An approach to creating a time-series dataset for news propagation: Ukraine-war case study

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

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

ABSTRACT

An efficient technique to comprehend news spreading can be achieved through the automation of machine learning algorithms. These algorithms perform the prediction and forecasting of news dissemination across geographical barriers. Despite the fact that news regarding any events is generally recorded as a time-series due to its time stamps, it cannot be seen whether or not the news time-series is propagating across geographical barriers. In this article, we explore an approach for generating time-series datasets for news dissemination that relies on Chat-GPT and sentence-transformers. The lack of comprehensive, publicly accessible event-centric news databases for use in time-series forecasting and prediction is another limitation. To get over this bottleneck, we collected a news dataset consisting of 1 year and 3 months related to the Ukraine war using Event Registry. We also conduct a statistical analysis of different time-series (propagating, unsure, and not-propagating) of different lengths (2, 3, 4, 5, and 10) to document the prevalence of geographical barriers. The dataset is publicly available on Zenodo.

KEYWORDS

news propagation, time-series dataset, geographical barriers, Ukraine-war

1 INTRODUCTION

The process of information traveling from a sender to a set of receivers via a carrier is commonly referred to as propagation [3]. News propagate over time by different publishers about an event. It implicitly raises a few thoughts in our mind, such as: 1) There will be some news articles propagating similar information over time; 2) some news articles will be of a unique category that eventually will not be propagating or propagating across geographical barriers by a few publishers.
News streaming is classified into events where a relevant set of news is clustered and represented as an event [8, 9]. And there is a starting and ending time for an event, which is calculated by the publication time of the first and last news article. Hence, an event consists of a set of news articles, and these news articles follow a certain pattern based on hidden properties including cultural, economical, political, linguistic, and geographical [17].

Information Society 2023, 10 October 2023, Ljubljana, Slovenia

Moreover, news spreading comes across many barriers due to different reasons, including cultural, economic, political, linguistic, or geographical, and these reasons depend upon the type of news, such as sports, health, science, etc. [18]. For instance, it is more likely that the news spreading relating to the FIFA World Cup crosses cultural barriers since it involves multiple cultures. Similarly, news spreading relating to the Sri-Lankan economic crisis and the Ukraine-war probably comes across economic and geographical barriers since these events involve multiple stances from the international community; Eid celebrations and Christmas are likely to come across religious barriers; US elections are likely to come across political barriers [17].
The identification of news spreading patterns while crossing barriers can be useful in the context of numerous real-world applications, such as trend detection and content recommendations for readers and subscribers. To perform the classification of news published across barriers (geographical, cultural, economic, etc.) and, in that attempt, to recommend and identify trends of news spreading belonging to different categories, some methodological considerations are necessary.
In this paper, we introduce an approach to creating a time-series dataset for news propagation. While previous work has focused on creating events from collections of news articles [9, 16], we focus on creating propagation time-series. We take the Ukraine-war as an example to be researched in the propagation analysis across geographical barriers.
Following are the main scientific contributions of this paper:

(1) We present an approach to creating a time-series dataset for news propagation.
(2) A dataset for forecasting and predicting news propagation, that has been labeled with the assistance of Chat-GPT and sentence transformers.

The remainder of the paper is structured as follows. Section 2 describes the related work on barriers to news spreading, time-series datasets for news propagation, and topic modeling. Section 3 presents the proposed approach. We discuss the dataset construction and annotation guidelines in Section 4. The evaluation details and statistical analysis is explained in Section 5, while Section 6 concludes the paper and outlines areas of future work.

2 RELATED WORK

In this section, we review the related literature about geographical barriers to news spreading, time-series datasets for news propagation, and topic modeling.

