrix factorization method to extract latent factors that determine the selection of the event by an outlet. They combined the method with a BPR optimization scheme developed by Rendle et al.[8]. They used the events derived from the GDELT dataset and arranged the outlets in rows and their reported events in columns to form a matrix. Each cell value of the resulting matrix describes the selection/rejection of the event by the outlet.

For the bias analysis, they chose affiliation, ownership, and geographic proximity of the different outlets as the major factors. They suggest that each outlet follows its own latent preferences structure which facilitates the outlet to rank events. They also suggested that events should be selected such that the selected list should be diverse and should include a wide range of actively reported events. They thus adopted the method of Maximum Marginal Relevance which facilitates ranking based on the relevance and diversity of the events. It is discovered that event selection favors the most discussed topics rather than the unique ones.

F. Hamborg et al. [6] uses a matrix similar to the one created by Bourgeois et al.[3] Each cell in the matrix represent the most representative topic of the article reported by one country about the other. By spanning the matrix through outlets and topics in a region, the bias can be examined. They used a collection of 1.6 million articles from more than 100 countries over a two-month span from the Europe Media Monitor (EMM)[1] as their dataset.

Authors in [6] aggregates the related articles and then outsource the task of bias identification to the users, forcing them to determine the bias on their own. While the rest of the existing work analyzes the selection bias, it certainly does not present an automated approach suited to the outlet prediction task, unlike our work.

3 DATA DESCRIPTION

3.1 Raw Data Source

Event Registry[2] [7] monitors, collects, and provides news articles from news outlets around the world. It also aggregates them into clusters that are referred to as events. Each event is then annotated with several metadata such as unique id to track the event coverage, categories to which it may belong, geographical location, sentiment, etc. As a result, its large-scale temporal coverage can be used effectively to study the event selection process of news outlets.

3.2 Dataset

For our experiments, we first selected the top three news outlets based on Alexa Global Rankings [3]. We then used the Event Registry API to collect all news events reported in English between January 2019 and May 2020. We excluded events that were not covered by any of the selected outlets. We ended up with 51, 409 events for which we extracted basic information such as event id, title, summary, and source. Since the event coverage by these outlets is not uniform, which can be visualized in Figure 1, we used a stratified split to mimic this imbalance across the generated train-valid-test sets.

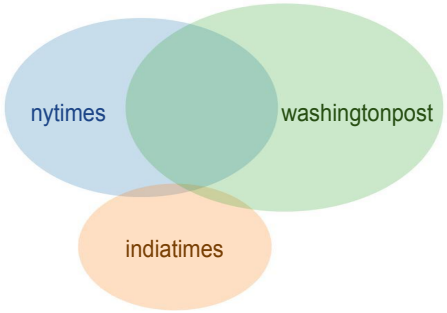

Figure 1: Distribution of event coverage by the outlets.

4 MATERIALS AND METHODS

4.1 Problem Modeling

For an event E and its associated pair (T, S), the task is to generate a list of outlets O expected to cover E. Here T is the event title and S is a short summary of the event as provided by the Event Registry. Mathematically, the task can be formulated as,

$$O = f(T, S, \alpha) \tag{1}$$

where, f is the outlet prediction function and α denotes the model parameters. O can have a well-thought-out variable length response generated from the list unique outlets O^l. For this work, $|O^l| = 3$.

4.2 Methodology

We extract feature vectors from T and S. We fuse them together to create a fused vector which is then passed through several layers to finally generate O. Figure 2 illustrates the entire prediction process. We further outline these tasks with more details in the following subsections.

4.2.1 Feature Extraction and Fusion. We used Google's *Universal Sentence Encoder* [4] *(USE)* to extract 128-dimensional feature vectors T' and S'. For feature fusion, we concatenated T' and S' and applied $tanh$ activation to generate F. We then used batch-normalization to increase the stability of the network and for regularization.

$$F = BN(tanh(T' \oplus S')) \tag{2}$$

In Eq 2, BN and $\oplus$ represents batch-normalization and concatenation respectively.

[1] https://ec.europa.eu/knowledge4policy/
[2] https://eventregistry.org
[3] https://www.alexa.com/topsites/category/Top/News/Newspapers
[4] https://tfhub.dev/google/universal-sentence-encoder/

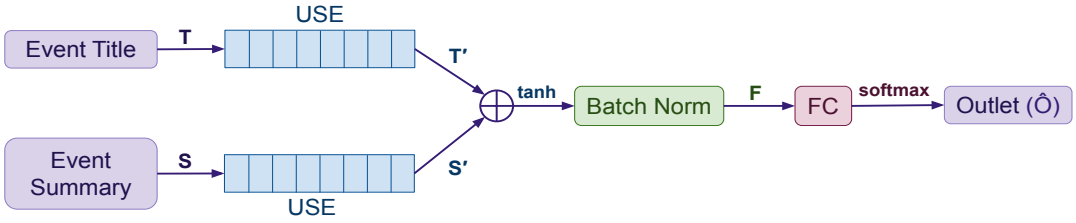

Figure 2: Outlet prediction process.

4.2.2 Outlet Prediction. We solve the problem using a multi-label classification model for which we create a separate outlet-index dictionary for outlets $D = \{o_1 : 1,\ o_2 : 2 \ldots\ o_n : n\}$, where n is the total number of unique outlets in O^l. To predict the list of outlets we pass F to the fully-connected layer (FC) having $softmax$ activation with n output neurons. Since an event can be covered by more than one outlet, we formulate the recursive prediction procedure as,

$$\hat{o} = \mathcal{P}(o_i|F, \hat{o}_{i-1}, b) = softmax(Fw_i + b_i) \tag{3}$$

$$= \frac{e^{Fw_i+b_i}}{\sum_{j=1}^{n} e^{Fw_j+b_j}} \tag{4}$$

where, $\hat{o}$ is the probability of selecting the i^{th} outlet (o_i) given F, bias (b), and the set of probabilities of previously predicted outlets ($\hat{o}_{i-1}$), and w is the weight. We use categorical cross entropy as the loss function as follows:

$$\mathcal{L}(o, \hat{o}) = -\sum_{j=1}^{n}\sum_{i=1}^{x}(o_{ij} * \log(\hat{o}_{ij})) \tag{5}$$

In Eq (5), for i^{th} outlet in the output sequence of length x, o_{ij} and $\hat{o}_{ij}$ denotes the actual and predicted probability of selecting the j^{th} outlet from D.

4.2.3 Hyper-parameters. We used Categorical accuracy[5] as the metrics to calculate the mean accuracy rate for multilabel classification problems across all the predictions. We consider a batch of size 128 and number of epocs as 100 for training. To optimize the weights during training we use Adam optimizer.

5 EXPERIMENTAL EVALUATION

5.1 Baselines

We use the following well-known and simplified methods as our baseline models.

- **Uniform:** Generate predictions randomly using a uniform distribution.
- **Stratified:** Generates predictions by respecting the class distribution of the training set.

5.2 Evaluation Metric

We aim to predict the list of outlets in this work. However, it is not necessary to predict the sequence in which outlets appear on this list. This is explained with an example given in Table 1. In other cases, a combination of correct and incorrect outlets may be predicted by the model.

We used the following metrics to evaluate the effectiveness of our model where, $\hat{o}$ is the predicted outlet, o is the true outlet, and N is the total number of instances.

[5]https://github.com/keras-team/keras/blob/master/keras/metrics.py

Table 1: Multiple correct predictions.

indiatimes nytimes washingtonpost
indiatimes washingtonpost nytimes

- **Subset Accuracy (a):** It measures the percentage of instances in which all of the outlets are correctly classified.

$$\text{Subset Accuracy } (a) = \frac{1}{N}\sum_{i=1}^{N}(\hat{o}_i - o_i) \tag{6}$$

- **Hamming Loss (ℓ):** It measures the fraction of the incorrectly predicted outlet to the total number of outlets. Since it is a loss function, its ideal value is 0.

$$\text{Hamming Loss } (\ell) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{o}_i \cap o_i}{\hat{o}_i \cup o_i}\right| \tag{7}$$

5.3 Results and Analysis

Table 2 shows the comparison of our model with the baseline models in terms of subset accuracy and hamming loss.

Table 2: Comparison between the baseline models and our proposed model.

	Subset Accuracy	Hamming Loss
Uniform	0.140	0.526
Stratified	0.286	0.422
Ours	**0.546**	**0.275**

Quantitative analysis of the experimental results shows that, our model outperforms the Uniform and Stratified models by a margin of 0.41 and 0.26 points for subset accuracy and by 0.25 and 0.15 points for hamming loss respectively. The performance difference is clearly visible in Figure 3.

The intersection that we find among the different outlet pairs differs considerably as evident in Figure 1. This can be best seen by assessing the conditional probability of an event covered by an outlet given that it is covered by another outlet as listed in Table 3. For example, we can note that the $P(washington|nytimes) = 0.492$ which is quite high and indicates that *washingtonpost* tends to cover most of the events covered by *nytimes*. It is also interesting to note that *indiatimes* do not follow *washingtonpost* or *nytimes*, and vice versa.

6 CONCLUSIONS AND FUTURE WORK

It is important for a journalist to know which event is worthy enough to be published. Even readers would be interested to know

Table 3: Conditional probability of an event to be covered by an outlet, provided it is covered by another outlet.

P(x|y)	nytimes	indiatimes	washingtonpost
nytimes	1.000	0.067	0.364
indiatimes	0.034	1.000	0.023
washingtonpost	0.492	0.063	1.000

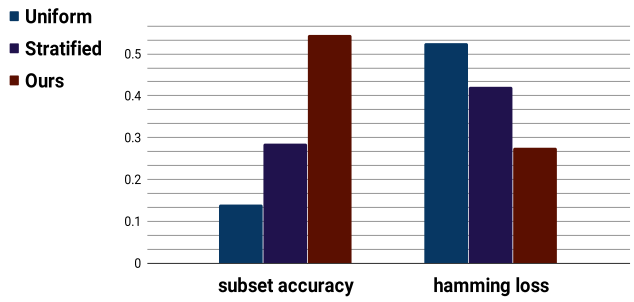

Figure 3: Comparison between the baseline models and our proposed model.

the outlets that are going to cover the event of their interest. Yet it is certainly not an automated approach, therefore in this work, we propose an approach to address the outlet prediction task given the event title and description. We also find that even in its simplest form, our model is capable of predicting the outlet. In the future, we intend to enhance our proposed model to better predict the outlets and to work in a cross-lingual setting. We plan to include a few more metadata provided by Event Registry (refer Section 3.1) along with Wikipedia concepts. We also plan to analyze the speed of reporting, time-span, and importance given to the events by the outlets. In addition, we will also be looking into how the outlets change their coverage style over time.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Brent H Baker, Tim Graham, and Steve Kaminsky. 1994. *How to identify, expose & correct liberal media bias.*

[2] Matthew Barnidge, Albert C Gunther, Jinha Kim, Yangsun Hong, Mallory Perryman, Swee Kiat Tay, and Sandra Knisely. 2020. Politically motivated selective exposure and perceived media bias, 82–103.

[3] Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. 2018. Selection bias in news coverage: learning it, fighting it. In *Companion Proceedings of the The Web Conference 2018*, 535–543.

[4] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, 13042–13054.

[5] Zihao Fu. 2019. An introduction of deep learning based word representation applied to natural language processing. In *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. IEEE, 92–104.

[6] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2018. Bias-aware news analysis using matrix-based news aggregation, 1–19.

[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] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (UAI '09). AUAI Press, Montreal, Quebec, Canada, 452–461. ISBN: 9780974903958.

[9] Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. PhD thesis. NUI Galway.

[10] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18.

[11] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 1679–1684.

[12] Rune J Sørensen. 2019. The impact of state television on voter turnout. *British Journal of Political Science*, 257–278.

MultiCOMET – Multilingual Commonsense Description

Adrian Mladenic Grobelnik
Artificial Intelligence Laboratory
Jozef Stefan Institute
Ljubljana Slovenia
adrian.m.grobelnik@ijs.si

Dunja Mladenic
Artificial Intelligence Laboratory
Jozef Stefan Institute
Ljubljana Slovenia
dunja.mladenic@ijs.si

Marko Grobelnik
Artificial Intelligence Laboratory
Jozef Stefan Institute
Ljubljana Slovenia
marko.grobelnik@ijs.si

ABSTRACT

This paper presents an approach to generating multilingual commonsense descriptions of sentences provided in natural language. We have expanded on an existing approach to automatic knowledge base construction in English to work on different languages. The proposed approach has been utilized to develop MultiCOMET, a publicly available online service for generating multilingual commonsense descriptions. Our experimental results show that the proposed approach is suitable for generating commonsense description for natural languages with Latin script. Comparing performance on Slovenian sentences to the English original, we have achieved precision as high as 0.7 for certain types of descriptors.

CCS CONCEPTS

•CCS Information systems Information retrieval Document representation Content analysis and feature selection

KEYWORDS

deep learning, commonsense reasoning, multilingual natural language processing

1 Introduction

As artificial intelligence systems are becoming better at performing highly specialized tasks, sometimes outperforming humans, they are unable to understand a simple children's fairy tale due to their inability to make commonsense inferences from simple events. With recent breakthroughs in the area of deep learning and overall increases in computing power, it has enabled us to model commonsense inferences with deep learning models. In our research, we expand on the approach to automatic generation of commonsense descriptors proposed in COMET [1] by applying their deep learning models to languages other than English.

The approach presented in COMET tackles automatic commonsense completion with the development of generative models of commonsense knowledge, and commonsense transformers that learn to generate diverse commonsense descriptions in natural language [1].

Our research hypothesis is that the approach proposed by COMET [1] can be expanded to Latin script languages other than English. To test this claim, we have trained our own deep learning model on the original training data, and another model on the data translated into another natural language.

The main contributions of this paper are (1) a new multilingual approach to annotating natural language sentences with commonsense descriptors, (2) implementation of the proposed approach that is made publicly available as an online service MultiCOMET http://multicomet.ijs.si/ (illustrated in Figure 4), (3) evaluation of the proposed approach on the Slovenian language. An additional contribution is the publicly available source code [3] allowing users to train their own models for other natural languages.

The rest of this paper is organized as follows: Section 2 provides a data description. Section 3 describes the problem and the algorithm used. Section 4 exhibits our experimental results. The paper concludes with discussion and directions for the future work in Section 5.

2 Data Description

One might say the only way for AI to learn to perform commonsense reasoning, is to learn from humans. Following the approach proposed by COMET [1], we used data from the ATOMIC [2] dataset. The ATOMIC dataset consists of over 24,000 sentences containing common phrases manually labelled by workers on Amazon Turk. For each sentence the workers were asked to assign open-text values to nine descriptors which capture nine if-then relation types to distinguish causes vs. effects, agents vs. themes, voluntary vs. involuntary events and actions vs. mental states [2] as described in ATOMIC.

The following are the nine descriptors and their explanations:

xIntent – Because PersonX **wanted…**

xNeed – Before, PersonX **needed…**

xAttr – PersonX is **seen as…**

xReact – As a result, PersonX **feels…**

xWant – As a result, PersonX **wants…**

xEffect – PersonX **then…**

oReact – As a result, others **feel…**

oWant – As a result, others **want…**

oEffect – Others **then…**

The dataset contains almost 300,000 unique descriptor values for the listed nine descriptors. An example of a labeled sentence is shown in Figure 3.

In order to test the proposed approach, we implemented it for the Slovene language. We have translated the sentences from the ATOMIC dataset to Slovene, keeping the descriptor values in English. The translation was done using Google Cloud's Translation API [4].

3 Problem Description and Algorithm

The problem we are solving is predicting the most likely values for each tag in the ATOMIC [1] dataset, given an input sentence in a Latin script language. Following the proposal in COMET, we are addressing the following problem:

Given a training knowledge base of natural tuples in the $\{s, r, d\}$ format, where s is the sentence, r is the relation type and d represents the relation values. The task is to generate d given s and r as inputs.

Figure 1 depicts our approach to solving this problem. The system takes labelled sentences as input, translates them to the targeted Latin language and trains a deep learning model capable of labelling previously unseen sentences with values for nine descriptors capturing the nine predefined relation types as described in Section 2.

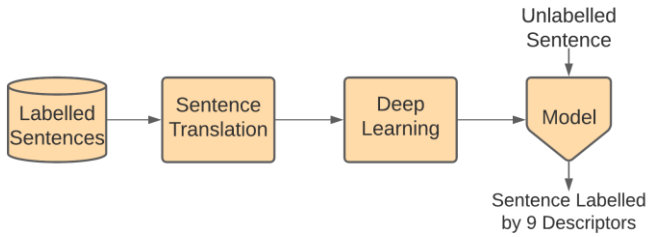

Figure 1: Architecture of the proposed approach

4 Experimental Results

Prior to training the model, we split the ATOMIC dataset into train, test and development sets identical to those used in COMET [1]. In our evaluation we used 100 sentences from the test set.

Our deep learning models are trained on the ATOMIC [2] dataset. We have trained one model on the original dataset in English, and another model on an automatically translated dataset to Slovene. Both models were trained under the same parameter settings: batch size=6, iterations=50000, maximum number of input features = 50.

To evaluate the performance of the proposed approach, we compared the predictions of the model trained on Slovene sentences with the predictions of the English model. As the performance metrics, we took the top 5 predicted values for each descriptor and checked their overlap. By taking the English predictions as the ground truth, we are measuring the precision of our model by the number of identical descriptor values. Note that we were strict in our comparisons, for instance "to stay away from people" and "to get away from others" do not count in overlap.

Experimental results show there is considerable difference in performance between the nine descriptors. The best performing descriptor was xReact, where precision@5 was 0.716, followed by oReact and oWant with precisions@5 of 0.706 and 0.468 respectively. The worst performing descriptor was xWant, with a precision@5 of 0.21 (see Table 1).

Descriptor	Precision
xIntent	0.324
xNeed	0.352
xAttr	0.438
xReact	0.716
xWant	0.210
xEffect	0.456
oReact	0.706
oWant	0.468
oEffect	0.310
Average	0.442

Table 1: Experimental results on the nine descriptors, showing precision of the top 5 predictions.

The best performing descriptor was xReact (representing the relation: As a result, PersonX feels). This was likely due to the fact that most predicted values were only one word long for both models, making it considerably easier for their predictions to overlap.

The worst performing descriptor was xWant (representing the relation: As a result, PersonX wants), this could be attributed to the fact that the most predicted values were at least 3-4 words in length, greatly decreasing the likelihood of overlap. Another reason for such low precision could be our strict overlap comparisons.

	Original	Translated/Predicted
Sentence	PersonX looks PersonY ___ in the face	PersonX izgleda PersonY ___ v obraz
xReact Values	nervous	**satisfied**
	happy	**happy**
	satisfied	attractive
	powerful	proud
	confident	angry

Table 2: One of the worst performing test sentences for xReact

Table 2 shows the predicted values of one of the worst performing sentences for the xReact descriptor. Note the sentence "PersonX looks PersonY ___ in the face" can refer to "Bob looks Mary slowly in the face" or "Adrian looks Anna kindly in the face" or something

else. The columns in Table 2 and Table 3 labelled “Original” show the original English sentence and its predicted descriptor values. The columns labelled “Translated/Predicted” show the sentence translated into Slovene and its predicted descriptor values.

Table 3 shows the predicted values of one of the worst performing sentences for the xWant descriptor. We can see that there are no common predictions between the two models. Note the sentence “PersonX avoids every ___” can refer to “Marko avoids every car on the road” or “Dunja avoids every boring event” or something else.

	Original	Translated/Predicted
Sentence	PersonX avoids every ___	PersonX se izogiba vsakemu ___
xWant Values	to stay away from people	to get away from others
	to avoid trouble	to make sure they are ok
	to stay away	to get away from the situation
	to not get caught	to be alone
	to not be noticed	to make a decision

Table 3: One of the worst performing test sentences for xWant

While Tables 2 and 3 show the model’s outputs for a single descriptor, Figure 3 shows the full output of the model, given an example sentence “Mojca je pojedla odličen sendvič” (Mary ate an excellent sandwich). Figure 2 shows a close-up of the output of Figure 3. The images in Figures 2 and 3 were taken directly from the interface of our online service MultiCOMET [5].

