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.

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

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.

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.

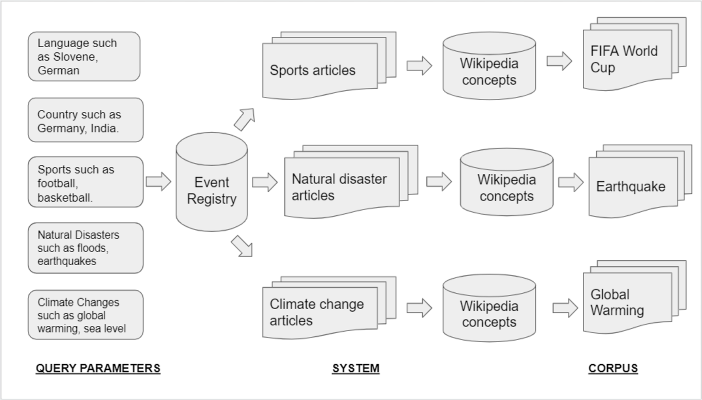
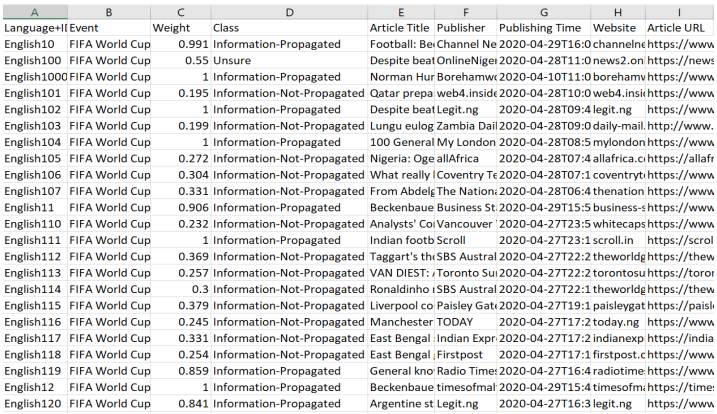

Language such as Slovene, German
Country such as Germany, India.
Sports such as football, basketball.
Natural Disasters such as floods, earthquakes
Climate Changes such as global warming, sea level
Event Registry
Sports articles
Natural disaster articles
Climate change articles
Wikipedia concepts
Wikipedia concepts
Wikipedia concepts
FIFA World Cup
Earthquake
Global Warming
QUERY PARAMETERS
SYSTEM
CORPUS

Figure 1: Data collection methodology



A	B	C	D	E	F	G	H	I
Language+II	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	OnlineNiger	2020-04-28T11:0	news2.on	https://news
English1000	FIFA World Cup	1	Information-Propagated	Norman Hur	Borehamw	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://toror
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 Times	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

[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/

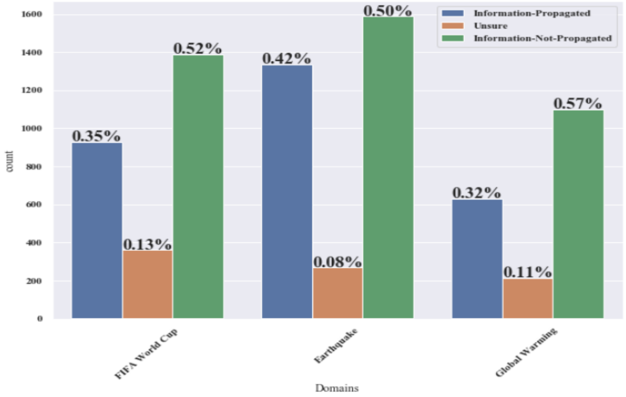


Figure 3: Class distribution for all domains

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.