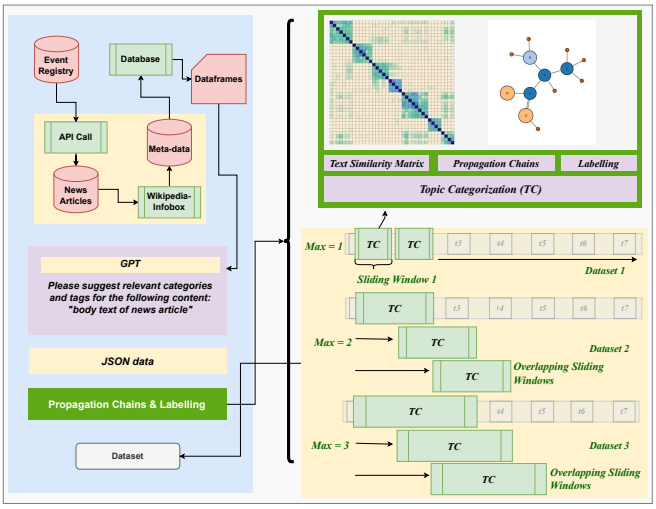

Figure 1: An overview of the proposed approach. To create the propagation time-series, it calculates the semantic similarity across news utilizing sentence transformers, and to evaluate the labeling process of the news, it utilizes a summary of the news articles generated by Chat-GPT.

2.1 Geographical barrier

Sittar reported that the geographical size of a news publisher's country is directly proportional to the number of publishers and articles reporting on the same information [17]. It is also reported that, based on some factors, the media targets specific foreign and regional events. For example, the spreading of news related to specific events may tilt toward developed countries such as the United Kingdom, the U.S.A., or Russia. Also, in the past, geographical representation of entities and events has been extensively utilized to detect local, global, and critical events [10, 20, 19, 2]. It has been said that countries with close distance share culture and language up to a certain extent, which can further reveal interesting facts about shared tendencies in information spreading [12, 11]. Given the difficulty of gathering longitudinal data, relatively little news flow research has systematically examined whether and to what extent foreign nation visibility and the factors that influence it have changed over time. Specifically, scholarship has typically only addressed why some countries get more news coverage than others at a specific point in time, not how and why the focus shifts over time from one country to another [5]. In this context, we propose an approach to collecting data to analyze the news spreading across geographical barriers.

2.2 Time-series datasets

News propagation can be represented in the form of a time-series [17]. The properties of cascading time-series can tell us the relationship between the time and size of cascading. It further answers which events last over a longer period with large communities across different languages. A time-series dataset can be used to understand evolving discussions over time. Different studies have utilized time-series datasets, such as [1] investigates how different discussions evolved over time and the spatial analysis of tweets related to COVID-19. [14] identifies how the discussions evolved over time in top newspapers belonging to three different continents (Europe, Asia, and North America) and nine different countries (UK, India, Ireland, Canada, the U.S.A., Japan, Indonesia, Turkey, and Pakistan). It uses spatio-temporal topic modeling and sentiment analysis. Different classification or mining tasks are proposed using time-series datasets. [6] has proposed the task of predicting stock market values such as price or volatility based on the news content or derived text features. Similarly, to forecast the values, a set of final classes is already defined, such as up meaning an increase in price, down meaning a decrease in price, and balanced meaning no change in price. Also, the same technique has been applied to predict price trends (incline, decline, or flat) immediately after press release publications. Also, Good news articles are categorized as inclines if the stock price relevant to the given article has increased with a peak of at least three points from its original value at the publication time [13].

2.3 Topic modeling

Generally, to find out the most important topics inside an event, multiple solutions have been proposed, including pooling based LDA and BERTopic. Unlike simple static topic modeling, pooling-based techniques assume that the data is partitioned on a time basis, e.g., hourly or daily. Pooling-based techniques are mostly applied to social media, where documents or tweets are partitioned based on hashtags and authors. BERTopic leverages transformers and TF-IDF to create dense clusters, allowing for easily interpretable topics while keeping important words in the topic descriptions. Therefore, the result is a list of topics ranked according to their importance.

The topic modeling techniques are performing surprisingly well. The relation of such topics to their hidden characteristics, such as cultural, economical, and political, has been analyzed in many studies because understanding its dynamics can help governments disseminate information effectively [4, 17, 14, 15]. It has changed rapidly in recent years with the emergence of social media, which provides online platforms for people worldwide to share their thoughts, activities, and emotions and build social relationships [7]. Over the years, scholars have studied the relationship between the news prominence of a country and its physical, economic, political, social, and cultural characteristics [11]. Communication scholars have long been interested in identifying the key determinants of what makes foreign countries newsworthy and why some countries are considered more newsworthy than others [5].