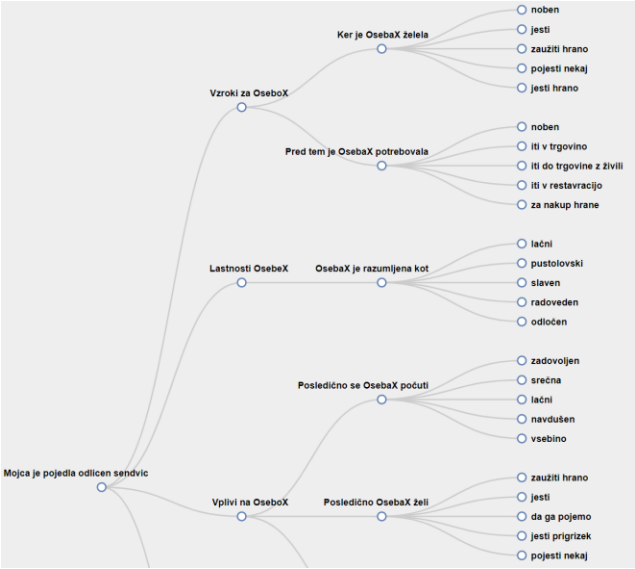

Figure 2: Close-up of predicted descriptor values generated for an example Slovene sentence

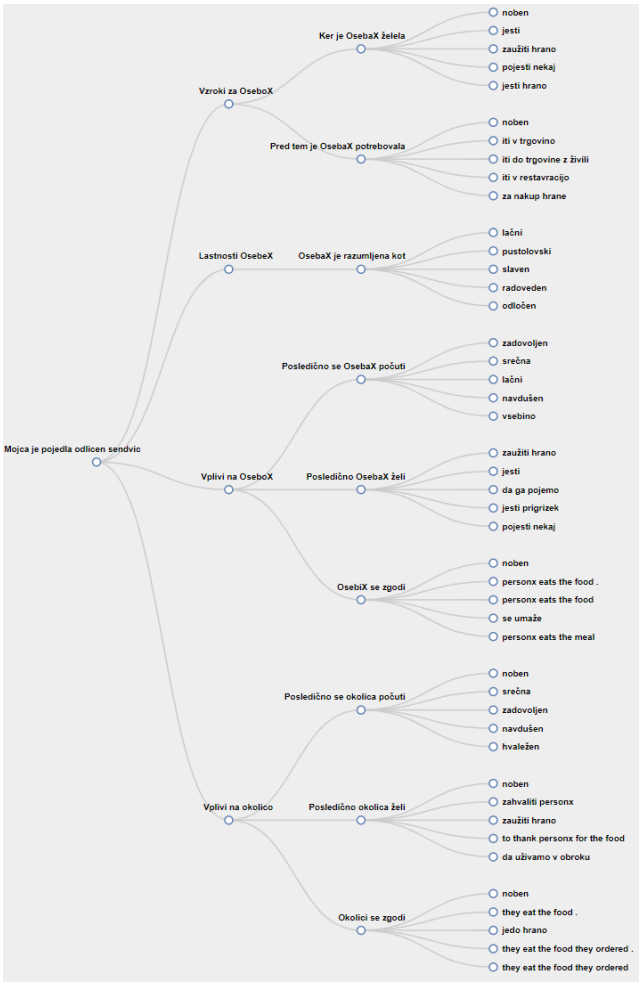

Figure 3: Full tree of predicted descriptor values generated for an example Slovene sentence

For the sentence “Mojca je pojedla odličen sendvič” (Mary ate an excellent sandwich) depicted in Figures 2 and 3, here is a potential English interpretation of the Slovenian output of the model:

Mary was hungry (xAttr) and wanted to eat food (xIntent). To do that, she needed to go to the restaurant (xNeed). At the restaurant, other people were also eating food (oEffect). As a consequence of eating the sandwich, Mary’s clothes got dirty (xEffect). Mary feels impressed (xReact) and wants to eat something else (xWant). The restaurant is grateful (oReact) for Mary’s visit and wants to thank Mary (oWant).

The MultiCOMET online service is a publicly available implementation of our proposed approach, shown in Figure 4. At the time of writing, MultiCOMET only supports English and Slovene.

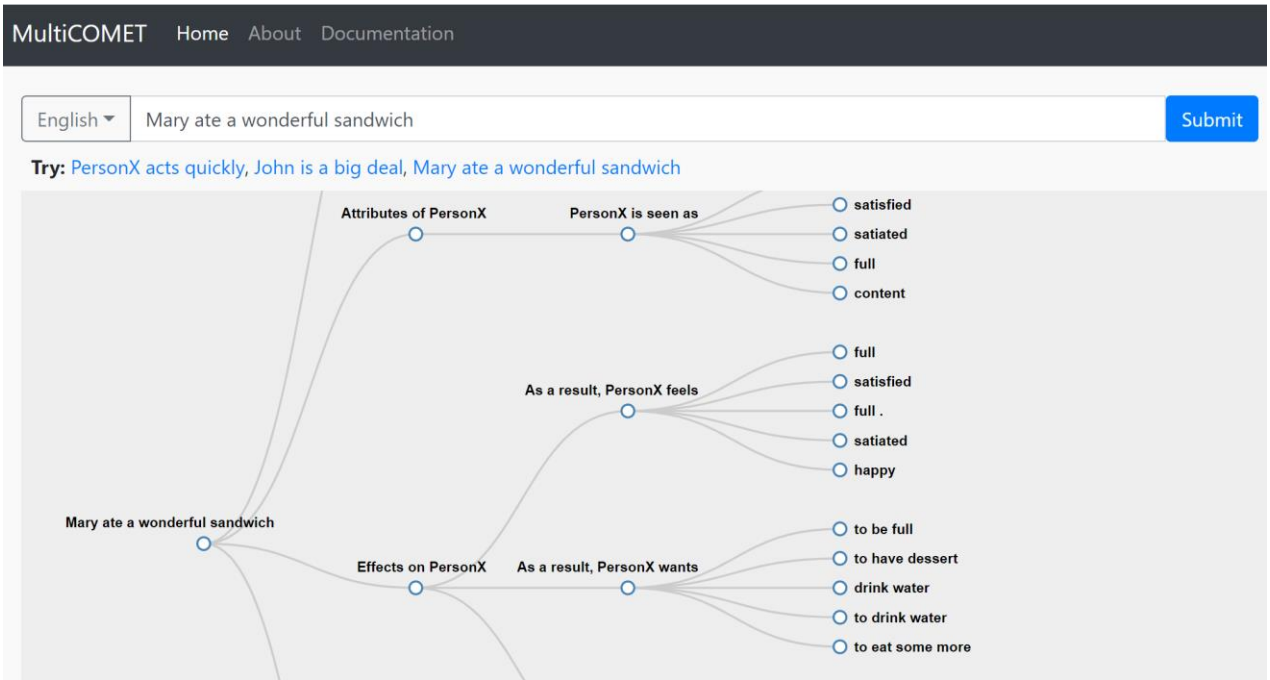

Figure 4: Illustrative example of MultiCOMET after submitting a query "Mary ate a wonderful sandwich."

5 Discussion

In our research we expanded on an existing monolingual approach and proposed a new approach to generating multilingual commonsense descriptions from natural language. In order to implement our approach, we built on an existing library, implementing the approach proposed by COMET [1]. Our experimental results show that we are getting meaningful values for the descriptors. Experimental comparison of the predicted descriptor values of the Slovene and English models show an average precision of 0.44, given our strict comparison methodology. We noted the precision values ranged from 0.716 to 0.210 across different descriptors.

Based on our literature review (September 2020), none of the articles citing the original COMET [1] paper expanded their approach to include other languages. The most similar work we found in the literature combining commonsense and multilinguality was [6] where the authors were extending the SemEval Task 4 solution using machine translation.

The possible direction for future work includes improving the quality of the translated sentences from ATOMIC by manual translation to improve the precision of the models. Another possible direction would be to evaluate the performance of our models on a larger number of sentences to increase the reliability of the results.

After testing the proposed multilingual approach on the Slovene language, we intend to expand our coverage to other Latin script languages including Croatian, Italian and French.

ACKNOWLEDGMENTS

The research described in this paper was supported by the Slovenian research agency under the project J2-1736 Causalify and co-financed by the Republic of Slovenia and the European Union under the European Regional Development Fund. The operation is carried out under the Operational Programme for the Implementation of the EU Cohesion Policy 2014–2020.

REFERENCES

[1] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, Yejin Choi. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. Allen Institute for Artificial Intelligence, Seattle, WA, USA. Paul G. Allen School of Computer Science & Engineering, Seattle, WA, USA. Microsoft Research, Redmond, WA, USA.

[2] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, Yejin Choi. (2019). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA. Allen Institute for Artificial Intelligence, Seattle, USA.

[3] MultiCOMET GitHub https://github.com/AMGrobelnik/MultiCOMET Accessed 31.08.2020

[4] Google Cloud's Translation API Basic https://cloud.google.com/translate Accessed 31.08.2020

[5] MultiCOMET http://multicomet.ijs.si/ Accessed 31.08.2020

[6] Josef Jon, Martin Fajcik, Martin Docekal, Pavel Smrz. (2020). BUT-FIT at SemEval-2020 Task 4: Multilingual commonsense. arXiv. https://arxiv.org/pdf/2008.07259.pdf

A Slovenian Retweet Network 2018-2020

Bojan Evkoski
Jožef Stefan International Postgraduate School,
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
Bojan.Evkoski@ijs.si

Igor Mozetič &
Nikola Ljubešić &
Petra Kralj Novak
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

As the popularity of social media has been growing steadily since the beginning of their era, the use of data from these platforms to analyze social phenomena is becoming more and more reliable. In this paper, we use tweets posted over a period of two years (2018-2020) to analyze the socio-political environment in Slovenia. We use network analysis by applying community detection and influence identification on the retweet network, as well as content analysis of tweets by using hashtags and URLs. Our study shows that Slovenian Twitter users are mainly grouped in three major socio-political communities: Left, Center and Right. Although the Left community is the most numerous, the most influential users belong to the Right and Center communities. Finally, we show that different communities prefer different online media to inform themselves, and that they also prioritize topics differently.

Keywords

Complex networks, Twitter, community detection, influencers

1. INTRODUCTION

Since the rise of the social networks, their data has been extensively used in social analysis. As the popularity of these platforms continues to grow daily, using them as a proxy to analyze specific phenomena is becoming more and more reliable. Their popularity, accessibility and availability made them the go-to way to share one's opinion, support another and even get in conflict with an opposing one. Recently, with the targeted advertising advancements, social media became the most important cultural and political battlefront.

In this paper, the country of interest is Slovenia and the proxy is Twitter data. By following the methodology developed in [3, 2, 4, 8], we address the following questions:

- Are there groups of densely connected Twitter users in the Slovenian retweet network 2018-2020?
- Who are the leading influencers in these groups?
- What is the content of the tweets in these groups and how much does it overlap?

This paper is organised as follows. In Section 2, the data acquisition process and the collected Twitter data are presented. Section 3 discusses the communities in the retweet network and their properties. Section 4 covers the notion of influencers and identifies the main influencers in the Slovenian retweet network. Section 5 investigates the content of the tweets in terms of hashtags and URLs. We draw conclusions in Section 6.

2. DATA

We acquired 5,147,970 tweets in the period from January 2018 to January 2020 with the TweetCat tool [6], built specifically for collecting Twitter data written in "smaller" languages. The tool identifies users tweeting in the focus language by searching for most common words in that language through the Twitter Search API, and collects these users' tweets through the whole data collection period. On average, the dataset containis around 8,000 tweets per day, with the three highest volume peaks on March 13, 2018 (11,556 tweets, the resignation of Slovenia's PM, Miro Cerar), June 1, 2018 (13,506 tweets, the last day of the 2018 Slovenian parliamentary elections campaign), and May 9, 2019 (12,381 tweets, Eurovision semi-final in which Slovenia had a successful run). The variation of the daily volume of tweets is affected by many phenomena, but the more evident are: a weekly seasonality with high volumes on working days and low volumes on weekends, extraordinary periods for the country (e.g. the 2018 Slovenian parliamentary elections campaign, boosting average daily tweets by around 2,000), and holidays (e.g. 2018 and 2019 Easters as local minima with 5,174 and 4,887 tweets, respectively).

3. COMMUNITY DETECTION

We used the collected tweets to construct a retweet network for the purpose of community detection. A retweet network is a directed weighted graph, where nodes represent Twitter users and edges represent the retweet relations. An edge from node (user) A to node B exists if B retweeted A at least once, indicating the information spread from A to B, or A influenced B. Note that retweeting a retweet is actually retweeting the original tweet (source), thus ignoring all intermediate retweets. The weight of an edge is the number of times user B retweeted user A. We removed all self-retweets, since they did not provide us additional information for community and influence detection. Consequently, we formed a network with 10,876 users (94% of all users) and 1,576,792 retweets (92% of all retweets).

This network can be simplified if the direction of the edges is ignored, meaning that two users are linked if one retweets the other while the source and destination are irrelevant. It turns out that such undirected retweet graphs between Twitter users are useful to detect communities of like-minded users who typically share common views on specific topics.

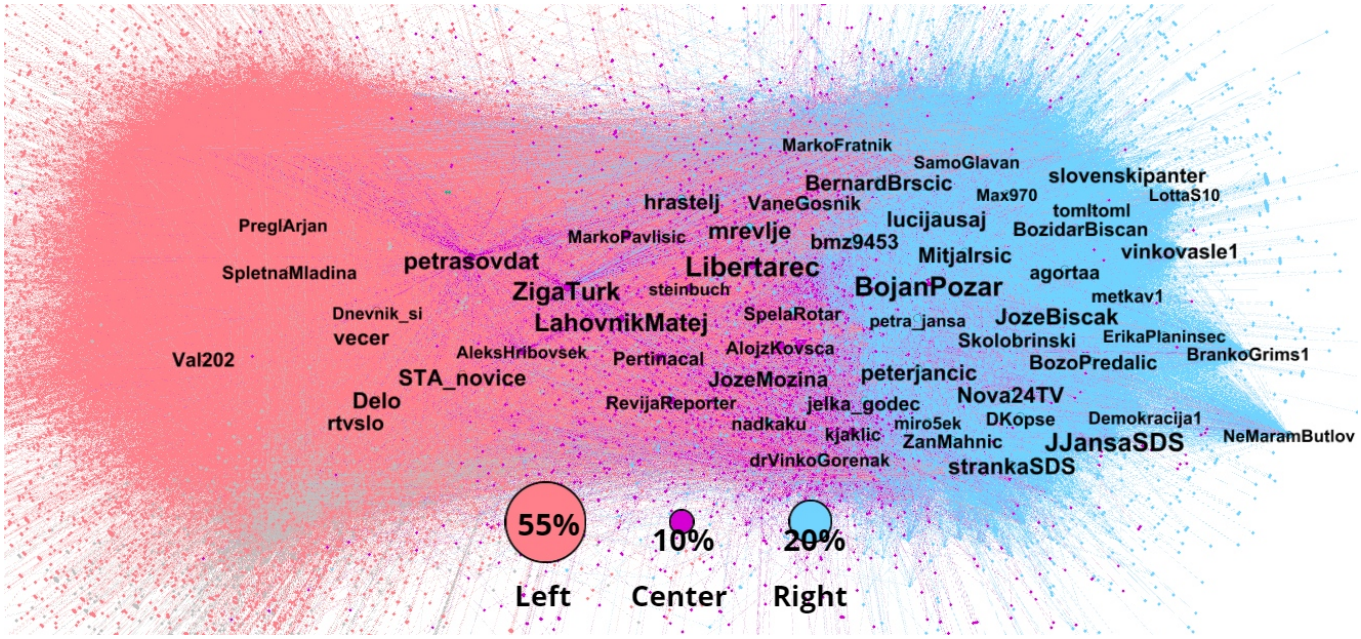

Figure 1: The Slovenian retweet network (2018-2020) colored according to the detected communities, with shares of the total number of users. The label size of a node corresponds to the number of unique users that retweeted it. Only nodes with at least 700 unique retweeters are included.

In complex networks, a community is defined as a subset of nodes that are more closely connected to each other than to other nodes. For the purpose of this paper, we apply a standard algorithm for community detection, the Louvain method [1]. The method partitions the nodes into communities by maximizing modularity (which measures the difference between the actual fraction of edges within the community and such fraction expected in a randomized graph with the same degree sequence) [7]. Modularity values range from -0.5 to 1.0, where a value of 0.0 indicates that the edges are randomly distributed, and larger values indicate a higher community density.

We ran the Louvain method (resolution = 1.05) on our undirected retweet network resulting in 183 communities with a modularity value of 0.382, which indicates a strong connectedness within communities. Only the three largest communities each have more than 5% of all users, while combined they contain 85% of all users. The three main detected communities are presented in Fig. 1. We observe the following:

- The three largest communities are labeled as Left, Center and Right with 55%, 20% and 10% as their respective shares of all users. The labeling of the communities does not necessarily represent their political orientation.
- The Left community, even though the largest, contains the smallest number of users with more than 700 unique retweeters.
- The Left community is well separated from the Center and the Right communities, which are more tightly interlinked.

We performed an exploratory data analysis and calculated the community properties presented in Table 1, to compare the communities. Most of the properties are normalized by the user to ease the comparison between communities.

- Nodes – unique users count
- Central user – user with most retweets
- Central user retweets – times the central user is retweeted
- Central user retweeters – unique users retweeting the central user
- HHI ($n = 50$) – Herfindahl–Hirschman index [9] measures the distribution of influence of the top n influential users. Higher value reflects the community influence concentrated only in few influential users, while lower value indicates more dispersed and balanced influence distribution.
- Edges in/node – edges remaining in the community per user (source and destination in the same community)
- Edges out/node – edges going out of the community per user (destination in a different community)
- Weighted edges in/node – weighted edges remaining in the community per user
- Weighted edges out/node – weighted edges going out of the community per user
- Out/In ratio – "Edges out" divided by "Edges in"
- Weighted out/in ratio – "Weighted edges out" divided by "Weighted edges in"

4. INFLUENCERS

We use two simple, but powerful metrics to detect influencers in the retweet network: the weighted out-degree and the Hirsch index (h-index) [5]. Both metrics are calculated from the number of retweets, thus known as retweet influence metrics, indicating the ability of a user to post content of interest to others.

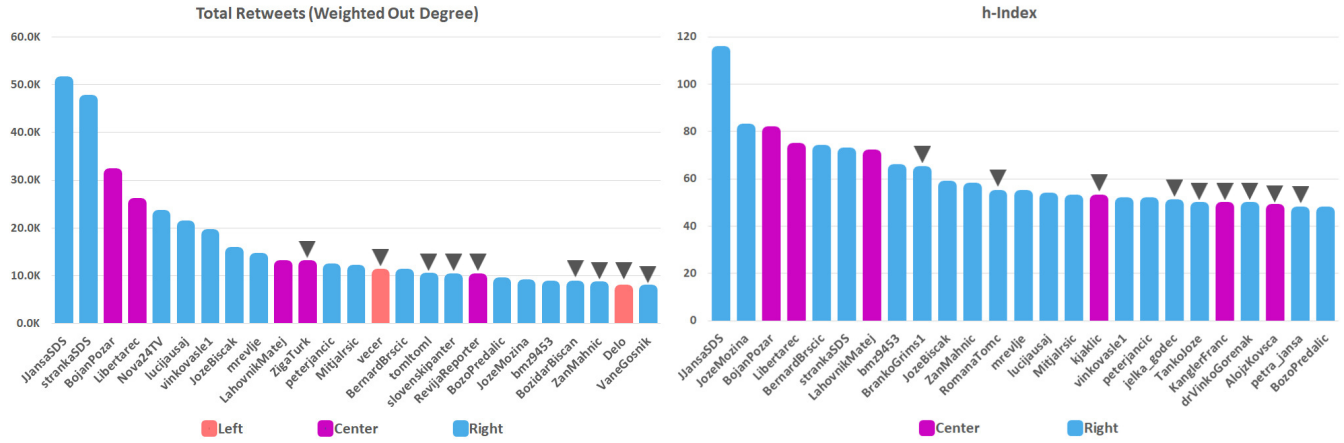

Figure 2: Weighted out-degree (total retweets) and h-index comparison. Both charts include the top 25 most influential Slovenian Twitter users according to their respective metric. Bar colors represent the community of a user. Triangles point to users exclusive to one of the charts.

Table 1: Community properties

	Left	Center	Right
Nodes	7,030	1,223	2,519
Central user	vecer	BojanPozar	JJansaSDS
Central user retweets	10,398	31,432	50,688
Central user retweeters	973	1,325	1,242
HHI ($n = 50$)	0.031	0.066	0.042
Edges in/node	19.32	14.53	69.30
Edges out/node	4.47	37.11	13.19
Weighted edges in/node	52.91	83.68	308.33
Weighted edges out/node	6.95	119.42	36.14
Out/In ratio	0.23	2.55	0.19
Weighted Out/In ratio	0.13	1.43	0.12

Weighted out-degree is simply the total number of retweets of a particular user, while the h-index is an author-level bibliometric indicator that measures the scientific output of a scholar by quantifying both the number of publications (i.e., productivity) and the number of citations per publication (i.e., citation impact). Adapted to a Twitter network, it would be described as: a user with an index of h has posted h tweets and each of them was retweeted at least h times.

Let RT be the function indicating the number of retweets for each original tweet. The values of RT are ordered in decreasing order, from the largest to the lowest, while i indicates the ranking position in the ordered list. The *h-index* is then defined as follows:

$$h\text{-}index(RT) = \max_i \min(RT(i), i)$$

The top 25 most influential users by weighted out-degree and h-index are shown in Fig. 2. The two metrics provide fairly similar results (they differ only in 9 users). Both results confirm the already visible phenomena from the previous observations: The Right community has the most influential users, while the Left community, even though the biggest, does not have nearly as popular users as the ones from the other two communities.

5. CONTENT ANALYSIS

We refer to content analysis in terms of getting knowledge from the text of the tweets. In this paper, we perform two kinds of content analysis: domain URLs and hashtags.

For domain URLs, we filtered the 2,297,008 tweets which contain a URL. Then, we extracted the domain part of the URLs and removed the domains with no specific meaning for Slovenia's content analysis (e.g. social networks: twitter.com, facebook.com, instagram.com, etc., and URL shorteners: ift.tt, bit.ly, ow.ly, etc.). This results in 512,308 tweets (approximately 22% of all the tweets with links). The most frequently occurring domains are owned by Slovenian media with nova24tv.si, rtvslo.si and delo.si as the top three URL domains with 23,879, 20,210 and 17,360 occurrences respectively. If instead of the total number of occurrences we count only the unique number of users which posted a domain URL, the top three domains are rtvslo.si, siol.net and delo.si with 2,802, 2,193 and 2,186 unique users respectively.

For the hashtag analysis, we filtered only tweets which contain a hashtag, ending up with 701,266 tweets. The top three hashtags are the following: #volitve2018 (the 2018 Slovenian parliamentary elections), #plts (the Slovenian First Football League) and #sdszate (Slovenian Democratic Party hashtag, meaning: SDS for you) with 9,845, 9,318 and 7,308 occurrences respectively. If we count only the unique number of users using a particular hashtag, the results for the top three Slovenian hashtags are as follows: #volitve2018 with 2,473, #slovenija with 1,611 and #fakenews with 1,343 users.

To see these results in the context of communities, we look at the tweets authored by members of the three largest communities, resulting in 84% of the tweets with relevant domain URLs and 83% of the tweets with relevant hashtags. We summed the domain URL counts, while grouping them by the community in which their user belongs. We applied the same procedure to the hashtags. Finally, we filtered the top eight domain URLs and hashtags for each community and put them on a single Sankey diagram in Fig. 3. Even though overlaps exist, the most popular hashtags and media very much differ from community to community, meaning that all three main communities prioritize topics differently and they inform themselves via different media.

