Unsupervised Misinformation Detection in News Articles Using Semantic Similarity: A Case Study on the Russia-Ukraine War Dataset

Nina Khairova
Umeå University,
Umeå, 90187,
Sweden
ninakh@cs.umu.se

Andrea Galassi
Bogdan Ivasiuk
University of Bologna,
Viale Risorgimento 2,
Bologna, Italy
a.galassi@unibo.it
bogdan.ivasiuk@studio.unibo.it

Fabrizio Lo Scudo
University of Calabria
via Bucci, Rende,
87036, Italy

Ivan Redozub
National Technical
University "Kharkiv
Polytechnic Institute",
Kharkiv, Ukraina
ivan.red707@gmail.com

Abstract

To tackle the problem of misinformation detection in news articles in cases where annotated datasets are absent, we suggest applying an unsupervised machine learning approach. We created the Russia-Ukraine War (RUWA) dataset with over 16,500 news articles on key events of the Russian invasion of Ukraine, from February to September 2022, sourced from outlets in the USA, EU, Asia, Ukraine, and Russia. On this dataset, we evaluate the potential of using semantic similarity measures to detect misinformation in news articles effectively.

1 Introduction

Currently, misinformation has become one of the most significant challenges facing modern society. This issue has been further exacerbated by the Russian invasion of Ukraine in February 2022. In January 2023, the European Council officially recognized misinformation, particularly that propagated by Russia, as a "long-term challenge for European democracies and societies"[1]. In the context of ongoing information warfare and propaganda, social networks and news articles serve as strategic tools to influence and manipulate public opinion. Therefore solving the problem of misinformation and disinformation detection is crucial for the future of free and democratic societies. However, the application of AI approaches for misinformation identification is significantly impeded by the scarcity of true/false annotated datasets. One of the reasons for this limitation arises due to the nuanced and multifaceted nature of misinformation in news articles across different languages, rendering definitive categorization as true or false a challenging task. To address the issue of misinformation detection (MD) in situations where obtaining annotated datasets is complicated—such as during an ongoing war with limited fact-checking—we hypothesize that using an unsupervised machine learning approach will be essential.

Our study aims to evaluate the effectiveness of using unsupervised machine learning approaches, such as semantic similarity measures, to detect misinformation in news articles. We consider creating and handling unannotated datasets containing news articles covering the events of the Russia-Ukraine war. We will compare articles from various global outlets based on a few hypotheses.

Primarily, we propose that news shared by media outlets in the two nations actively involved in the conflict is likely to display considerable differences. Information variations may be significant, even leading to conflicting accounts of events, such as the acknowledgment or denial of incidents like residential area bombings or civilian victims. Consequently, we can expect that the semantic similarity coefficient between texts from Russian and Ukrainian outlets should be minimal. Furthermore, we hypothesize that the semantic similarity coefficients among articles covering a specific event from various outlets, excluding one or two websites, are generally high. However, when comparing the semantic similarity of these one or two specific websites with all others, we can observe a significant divergence. This discrepancy suggests that these specific websites are likely to be untrustworthy.

2 Background

2.1 Approaches for automatic misinformation detection

Most current studies on misinformation detection employ Machine Learning (ML) or Deep Learning (DL) techniques (Rastogi and Bansal, 2023). Typically, MD processing involves four main steps: data source selection, data collection, data cleaning,

[1] https://www.consilium.europa.eu/en/documents-publications/library/library-blog/posts/the-fight-against-pro-kremlin-disinformation/

and the application of classification or clustering techniques. In the case of ML approaches an additional step for feature extraction is included.

Most research focuses on specific types of data sources, often concentrating either on misinformation detection in social media posts (Islam et al., 2020) or fake news articles on news websites (Reis et al., 2019). The selection of a specific data source type has an impact on the features that can be utilized by ML models. For instance, features relevant to the propagation properties of information can be extracted specifically from the social media context. This feature group includes user profiles and various aspects of user demographics, such as age, number of tweets or posts that the user has authored, and the average number of followers, etc. (Jarrahi and Safari, 2023).

However, obtaining the propagation features from news articles on websites is nearly impossible. For misinformation detection in these data sources, style-based or knowledge-based features are typically extracted (Zhou and Zafarani, 2020).

Mostly style-based methods aim to identify fake news by analyzing the manipulative elements present in the writing style of news content. The extraction of style-based features relies on the assumption that information created to intentionally deceive the public must sound 'more persuasive' compared to text without such intentions (Potthast et al., 2017). However, this assumption may not hold true for official news websites.

Utilizing knowledge-based features for classification and clustering tasks in MD requires effective fact-checking, which is challenging due to significant bias and the fog of war during the ongoing conflict.