3 APPROACH

This research article presents an approach to creating a time-series dataset for news propagation across geographical barriers, as shown in Figure 1. In the first step, we call an API that extracts the news articles from the Event Registry belonging to Ukraine-war. In the second step, we extract meta-data related to news publishers via searching for the news publishers on Google and extracting their Wikipedia links. Using these links, we obtain the necessary information from Wikipedia-Infobox [17]. We use the Bright Data service to crawl and parse Wikipedia-Infoboxes. In the third step, we perform the summarization of news articles. In the last step, we create a propagation time-series and perform labeling of

the time-series. To calculate the semantic similarity, we utilize monolingual sentence transformers. Since the propagation of information can be captured in the form of time-series we create time-series of different lengths, such as 2, 3, 4, 5, and 10. To evaluate the labeling process, we manually compare the summary generated by Chat-GPT (see Section 5).

4 DATASET CONSTRUCTION

We collected the news articles reporting on the Ukraine-war. Since Russia invaded Ukraine on February 24, 2022, in an escalation of the Russo-Ukrainian War, we fetched news articles that were published between January 2022 and March 2023. The dataset consists of 61261 news articles. Each news article consists of a few attributes: title, body text, name of the news publisher, date, and time of publication.

4.1 Semantic similarity

We calculate the cosine similarity between dense vector generated by sentence transformers. Sentence Transformers is a Python framework for state-of-the-art sentence, text, and image embeddings. Cosine similarity varies between zero and one; zero means no similarity, and one means maximum similarity, i.e., a duplicate article.

4.2 Chat-GPT Summarizing

Since manual evaluation of propagation time-series is difficult because of the length of the news articles, we utilized Chat-GPT to get the tags, categories, and summary representing the whole article. Summarizing a text is one of the many tasks ChatGPT is extremely good at. We can give it a piece of content and ask for a summary. By customizing our prompts, we can get ChatGPT to create much more than a plain summary. We have used the OpenAI API with the Python library. We used the following prompt to fetch the summary of the text, categories, and tags: "Please summarize the text and suggest relevant categories and tags for the following content: articleText:". articleText is a variable representing the text of a news article.

4.3 Annotations of time-series

We created three types of time-series recursively and annotated them based on a threshold of semantic similarity, as shown in Algorithm **??**. The threshold to decide the type of propagation time-series has been set by manually analyzing the similarity and summary of news articles. We set three thresholds for all three types of labels (propagating, unsure, and not-propagating). For instance, the time-series with greater or equal to 0.7 similarity were labeled "Propagating", the time-series with greater or equal to 0.5 similarity were labeled "Unsure", and the time-series with less than 0.5 similarity were labeled "Not-propagating". This criteria has been followed for the minimum length of a time-series (2). However, for the length of a time-series greater than 2, we count the number of pairs with each label, and then the time-series is labeled as one with the highest count. If two labels have the same highest count, then we give priority to the "Propagating" label over "Unsure" and "Unsure" over "Not-Propagating". The Algorithm **??** takes five parameters, such as the start and end of the data-frames, a copy of the data-frames, length of the time-series, and an array. The statistics about the propagation time-series are presented in Figure 2.

To annotate the propagation time-series across geographical barriers, we consider the label "Propagating" for a pair of news articles if the pair is published from two different countries; otherwise, we label it "Not-Propagating". We repeat this process for all lengths of news articles. The statistics after applying this guideline are presented in Figure 3.

5 STATISTICAL ANALYSIS AND EVALUATION

The statistics about the propagation time-series without taking geographical barriers into account are presented in bar chart 2. The number of time-series with the label "Propagating" is higher than the "Unsure", and "Not-Propagating" labels when the length of the time-series is 3 or 5, whereas in the other three cases (2, 4, and 10), the number of time-series is equal for all three labels. The statistics of the propagation time-series that are generated after taking the geographical location of the news publisher into account are presented in bar chart 3. The number of propagation time-series with "Propagated" and "Unsure" labels reduced to almost 40% whereas the number of propagation time-series with the "Not-propagated" label increased significantly.