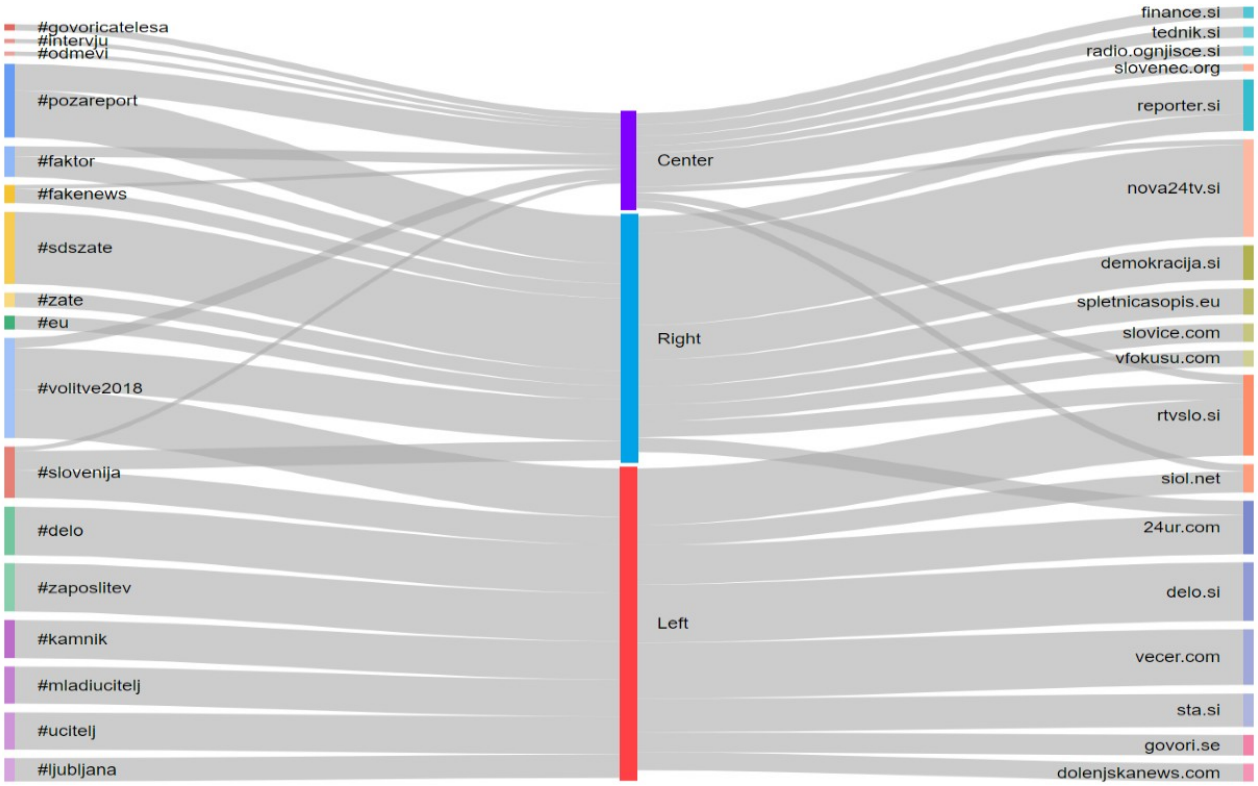

Figure 3: A Sankey diagram depicts the use of the eight most common hashtags (left-hand side) and URLs (right-hand side) by the three largest detected communities.

6. CONCLUSIONS

In this paper we explored the Slovenian twitter network from January 2018 until January 2020. We applied community detection, identifying three main communities: Left, Center and Right. We identified the most influential and the central users of each community by calculating the weighted out-degree and the h-index of the nodes. We used the Herfindahl–Hirschman index to estimate the distribution of influence within the top communities in the network. Finally, by analysis of hashtags and URL domains in tweets, we discovered the most popular topics for Slovenians as well as the most referred Slovenian media on Twitter. We showed that users from different communities prioritize different topics and use different media to inform themselves.

7. ACKNOWLEDGMENTS

The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P2-103 and P6-0411), and the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263).

8. REFERENCES

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] D. Cherepnalkoski, A. Karpf, I. Mozetič, and M. Grčar. Cohesion and coalition formation in the European Parliament: Roll-call votes and Twitter activities. *PLoS ONE*, 11(11):e0166586, 2016.

[3] D. Cherepnalkoski and I. Mozetič. Retweet networks of the European Parliament: Evaluation of the community structure. *Applied Network Science*, 1(1):2, 2016.

[4] M. Grčar, D. Cherepnalkoski, I. Mozetič, and P. Kralj Novak. Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks*, 4(1):6, 2017.

[5] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, pages 16569–16572, 2005.

[6] N. Ljubešić, D. Fišer, and T. Erjavec. TweetCaT: a tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2279–2283, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[7] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[8] P. K. Novak, L. D. Amicis, and I. Mozetič. Impact investing market on twitter: influential users and communities. *Applied Network Science*, 3(1):40, 2018.

[9] G. J. Werden. Using the Herfindahl–Hirschman index. In L. Phlips, editor, *Applied Industrial Economics*, number 2, pages 368–374. Cambridge University Press, 1998.

Toward improved semantic annotation of food and nutrition data

Lidija Jovanovska
Jožef Stefan International Postgraduate School &
Jožef Stefan Institute
Ljubljana, Slovenia
lidija.jovanovska@ijs.si

Panče Panov
Jožef Stefan Institute &
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
pance.panov@ijs.si

ABSTRACT

This paper aims to provide a critical overview of the state-of-the-art vocabularies used for semantic annotation of databases and datasets in the domain of food and nutrition. These vocabularies are commonly used as a backbone for creating metadata that is usually used in search. Furthermore, the paper aims to provide a summary of ICT technologies used for storing food and nutrition datasets and searching digital repositories of such datasets. Finally, the results of the paper will provide a roadmap for moving towards FAIR (findable, accessible, interoperable, and reusable) food and nutrition datasets, which can then be used in various AI tasks.

KEYWORDS

ontologies, semantic technologies, data mining, food and nutrition

1 INTRODUCTION

Today more than ever before in history, we live in an age of information-driven science. Vast amounts of information are being produced daily as a result of new types of high-throughput technology in all walks of life. Consequently, the quantity of available scientific information is becoming overwhelming and without its proper organization, we would not be able to maximize the knowledge we harvest from it. Namely, research groups carry out their research in different ways, with specific and possibly incompatible terminologies, formats, and computer technologies. To tackle these issues, researchers have developed high-level knowledge organization systems (KOS), such as ontologies, which constitute the core of the semantic web stack. Throughout the years, an abundance of ontologies has been developed and released, slowly expanding from the biomedical sciences to the fields of information science, machine learning, as well as the domain of food and nutrition science.

There is an old, yet simple saying which goes: "You are what you eat". As the world becomes more globalized and food production grows massively, it is becoming increasingly difficult to track the farm-to-fork food path. In the last few decades, digital technology has been profoundly affecting many health and economic aspects of food production, distribution, and consumption. Issues regarding food safety, security, authenticity as well as conflicts arising from biocultural trademark protection are issues that were further enhanced by the lack of a centralized food data repository without which there is a great difficulty in achieving cross-cultural and expert consensus. [1]

In this paper, we will briefly go through the fundamental components of the Semantic Web technologies, as well as the standards for the development of high-level KOS (Section 2). Next, we provide a critical overview of the most significant semantic resources in the domain of food and nutrition (Section 3). Finally, we present a proposal for the design and implementation of a broad ontology that would allow us to harmonize and integrate reference vocabularies and ontologies from different sub-areas of food and nutrition (Section 4).

2 BACKGROUND

The goal of the Semantic Web is to make Internet data machine-readable by enhancing web pages with semantic annotations. Linked data is built upon standard web technologies, also including semantic web technologies in its technology stack [11]. **Resource Description Framework (RDF)** allows the representation of relationships between entities using a simple subject-predicate-object format known as a triple. The triples form an RDF database — called a triplestore — which can be populated with RDF facts about some domain of interest. **RDF Schema (RDFS)** was developed immediately after the appearance of RDF as a set of mechanisms for describing groups of related resources and the relationships between them. **Simple Protocol and RDF Query Language (SPARQL)** is the query language for querying RDF triples stored in RDF triplestores.

The Web Ontology Language (OWL) is based on Description Logics, a family of logics that are expressively weaker than First Order Logic, but enjoy certain computational properties advantageous for purposes such as ontology-based reasoning and data validation. Most of the ontologies used today are represented in the OWL format.

All the semantic technologies operate on top of various KOS. A KOS is intended to encompass all types of schemes for organizing information and promoting knowledge management [7]. One example of a KOS is a thesaurus as a structured, normalized, and dynamic vocabulary designed to cover the terminology of a field of specific knowledge. It is most commonly used for indexing and retrieving information in a natural language in a system of controlled terms. When looking at the expressiveness of a KOS, a thesaurus is on the lower side of the scale. On the other side, ontologies enjoy greater expressiveness than thesauri due to the inclusion of description logics. Arp, Smith, and Spear define the term ontology as "A representation artifact, comprising a taxonomy as proper part, whose representations are intended to designate some combination of universals, defined classes, and certain relations between them" [1].

Information Society 2020, 5–9 October, 2020, Ljubljana, Slovenia

[1] https://www.nature.com/scitable/knowledge/library/food-safety-and-food-security-68168348/, accessed 22/04/2020

The Open Biomedical Ontologies (OBO) Foundry applies the key principles that ontologies should be open, orthogonal, instantiated in a well-specified syntax, and designed to share a common space of identifiers. Open means that the ontologies should be available for use without any constraint or license and also receptive to modifications proposed by the community. Orthogonal means that they ensure the additivity of annotations and compliance with modular development. The proper and well-specified syntax is expected to support algorithmic processing and the common system of identifiers enables backward compatibility with legacy annotations as the ontologies evolve [17].

The FAIR guiding principles for scientific data management and stewardship were conceived to serve as guidelines for those who wish to enhance the reusability and invaluableness of their data holdings [19]. The power of these principles lies in the fact that they are simple and minimalistic in design and as such can be adapted to various application scenarios. *Findability* ensures that a globally unique and persistent identifier is assigned to the data and the metadata which describes the data. *Accessibility* ensures that the data and the metadata can be retrieved by their identifier using a standardized communications protocol. *Interoperability* ensures that data, as well as metadata, use a formal, accessible, and shared language for knowledge representation. *Reusability* ensures that data and metadata are accurately described, released with a clear and accessible license, have detailed provenance, and meet domain-relevant community standards.

3 CRITICAL OVERVIEW OF FOOD AND NUTRITION SEMANTIC RESOURCES

In this section, we provide a critical overview of the most relevant KOS in the field of food and nutrition. We start by describing LanguaL [8], a thesaurus that serves as a foundation for most of the ontologies in this domain. We are more focused on analyzing ontologies which belong to different sub-spheres of the food and nutrition domain. Namely, FoodOn [4], as a more general food description ontology, ONS [18], relevant in the field of nutritional studies and ISO-Food [6], relevant in the field of annotating isotopic data acquired from food samples.

LanguaL [8] is a thesaurus used for describing, capturing, and retrieving data about food. Since 1996, it has been used to index numerous European Union (EU) and US agency databases, among which, the US Department of Agriculture (USDA) Nutrient Database for Standard Reference and 30 European Food Information Resource (EuroFIR) databases. Food ingredients are represented with indexing terms, preferably in the form of a noun or a phrase. The thesaurus also includes precombined terms which are food product names to which facet terms have been assigned. There are 4 main facets in LanguaL: A (Product Type), B (Food Source), C (Part of Plant or Animal), and E (Physical State, Shape, or Form). Other food product description facets include chemical additive, preservation or cooking process, packaging, and standard national and international upper-level product type schemes.

The LanguaL thesaurus complies with the FAIR guidelines. The completeness of LanguaL's indexing is to a large extent assured by the Langual Food Product Indexing (FPI) software, which verifies that all facets have been indexed for each food in the list [8]. It is available online[2] and can be queried using a food descriptor or synonym. Its interoperability and reusability are eminent as it represents a cornerstone in the development of more sophisticated ontologies, such as FoodOn. Even though the OBO Foundry principles apply only to ontologies, we can use the more general ones as evaluation criteria for the LanguaL thesaurus. For instance, as previously mentioned, the thesaurus is open, made available in an accepted concrete syntax, versioning is ensured, textual definitions are available for all the terms and a sufficient amount of documentation is provided.

FoodOn [4] is an open-source, comprehensive ontology composed of term hierarchy facets that cover basic raw food source ingredients, process terms for packaging, cooking, and preservation, and different product type schemes under which food products can be categorized. FoodOn is applicable in several use-cases, such as personalized foods and health, foodborne pathogen surveillance and investigations, food traceability and food webs, and sustainability. FoodOn echoes most of LanguaL's plant and animal part descriptors —– both anatomical (arm, organ, meat, seed) and fluid (blood, milk) —– but reuses existing Uberon [12] and Plant Ontology [10] term identifiers for them. Multiple component foods are more challenging because LanguaL provides no facility for giving identifiers to such products.

Building on top of this, FoodOn allows food product terms like lasagna noodle to be defined directly in the ontology, and allows them to reference component products through various relations which do not exist in LanguaL, such as: "has ingredient", "has part", "composed primarily of". As a suggestion, these relations can all be represented with a single relation "has ingredient" and the quantity can be expressed explicitly when annotating the objects. All of the ontology terms have unique identifiers and the ontology is accessible and can be searched via The European Bioinformatics Institute (EMBL-EBI) and its Ontology Lookup Service (OLS).[3] The ontology itself is open-source and is a member of the OBO Foundry. It also includes the upper-level Basic Formal Ontology (BFO) [1]. The adherence to BFO proves useful in the case of aligning ontologies covering different domains because they share the same top-level.

ONS [18] is the first systematic effort to provide a solid and extensible ontology framework for nutritional studies. ONS was built to fill the gap between the description of nutrition-based prevention of disease and the understanding of the complex impact nutrition has on health. Its structure consists of 3334 terms imported from already existing ontologies and 100 newly defined terms. The usability of ONS was tested in two scenarios: an observational study, which aims at developing novel and affordable nutritious foods to optimize the diet and reduce the risk of diet-related diseases among groups at risk of poverty, and an intervention study represented by the impact of increasing doses of flavonoid-rich and flavonoid-poor fruit and vegetables on cardiovascular risk factors in an "at risk" group study.

The development of ONS followed FAIR principles and as a result, it has been published in the FAIR-sharing database.[4] Before defining new terms, the developers of ONS have ensured that they are not yet defined, with the use of the ONTOBEE web service. Terms that were already defined were imported using the ontology reuse service — ONTOFOX [20]. In compliance with the OBO Foundry principles, the ONS has been developed to be interoperable with other ontologies, as it has been formalized

[2]https://www.langual.org, accessed 22/04/2020

[3]https://www.ebi.ac.uk/ols/ontologies/FoodOn, accessed 22/04/2020

[4]https://fairsharing.org/bsg-s001068/, accessed 22/04/2020

using the latest OWL 2 Web Ontology Language and RDF specifications and edited using Protégé [13] and the Hermit reasoner for consistency checking. It is also accessible, under the Creative Commons license (CC BY 4.0), published on GitHub and at NCBO BioPortal. Moreover, this ensured the adoption of a well-defined and widely adopted structure for the top and mid-level classes and principally the adherence to BFO as upper-level ontology.

ISO-Food is an ontology that was conceived to aid with the organization, harmonization, and knowledge extraction of datasets containing information about isotopes, that represent variants of a particular chemical element which differ in neutron number. To develop this ontology a mixed approach was used, a combination of both expert knowledge-driven (bottom-up) and data-driven (top-down) methods. Its main classes include Isotope, Sample, Location, Measurement, Article. The main class Isotope is connected to the rest of the classes with respective relations. The Food and Nutrient classes are linked to the RICHFIELDS ontology [5]. The ontology was further applied in a study for describing isotopic data, to annotate a data sample that consists of isotopic measurements of milk and potato samples.

The ISO-Food ontology can be accessed online via the BioPortal repository of biomedical ontologies.[5] It reuses terms from several ontologies, such as the concept Unit from the Units of Measurements Ontology (UO), the classes Food and Component from the RICHFIELDS ontology [5], the class Document from the Bibliographic Ontology (BIBO) [3].

4 PROPOSAL

Ontologies for data mining. To provide a suitable formalized representation of the outcomes of the research in the food and nutrition domain, as well as to suggest new ways to extract knowledge from the ever-abundant data produced in this field, we turn to ontologies that are used to formally represent the data analysis process. More specifically, we focus on the **OntoDM** ontology, which provides a unified framework for representing data mining entities. It consists of three modular ontologies: **OntoDM-core** [15] which represents core data mining entities, such as datasets, data mining tasks, algorithms, models and patterns, **OntoDT** [16] — a generic ontology of datatypes, and **OntoDM-KDD** [14] which describes the process of knowledge discovery.

The ontology defines top-level concepts in data mining and machine learning, such as data mining task, algorithm, and their generalizations, which denote the outputs of applying an implementation of an algorithm on a particular dataset. Starting with these general concepts, OntoDM also defines the components of the algorithms, such as distance and kernel functions, and other features they may contain. From the input and output data perspective, in this ontology, there is a hierarchical representation of data, from general concepts such as dataset to more specific concepts regarding its structure, such as the number of features, their role in a given task, concluding with the datatype of each attribute. These properties of OntoDM provide a complete formal representation of the data mining process from beginning to end.

Combining orthogonal domain ontologies. Our goal is to align the selected ontologies in the domain of food and nutrition with the OntoDM ontology of data mining to improve the semantic annotation of the food and nutrition domain datasets, as well as to formally represent data analysis tasks performed in the domain of food and nutrition (see Figure 1). In this way, we can also use the benefits of cross-domain reasoning. Since FoodOn, ONS, and OntoDM all use BFO as a main top-level ontology, they speak the same general language and are consequently, easier to align.

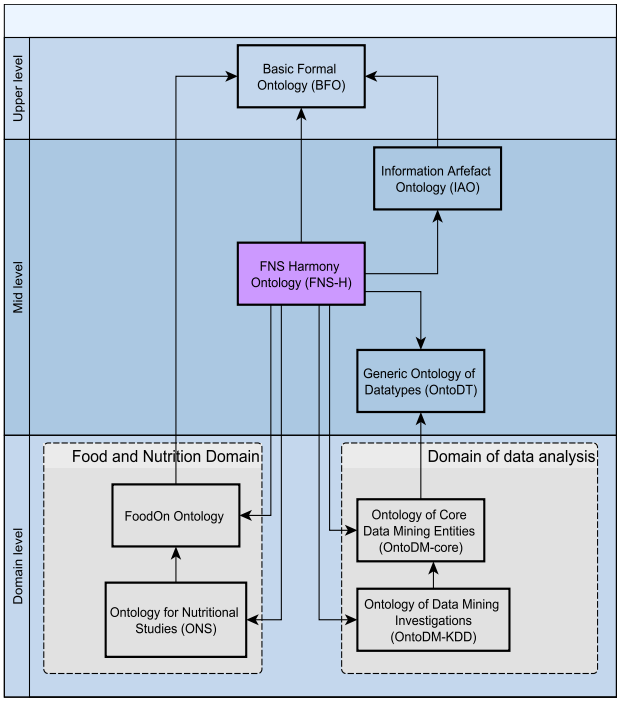

Figure 1: Diagram representing the alignment of the proposed ontology with the identified relevant upper-level and domain ontologies.

Towards the FNS Harmony ontology. In the context of the H2020 project FNS Cloud[6] (food, nutrition, security) the goal is to develop an infrastructure and services to exploit food, nutrition and security data (data, knowledge, tools – resources) for a range of purposes. To support the different functionalities required by the cloud platform, we started with the development of the FNS-Harmony (FNS-H). The application ontology would allow us to harmonize and integrate the different reference vocabularies and ontologies from different sub-areas of food and nutrition, as well as ontologies representing the domain of data analysis.

Initial ontology development. The development of FNS-H, which is intended to bridge the gap between the field of data analysis and food and nutrition will be guided by common best practice principles for ontology development. The aim is to maximize the reuse of available ontology resources and simultaneously follow the Minimum Information to Reference an External Ontology Term (MIREOT) principles [2]. In the first phase, we will integrate the FoodOn ontology and the ONS ontology with the OntoDM suite of ontologies. With this integration, we will be able to (1) define domain-specific data types for the domain of food and nutrition by extending OntoDT generic data types; (2) define food and nutrition analysis pipelines for the domain of food and nutrition by extending OntoDM-core, and (3) define

[5]http://bioportal.bioontology.org/ontologies/ISO-FOOD, accessed 22/04/2020

[6]https://www.fns-cloud.eu/

food and nutrition knowledge discovery scenarios by extending OntoDM-KDD ontology.

The development of the ontology already started in a top-down fashion, it is expressed in OWL2 and being developed using the Protégé ontology development tool. Aspiring to maximize accessibility, the ontology will be available for access on a GitHub repository, [7] as well as via BioPortal. In the current stage of development, an initial set of higher-level domain terms, data types, data formats, data provenance metadata, lists of external ontologies and vocabularies were extracted from the literature and FNS-Cloud project documents.

In the next steps, we will first align the extracted terms with the BFO ontology and then integrate them with domain terms from the domain ontologies based on BFO, such asFoodOn, and ONS, at the first instance, as well as with the OntoDM set of ontologies. Other potentially relevant ontologies include the Ontology for Biomedical Investigations (OBI), Ontology of Biological and Clinical Statistics (OBSC), Ontology of Chemical Entities of Biological Interest (ChEBI), Ontology of Statistical Methods (STATO), and others. To achieve integration of different ontological resources, we will use the ROBOT tool [9] that supports the automation of a large number of ontology development tasks and helps developers to efficiently produce high-quality ontologies.