2.2 Existing dataset

Due to the resource-intensive and laborious nature, problems of scalability, and subjectivity of true/false annotated datasets building, their availability is dramatically limited (Murayama, 2021), (D'Ulizia et al., 2021). Existing labeled datasets primarily focus on political news and are annotated through manual efforts (Wang, 2017) or by leveraging fact-checking websites like PolitiFact or GossipCop (Shu et al., 2020). In certain instances, authentic news sources were selected from a designated group of reliable outlets, whereas fake news sources were drawn from known fake news lists, such as "Insider's Zimdars Fake News list" (Janicka et al., 2019). Another annotation approach for the fake news dataset involved the AMT dataset (Potthast et al., 2017), which comprises 480 articles annotated as either fake or true. In this dataset, fake news articles were intentionally crafted by journalists, while genuine news pieces were sourced from various domains. The datasets focusing on fake news related to conflicts or wars exhibit a distinct nature. For instance, the FA-KES dataset (Salem et al., 2019) encompasses 804 news articles related to the Syrian war gathered from sources like Reuters, Etilaf, and others. To determine the veracity of the information the obtained data was compared with information from the Syrian Violation Documentation Center (VDC), which meticulously records all deaths during specific events. In the last two years, several researchers have addressed the issue of dataset collection from social networks, primarily from Twitter, in the specific context of propaganda and fake news detection related to the Russian invasion of Ukraine (Geissler et al., 2023), (Haq et al., 2022). However, while several significant studies have addressed the challenges of misinformation detection in content related to the events of the on-going Russia-Ukraine war, a well-annotated true/false dataset is still absent.

In our study, we first focus on creating an unannotated dataset containing news articles about the events of the Russia-Ukraine war from various global outlets. We then evaluate the effectiveness of using unsupervised machine learning approaches, such as semantic similarity measures, to detect misinformation in the dataset.

3 Methodology

3.1 Data Collection

We created the RUWA (Russian-Ukraine WAr) dataset (Khairova et al., 2024), which compiled news articles covering key events related to the Russia-Ukraine war[2]. To ensure a balanced representation of journalistic perspectives, we sourced texts from reputable global outlets spanning various world regions. These include BBC, Euronews, and The Guardian (European region); NBC News, CNN, and Bloomberg (USA region); Ukrinform and Censor.net (Ukraine); and Russia Today, Newsfront.info (Russia), as well as Al Jazeera and Reuters.

To mitigate the risk of creating a topic detection model instead of a misinformation detection model, we identified nine widely acknowledged events of

[2] https://github.com/ninakhairova/dataset_RUWA

the 2022 year of the Russian-Ukraine war, such as 'The Beginning of the War', 'Bucha Massacre', and so on, and classified all articles regarding the nine topics.

The selection of articles for each event adhered to predefined criteria, including the publication time interval and keyword lists. The time interval typically spanned from the date of the specific event and extended three to four weeks thereafter. This approach aligns with the common pattern in media, where dedicated coverage of a particular event tends to last no more than two to three weeks.

Table 1 shows the distribution of approximately 16,000 obtained articles across various websites and war-related topics. The columns Definition and Definition source contain the definition of each considered event and the website from which the definition was obtained, respectively.

3.2 Data Analysis

As previously mentioned, obtaining information with complete certainty about events during an ongoing war is virtually impossible. Any narrative or description of an event inherently carries potential bias and can reflect the subjective perspective. Consequently, the creation of a true/false annotated dataset covering the Russia-Ukraine war poses significant challenges due to the inherent subjectivity and variability in how events are reported and interpreted.

Our approach involved constructing the events-aligned RUWA dataset, followed by the application of unsupervised machine learning methods to address semantic similarity tasks.

The study involves three types of experiments for detecting semantic similarity: (1) comparing the full texts of the articles, (2) analyzing article headings, and (3) comparing semantically meaningful sentences within the articles. To assess the similarity of semantically significant sentences from various sources we utilize keywords associated with the event under consideration or verbs representing the actions linked to specific events. Compiling these lists for each event, we relied on the existing list of words associated with the Russian-Ukrainian war from (Solopova et al., 2023) and added verbs extracted from the articles covering each specific event.

For linguistic preprocessing, we employed stemming and stop-word removal. Additionally, we eliminated numerous specific symbols commonly found in web-wrapped texts. To generate pre-trained vectors, we employed two types of language models (LM), based on Spacy and FastText. In contrast to other language models, FastText successfully predicted subwords and character n-grams. Therefore, FastText handled texts extracted from web pages, which contain breaks due to inserted images, links, quotes, and ads, more efficiently.

4 Results and findings

We evaluated semantic similarity among every pair of outlets across nine topics by comparing full texts, article headings, and selected sentences from the articles. To avoid building a topic model instead of a misinformation detection model, each of the nine topics was examined individually.

We observed that evaluating the semantic similarity of headlines encountered challenges, particularly when dealing with distributive semantic similarity scores. Even headlines from articles covering the same event and belonging to the same outlet yield relatively low similarity values. Several factors contribute to this outcome. Primarily, the efficacy of comparing article titles is significantly influenced by the number of articles published by each outlet for a specific event. The RUWA dataset, however, is not well-balanced across events. In certain cases, a website may have produced only a few articles related to a particular event, impacting the reliability of the semantic similarity headlines assessment. Furthermore, each headline frequently not only neutrally conveys or describes an event but also mirrors the subjective perspectives and sentiments of certain authors.