For the evaluation of the dataset, we have checked the summary, including categories and tags of articles for a specific label, manually. We randomly selected 50 time-series of different lengths for all three types of labels. According to the manual evaluation, the propagation time-series with the "Propagating" label followed almost one or two themes of discussion for all the news articles in a chain. For instance, the following topics have appeared in the propagation time series of length 5: 1) "The United States will be sanctioning Russian President Vladimir Putin; 2) "the national team of the Polish FA will not play against Russia; 3) the Polish Football Association will not play its World Cup qualifying match against Russia; 4) "the Polish Football Association has refused to play a World Cup against Russia; 5) "the Polish national team does not intend to play-off match against Russia". On the contrary, propagation time-series with "Not-Propagating" labels discussed always different points of view about the Ukraine-war. For example, the following topics have appeared in the propagation time-series of length 5: 1) "a resolution passed against Russia in the United Nations"; 2) "Canadian president urges to impose sanctions against Russia"; 3) "the UN Security Council has voted on a US-led draft resolution; 4) "President Trump is inviting Russian President Vladimir Putin to come to Washington; and 5) "India abstained from the vote on the draft resolution". However, in the case of propagation time-series with "Unsure" labels, there were three or four sub-topics discussing the Ukraine-war.

Evaluation results show that as the window size increased to capture the information propagation, the noise of overlapping topics also increased. Similarly, this overlapping window presented sub-topics that overlapped at the time of publication.

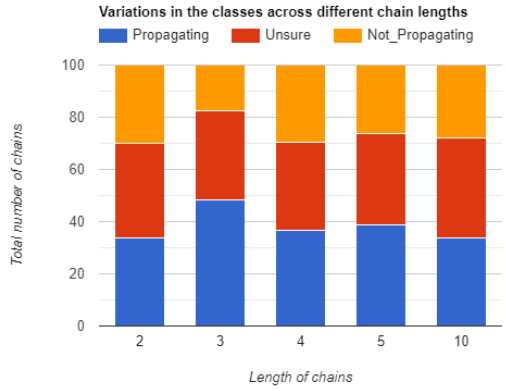

Figure 2: The bar chart shows the statistics about the propagation time-series of different lengths (2, 3, 4, 5, 10) that has been labelled as "Propagating", "Unsure", and "Not-Propagating". The x-axis shows the length of time-series, the y-axis shows the count of the propagation time-series.

.

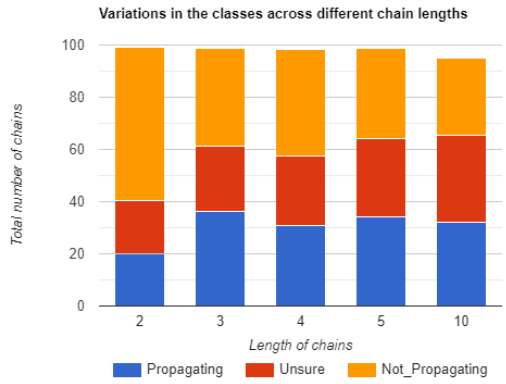

Figure 3: The bar chart shows the statistics about the propagation time-series after applying the condition of the location of a news publisher. Each bar presents three types of propagation time-series that has been labelled as "Propagating", "Unsure", and "Not-Propagating". The x-axis shows the length of time-series, the y-axis shows the count of the propagation time-series.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an approach to creating a time-series dataset. The goal of this work was to investigate the length of the propagation time-series for news propagation. In the future, we plan to utilize the same approach for different events. Moreover, currently, geographical barriers have been analyzed. In the future, we would like to extend the barriers to political, economic, and cultural barriers and find patterns of news propagation. Also, we would like to perform prediction and forecasting on the labeled time-series dataset. We would like to perform experiments with classical time-series classification methods, deep learning, transformer-based methods, and large language models (LLMs).