5 CONCLUSION

In this paper, we provided an overview of the most relevant knowledge organization systems in the domain of food and nutrition. We started with the LanguaL food thesaurus that served as a foundation for the development of the more sophisticated ontologies — FoodOn, used for a multi-faceted description of various foods; ONS, used for observational and interventional nutrition studies; ISO-Food for the studies of isotopic data in foods. Next, we assessed the selected vocabularies with respect to the FAIR principles and OBO Foundry guidelines for scientific data management. All of the selected vocabularies showed compliance with these accomplishment criteria, with only minor suggestions for improvement provided from our side. Finally, in our proposal, we lay down the foundations of a new ontology which would connect data mining concepts in the domain of food and nutrition using domain ontologies (FoodOn, ONS) with ontologies for datatypes, data mining, and knowledge discovery in databases (OntoDT, OntoDM-core, OntoDM-KDD). By doing so, we can provide richer semantic annotation and discover new scenarios of harvesting knowledge from the food and nutrition data.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency through the grant J2-9230, as well as the European Union's Horizon 2020 research and innovation programme through grant 863059 (FNS-Cloud, Food Nutrition Security).

REFERENCES

[1] Robert Arp, Barry Smith, and Andrew D Spear. 2015. *Building ontologies with basic formal ontology*. Mit Press.

[2] Mélanie Courtot, Frank Gibson, and Allyson L Lister et al. 2011. Mireot: the minimum information to reference an external ontology term. *Applied Ontology*, 6, 1, 23–33.

[3] Bojana Dimić Surla, Milan Segedinac, and Dragan Ivanović. 2012. A bibo ontology extension for evaluation of scientific research results. In *Proceedings of the Fifth Balkan Conference in Informatics*, 275–278.

[4] Damion M Dooley, Emma J Griffiths, and Gurinder S Gosal et al. 2018. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2, 1, 1–10.

[5] Tome Eftimov, Gordana Ispirova, and Peter Korosec et al. 2018. The richfields framework for semantic interoperability of food information across heterogenous information systems. In *KDIR*, 313–320.

[6] Tome Eftimov, Gordana Ispirova, and Doris Potočnik. 2019. Iso-food ontology: a formal representation of the knowledge within the domain of isotopes for food science. *Food chemistry*, 277, 382–390.

[7] Heather Hedden. 2016. *The accidental taxonomist*. Information Today, Inc.

[8] Jayne D Ireland and A Møller. 2010. Langual food description: a learning process. *European journal of clinical nutrition*, 64, 3, S44–S48.

[9] Rebecca C Jackson, James P Balhoff, and Eric Douglass. 2019. Robot: a tool for automating ontology workflows. *BMC bioinformatics*, 20, 1, 407.

[10] Pankaj Jaiswal, Shulamit Avraham, and Katica Ilic et al. 2005. Plant ontology (po): a controlled vocabulary of plant structures and growth stages. *Comparative and functional genomics*, 6, 7-8, 388–397.

[11] Brian Matthews. 2005. Semantic web technologies. *E-learning*, 6, 6, 8.

[12] Christopher J Mungall, Carlo Torniai, and Georgios V Gkoutos et al. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13, 1, R5.

[13] Mark A Musen. 2015. The protégé project: a look back and a look forward. *AI matters*, 1, 4, 4–12.

[14] Panče Panov, Larisa Soldatova, and Sašo Džeroski. 2013. Ontodm-kdd: ontology for representing the knowledge discovery process. In *International Conference on Discovery Science*. Springer, 126–140.

[15] Panče Panov, Larisa Soldatova, and Sašo Džeroski. 2014. Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, 28, 5-6, 1222–1265.

[16] Panče Panov, Larisa N Soldatova, and Sašo Džeroski. 2016. Generic ontology of datatypes. *Information Sciences*, 329, 900–920.

[17] Barry Smith, Michael Ashburner, and Cornelius Rosse et al. 2007. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25, 11, 1251–1255.

[18] Francesco Vitali, Rosario Lombardo, and Damariz Rivero et al. 2018. Ons: an ontology for a standardized description of interventions and observational studies in nutrition. *Genes & nutrition*, 13, 1, 12.

[19] Mark D Wilkinson, Michel Dumontier, and IJsbrand Jan Aalbersberg et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

[20] Zuoshuang Xiang, Mélanie Courtot, and Ryan R Brinkman et al. 2010. Ontofox: web-based support for ontology reuse. *BMC research notes*, 3, 1, 175.

[7] https://github.com/panovp/FNS-Harmony

Absenteeism prediction from timesheet data: A case study

Peter Zupančič
1A Internet d.o.o.
Naselje nuklearne elektrarne 2
Krško, Slovenia
peter.zupancic91@gmail.com

Biljana Mileva Boshkoska
Faculty of Information Studies in
Novo mesto, Ljubljanska cesta 31a,
Novo mesto, Slovenia
Jožef Stefan Institute, Jamova cesta
39, Ljubljana, Slovenia
biljana.mileva@fis.unm.si

Panče Panov
Jožef Stefan Institute and
Jožef Stefan International
Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
pance.panov@ijs.si

ABSTRACT

Absenteeism, or employee absence from work, is a perpetual problem for all businesses, given the necessity to replace an absent worker to avoid a loss of revenue. In this paper, we focus on the task of predicting worker's absence based on historical timesheet data. The data are obtained from MojeUre, a system for tracking and recording working hours, which includes timesheet profiles of employees from different companies in Slovenia. More specifically, based on historical data for one year, we want to predict, under (which) certain conditions, if an employee will be absent from work and for how long (e.g., a week, a month). In this respect, we compare the performance of different predictive modeling methods by defining the prediction task as a binary classification task and as a regression task. Furthermore, in the case of one week ahead prediction, we test if we can improve the predictions by using additional aggregate descriptive attributes, together with the timesheet profiles.

KEYWORDS

Absenteeism at work, absence prediction, predictive modeling, timesheet data, human resource management

1 INTRODUCTION

Companies strive to have better predictive accuracy in their day to day operations, with the main goal of improving the productivity of the human resources (HR) department and hence obtaining higher profits and lower HR expenditures. They obtain information and insight from the large collections of human resource management (HRM) data that each employer owns, to support day to day operations and decision making, as well as, to comply to the national and international legislation.

The new era of HR executives is moving from settling on receptive choices exclusively taking into account reports and dashboards towards connecting business information and human asset information to foresee future results which will bring changes. Having such data enables them to detect patterns and trends, anticipate events and spot anomalies, forecast using what-if simulations and learn of changes in employee behaviour so that employee can take actions that lead to desired business outcomes. The purpose of HRM is measuring employee performance and engagement, studying workforce collaboration patterns, analyzing employee churn and turnover and modelling employee lifetime value [1].

Information society '20, October 5–9, 2020, Ljubljana, Slovenia

In this paper, we address the task of absenteeism prediction from time sheets data. More specifically, based on data that we get from MojeUre time attendance register system, we want to build a predictive model to predict if or for how many days an employee would be absent. In this case, we are considering one-week-ahead prediction from workers profiles and one year historical time sheets data. To predict if an employee will be absent in a given week, we employee the task of binary classification, which can be addressed by using a large number of binary classification methods. On the other hand, to predict the number of days an employee would be absent in a given week, we employee regression, which can be addressed by using regression methods. Furthermore, we observe and discuss how adding of aggregate attributes influences the prediction power if used together with the timesheet profiles.

2 DATA

In this section, we present the MojeUre system and then describe the structure of the raw data, as well as the process of data cleaning. Then we present the structure of the dataset, used for learning the predictive and the aggregate attributes, we constructed in order to test if they would improve the predictive power of the predictive models.

2.1 MojeUre system

The MojeUre system (https://mojeure.si) was developed to support the process of planning workers schedules, as well as for recording work attendance and absenteeism. In addition to the easy recording of the working hours of employees by a company, the system also provides access to each employee's own working hours, vacation control, sick leave, travel orders, etc. The system can be accessed using the web or by using a mobile application.

The entry of working hours is done either through a web application or a mobile application. In the case the company also wants to invests into a working time registrar, this can be done through the registrar where the employee has a personalized card for clock-in or clock-out (for example usage of break, such as a lunch break, a private break, etc.). The system allows different types of registered hours to be entered in the system in a single day.

All data used in the paper was obtained from the electronic system for recording working hours. There are currently more than 150 different companies that use the system for registering workers attendance. The basic function of the system is to record the arrivals and departures of an employee at work and to record the various types of employee absence, such as sick leave and vacation leave. In addition, the system covers other absences such as paternity leave, maternity leave, part-time leave, study leave, student leave, etc.

In this paper, we use data from the MojeUre system for the year 2019 and we have timesheet attendance data for all 52 weeks. The data instances are composed of three types of attributes: (1) attributes describing workers profiles (See Table 1), (2) attributes describing timesheets absence profiles of each worker (See Table 2), and (3) attributes that are aggregates from timesheets profiles constructed using domain knowledge (more details about the attributes is provided in Section 2.2). The timesheets attributes composing the absence profile of each worker are calculated based on the logged presence and absence logging data aggregated on the week level.. The entire dataset for the whole year consists of 232 different attributes and 2363 employees which are defined as each row.

Table 1: Workers profile attributes

Attribute name	Type	Description
EmployeeID	numeric	Unique employee identifier.
WorkHour	numeric	Data indicating how many hours per day an employee is employed by contract.
CompanyType	nominal	Company type by specific categories.
EmploymentYears	numeric	Describes how many years the person has been employed by the current company.
JobType	nominal	Describes type of job (e.g. permanent, part-time).
Region	nominal	The region in which the employee's company is located.

Table 2: Timesheet absence profile attributes

Attribute name	Type	Description
WeekWNYTotal	numeric	The number of all absences in a given week, including the sum of sick leave and (vacation) leave.
WeekWNY VacationLeave	numeric	The number of absences with type vacation leave in a given week.
WeekWNY SickLeave	nominal	The number of absences with type sick leave in a given week.
WeekWNY Absence	nominal	Value tells if employee was absent at least 1 day in whole week.

2.2 Data prepossessing and feature engineering

Feature Engineering is an art (Shekhar A, 2018) and involves the process of using domain knowledge to create features with the goal to increase the predictive power of machine learning algorithms. In this section, we describe the newly constructed attributes using domain knowledge. Furthermore, we present the process of data cleaning. Before cleaning, the original dataset contains 2087 instances of individual employees. The engineered aggregate attributes using domain knowledge from timesheets profiles are presented in Table 3.

Table 3: Attributes representing the workers profiles

Attribute name	Type	Description
VacationLeave TotalDays	numeric	Total days of vacation leave for all weeks, which are defined in the timesheets data used for the descriptive attribute space.
SickLeave TotalDays	numeric	Total days of sick leave for all weeks, which are defined in the timesheets data used for the descriptive attribute space.
ShortTerm VacationLeave3	numeric	A count of how many times an employee was at vacation leave for at least 3 days per week.
LongTerm VacationLeave5	numeric	A count of how many times an employee was on vacation leave for at last 5 days per week.
ShortTerm SickLeave3	numeric	A count of how many times an employee was on sick leave for at least 3 days.
LongTerm SickLeave5	numeric	A count of how many times an employee was on sick leave for at least 5 days.
WinterVacation LeaveAbsence	numeric	The number of vacation leave days that were used in winter.
SpringVacation LeaveAbsence	numeric	The number of vacation leave days that were used in spring.
SummerVacation LeaveAbsence	numeric	The number of vacation leave days that were used in summer.
AutumnVacation LeaveAbsence	numeric	The number of vacation leave days that were used in autumn.
WinterSickLeave Absence	numeric	The number of sick leave days that were used in winter.
SpringSick LeaveAbsence	numeric	The number of sick leave days that were used in spring.
SummerSick LeaveAbsence	numeric	The number of sick leave days that were used in summer.
AutumnSick LeaveAbsence	numeric	The number of sick leave days that were used in autumn.
WinterVacation LeaveHoliday	numeric	The number of vacation leave days that were used in winter during school holidays.
SpringVacation LeaveHoliday	numeric	The number of vacation leave days that were used in spring during school spring holidays.
SummerVacation LeaveHoliday	numeric	The number of vacation leave days that were used in summer during school summer holidays.
AutumnVacation LeaveHoliday	numeric	The number of vacation leave days that were used in autumn during school holidays.

The period we are considering in our analysis is one year, that is composed of 52 weeks. For construction of the aggregate attributes, we have defined our seasons by weeks, defined as follows: (1) the winter season is defined from week 51 in the previous year to week 12 in the New year; (2) the spring season is defined from week 13 to week 25; (3) the summer season is defined from week 26 week to week 39; and (4) the autumn season is defined from week 40 week to week 49.

In addition, we also defined the school holidays by weeks, which are defined as follows: (1) the winter holidays are defined from week 7 to 8; (2) the spring holidays are defined from week 18 to 19; (3) the summer holidays are defined from week 26 to week 35; and (4) the autumn holidays are defined from week 44 to week 45.

After we cleaned up the initial dataset, we obtained a smaller number of dataset instances. This resulted in a dataset with 961 distinct rows or more precisely different employees. The main control statement for the data cleaning was a test if an employee has less than one VacationLeaveTotalDays in the defined period. This would mean that: (1) an employee that fulfills this condition doesn't work any more in company; or (2) the company doesn't use recording system anymore; or (3) the employee is student and for students the vacation leave days are not recorded as they are usually paid per working hour only.

The most of employees in the dataset are working in company type called "Izobraževanje, prevajanje, kultura, šport" (Education, translation services, culture, sports). In addition, most of the employees are coming from the region "Osrednjeslovenska" (Central Slovenia region). The largest number of absence vacation leave or holiday leave was in week 52, which is the last week in year 2019 which is expected.

Descriptive attributes		**Target attribute**
Worker profile	Timesheet absence binary profile 1-(K-1) week	Week K Absence

(a) Without aggregate attributes

Descriptive attributes			**Target attribute**
Worker profile	Timesheet absence binary profile 1-(K-1) week	Timesheet absence aggregates 1-(K-1) week	Week K Absence

(b) With aggregate attributes

Figure 1: The structure of the data instances used for learning predictive models

3 DATA ANALYSIS SCENARIOS AND EXPERIMENTS

Research question. In general, in this paper we want to perform one-week ahead prediction of employee absence, using worker profile data, historical timesheet data aggregated on a week level, as well as aggregated attributes described in the previous section. We explore the task of predicting employee absence both as a binary classification task and as a regression task. In the experiments, we want to test if and how the aggregates attributes influence the predictive power of the built models both for the case of binary classification and regression.

Tasks. In the binary classification task, we want only to predict if an employee will be absent in a given week. For this case, we use the boolean attribute *WeekWNYAbsence* as a target attribute (WNY is the identifier of the target week). In the regression task, we want to predict the number of absence days. For this case, we use one of the following numeric attributes as targets *WeekWNYTotal* (for predicting the total number of absence days), *WeekWNYVacationLeave* (for predicting the number of vacation leave days), or *WeekWNYSickLeave* (for predicting the number of sick leave days).

Construction of the experimental datasets For the purpose of analysis, we construct two types of datasets: (1) the first type contain worker profile and timesheet absence profiles as descriptive attributes (see Figure 1a); and (2) the second type includes also timesheets absence aggregates (see Figure 1b).

In order to perform analysis, we need to properly construct the datasets used for learning predicting models. For example, if we want to predict workers absence for week 15, we use historical timesheets data from week 1-14 together with the aggregates calculated on this period as descriptive attributes.

We decided to split the year consisting of 52 weeks in four quarters (Q1: W1-W13, Q2: W14-W26, Q3:W27-W39, Q4:W40-W52), each containing 13 weeks. The absence data for the first 12 weeks were used as historical timesheet profiles, out of which the aggregate attributes were calculated. The absence of the 13^{th} week was used a target attribute. For each quarter, we constructed two different variants of datasets, one containing the aggregate attributes and the other without the aggregate attributes. This procedure was done for both tasks: binary classification and regression.

Experimental setup. For our paper, we used Weka as main software [2] to execute predictive modelling experiments. WEKA is an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that one can develop machine learning techniques and apply them to real-world data mining problems. In the experiments, for all methods we used the default method settings from Weka mining software. The evaluation method used was 10 fold cross-validation.

Methods. Here, we used different predictive methods implemented in the WEKA software with different settings. For the regression task, we compare the performance of the following methods Linear regression (LR), M5P (both regression and model trees)[3], RandomForest (RF) [4] with M5P trees as base learners, Bagg (Bag) [5] having M5P trees as base learners, IBK (nearest neighbour classifier with different number of neighbours) [6] and SMOreg (support vector regression) [7].

For binary prediction, we compare the performance of the following methods: jRIP (decision rules) J48 (decision trees) RandomForest (RF), Bagging (Bagg) having J48 trees as base learners, RandomSubSpace (RS) [8] having J48 trees as base learners, SMO (support vector machines) [9], and IBK (nearest neighbour classifier with different number of neighbours).

Evaluation measures. To answer our research question for the case of regression, we use several measures for regression analysis, such as: Mean Absolute Error (MAE), Root mean squared error (RMSE), and Correlation coefficient (CC).

For the case of classification, we use several measures for classification analysis, such as: the percentage of correctly classified instances (classification accuracy), precision, and recall.

Table 4: Predictive performance results. The bold value denotes the highest value when we compare datasets with (A) or without (NA) added aggregate attributes. The gray cells denote the best performing method for each dataset.

(a) Performance results for the regression task - RMSE measure (less is better)

Dataset	LR	MP5	M5P-R	RF	Bagg	IBK(K=1)	IBK(K=3)	IBK(K=7)	SMOreg
Q1-A	0.789	0.692	0.775	**0.688**	**0.64**	0.804	**0.687**	0.734	0.681
Q1-NA	**0.723**	**0.674**	**0.767**	0.729	0.647	**0.798**	0.693	**0.724**	**0.659**
Q2-A	1.692	**1.369**	1.422	**1.412**	1.438	1.894	**1.476**	**1.382**	1.617
Q2-NA	**1.44**	1.382	**1.396**	1.457	**1.379**	**1.752**	1.506	1.425	**1.497**
Q3-A	0.942	**0.919**	0.976	0.999	0.935	1.409	1.074	**1.015**	**0.963**
Q3-NA	**0.911**	0.929	**0.956**	**0.968**	**0.927**	**1.223**	**1.046**	1.017	0.969
Q4-A	**0.977**	**0.947**	**0.961**	**0.923**	**0.922**	1.222	**1.029**	1.005	**0.984**
Q4-NA	0.992	0.985	0.976	1.024	0.975	**1.186**	1.066	**0.999**	1.007

(b) Performance results for the classification task - Accuracy in% (more is better)

Dataset	JRip	j48	RF	Bagg	RS	SMO	IBK(K=1)	IBK(K=3)	IBK(K=7)
Q1-A	**87.429**	**90.810**	90.357	**90.833**	89.881	**92.762**	87.452	**91.810**	90.810
Q1-NA	**87.429**	**90.810**	**90.381**	89.857	**90.357**	90.833	**89.429**	**91.810**	**90.833**
Q2-A	63.645	**68.879**	**65.751**	65.419	**66.736**	69.200	58.153	**64.347**	**68.842**
Q2-NA	**66.466**	68.177	67.118	**66.441**	66.429	**66.773**	**65.049**	62.291	67.463
Q3-A	**84.429**	**84.404**	**83.288**	83.061	84.409	**86.677**	77.182	82.616	**85.333**
Q3-NA	83.737	83.520	82.379	**83.737**	**84.864**	86.449	**81.263**	**85.101**	84.879
Q4-A	**71.130**	67.277	**72.150**	**70.460**	**70.305**	**70.452**	**69.627**	**70.644**	**70.302**
Q4-NA	70.455	**68.266**	66.774	67.441	69.791	69.466	66.093	67.610	68.960

4 RESULTS AND DISCUSSION

Regression task[1]**.** In Table 4a, we present the results for RMSE measure. It indicates how close the observed data points are to the model's predicted values, and lower values indicate better fit. From the results, we can observe that in general Bagging of M5P trees obtains the best performance. Predicting absence in week 13 from Q1 is generally better without using aggregate attributes. We have similar behaviour for predicting absence in week 26 (Q2) and week 39 (Q3). Predicting absence for the last week in the year from Q4 is generally better done using additional aggregate attributes. If we consider MAE, the best performing method is SMOreg, and for Q1, Q2 better results are obtained without the use of aggregate attributes, opposite to the Q3 and Q4. Finally, if we consider CC the best performing method is Bagging, and for Q1 and Q4 better results are obtained without using aggregate attributes, opposite to Q2 and Q3.

Classification task[2]**.** In Table 4b, we present the results for accuracy. From the results, we can observe that in general SMO obtains the best performance. For Q1, we obtain better results if we do not include aggregate attributes. For Q2, Q3 and Q4 the best results are obtained by using the additional aggregate attributes. If we consider precision the best performing methods are SMO and JRip, while for recall the best performing method is IBK using 7 nearest neighbours.

5 CONCLUSION AND FUTURE WORK

The main goal of the paper was to test if adding additional timesheet aggregate attributes can influence the predictive power in the case of one-week ahead absenteeism prediction from timesheet data. The research was performed on data from year 2019, collected by the MojeUre work attendance register system. We used various predictive modelling methods formulating the prediction task as regression (predicting the number of absent days in a week) and classification (predicting if an employee will be absent in a given week). To see the difference in performance, we performed experiments on datasets constructed on different quarters of the year. The best prediction method in the case of regression is Bagging and in general we could say that predictions are slightly better if we don't use aggregate attributes. The best method in the case of classification is SMO. Again almost same results with using or not using external aggregate attributes.

In future work, we plan to perform selective analysis of absenteeism using the same data based on different criteria, such as seasonality, closeness to holidays (before, after), critical weeks for certain professions etc. In addition, we plan to perform regional analysis and workers domain analysis which is based on company type. Moreover, more insight into absence patterns will be available after collecting several years of attendance data for each employee. Finally, we plan to compare the different granularity of prediction (day - based vs. week - based vs. half a month based vs. month based analysis).