We obtained the best results in the third experiment by considering the topic-specific parts of the texts and steering clear of broad or generalized content in the articles. This approach allowed us to generate more specific and directly relevant texts that are closely tied to the subject of the event. For instance, Table 2 demonstrates the semantic similarity for texts obtained by concatenating all sentences containing specific verbs related to the sinking of the warship Moskva.

Almost all nine topics in the final experimental group, which involved additional knowledge regarding actions specified by concrete verbs, provided a clear confirmation of our initial hypothesis. The experiment indicates that the semantic similarity coefficient is notably lower between established outlets from countries engaged in the war on opposing sides.

Topic	Description	Definition Source	Article count
Azovstal	Russia says Azovstal siege is over, in full control of Mariupol	Al Jazeera	1,816
Beginning	NATO officials say Russian attack of Ukraine has begun	CBS News	6,490
Bucha	Killing of civilians in Bucha and Kyiv condemned as 'terrible war crime'	Guardian	1,429
Nuclear Plant	Evacuations from Zaporizhzhia renew concerns for nuclear power plant safety	CNN	3,373
Prisoners	'Absolute evil': inside the Russian prison camp where dozens of Ukrainians burned to death	Guardian	578
Railway	'Ukraine missile attack: Dozens killed at Kramatorsk railway station	Al Jazeera	1,466
Moskva Sinking	Russia is losing the battle for the Black Sea	Economist	175
Kremenchug Supermarket	Russian missile strike kills 16 in a shopping mall, Ukraine says	Reuters	436
Mariupol Theatre	Russia bombs theater where hundreds sought shelter and 'children' was written on grounds	CNN	761
Total			16,526

Table 1: RUWA articles distribution by the outlets and the nine topics

	The guardian	Reuters	Aljazeera	CensorNet	CNN	Ukrinform	RT
The guardian	-	16.8%	40.9%	18.6%	24.7 %	25.2%	17.6%
Reuters	16.8%	-	14.7%	17.6%	10,1%	7.0%	19.4%
Aljazeera	40.9%	14.7%	-	15.5%	24.6	21.5%	16.4 %
CensorNet	18.6%	17.6%	15.5%	-	7.2%	9.3%	8.1%
CNN	24.7%	10.1%	24.6%	7.2%	-	42.8%	9.7%
Ukrinform	25.2%	7.0%	21.5%	9.3%	42.8%	-	11.5%
RT	17.6%	19.4%	16.4%	8.1%	9.7%	11.5%	-

Table 2: Semantic similarity scores are calculated for texts created by concatenating all sentences containing specific verbs related to a particular event, such as the sinking of the warship Moskva

In our assessment, this finding not only underscores the distinctiveness and divergence in the reporting styles and perspectives of news outlets representing countries with conflicting interests in the ongoing war but also suggests the potential dissemination of misinformation by one country regarding a specific event.

4.1 Conclusion

In our study, we introduced an innovative dataset focused on the Russian-Ukrainian war. This RUWA dataset involves above 16,500 web news articles from established world outlets, covering nine significant events of the Russian invasion of Ukraine that occurred from February to September 2022. The dataset offers a comprehensive view of diverse journalistic narratives surrounding the Russian-Ukrainian war, providing valuable support for future research.

Furthermore, our research contributes to illustrating how unsupervised machine learning approaches, such as semantic similarity scores, can offer insights into potential misinformation within news coverage of widely reported events across various outlets. We critically examined the pros and cons of multiple methods for assessing the semantic similarity of news articles discussing the same event across diverse reputable news outlets. Additionally, we showed that while relying solely on semantic similarity analysis may not be enough for effective misinformation detection, it offers valuable insights that can be synergistically combined with other techniques to enhance overall accuracy in detection.

References

Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

Dominique Geissler, Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2023. Russian propaganda on social media during the 2022 invasion of ukraine. *EPJ Data Science*, 12(1):35.

Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Tristan Braud, and Pan Hui. 2022. Twitter dataset for 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.02955*.

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):82.

Maria Janicka, Maria Pszona, and Aleksander Wawer. 2019. Cross-domain failures of fake news detection. *Computación y Sistemas*, 23(3):1089–1097.

Ali Jarrahi and Leila Safari. 2023. Evaluating the effectiveness of publishers' features in fake news detection on social media. *Multimedia Tools and Applications*, 82(2):2913–2939.

Nina Khairova, Andrea Galassi, Fabrizio Lo Scudo, Bogdan Ivasiuk, and Ivan Redozub. 2024. Unsupervised approach for misinformation detection in russia-ukraine war news. In *CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024)*, volume 4. CEUR-WS.

Taichi Murayama. 2021. Dataset of fake news detection and fact verification: a survey. *arXiv preprint arXiv:2111.03299*.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Shubhangi Rastogi and Divya Bansal. 2023. A review on fake news detection 3t's: typology, time of detection, taxonomies. *International Journal of Information Security*, 22(1):177–212.

Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

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

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Veronika Solopova, Oana-Iuliana Popescu, Christoph Benzmüller, and Tim Landgraf. 2023. Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts. *Datenbank-Spektrum*, 23(1):5–14.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.