ACKNOWLEDGMENTS

The research described in this paper was supported by the Slovenian research agency under the project J2-1736 Causalify, National grants (CRP V2-2272; V5-2264; CRP V2-2146) and by the EU's Horizon Europe Framework under grant agreement number 101095095.

REFERENCES

[1] Iyad AlAgha. 2021. Topic modeling and sentiment analysis of twitter discussions on covid-19 from spatial and temporal perspectives. *Journal of Information Science Theory and Practice*, 9, 1, 35–53.

[2] Simon Andrews, Helen Gibson, Konstantinos Domdouzis, and Babak Akhgar. 2016. Creating corroborated crisis reports from social media data through formal concept analysis. *Journal of Intelligent Information Systems*, 47, 2, 287–312.

[3] Firdaniza Firdaniza, Budi Nurani Ruchjana, Diah Chaerani, and Jaziar Radianti. 2021. Information diffusion model in twitter: a systematic literature review. *Information*, 13, 1, 13.

[4] Guoyin Jiang, Saipeng Li, and Minglei Li. 2020. Dynamic rumor spreading of public opinion reversal on weibo based on a two-stage spnr model. *Physica A: Statistical Mechanics and its Applications*, 558, 125005.

[5] Timothy M Jones, Peter Van Aelst, and Rens Vliegenthart. 2013. Foreign nation visibility in us news coverage: a longitudinal analysis (1950-2006). *Communication Research*, 40, 3, 417–436.

[6] Abdullah S Karaman and Tayfur Altiok. 2004. An experimental study on forecasting using tes processes. In *Proceedings of the 2004 Winter Simulation Conference, 2004.* Vol. 1. IEEE.

[7] Sanjay Kumar, Muskan Saini, Muskan Goel, and BS Panda. 2021. Modeling information diffusion in online social networks using a modified forest-fire model. *Journal of intelligent information systems*, 56, 2, 355–377.

[8] Haewoon Kwak and Jisun An. 2016. Two tales of the world: comparison of widely used world news datasets gdelt and eventregistry. In *Proceedings of the International AAAI Conference on Web and Social Media* number 1. Vol. 10, 619–622.

[9] 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.

[10] Mauricio Quezada, Vanessa Peña-Araya, and Barbara Poblete. 2015. Location-aware model for news events in social media. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 935–938.

[11] Elad Segev. 2015. Visible and invisible countries: news flow theory revised. *Journalism*, 16, 3, 412–428.

[12] Elad Segev and Thomas Hills. 2014. When news and memory come apart: a cross-national comparison of countries' mentions. *International Communication Gazette*, 76, 1, 67–85.

[13] Sadi Evren Seker, MERT Cihan, AL-NAAMİ Khaled, Nuri Ozalp, and AYAN Ugur. 2013. Time series analysis on stock market for text mining correlation of economy news. *International Journal of Social Sciences and Humanity Studies*, 6, 1, 69–91.

[14] Abdul Sittar, Daniela Major, Caio Mello, Dunja Mladenić, and Marko Grobelnik. 2022. Political and economic patterns in covid-19 news: from lockdown to vaccination. *IEEE Access*, 10, 40036–40050.

[15] Abdul Sittar and Dunja Mladenic. 2021. How are the economic conditions and political alignment of a newspaper reflected in the events they report on? In *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 201–208.

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

[17] Abdul Sittar, Dunja Mladenić, and Marko Grobelnik. 2022. Analysis of information cascading and propagation barriers across distinctive news events. *Journal of Intelligent Information Systems*, 58, 1, 119–152.

[18] Abdul Sittar, Dunja Mladenić, and Marko Grobelnik. [n. d.] Profiling the barriers to the spreading of news using news headlines. *Frontiers in Artificial Intelligence*, 6, 1225213.

[19] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. 2011. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2541–2544.

[20] Hong Wei, Jagan Sankaranarayanan, and Hanan Samet. 2020. Enhancing local live tweet stream to detect news. *GeoInformatica*, 1–31.