[1]Complete results for regression are presented at the following URL https://tinyurl.com/yyp85vfr

[2]Complete results for classification are presented at the following URL https://tinyurl.com/y6o6h6d8

ACKNOWLEDGMENTS

We thank the company 1A Internet d.o.o., which provided us access to the data which were used in our research. Panče Panov is supported by the Slovenian Research Agency grant J2-9230.

REFERENCES

[1] Malisetty, S., Archana, R. V., & Kumari, K. V. (2017). *Predictive analytics in HR management.*, Indian Journal of Public Health Research & Development, 8(3), 115-120.
[2] Witten, I. H., & Frank, E. (2002). *Data mining: practical machine learning tools and techniques with Java implementations.*, Acm Sigmod Record, 31(1), 76-77.
[3] Ross J. Quinlan. *Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence.*, Singapore, 343-348, 1992.
[4] Leo Breiman (2001). *Random Forests.*, Machine Learning. 45(1):5-32.
[5] Leo Breiman (1996). *Bagging predictors.*, Machine Learning. 24(2):123-140.
[6] D. Aha, D. Kibler (1991). *Instance-based learning algorithms.*, Machine Learning. 6:37-66.
[7] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy. *Improvements to the SMO Algorithm for SVM Regression.*, In: IEEE Transactions on Neural Networks, 1999.
[8] Tin Kam Ho (1998) *The Random Subspace Method for Constructing Decision Forests.*, IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(8):832-844. URL http://citeseer.ist.psu.edu/ho98random.html.
[9] J. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization.*, In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.

Monitoring COVID-19 through text mining and visualization

M.Besher Massri
Jožef Stefan Institute, Slovenia
besher.massri@ijs.si

Joao Pita Costa
Quintelligence, Slovenia
joao.pitacosta@quintelligence.com

Andrej Bauer
University of Ljubljana, Slovenia
andrej.bauer@andrej.com

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

Janez Brank
Jožef Stefan Institute, Slovenia
janez.brank@ijs.si

Luka Stopar
Jožef Stefan Institute, Slovenia
luka.stopar@ijs.si

ABSTRACT

The global health situation due to the SARS-COV-2 pandemic motivated an unprecedented contribution of science and technology from companies and communities all over the world to fight COVID-19. In this paper, we present the impactful role of text mining and data analytics, exposed publicly through IRCAI's Coronavirus Watch portal. We will discuss the available technology and methodology, as well as the ongoing research based on the collected data.

KEYWORDS

Text mining, Data analytics, Data visualisation, Public health, Coronavirus, COVID-19, Epidemic intelligence

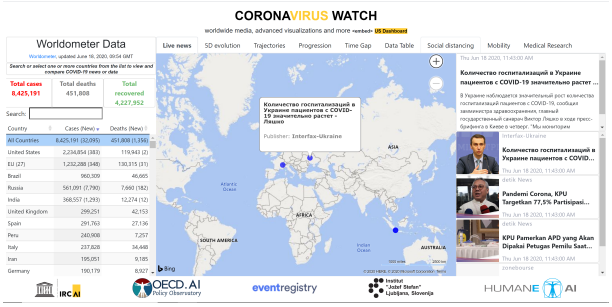

Figure 1: Coronavirus Watch portal

1 INTRODUCTION

When the World Health Organization (WHO) announced the global COVID-19 pandemic on March 11th 2020 [25], following the rising incidence of the SARS-COV-2 in Europe, the world started reading and talking about the new Coronavirus. The arrival of the epidemic to Europe scaled out the news published about the topic, while public health institutions and governmental agencies had to look for existing reliable solutions that could help them plan their actions and the consequences of these.

Technological companies and scientific communities invested efforts in making available tools (e.g. the GIS [1] later adopted by the World Health Organisation (WHO)), challenges (e.g. the Kaggle COVID-19 competition [13]), and scientific reports and data (e.g. the repositories medRxiv [15] and Zenodo [27]).

In this paper we discuss the Coronavirus Watch portal [12], made available by the UNESCO AI Research Institute (IRCAI), comprehending several data exploration dashboards related to the SARS-COV-2 worldwide pandemic (see the main portal in Figure 1). This platform aims to expose the different perspectives on the data generated and trigger actions that can contribute to a better understanding of the behavior of the disease.

2 RELATED WORK

The many platforms that have been made publicly available over the internet to monitor aspects of the COVID-19 pandemics are mostly focusing on data visualization based on the incidence of the disease and the death rate worldwide (e.g., the CoronaTracker [3]). The limitations of the available tools are potentially due to the lack of resolution of the data in aspects like the geographic location of reported cases, the commodities (i.e., other diseases that also influence the death of the patient), the frequency of the data, etc. On the other hand, it was not common to monitor the epidemic through the worldwide news (with some exceptions as the Ravenpack Coronavirus News Monitor [21]).

The Coronavirus Watch portal suggests the association of reported incidence with worldwide published news per country, which allows for real-time analysis of the epidemic situation and its impact on public health (in which specific topics like mental health and diabetes are important related matters) but also in other domains (such as economy, social inequalities, etc.). This news monitoring is based on state-of-the-art text mining technology aligned with the validation of domain experts that ensures the relevance of the customized stream of collected news.

Moreover, the Coronavirus Watch portal offers the user other perspectives of the epidemic monitoring, such as the insights from the published biomedical research that will help the user to better understand the disease and its impact on other health conditions. While related work was promoted in [13] in relation with the COVID-19, and is offered in general by MEDLINE mining tools (e.g., MeSH Now [16]), there seems to be no dedicated tool to the monitoring and mining of COVID-19 - related research as that presented here.

3 DESCRIPTION OF DATA

3.1 Historical COVID-19 Data

To perform an analysis of the growth of the coronavirus, we need to use the historical data of cases and deaths. This data is retrieved from a GitHub repository by John Hopkins University[4]. The data source is based mainly on the official data from the World Health Organization (WHO)[24] along with some other sources, like the Center for Disease and Control[2], and Worldometer[26], among others. This data provides the basis for all functionality that depended on the statistical information about COVID-19 numbers.

Information society '20, October 5–9, 2020, Ljubljana, Slovenia

3.2 Live Data from Worldometer

Apart from historical data, live data about the COVID-19 number of cases, deaths, recovered, and tests are retrieved from the worldometer website. Although the cases might not be as official as the one provided by John Hopkins University (which is based on WHO data), this source is updated many times per day providing the latest up-to-date data about COVID-19 statistics at all times.

3.3 Live News about Coronavirus

The live news is retrieved from Event Registry [10], which is a media-intelligence platform that collects news media from around the world in many languages. The service analyzes news from more than 30,000 news, blogs, and PR sources in 35 languages.

3.4 Google COVID-19 Community Mobility Data

Google's Community Mobility [11] data compares mobility patterns from before the COVID-19 crisis and the situation on a weekly basis. Mobility patterns are measured as changes in the frequency of visits to six location types: Retail and recreation, Grocery and pharmacy, Parks, Transit stations, Workplaces, and Residential. The data is provided on a country level as well as on a province level.

3.5 MEDLINE: Medical Research Open Dataset

The MEDLINE dataset [14] contains more than 30 million citations and abstracts of the biomedical literature, hand-annotated by health experts using 16 major categories and a maximum of 13 levels of deepness. The labeled articles are hand-annotated by humans based on their main and complementary topics, and on the chemical substances that they relate to. It is widely used by the biomedical research community through the well-accepted search engine PubMed [19].

4 CORONAVIRUS WATCH DASHBOARD

The main layout of the dashboard displayed in figure 1 consists of two sides. It is split into the left table of countries, where a simple table of statistics is provided about countries along with the total numbers of cases, deaths, and recovered. On the right side, there is a navigation panel with tabs, each representing a functionality. Each functionality answers some questions and provides insights about a certain type of data.

4.1 Coronavirus Data Table

The data table functionality is a simple table that shows the basic statistics about the new coronavirus. It's taken from Worldometer as it's the most frequently updated source for coronavirus. The data table comes in two forms, one that is a simplified version which is the table on the left, and one contains the full information in a separate tab.

4.2 Coronavirus Live News

The second functionality is a live news feed about coronavirus from around the world. The feed comes from Event Registry, which is generated by querying for articles that are annotated with concepts and keywords related to coronavirus. The user can check for a country's specific news (news source in that country) by clicking on the country name on the left table. As seen in figure 1.

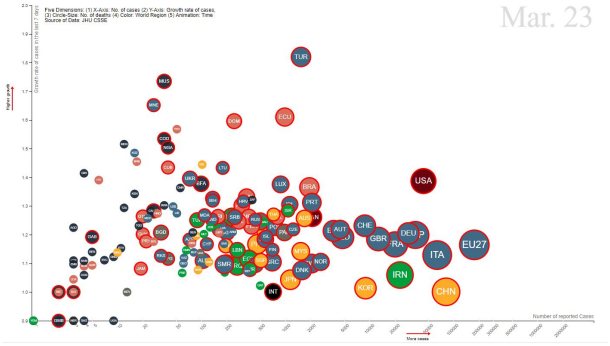

Figure 2: A snapshot of the 5D Visualization on March 23rd. Countries that were at the peak in terms of growth are shown high up like Turkey. Whereas countries that mostly contained the virus are shown down like China.

4.3 Statistical Visualizations

The following set of visualization all aims at displaying the statistics about COVID-19 cases and deaths in a visual format. While they all provide countries comparison, each one focus on different perspective; Some are more complex and focus on the big picture (5D evolution), and some are simple and focus on one aspect (Progression and Trajectory). Besides, all of them have configuration options to tweak the visualization, like the ability to change the scale of the axes to focus on the top countries or the long tale. Or a slider to manually move through the days for further inspection. Furthermore, the default view compares all the countries or the top N countries, depending on the visualization. However, it's possible to track a single country or a set of countries and compare them together for a more focused view. This is done by selecting the main country by clicking on it on the left table and proceeding to select more countries by pressing the ctrl key while clicking on the country.

4.3.1 5D Evolution. 5D Evolution is a visualization that displays the evolution of the virus situation through time. It is called like that since it encompasses five dimensions: x-axis, y-axis, bubble size, bubble color, and time, as seen in figure 2. By default, it illustrates the evolution of the virus in countries based on N. cases (x-axis), The growth factor of N. Cases (y-axis), N. Deaths (bubble size), and country region (bubble color) through time. In addition, a red ring around the country bubble is drawn whenever the first death appears. The growth rate represents how likely that the numbers are increasing with respect to the day before. A growth rate of 2 means that the numbers are likely to double in the next day. The growth rate is calculated using the exponential regression model. At each day the growth rate is based on the N. cases from the previous seven days. The goal of this visualization to show how countries relate to each other and which are exploding in numbers and which ones managed to "flatten the curve", since flattening the curve means less growth rate. It's intended to be one visualization that gives the user a big picture of the situation.

4.3.2 Progression. The progression visualization displays the simple Date vs N. cases/deaths line graph. It helps to provide a simplistic view of the situation and compare countries based on the raw numbers only. The user can display the cumulative

numbers where each day represents the numbers up to now, or daily where at each date the numbers represent the cases/deaths on that day only.

4.3.3 Trajectory. While the progress visualization displays the normal date vs N. cases/deaths, this visualization seeks to compare how the trajectory of the countries differ starting from the point where they detect cases. This visualization helps to compare countries' situations if they all start having cases on the same date. The starting point has been set to the day the country reaches 100 cases, so we would compare countries when they started gaining momentum.

4.4 Time Gap

The time gap functionality tries to estimate how the countries are aligned and how many days each country is behind the other, whether that is in the number of cases or deaths. This assumes that the trajectory of the country will continue as it with taking much more strict/loose measurements, which is a rough assumption. It helps to estimate how bad or good the situation in terms of the number of days. To see the comparison, a country has to be selected from the table on the left. However, not all countries are comparable as they have very different trajectories or growth rates.

The growth of each country is represented as an exponential function, the base is calculated using linear regression on the log of the historical values (that is, exponential regression). Based on that, the duplication N. days, or the N. days the number of cases/deaths will double is determined. two countries are comparable if they have a reasonable difference is the base or doubling factor. If they are comparable, we see where the country with the smaller value fits in the historical values of the country with the larger numbers, with linear interpolation if the number is not exact, hence the decimal values.

4.5 Mobility

The mobility visualization is based on google community mobility data that describe how communities in each country are moving based on 6 parameters: Retail and recreation, Grocery and pharmacy, Parks, Transit stations, Workplaces, and Residential. The data is then reduced to 2-dimensional data while keeping the Euclidean proximity nearly the same. The visualization can indicate that the closer the countries are on the visualization, the similar the mobility patterns they have. The visualization uses the T-SNE algorithm for dimensionality reduction [23], which reduces high dimensional data to low dimensional one while keeping the distance proximity between them proportionally the same as possible. The algorithm works in the form of iterations, at each iteration, the bubbles representing the country are drawn. We used those iterations to provide animation to the visualization.

4.6 Social Distancing Simulator

The Social Distancing simulator is displayed in figure 3. Each circle represents a person who can be either healthy (white), immune (yellow), infected (red), or deceased (gray). A healthy person is infected when they collide with an infected person. After a period of infection, a person either dies or becomes permanently immune. Thus the simulation follows the Susceptible-Infectious-Recovered-Deceased (SIRD) compartmental epidemiological model.

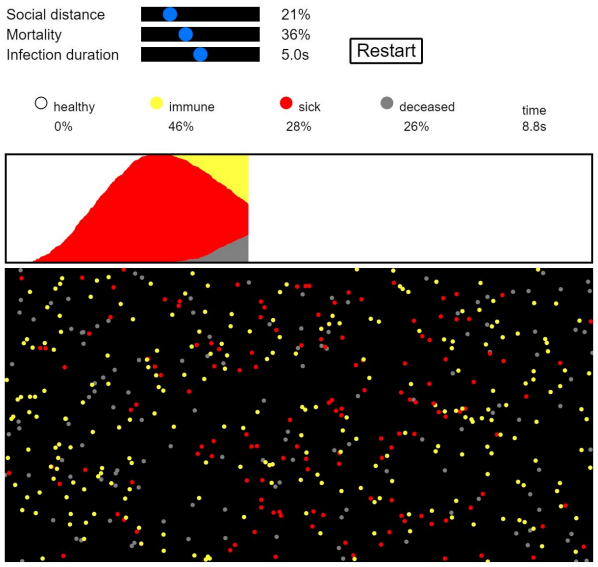

Figure 3: A snapshot of the Social Distancing Simulator. The canvas show a representation of the population. with red dots representing sick people, yellow dots representing immunized people, and grey dots represent deceased people.

The simulator is controlled by three parameters. First, Social distancing that controls to what extent the population enforces social distancing. At 0% there is no social distancing and persons move with maximum speed so that there is a great deal of contact between them. At 100% everyone remains still and there is no contact at all. Second, mortality is the probability that a sick person dies. If you set mortality to 0% nobody dies, while the mortality of 100% means that anybody who catches the infection will die. Finally, infection duration determines how long a person is infected. A longer time gives an infected person more opportunities to spread the infection. Since the simulation runs at high speed, time is measured in seconds.

4.7 Biomedical Research Explorer

To better understand the disease, the published biomedical science is the source that provides accurate and validated information. Taking into consideration a large amount of published science and the obstacles to access scientific information, we made available a MEDLINE explorer where the user can query the system and interact with a pointer to specify the search results (e.g., obtaining results on biomarkers when searching for articles hand-annotated with the MeSH class "Coronavirus").

To allow for the exploration of any health-related texts (such as scientific reports or news) we developed an automated classifier [5] that assigns to the input text the MeSH classes it relates to. The annotated text is then stored in Elasticsearch [18], from where it can be accessed through Lucene language queries, visualized over easy-to-build dashboards, and connected through an API to the earlier described explorer (see [8], [20] and [17] for more detail).

The integration of the MeSH classifier with the worldwide news explorer Event Registry allows us to use MeSH classes in the queries over worldwide news promoting an integrated health news monitoring [9] and trying to avoid bias in this context [7]. An obvious limitation is a fact that the annotation is only

available for news written in the English language, being the unique language in MEDLINE.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented the coronavirus watch dashboard as a use-case of observing pandemic. However, this methodology can be applied to other kinds of diseases given the availability of similar data. For further development, we plan to implement a local dashboard for other countries as well which would provide local data in the local language. In addition, given the existence of more than seven months of historical data, we would like to build some predictive models to predict the number of cases/deaths in the next few days.

Moreover, we are using the StreamStory technology [22] in order to: (i) compare the evolution of the disease between countries by comparing their time-series of incidence; (ii) investigate the correlation between the incidence of the disease with weather conditions and other impact factors; and (iii) analyze the dynamics of the evolution of the disease based on incidence, morbidity, and recovery. This technology allows for the analysis of dynamical Markov processes, analyzing simultaneous time-series through transitions between states, offering several customization options and data visualization modules.

Furthermore, following the work done in the context of the Influenza epidemic in [6], we are using Topological Data Analysis methods to understand the behavior of COVID-19 throughout Europe. In it, we examine the structure of data through its topological structure, which allows for comparison of the evolution of the epidemics within countries through the encoded topology of their incidence time series.

ACKNOWLEDGMENTS

The first author has been supported by the Knowledge 4 All foundation and the H2020 Humane AI project under the European research and innovation programme under GA No. 761758), while the second author was funded by the European Union research fund 'Big Data Supporting Public Health Policies', under GA No. 727721. The third author acknowledges that this material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0326.

REFERENCES

[1] ArcGIS. 2020. ArcGIS who covid-19 dashboard. https://covid19.who.int/. (2020).
[2] CDC. 2020. Center for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/index.html. (2020).
[3] CoronaTracker. 2020. CoronaTracker. https://www.coronatracker.com/analytics/. (2020).
[4] CSSE. 2020. Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university. https://github.com/CSSEGISandData/COVID-19. (2020).
[5] J. Pita Costa et al. 2020. A new classifier designed to annotate health-related news with mesh headings. *Artificial Intelligence in Medicine*.
[6] J. Pita Costa et al. 2019. A topological data analysis approach to the epidemiology of influenza. In *Proceedings of the Slovenian KDD conference*.
[7] J. Pita Costa et al. 2019. Health news bias and its impact in public health. In *Proceedings of the Slovenian KDD conference*.
[8] J. Pita Costa et al. 2020. Meaningful big data integration for a global covid-19 strategy. *Computer Intelligence Magazine*.
[9] J. Pita Costa et al. 2017. Text mining open datasets to support public health. In *WITS 2017 Conference Proceedings*.
[10] EventRegistry. 2020. Event Registry. https://eventregistry.org. (2020).
[11] Google. 2020. Google COVID-19 Community Mobility Report. https://www.google.com/covid19/mobility/. (2020).
[12] IRCAI. 2020. IRCAI coronavirus watch portal. http://coronaviruswatch.ircai.org/. (2020).
[13] Kaggle. 2020. Kaggle covid-19 open research dataset challenge. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. (2020).
[14] MEDLINE. 2020. MEDLINE description of the database. https://www.nlm.nih.gov/bsd/medline.html. (2020).
[15] medRxiv. 2020. medRxiv covid-19 sars-cov-2 preprints from medrxiv and biorxiv. https://connect.medrxiv.org/relate/content/181. (2020).
[16] MeSHNow. 2020. MeSHNow. https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/MeSHNow/. (2020).
[17] MIDAS. 2020. MIDAS COVID-19 portal. http://www.midasproject.eu/covid-19/. (2020).
[18] Elastic NV. 2020. Elasticsearch portal. https://www.elastic.co/. (2020).
[19] PubMed. 2020. PubMed biomedical search engine. https://pubmed.ncbi.nlm.nih.gov/. (2020).
[20] Quintelligence. 2020. Quintelligence COVID-19 portal. http://midas.quintelligence.com/. (2020).
[21] Ravenpack. 2020. Ravenpack coronavirus news monitor. https://coronavirus.ravenpack.com/. (2020).
[22] Luka Stopar. 2020. StreamStory. http://streamstory.ijs.si/. (2020).
[23] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9, (November 2008), 2579–2605.
[24] WHO. 2020. WHO Coronavirus portal. https://www.who.int/emergencies/diseases/novel-coronavirus-2019. (2020).
[25] WHO. 2020. World Health Organization who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. (2020).
[26] WorldoMeters. 2020. WorldoMeters. https://www.worldometers.info/coronavirus/. (2020).
[27] Zenodo. 2020. Zenodo coronavirus disease research community. https://zenodo.org/communities/covid-19/. (2020).

Usage of Incremental Learning in Land-Cover Classification

Jože Peternelj
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
joze.peternelj@ijs.si

Beno Šircelj
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
beno.sircelj@ijs.si

Klemen Kenda
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
klemen.kenda@ijs.si

ABSTRACT

In this paper we present a comparison of a variety of incremental learning algorithms along with traditional (batch) learning algorithms in an earth observation scenario. The approach was evaluated with the earth observation data set for land-cover classification from Europe Space Agency's Sentinel-2 mission, the digital elevation model and the ground truth data of land use and land cover from Slovenia. We show that incremental algorithms can produce competitive results while using less time than batch methods.

Keywords

remote sensing, earth observation, incremental learning, machine learning, classification

1. INTRODUCTION

Land cover classification is one of the common and well researched tasks of machine learning (ML) in the Earth Observation (EO) community [1]. The challenge is to classify land into different types based on remote sensing data such as satellite images, radar data, information on weather [12] and altitude. The most commonly used data are satellite images, which may vary in acquisition period, resolution or wavelength. A plethora of algorithms have explored the potential of using a single-date image [3] and even time series of images for the task [11, 13]. Extensive work with state-of-the-art accuracy was performed using methods of deep learning [14]. The latter report a high computational effort in the learning and forecasting phase, which reduces their potential for continuous tasks requiring a timely response. There have also been efforts to reduce learning and prediction times using intelligent feature selection [6, 7]. To the best of our knowledge, no cases have been reported where stream models have been used in an EO scenario. The primary purpose of incremental learning would be to reduce the computational cost of classification, regression, or clustering techniques, which, when dealing with large data provided by Sentinel 2 and other sources, can be a significant cost to organizations trying to extract knowledge from that data. One of the advantages of incremental learning is that it is not necessary to load all the data into memory at once when creating a model. We only need to store the model and the part of the data we are processing. This could be especially useful in various EO scenarios, as the data from Copernicus services is estimated to exceed 150PB.

2. DATA

2.1 EO data

The Earth observation data were provided by the Sentinel 2 mission of the EU Copernicus programme, whose main objectives are land monitoring, detection of land use and land changes, support for land cover creation, disaster relief support and monitoring of climate change [2]. The data comprise 13 multi-spectral channels in the visible/near- infrared (VNIR) and short wave infrared (SWIR) spectral range with a temporal resolution of 5 days and spatial resolutions of 10m, 20m and 60m [8]. The Sentinel's Level-2A products (surface reflections in cartographic geometry) were accessed via the services of SentinelHub[1] and processed using `eo-learn`[2] library. Additionally, a digital elevation model for Slovenia (EU-DEM) with 30m resolution[3] was used.

2.2 LULC data

LULC (Land Use Land Cover) data for Slovenia is collected by the Ministry of Agriculture, Forestry and Food and is publicly available [10]. The data is provided in shapefile format, with each polygon representing a patch of land marked with one of the LULC classes. Originally there were 25 classes, but we introduced a more general dataset by grouping similar classes together. The frequencies of 8 newly grouped classes are shown in Figure 1.

2.3 Feature Engineering

The EO data were collected for the whole year. 4 raw band measurements (red, green, blue - RGB and near-infrared - NIR) and 6 relevant vegetation- related derived indices (normalized differential vegetation index - NDVI, normalized differential water index - NDWI, enhanced vegetation index - EVI, soil-adjusted vegetation index - SAVI, structure intensive pigment index - SIPI and atmospherically resistant vegetation index - ARVI) were considered. The derived indices are based on extensive domain knowledge and are used for assessing vegetation properties. One example is the NDVI index, which is an indicator of for vegetation health and biomass. Its value changes during the growth period of the plants and differs significantly from other unplanted

[1] `https://www.sentinel-hub.com/`
[2] `https://github.com/sentinel-hub/eo-learn`
[3] `https://www.eea.europa.eu/data-and-maps/data/eu-dem#tab-original-data`

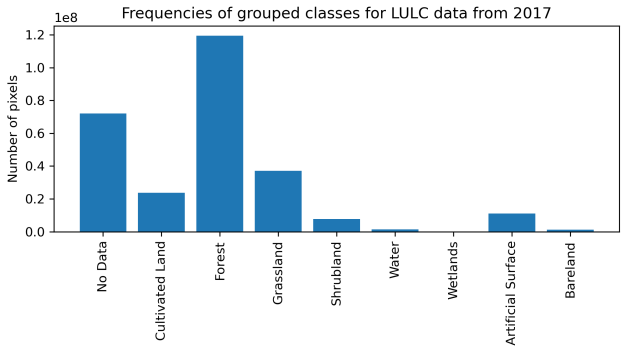

Figure 1: Frequencies of grouped classes for LULC data from 2017 show that the new simplified classification preserves the most common classes separated and merges the less common classes. Classes with the lowest frequencies were selected for oversampling.

areas. The NDVI is calculated as:

$$NDVI = \frac{NIR - red}{NIR + red}$$

Timeless features were extracted based on Valero et al. [11]. These features can describe the three most important crop stages: the beginning of greenness, the ripening period and the beginning of senescence [11, 13]. Annual time series have different shapes due to the phenological cycle of a crop and characterize the development of a crop. With timeless features, they can be represented in a condensed form.

For each pixel, 18 features per each of 10 time series were generated. From elevation data, the raw value and maximum tilt for a given pixel were calculated as 2 additional features. In total 182 features were constructed. From these features only a Pareto-optimal subset of 9 features was selected [6].

3. METHODOLOGY

Classification accuracy **(CA)** and **F1** score were calculated for **11** different ML methods, **6 batch** learning methods and **5 incremental** learning methods. All incremental learning methods are available in the `ml-rapids` (MLR)[4] library which has been developed in order to support the use of incremental learning techniques within eo-learn [4] library.

Hoeffding Tree (incremental)

Hoeffding tree (HT) is an incremental decision tree that can learn from massive streams. It assumes that the distribution of generating examples does not change over time. The Hoeffding tree begins as an initially empty leaf. Each time the new example arrives, the algorithm sorts it down the tree (it updates the internal nodes statistics) until it reaches the leaf. When it reaches the leaf, it updates the leaf statistics of all unused attributes. It then takes the best (A) and second-best (B) attributes based on standard deviation and calculates the ratio of their reductions. To find the best attribute to split a node the Hoeffding bound is used. First algorithm

[4]`https://github.com/JozefStefanInstitute/ml-rapids`

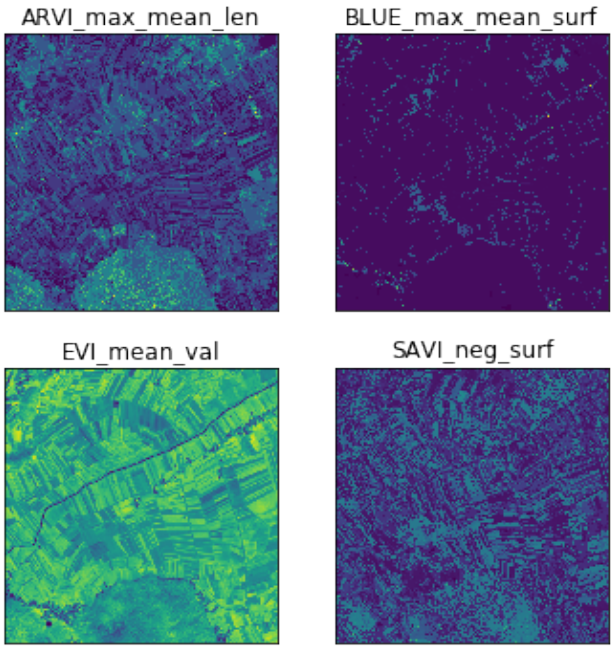

Figure 2: Example of some of the timeless features. `ARVI_max_mean_len` shows the length of maximum mean value in a sliding temporal neighbourhood of ARVI index. `BLUE_max_mean_surf` shows the surface of the flat interval area containing the peak using the blue raw band. `EVI_mean_val` shows mean value of EVI index and `SAVI_neg_sur` shows the maximum surface of the first negative derivative interval of SAVI index.

checks if the ratio is less than $1 - \epsilon$, where $\epsilon = \sqrt{\log \frac{1/\delta}{2n}}$ and $1 - \delta$ is desired confidence. If the ratio is small enough, meaning that attribute A is really better than attribute B, then the algorithm divides the node by that attribute.

Bagging of HT (incremental)

Given a standard training set D of size n, bagging generates m new training sets D_i, each of size n', by uniform sampling from D. Because the sampling is done with replacement, some observations can be repeated in each D_i. If $n' = n$, then for large n the set D_i is expected to have the fraction $(1 - 1/e)(\approx 63.2\%)$ of the unique examples of D, the rest being duplicates. Then, m HT models are fitted using the above m samples and combined by voting. To include a new sample, a random subset of models are selected according to Poisson distribution [9], and these models are updated with the sample in the same way as the HT model described above.

Naïve Bayes (incremental)

Naïve Bayes (NB) is a classification technique based on Bayes's Theorem. It lets us calculate the probability of data belonging to a given class, given prior knowledge. Bayes' Theorem is:

$$P(class|data) = \frac{P(data|class)\ times P(class)}{P(data)}$$

where $P(class|data)$ is the probability of class given the provided data. To add a new training instance, NB only needs to update relevant entries in its probability table.

Logistic Regression (incremental)
Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. A model with two predictors x_1 and x_2 and a binary variable Y, denoted by $p = P(Y = 1)$, which gives us the odds of the values belonging to the class p. The relationship between these terms can be modeled with the following equation:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

The parameters $\beta_0, \beta_1, \beta_2$ can be determined by stochastic gradient descend using logistic loss function.

Perceptron (incremental)
Perceptron is very similar to Logistic regression. It models a binary variable with the same activation function. The only difference is in the cost function that is used for gradient descend.

Batch learning methods
Batch learning methods learn from the whole training set and do not have to rely on heuristics (e.g. Hoeffding bound) or incremental approaches (like SGD) for building the model. The following batch methods have been tested: decision trees, gradient boosting (LGBM), random forest, perceptron, multi-layer perceptron, and logistic regression [5].

4. RESULTS

Results of the experiments are summarised in Figures 3, 4 and Table 1. Figures depict dependency of algorithm-specific F_1 score vs. its training and inference times. An ideal algorithm would be located in the top left corner, achieving full F_1 score with a training and inference time of 0. Any algorithm that has no other algorithm in its top-left quadrant (no algorithm is both more accurate and faster) belongs to a Pareto front, which means that this algorithm is optimal for a certain set of use-cases.

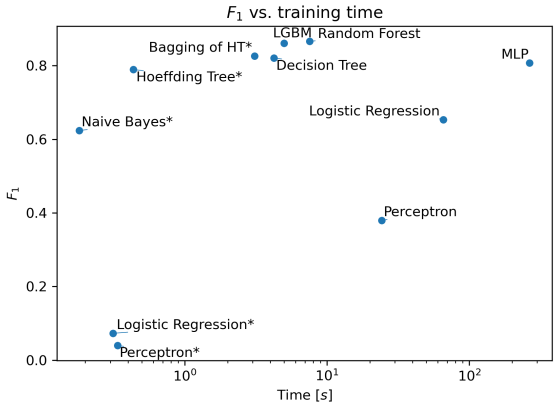

Figure 3: F1 score vs. training time of different models for predicting LULC classes. *Denotes incremental algorithms.

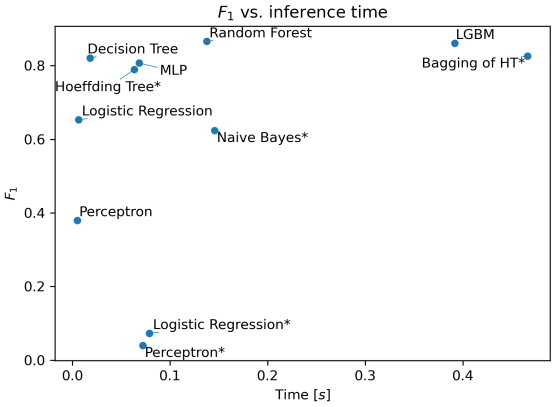

Figure 4: F1 score vs. inference time of different models for predicting LULC classes. *Denotes incremental algorithms.

We can observe that ml-rapid's Naïve Bayes, Hoeffding Tree, Bagging of HT, Decision Trees, LGBM and Random Forest belong to the Pareto optimal set of algorithms according to the training time and F1 score. Regarding inference times Logistic Regression, Decision Trees and Random Forest are the only Pareto optimal algorithms. The choice of algorithm depends on the available processing power and time. For a system that has a lot of time and resources available, it would be best to use Random Forest as it has the highest F1 score. In practice, this is not always feasible. For example, if the algorithm were used for an on-board system on the satellite, we could not afford to save all the data and would prefer to load only the model. With an incremental algorithm, the data could be collected, processed and discarded while the acquired knowledge would be stored in the model. Another preference for HT would be in a wrapper feature selection algorithm [6]. This type of algorithms do a lot of evaluations of the selected method. The main result is a subset of features that can later be used with other algorithms. The acquired set of features might be biased towards the method used, but the results would be obtained much faster.

From the confusion matrix of the HT algorithm shown in Figure 5, we can see that shrubland is often wrongly classified as forest, bareland or grassland and vice versa. This is mainly due to the unclear distinction between these classes (e.g. shrubland can be anything between bareland and forest) and poor ground truth data due to infrequent updates, low accuracy, and lack of detail (e.g. patch of land labeled as shrubland can also grassland and trees). The unclear distinction between certain classes may also explain confusion between wetlands and shrubland or wetlands and grassland, as wetlands may be covered with grass or shrubs. The lack of detail also contributes to misclassification between grassland and artificial surface, as not every small grassy area, such as park or lawn, is included in ground truth data. Finally, grass cultures, unused land overgrown by grass and rotation of crops are likely some of the reasons for confusion between cultivated land and grassland.

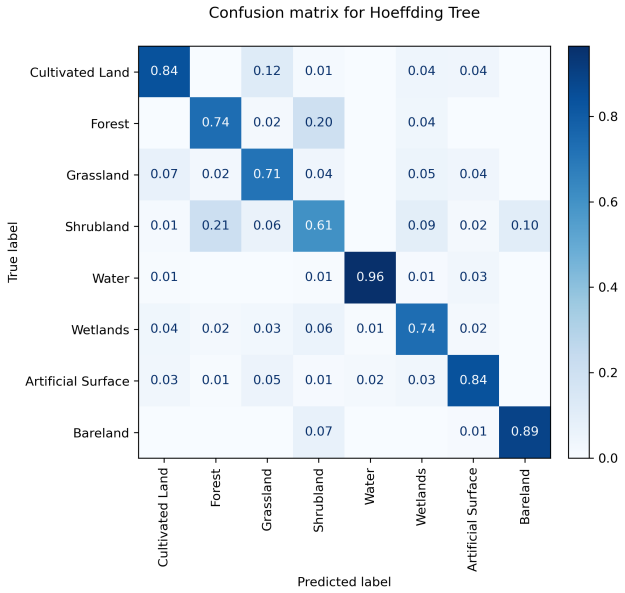

Figure 5: Confusion matrix of HT based model for predicting LULC classes.

	Training time	Inference time	CA	F1
LGBM	4.87	0.38	0.86	0.86
Decision Tree	4.18	0.02	0.82	0.82
Random Forest	7.53	0.14	0.87	0.87
MLP	264.67	0.07	0.81	0.81
Logistic Regression	63.50	0.01	0.67	0.65
Perceptron	24.05	0.01	0.45	0.38
Hoeffding Tree*	0.44	0.06	0.79	0.79
Bagging of HT*	3.07	0.46	0.83	0.83
Naïve Bayes*	0.18	0.15	0.64	0.62
Logistic Regression*	0.31	0.08	0.15	0.07
Perceptron*	0.33	0.07	0.14	0.04

Table 1: Comparison of models for predicting LULC classes. *Denotes incremental algorithms.

5. CONCLUSIONS

In our approach we have concentrated on effective processing. Our goal was to provide methods and workflows which can reduce the need for extensive hardware and processing power. Our goal was focused on use cases where a near state-of-the-art accuracy can be achieved with only a fraction of the processing power required by the state-of-the-art. We have researched stream mining algorithms. We have shown that these algorithms, even if they are not the most accurate or the fastest, take their place at the Pareto front in a multi-target environment, which means that some users might find them suitable for their needs and that they provide the best results for particular computational demand.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT program of the EC under project PerceptiveSentinel (H2020-EO-776115) and project EnviroLENS (H2020-DT-SPACE-821918).

7. REFERENCES

[1] D4.7 stream-learning validation report, May 2020. Perceptive Sentinel.

[2] Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment 120* (2012), 25–36.

[3] Gómez, C., White, J. C., and Wulder, M. A. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing 116* (2016).

[4] H2020 PereptiveSentinel Project. Eo-learn library. `https://github.com/sentinel-hub/eo-learn`. Accessed: 2019-09-06.

[5] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[6] Koprivec, F., Kenda, K., and Šircelj, B. Fastener feature selection for inference from earth observation data. *Entropy* (Sep 2020).

[7] Koprivec, F., Peternelj, J., and Kenda, K. Feature Selection in Land-Cover Classification using EO-learn. In *Proc. 22th International Multiconference* (Ljubljana, Slovenia, 2019), vol. C, Institut "Jožef Stefan", Ljubljana, pp. 37–40.

[8] Koprivec, F., Čerin, M., and Kenda, K. Crop Classification using Perceptive Sentinel. In *Proc. 21th International Multiconference* (Ljubljana, Slovenia, 2018), vol. C, Institut "Jožef Stefan", Ljubljana, pp. 37–40.

[9] Oza, N. C. Online bagging and boosting. In *2005 IEEE international conference on systems, man and cybernetics* (2005), vol. 3, Ieee, pp. 2340–2345.

[10] Slovenian ministry of agriculture. Mkgp - portal. `http://rkg.gov.si/`. Accessed: 2020-08-11.

[11] Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O., Dedieu, G., Bontemps, S., Defourny, P., and Koetz, B. Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions. *Remote Sensing 8(1)* (2016), 55.

[12] Čerin, M., Koprivec, F., and Kenda, K. Early land cover classification with Sentinel 2 satellite images and temperature data. In *Proc. 22th International Multiconference* (Ljubljana, Slovenia, 2019), vol. C, Institut "Jožef Stefan", Ljubljana, pp. 45–48.

[13] Waldner, F., Canto, G. S., and Defourny, P. Automated annual cropland mapping using knowledge-based temporal features. *ISPRS Journal of Photogrammetry and Remote Sensing 110* (2015).

[14] Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine 5*, 4 (2017), 8–36.

Predicting bitcoin trend change using tweets

Jakob Jelencic
Artificial Intelligence Laboratory
Jozef Stefan Institute and Jozef International Postgraduate School
Ljubljana, Slovenia
jakob.jelencic@ijs.si

ABSTRACT

Predicting future is hard and challenging task. Predicting financial derivative that one can benefit from is even more challenging. The idea of this work is to use information contained in tweets data-set combined with standard Open-High-Low-Close [OHLC] data-set for trend prediction of crypto-currency Bitcoin [XBT] in time period from 2019-10-01 to 2020-05-01. A lot of emphasis is put on text preprocessing, which is then followed by deep learning models and concluded with analysis of underlying embedding. Results were not as promising as one might hope for, but they present a good starting point for future work.

1. INTRODUCTION

Twitter is an American microblogging and social networking service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, but unregistered users can only read them. Users access Twitter through its website interface, through Short Message Service (SMS) or its mobile-device application software. Tweets were originally restricted to 140 characters, but was doubled to 280 for non-CJK languages in November 2017. People might post a message for a wide range of reasons, such as to state someone's mood in a moment, to advertise one's business, to comment on current events, or to report an accident or disaster [5].

Bitcoin is a cryptocurrency. It is a decentralized digital currency without a central bank or single administrator that can be sent from user to user on the peer-to-peer bitcoin network without the need for intermediaries. Bitcoin is known for its unpredictable price movements, sometimes even to 10% on the daily basis. Bitcoin also serve as an underlying asset for various financial derivatives, which means that one can profit from knowing the future price changes.

Tweets data offer a constant stream of new information about people beliefs about Bitcoin. Since Bitcoin is very volatile asset, without any real-world value, its value is mainly driven by people's trust in it. Which means that possible up or down trends could be predicted by understanding sentiment of people tweets related to Bitcoin and other cryptocurrencies. Tweets data-set is combined with classical Open-High-Low-Close [OHLC] data-set for 5 minute time periods. OHLC data-set contain information about opening and closing price of given time period, its maximum and minimum price during observed time period and sum of volume and number of transactions made [4]. This present additional information how the market is behaving at any given point.

In financial mathematics derivatives are usually modeled with some kind of stochastic process. Most commonly some form of Brownian motion is used. In theory increment in Brownian motion is distributed as $N(\mu, \Sigma)$ independent from previous increment. This implies that prediction of a real time price change of a derivative is not possible, so the target goal should be changed accordingly. Instead of predicting the impossible, the goal of this work is to predict a change in a trend. Trend is calculated with exponential moving average, application of it can be observed in Figure 1.

Definition: Exponential moving average:

$$EMA(TS, n) = \alpha \cdot (\sum_{i=0}^{n-1} (1-\alpha)^i TS_{n-i}),$$

$$\alpha = \frac{2}{n+1}.$$

Figure 1: Example of exponential moving average

	time	tweets	follow	friends	tw1	tw2	tw3	open	high	low	close	volume	trans	ama
211772	2019-10-02 11:50:00	Acquisition Marks Broadridge Financial's First Foray Into Crypto Services #CryptoCurrency #crypto #blockchain https://t.co/OcYrkU3QUf	12557	12094	1674	78778	9080	1.1	4.4	5.6	4.4	154179.3	78	-0.0005918
211777	2019-10-02 11:55:00	Stratis (Oct 02) #STRAT $STRAT #BTC $BTC https://t.co/NkYIIIWkDo 🚀Nash (NEX) about to Mo0n? ⟶ https://t.co/ibHB6cg51p √ https://t.co/xX6cMO4kYj	133665	139314	2450	1846824	5904	4.4	1.4	-0.7	1.4	167407.6	70	0.0169455
211782	2019-10-02 12:00:00	#bitcoin Price Risks Further Decline After Recovery Rally Stalls - CoinDesk #Prices #Markets https://t.co/SQtnAUUXGJ https://t.co/GA458MrfJk	51837	13150	7324	914865	2768	1.4	9.3	5.1	9.3	223545.8	104	0.9513366

Figure 2: Example of working dataset.

2. DATA DESCRIPTION

Collected tweets range from 01-10-2019 to 01-05-2020. We have filtered tweets by crypto-related hashtags. Originally tweets contained multilingual data, but only English one were extracted. Data-set still resulted in more than 5 000 000 tweets over a little more than a half year period. Dealing with such big data-set has proven to be too difficult of a task. But since a lot of tweets are just pure noise, this data-set can be reduced. Idea is to extract the tweets with the largest target audience. Since the data-set contain number of tweet's author friends and followers, we have extracted the tweets with maximum sum of both in a 5 minute period. Unfortunately, crypto world is relatively anonymous, so there is no Warren Buffet alike personalty, to whom we could gave extra weight.

Then we concatenated the reduced tweets with 5-minute OHLC data-set. Snapshot can be observed in Figure 2. Column names should be pretty self-explanatory, expect for "tw1","tw2","tw3", which stands for metadata information about tweets and "ama", which stand for current movement of trend. Continuous features are then normalized, "ama" is shifted one step into the future so it forms the target variable. Regression task has the most success with predictions.

3. TWEETS PROCESSING

Aim of this chapter is to focus on processing tweets. Tweets differ from regular text data, since many of them consist hyperlink, hashtags, abbreviations, grammar mistakes and so on. This excludes any pre-build preprocessing tools, like the one available in deep learning library Tensorflow [1] which is used for building deep learning models. In the Figure 2 we can see an example of some tweets. The cleaning process was executed in the same order as it is stated below. For each tweet the following process was executed:

- Escape characters were removed.
- Tweet was split by " ".
- All non alphanumeric characters were removed, including "#".
- All characters were converted to lower case.
- Usual stop-words were removed.

At this point data-set contain over 200000 different tokens, which is way to sparse for so limited data-set. At this point empirical cumulative distribution function was calculated and all tokens that have less than 50 appearances were removed. The dictionary size is now 2150.

Another thing to consider is how to process numbers that appear in between text. Obviously a separate token for each number is not acceptable, since it would negate all the work it was done so far. The following function was applied to process numbers. 5 more tokens were created and then numbers from a certain interval were assigned corresponding token.

- Small number: $X < 1000$.
- Medium number: $X \in [1000, 10000)$.
- Semi big number: $X \in [10000, 100000)$.
- Big number: $X \in [100000, 1000000)$.
- Huge number: $X \geq 1000000$.

Additional masking token were assigned for missing data. This wrap up dictionary, final length of dictionary is 2156.

Last thing in processing tweets is to handle their length. Not all tweets have the same length. One idea is to take the maximum length of all tweets, then mask the others so they all have the same length. Unfortunately this would take a lot of unnecessary space, which is a problem. Also long tweets does not mean informative tweet. In Figure 3 is plotted the empirical cumulative distribution function of tweets' length.

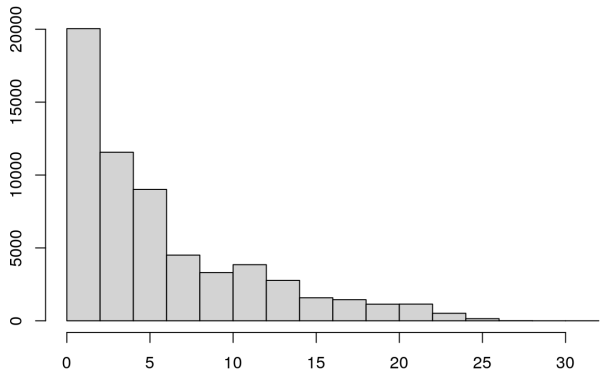

Figure 3: Histogram of tweets' length.

No additional manipulation of tokens were done. It is known that tokens "bitcoin" and "btc" means the same, and they could be join into one token, but they are left intact and the deep learning model will decide either they are the same or not.

4. DEEP LEARNING MODELS

Obvious choice for text models are recurrent neural networks, more specifically Long-Short-term-Memory [LSTM] recurrent networks [2]. They are usually combined with embedding layers, which transform singular token to vector of arbitrary size [6].

Since the task at hand is predicting the future, there is no good benchmark metric or model which could serve as a threshold for our model performance. So in order to see if the tweets can contribute anything, we have decided to build a shallow neural network of just OHLC data which would serve as a benchmark model. 80% of the data-set was taken as a training set, remaining was left out for validation. Split was the same in both models. Both time we used Adam optimizer [3] and mean-squared error [MSE] as a loss function. Training was stopped as soon as validation loss did not improve for 10 epochs. Batch size was 256.

Structure of a benchmark model:

- Input dense layer with 32 neurons.
- Stacked dense layer with 32 neurons.
- Stacked dense layer with 32 neurons.
- Output dense layer with 1 neuron.

Structure of a tweets model:

- Input embedding layer of size 64 (tweets).
- Stacked LSTM layer with 128 neurons.
- Stacked LSTM layer with 128 neurons.
- Second input layer with 64 neurons (OHLC).
- Concatenation.
- Stacked dense layer with 64 neurons.
- Output dense layer with 1 neuron.

Loss process of benchmark model can be observed in Figure 4, while loss process of tweets model can be observed in Figure 5. Orange color represent training set, while blue validation set. It is clear that the tweets model behaved a lot worse on training set than benchmark model, but on test set it has slightly lower MSE (benchmark: 13.78, tweets: 13.74). This implies that there is a lot of reserve in fitting of the tweets model, since the difference between the train and validation loss is so big. That is good since otherwise it seems that tweets do not contribute much for prediction. It is also worth noting that tweets model took way longer to learn, around 380 epochs compared to benchmark's model 40.

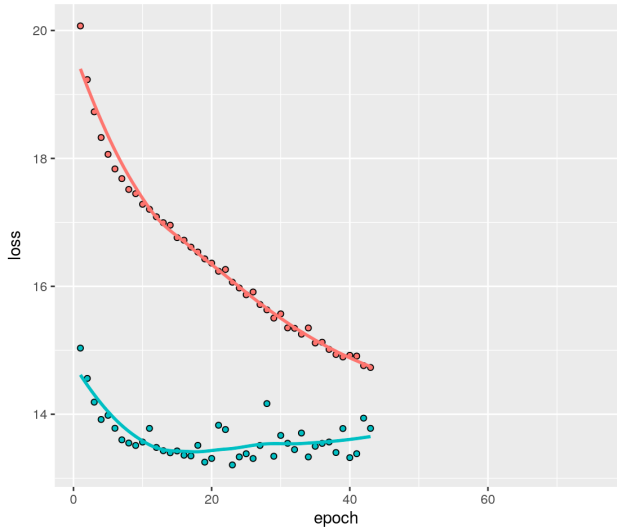

Figure 4: Loss process of benchmark model.

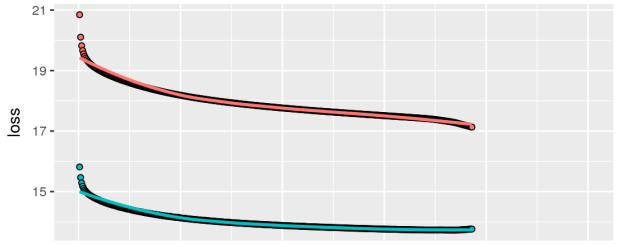

Figure 5: Loss process of tweets model.

5. ANALYSIS OF UNDERLYING EMBEDDING MATRIX

We have extracted underlying embedding matrix from tweets model. Since the model tried to minimize mean-squared error

Figure 6: TSNE projection of embedding matrix.

[MSE] of predicted trend and actual trend, the embedding matrix accordingly to MSE derivative. For analysis we will use cosine similarity as a metric. If 2 words are close in the embedding matrix, this does not mean that they are semantically similar in concept of everyday language, but it means that they are similar in concept of Bitcoin trend prediction. For example if model converged perfectly, and tokens "bitcoin" and "eth" have cosine similarity near 1, that would mean that they both have similar impact on Bitcoin trend. Which is not so hard to believe since it is known that all crypto-currencies are heavily correlated with one another. On Table 1 it can be seen cosine similarity of some of the most common tokens in the dictionary.

Table 1: Cosine similarity pairs of most common tokens.

Tokens Pair	Similarity
bitcoin, crypto	0.472
blockchain, entrepreneur	0.561
crypto, cryptocurrency	0.519
cryptocurrency, blockchain	0.560
volume, social media	0.508
ethereum, blockchain	0.557

We cannot be completely satisfied with results, but for such limited data-set they are not that bad. As it is with any embedding evaluation, it comes to certain amount of subjectivity what is good and what is not.

In order to gain the better perspective of obtained embedding we did a T-distributed stochastic neighbor embedding projection to 2 dimension and plotted 100 nearest pairs. Projection can be observed in Figure 6.

6. CONCLUSION

While the obtained model cannot be served as production model for automatic trading, it presents a nice future work opportunity. We will continue to collect tweets, and hopefully with time build a more accurate data-set and with some hyper-tuning of tweets models achieve improved prediction.

7. ACKNOWLEDGMENTS

This work was financially supported by the Slovenian Research Agency.

8. REFERENCES

[1] *TensorFlow*. `https://www.tensorflow.org/`.

[2] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[3] D. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2014. `https://arxiv.org/abs/1412.6980`.

[4] J. J. Murphy. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. New York Institute of Finance Series. New York Institute of Finance, 1999.

[5] R. Nugroho, C. Paris, S. Nepal, J. Yang, and W. Zhao. A survey of recent methods on deriving topics from twitter: algorithm to evaluation. *Knowledge and Information Systems*, pages 1–35, 2020.

[6] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Series in Artificial Intelligence. Prentice Hall, Upper Saddle River, NJ, third edition, 2010.

Large-Scale Cargo Distribution

Luka Stopar, PhD
Researcher
Jozef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenija
luka.stopar@ijs.si

Luka Bradesko, PhD
Researcher
Jozef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenija
luka.bradesko@ijs.si

Tobias Jacobs, PhD
Senior Researcher
NEC Laboratories Europe GmbH
Kurfürsten-Anlage 36
69115 Heidelberg
tobias.jacobs@neclab.eu

Azur Kurbašić
Researcher
Jozef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenija
azurkurbasic@gmail.com

Miha Cimperman, PhD
Researcher
Jozef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenija
miha.cimperman@ijs.si

ABSTRACT
This study focuses on the design and development of methods for generating cargo distribution plans for large-scale logistics networks. It uses data from three large logistics operators while focusing on cross border logistics operations using one large graph.

The approach uses a three-step methodology to first represent the logistic infrastructure as a graph, then partition the graph into smaller size regions, and finally generate cargo distribution plans for each individual region. The initial graph representation has been extracted from regional graphs by spectral clustering and is then further used for computing the distribution plan.

The approach introduces methods for each of the modelling steps. The proposed approach on using regionalization of large logistics infrastructure for generating partial plans, enables scaling to thousands of drop-off locations. Results also show that the proposed approach scales better than the state-of-the-art, while preserving the quality of the solution.

Our methodology is suited to address the main challenge in transforming rigid large logistics infrastructure into dynamic, just-in-time, and point-to-point delivery-oriented logistics operations.

Keywords
Logistics, graph construction, vehicle routing problem, spectral clustering, optimization heuristics, discrete optimization.

1. INTRODUCTION
The complexity of operations in the logistics sector is growing, so is the level of digitalization of the industry. With data driven logistics, dynamic optimization of basic logistics processes is at the forefront of the next generation of logistics services.

Finding optimal routes for vehicles is a problem which has been studied for many decades from a theoretical and practical point of view: see [2] for a survey. The most prominent case is the Traveling Salesperson Problem (TSP), where the shortest route for visiting *n* locations using a single vehicle has to be determined. What is typically associated with the Vehicle Routing Problem (VRP) is a generalization of TSP where multiple vehicles are available. This class of routing problems is notoriously hard; it not only falls into the class of NP-complete problems, but also in practice it cannot be solved optimally even for moderate instance sizes.

Nevertheless, due to its practical importance, many heuristics and approximation algorithms for the vehicle routing problem have been proposed. Bertsimas et al. propose to an integer programming based formulation of the Taxi routing problem and present a heuristic based on a max-flow formulation, applied in a framework which allows to serve 25,000 customers per hour. A heuristic based on neighborhood search has been presented by Kytöjoki et al. in [4] and evaluated on instances with up to 20,000 customers. A large number of natural-inspired optimization methods have been applied to VRP, including genetic algorithms [7], particle swarm optimization [8], and honey bees mating optimization [9].
The particular approach of partitioning the input graph for VRP has been proposed by Ruhan et al. [5]. Here k-means clustering is combined with a re-balancing algorithm to obtain areas with balanced number of customers. Bent et al. study the benefits and limitations of vehicle and customer based decomposition schemes [6], demonstrating better performance with the latter.

In this paper, we present a methodology for large-scale parcel distribution, by utilizing optimization methods with large graph clustering. The paper is structured as follows. In Section 2, we present the technical details of the proposed methodology. We explain the algorithms and data structures used in each of the steps and discuss the interfaces required to link the steps into a working system. In Section 3, we demonstrate the performance of our methodology on two real-world use cases and compare it to the state-of-the-art on synthetic datasets. Finally, in Section 4 we include key findings, summarizing the strengths and limitations of the proposed approach.

2. METHODOLOGY

2.1 Overview

In this section, we present the details of the proposed methodology for large-scale cargo distribution planning. The methodology, illustrated in Figure 1, uses a three-step, divide and conquer approach to cargo distribution, where we reduce the size of the optimization problem by (i) abstracting the physical infrastructure into a sparse graph representation, (ii) partitioning the graph into smaller chunks (i.e. regions) and (iii) planning the distribution in each region independently. This allows us to run the optimization on large graphs while producing better local results.

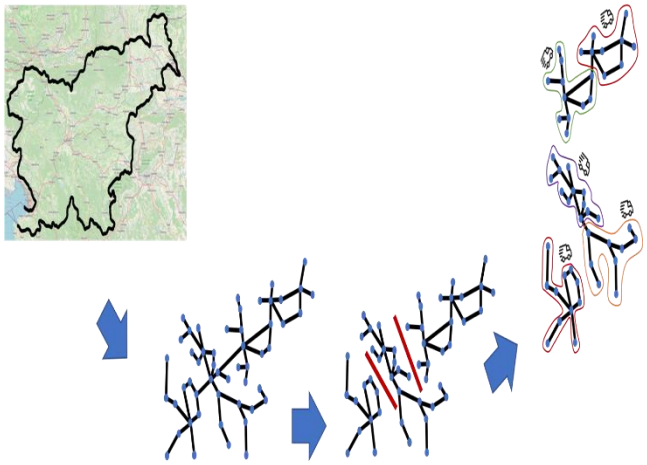

Figure 1: Three step methodology for logistics optimization.

Initially, we create a representation of the physical infrastructure as an abstract graph, representing each pickup and drop-off location as a node with edges as shortest connections on road in between.

Next, we partition the abstract graph with a spectral partitioning approach. The method is an adaptation of [10] to graphs, where we use the first k eigenvalues and eigenvectors of the graphs' Laplacian to construct the partitions. In each partition, we construct a distribution plan using an iterative search algorithm. From an initial solution, the algorithm constructs a linear search path by changing the position of a node in the distribution plan. To avoid local minima, it uses design-time blacklist rules which prevent the algorithm from oscillating in a local neighborhood. Each step is described in more details in the following sections.

2.2 Graph Construction

For graph construction, the Dijkstra SPF algorithm [11] was applied to identify neighbor relationships between the nodes in the OpenStreetMaps (OSM) dataset and construct the graph representation. By mapping post offices to the closest node on OSM, we tag the post office nodes for SPF search.

The search frontier is a baseline for the SPF procedure and represents the list of nodes whose graph neighbors are to be searched. The final graph is built by iterating with the SPF procedure through the list of all post offices in physical infrastructure (graph nodes), and consolidating results into final the sparse matrix – each iteration computes one row of the matrix.

2.3 Graph Partitioning

The partitioning step first represents the graph as a transition rate matrix $(Q)_{ij} = q_{ij}$, where q_{ij} represents the rate of going from node i to node j and is computed as the inverse minimal travel time (obtained from step 1) between the two nodes. With this approach, the rate of going from i to j is represented in terms of the number of possible trips that the driver can make between the two locations in one hour.

The algorithm works by approximating the minimal k-cut of the graph, removing its edges and thus reducing the graph to k disconnected components. We adapt a spectral partitioning algorithm introduced in [10] to graphs.

The algorithm first symmetrizes the transition rate matrix as $Q_s = \frac{1}{2}(Q + Q^T)$, to ensure real-valued eigenvalues, and computes its Laplacian:

$$L = I - diag\left(Q_s\vec{1}\right)^{-1}Q_s$$

Next, it computes the k eigenvectors of L, corresponding to the smallest k eigenvalues. It then discards the eigenvector corresponding to $\lambda_1 = 0$ and assembles eigenvectors $v_2, v_3, \dots, v_k$ corresponding to eigenvalues $\lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_k$ as columns of matrix V. The rows of V are then normalized and used as input to the *k-means* clustering algorithm which constructs the final partitions.

2.4 Vehicle Routing

The vehicle routing step uses *Tabu search* [12] to construct the distribution plan. Starting with an initial solution, *Tabu search* constructs a linear search path by iteratively improving the solution in a greedy fashion until a stopping criterion is met. To avoid converging to local minima, *Tabu search* blacklists recent moves and/or solutions for one or more iterations using design-time rules.

In each iteration, the search process generates new possible solutions by removing a node from its current route and placing it after one of the other nodes in the graph, possibly on a different route. To mitigate scaling problems associated with generating $O(n^2)$ possible moves in each step, the algorithm only considers a handful of moves. Specifically, the probability of considering placing node i after node j is proportional to the inverse of the Euclidean distance $d(i, j)$ between the nodes.

Like other local search algorithms, *Tabu search* starts from an initial feasible solution which is constructed using a construction-based heuristic algorithm. The heuristic procedure iteratively selects a node and places it after one of the other nodes in a way that minimizes the travel distance. The procedure iterates until all values are initialized.

3. DEMONSTRATION AND RESULTS

In this section, we demonstrate the effectiveness of the proposed methodology on two real-world use cases and compare the methodology to the state-of-the-art in vehicle routing. The first pilot included two national logistics operators, namely Hrvatska Posta (Croatia) and Posta Slovenije (Slovenia). As the main focus of future logistics in Europe is to operate as one large homogenous logistics infrastructure, the two infrastructures were considered as one logistics graph. The second pilot included Hellenic Post (Greece) graph representation and data.

In initial testing, simulated data were used for modelling parcel flow with graph abstraction, graph processing, and optimization responses. The final instances were constructed from real infrastructure data to test the functionalities. The results are presented in the following subsections.

3.1 Evaluation on Large Synthetic Graphs

We now demonstrate the scalability of the proposed methodology by comparing its performance to the performance of the baseline *Tabu search* algorithm on synthetic graphs of various sizes, comparing both algorithms' running time and the total travel time in the generated cargo distribution plan. Our results show that the proposed methodology enables fast generation of distribution plans on graphs of up to 10,000 nodes, while also improving the quality of the generated result.

We simulate the logistics infrastructure by generating random planar graphs representing the road network and drop-off locations. First, we generate a cluster of n drop-off locations by sampling a Gaussian distribution around k randomly chosen locations. Next, we connect the locations with Delaunay triangulation [13], resulting in a planar graph. We compute the distance between two locations using the Euclidean metric and assign a $50\ km/h$ speed limit to intra-city edges and a $90\ km/h$ speed limit to inter-city edges. Part of a synthetic graph with 10,000 nodes is shown in Figure **2** below.

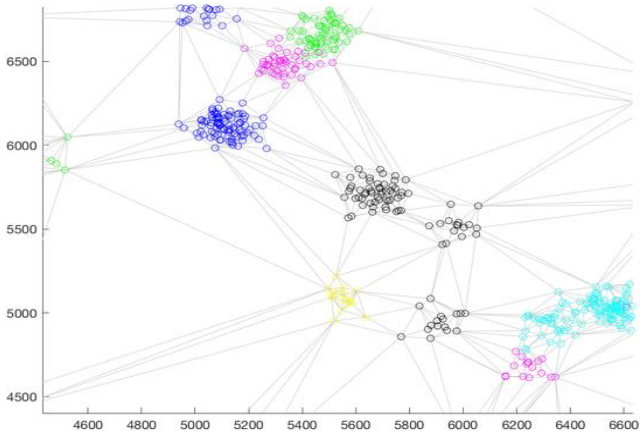

Figure 2: Representation of simulated graph with 10,000 nodes.

Table **1** summarizes the computation times of the proposed method along with the quality of the generated distribution plan and compares the results to *Tabu search* without prior clustering. We measure the quality of the generated distribution plan as the distance travelled by all vehicles according to the plan. In each row, we show the average of 10 trials on 10 different graphs.

Table 1: Comparison of efficiency of Tabu search and proposed methodology.

Graph Size	**Proposed Methodology**		**Tabu search**	
	Running Time	**Travel Distance [km]**	**Running Time**	**Travel Distance [km]**
1000	6.07min	64.7k	0.76min	85.5k
2000	10.07min	122.9k	2.98min	160.8k
5000	30.14min	259.2k	60.04min	428.2k
7000	39.29min	377.9k	166.79min	577.1k
10000	55.64min	552.2k	10.78h	845.1k

For the experiments we used a *Tabu* list with a length of 5% of the entities (locations) that the algorithm must check, and terminated the algorithm when there was no improvement in the solution for more than 10 seconds.

On large graphs, we see that the proposed methodology significantly reduces the computation time while preserving the quality of the result. The proposed methodology reduces the computation time on graphs larger than 5k nodes, providing a substantial saving of 91% on graphs with 10k nodes. We also observe that the quality of the output slightly improved when applying our divide-and-conquer methodology over *Tabu search.* The improvement ranges between 23% and 40% and is largely attributed to the significantly reduced search space in the partitions as compared to the entire graph.

3.2 Testing the instances on pilot use cases

The methods presented and tested on synthetic graphs were also tested on data from two pilot scenarios, namely Slovenian-Croatian post (Pošta Slovenije & Hrvatska Pošta) and Hellenic Post (Greece). In the pilot use cases, the analytical pipeline is used to process ad-hoc events in the logistics infrastructure. The ad-hoc events included were structured into three categories: new parcel request (ad-hoc order), event on distribution objects (vehicle break down) and events related to changes in border crossings – border closed (cross border event).

The instances built on simulated data were loaded with OpenStreetMaps data for abstraction of real infrastructure description into graph representation, as illustrated in Figure 4.

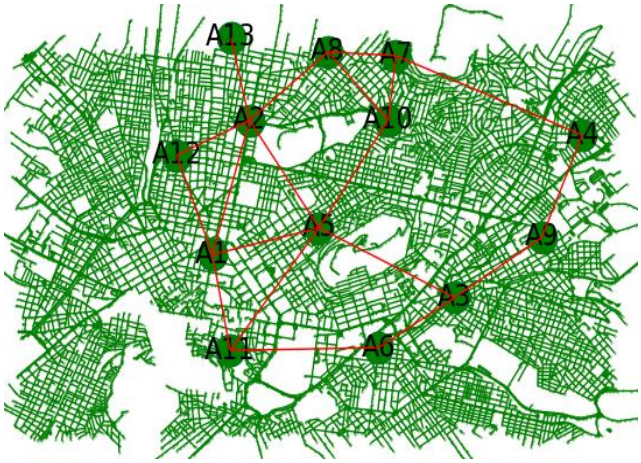

Figure 4: A region of Posta Slovenia graph representation, using OpenStreetMap.

A similar approach was used for the case of Hellenic Post, where the OSM data for the region of Greece were loaded into the graph abstraction instance. For traffic modelling of the vehicles, the SUMO simulator [14] was used with the regional map. For graph manipulations, the SIoT infrastructure was used to generate the social graph when an ad-hoc event was triggered. The social graph represented all entities (vehicles, etc.) in the infrastructure that are in the scope to be included in event processing. In this way, distribution objects were mapped to physical infrastructure for loading the objects into the graph representation for further optimization and distribution plan estimation

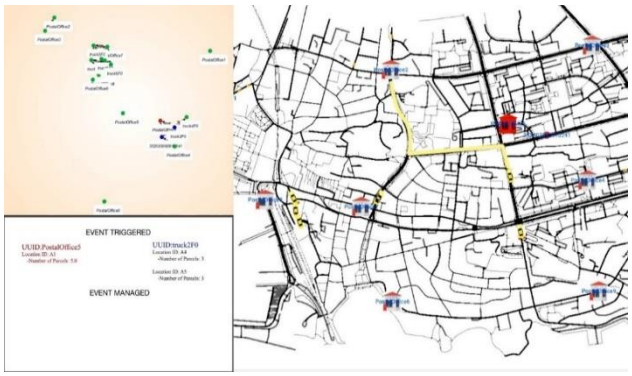

Figure 4: Processing ad-hoc order on a pilot scenario, using SUMO simulator.

An example of the social graph generation and ad-hoc event processing is presented in Figure 4, where a new ad-hoc request is processed by SIoT and analytical pipeline.

The results show that abstracting the logistics infrastructure and clustering the graph into regional structures enabled real-time processing of complex events in the logistics infrastructure. The response time for processing an ad-hoc event in regions of between 50 and 100 nodes was between 20 and 30 seconds. This is relatively fast compared to alternatively processing 1000 nodes or more

4. CONCLUSION

In this paper, we presented an approach for generating cargo distribution plans on large logistic infrastructures. Our results show that the proposed approach can scale to graphs of up to 10,000 nodes in practical time while preserving and even slightly improving the quality of the result.

Since the main use case of logistics is point-to-point regional delivery and just-in-time delivery, these new services are oriented exactly to regional logistics optimization. More importantly, the approach enables to process ad-hoc events, such as new parcel delivery requests, events related to distribution vehicles, or to infrastructure. The ad-hoc event processing includes manipulating the graph representation and running the optimization methods in real-time. Since our method clusters and regionalizes large graphs, such approach can enable real-time processing of events on large graphs, by limiting the changes to the affected regional parts of the infrastructure.

However, while our approach can be combined with several state-of-the-art methods, its main drawback remains the inability to generate inter-region routes, making it suitable only for local and last-mile distribution plans. Future work will focus on investigating the generation of inter-region plans and connecting multiple regions into one distribution plan. Some of the options include introducing border checkpoints where cargo can be handed over to vehicles of neighboring regions, using dedicated inter-region "highway" channels, and using dedicated vehicles for cross-region deliveries.

5. ACKNOWLEDGEMENTS

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 769141, project COG-LO (COGnitive Logistics Operations through secure, dynamic and ad-hoc collaborative networks).

6. REFERENCES

[1] European Commission. (2015). Fact-finding studies in support of the development of an EU strategy for freight transport logistics. Lot 1: Analysis of the EU logistics sector.

[2] Kumar, Suresh Nanda, and Ramasamy Panneerselvam. "A survey on the vehicle routing problem and its variants." (2012).

[3] Bertsimas, Dimitris, Patrick Jaillet, and Sébastien Martin. "Online vehicle routing: The edge of optimization in large-scale applications." Operations Research 67.1 (2019): 143-162.

[4] Kytöjoki, Jari, et al. "An efficient variable neighborhood search heuristic for very large scale vehicle routing problems." Computers & operations research 34.9 (2007): 2743-2757.

[5] He, Ruhan, et al. "Balanced k-means algorithm for partitioning areas in large-scale vehicle routing problem." 2009 Third International Symposium on Intelligent Information Technology Application. Vol. 3. IEEE, 2009.

[6] Bent, Russell, and Pascal Van Hentenryck. "Spatial, temporal, and hybrid decompositions for large-scale vehicle routing with time windows." International Conference on Principles and Practice of Constraint Programming. Springer, Berlin, Heidelberg, 2010.

[7] Razali, Noraini Mohd. "An efficient genetic algorithm for large scale vehicle routing problem subject to precedence constraints." Procedia-Social and Behavioral Sciences 195 (2015): 1922-1931.

[8] Marinakis, Yannis, Magdalene Marinaki, and Georgios Dounias. "A hybrid particle swarm optimization algorithm for the vehicle routing problem." Engineering Applications of Artificial Intelligence 23.4 (2010): 463-472.

[9] Marinakis, Yannis, Magdalene Marinaki, and Georgios Dounias. "Honey bees mating optimization algorithm for the vehicle routing problem." Nature inspired cooperative strategies for optimization (NICSO 2007). Springer, Berlin, Heidelberg, 2008. 139-148.

[10] Ng, Jordan, Weiss. "On Spectral Clustering: Analysis and an algorithm". Advances in Neural Information Processing Systems. MIT Press, 2001. 849-856.

[11] Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271, 1959

[12] Handbook of Combinatorial Optimization, Fred Glover, Manuel Laguna, Vol. 3, 1998

[13] Computational Geometry: Algorithms and Applications, Mark de Berg, Otfried Cheong, Marc van Kreveld, Mark Overmars, Third Edition, 2008

[14] http://sumo.sourceforge.net

Amazon forest fire detection with an active learning approach

Matej Čerin
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
matej.cerin@ijs.si

Klemen Kenda
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
klemen.kenda@ijs.si

ABSTRACT

Wildfires are a growing problem in the world. With climate change, the fires have a larger range an are harder to put down. Therefore it is important to find a way to detect and monitor fires in real-time. In this paper, we explain how we can use satellite images and combine it with knowledge of active learning to get accurate classifier for forest fires. To build the classifier we used active learning like approach. We train the classifier with one labeled image. Then used a classifier to classify the set of images. We manually inspected the images and relabeled wrongly classified examples and build a new classifier. In the paper, we show that in a few iteration steps we can get a classifier that can with good accuracy identify wildfires.

Keywords

remote sensing, earth observation, active learning, rain forest, wildfires, machine learning, feature selection, classification

1. INTRODUCTION

In last years wildfires are a growing problem for the world. Each year the number of forest fires around the world grow. In recent years we had growing number of fires in Amazon, Australia, Africa and Siberia. Because of high global warming and high temperatures, the wildfires have a bigger range and are also harder to put out. Forest fires are partially responsible for the air pollution [12], loss of habitat for animals. Amazon rain forest is also called the lungs of the world, because of oxygen production by the trees. The loos of forest also connects to a higher chance of floods and landslides [6]. Therefore the classification and monitoring of wildfires is an important task. It is important to know the time series of the spread of the fire. With that knowledge we can create models for future fire events, and to plan measures in case of wildfire.

The satellite images are a good source for observation of land type [5]. Therefore they could be used for monitoring forest fires. They can be detected on satellite images, but the area of Amazon is big and it would take a lot of time to manually label burned areas by forest fires. Therefore we should develop an algorithm that can detect fires.

There are already existing algorithms for fire detection using satellite images [6, 11], they inspect changes on satellite images to detect fires. Our solution to that problem is to use machine learning. Because we do not have prepared labeled data-set active learning like approach is our next candidate.

Active learning is the approach used when the labeled data are unavailable, and labeling data is too expensive or time-consuming. The algorithm starts with a small labeled data set and then use its predictions to train itself again. That way the algorithm can learn itself. Algorithms usually need additional input for some data points. In these cases, a human should label those data, and the algorithm can then correct its predictions. The active learning approach is used in many use cases (speech recognition, information extraction, classification, ...). Over the years, it proved to work relatively well [8].

In this paper we use active learning like approach to classify wildfires. By the principle of active learning approach, we label a small subset of data and then train the classifier. Then we manually check the classification results and correct the wrongly classified examples. We then use a new bigger data-set to train the new classifier. We continue with iterations until we are satisfied with the results. That way we can iteratively get a good classifier without labeling huge amounts of data.

2. DATA

2.1 Data Acquisition

In the article, we use data from ESA Sentinel-2 mission [3]. The sentinel-2 mission produces satellite images in 13 different spectral bands with wave lengths of ligt observed from approximately 440 nm to 2200 nm. The spatial resolution is between 10 and 60 m. It consists of two satellites that circle the earth with 180° phase. One point on the earth's surface is visited at least once every five days. In future we could use also use some other satellite data sources like available at www.planet.com [1]. Those data have revisit time of 1 day and might be even better candidate for accurate monitoring of wildfires.

To download data we use eo-learn library [9] that have integrated sentinel-hub[10] library used to access satellite data. Data were downloaded for the year 2019, with a spatial resolution of 30 m. The 30 m resolution was chosen because

burned areas usually extends through much bigger area than 30 m and a therefore higher resolution would not help us identify forest fires. But the processing of each image would take significantly more time than it did now.

2.2 Data Preprocessing

ESA already makes most of the preprocessing steps, like atmospheric reflectance or projection [4]. Therefore data is already clean and ready for use. For our experimentation purposes, we filtered out clouds for that purpose we used models available in eo-learn library.

In our experiments, we used all spectral bands, but the earth observation community developed many different indices that can be calculated from raw spectral bands and use them as a feature in our machine learning experiments. Indices that we used are NDVI, SAVI, EVI, NDWI, and NBR, defined in papers [7, 2]. As our feature vector we used all 13 raw bands and mentioned indices.

3. METHODOLOGY

In our experiments, we iteratively improved the classifier. In each iterative step, we looked at the images and determine if the classification was good or not. To do that most successfully we plotted the images in true color, where the burned area is usually dark, and if the fire is active the smoke is also visible. The other figure that we checked was image with RGB colors plotted Sentinel-2 bands 12, 11, and 3 (false color). Here most of the image is usually in shades of green. The burned area is dark gray color and the area currently burning is yellow or orange (Figure 2). With those two images, we have no problem checking if the area is burned or not.

We experimented with two different approaches. In the first approach, we evaluated the results of classification for each pixel and in the second experiment, we evaluated the average result for a bigger area determined with the clustering algorithm.

The classifier used in our experiment was logistic regression. We used it because it is quite an accurate classifier for earth observation and it can assess how strong the prediction is.

3.1 Experiment 1

First, we manually searched the area of the Amazon forest to find the first satellite image with a forest fire. Then we used that satellite image and labeled 270 pixels as fire area and 270 pixels as not fire area. We trained the logistic regression classifier and used it as our initial classifier in our iteration.

The iteration steps in our experiment were:
1. Use a classifier and classify pixels of a random images of the Amazon rain forest.

2. We took images that the classifier would classify with a forest fire. The images were classified as containing a burned area if at least 3 % of pixels on the image were classified as fire.

3. We checked those images and manually assigned them into two sets (true-positive and false-positive). We checked

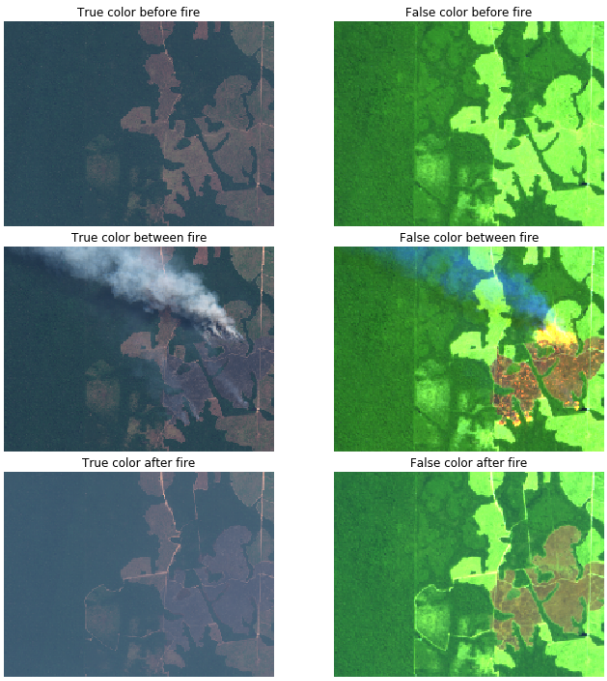

Figure 1: The Figure shows the true color and false-color images of the same area before, during and after the fire. These kinds of images can be used to manually determine burned areas.

only images, where the classifier classified fire. That is because we noticed that the classifier already, in the beginning, finds fire, but it picked up some other areas and objects as fire as well. Therefore we need to find those images and label them as not fire.

4. We used a false-positive set to add to data-set the pixels that the classifier classified wrongly and true positive examples to keep the data-set balanced. We chose in each iteration the two values for the probability of prediction in logistic regression. The first value was used to determine in false-positive images to find pixels that were classified with a probability above that value to add those pixels in the data set. And the second value was used to find pixels that contained forest fire. We changed those values because the algorithm is unreliable in the first iterations and low value in the images with fire would pick up a lot of noise in the data set. But with each iteration the algorithm became more reliable, therefore we could pick lower probability without much noise. The values are shown in the Table 1.

3.2 Experiment 2

The formation of the initial classifier and the first three steps in that experiment were the same as in the first experiment.

Additional steps in the experiment are:
4. For the evaluation of the classifier, we first made clustering with the K-Means algorithm to group similar pixels on each image. The idea of that step is to use a homogeneous group of pixels that probably represent the same ground cower. Those steps are useful because we noticed that K-

Iteration	FP	TP
Iteration 1	0.0	0.80
Iteration 2	0.4	0.70
Iteration 3	0.4	0.70
Iteration 4	0.5	0.60
Iteration 5	0.5	0.60
Iteration 6	0.5	0.50

Table 1: The table shows the values of the minimum average probability of a pixel being burned area for false-positive images (FP) and true-positive images (TP).

Means usually grouped fire areas in one or two clusters. We clustered the pixels in 6 clusters. That number was chosen because on most images that number split the area that way that clusters with fire were separated from not burned area. At the same time it did not split same ground types on too many clusters.

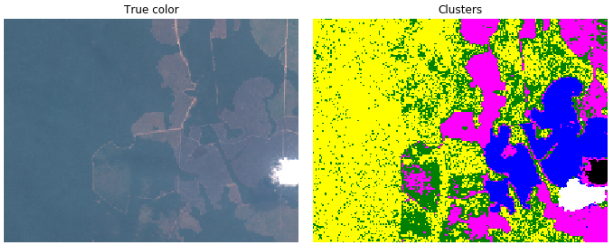

Figure 2: The figure shows how clustering groups different pixels. The burned area is all in one cluster.

5. Calculate the average probability of pixel representing forest fire for each cluster.

6. To choose what pixels to add in the data-set we once again determined two values. They defined above what average pixel probability should cluster have to add pixels from that cluster in the data set. The used values for each iteration are presented in Table 2.

Iteration	FP	TP
Iteration 1	-	0.75
Iteration 2	0.5	0.75
Iteration 3	0.5	0.60
Iteration 4	0.5	0.60
Iteration 5	0.5	0.60
Iteration 6	0.5	0.5

Table 2: The table shows the values of minimum average probability in the cluster for false-positive images (FP) and true-positive images (TP).

4. RESULTS

We tested the classifiers from each experiment on data set form the other experiment. To evaluate results we calculated F1 scores. The results are shown in Table 3.

The F1 scores are relatively high, but those data sets were constructed in a similar way, therefore the scores might be higher than they would be on real images. In both experiments we used random images from the area of amazon, therefore some images might be in both training and testing set.

	F1 score
Classifier from Experiment 1 predicting on data-set from Experiment 2	0.81
Classifier from Experiment 2 predicting on data-set from Experiment 1	0.78

Table 3: The F1 scores of classifiers.

Figure 3 depicts a time-lapse of a wildfire progress. We can see that there are some small noise pixels that are classified wrongly, but they are relatively rare.

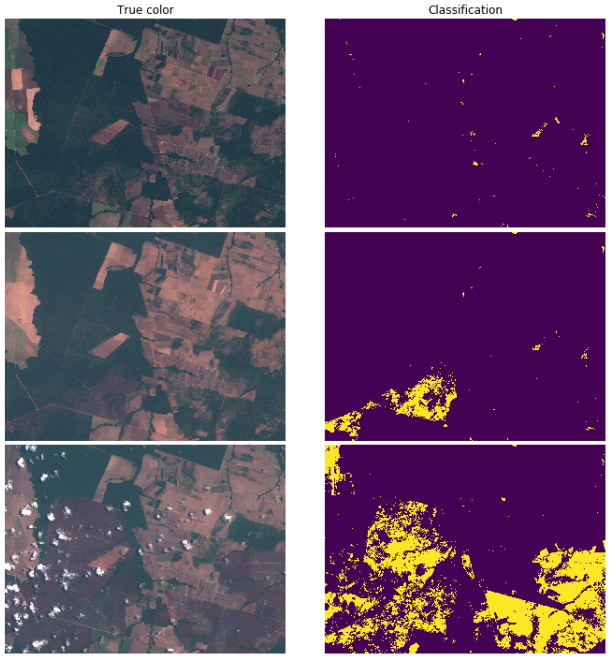

Figure 3: The sub-figures show the development of forest fire. On the left, we have true color satellite images and on the right, we have the classification result with our algorithm. yellow color depicts the burned area.

Another interesting thing to observe in our experiments is what the classifier learned and how it improved in each iteration. We noticed that in the first iterations of our experiments, the classifier did already find fire, but it also picked up many other areas as fire. One of the first improvements of the classifier was that it did not classify water areas (rivers and lakes) as fire. The other later improvements classifier were also some rocky areas. It also improved significantly in the agricultural areas, but in some cases, we could not train classifiers that there is no fire.

The classifier learned wrongly and we could not remove com-

pletely some agricultural areas and some roads in the cities. Most of the agricultural areas were classified correctly, but there were present some fields that no matter what we did were not classified correctly. This might be due to the fact that the field might be on the place that was previously burned and the algorithm still pick that up even though it was not visible from the imagery to us.

5. CONCLUSIONS

The approach with active learning seems promising and we can get relatively good classifiers in a short time. That way we could train a classifier for any classification task of satellite images. With that approach we do not need to check all images as we would if we would like to label all the data by hand. In the end, we get a relatively good classifier.

In this paper, we showed that it is possible in a relatively small number of iterations to get a good and reliable classifier of forest fires. Because satellite images are more accessible in last years than previously it could give us almost real-time insight in the Amazon rain forest.

In the feature one could use other satellite sources with better time-resolution to monitor wildfires. That way we could get more accurate view on the spread of fires.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT program of the EC under projects enviroLENS (H2020-DT-SPACE-821918) and PerceptiveSentinel (H2020-EO-776115). The authors would like to thank Sinergise for their contribution to EO-learn library along with all help with data analysis.

References

[1] https://www.planet.com/. Accessed 1 September 2020 .

[2] Bannari Abdou et al. “A review of vegetation indices”. In: *Remote Sensing Reviews* 13 (Jan. 1996), pp. 95–120. DOI: 10.1080/02757259509532298.

[3] ESA. https://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-2/Satellite_constellation. Accessed 13 August 2018.

[4] ESA. https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/processing-levels/level-2. Accessed 13 August 2018.

[5] Filip Koprivec, Matej Čerin, and Klemen Kenda. “Crop classification using PerceptiveSentinel”. In: (Oct. 2018).

[6] Rosa Lasaponara, Biagio Tucci, and Luciana Ghermandi. “On the Use of Satellite Sentinel 2 Data for Automatic Mapping of Burnt Areas and Burn Severity”. In: *Sustainability* 10 (Oct. 2018), p. 3889. DOI: 10.3390/su10113889.

[7] David Roy, Luigi Boschetti, and S.N. Trigg. “Remote Sensing of Fire Severity: Assessing the Performance of the Normalized Burn Ratio”. In: *Geoscience and Remote Sensing Letters, IEEE* 3 (Feb. 2006), pp. 112–116. DOI: 10.1109/LGRS.2005.858485.

[8] Burr Settles. “Active Learning Literature Survey”. In: (July 2010).

[9] Sinergise. https://github.com/sentinel-hub/eo-learn. Accessed 23 August 2019.

[10] Sinergise. https://github.com/sentinel-hub/sentinelhub-py. Accessed 14 August 2018.

[11] Mihai Tanase et al. “Burned Area Detection and Mapping: Intercomparison of Sentinel-1 and Sentinel-2 Based Algorithms over Tropical Africa”. In: *Remote Sensing* 12 (Jan. 2020), p. 334. DOI: 10.3390/rs12020334.

[12] G. R. van der Werf et al. “Global fire emissions estimates during 1997–2016”. In: *Earth System Science Data* 9.2 (2017), pp. 697–720. DOI: 10.5194/essd-9-697-2017. URL: https://essd.copernicus.org/articles/9/697/2017/.

Indeks avtorjev / Author index

IS 2020

Odkrivanje znanja in podatkovna skladišča • SiKDD

Data Mining and Data Warehouses • SiKDD

Dunja Mladenić, Marko Grobelnik