Automated Detection of Misinformation on Twitter During the 2022 Russo–Ukrainian Conflict

Sushma Kumari

University of Stuttgart, Germany

Abstract

Much research has been conducted on debunking and analyzing misinformation, often focusing on specific events. In 2022, when the Russo–Ukrainian conflict began, a large volume of misinformation circulated across online platforms. In response, multiple fact-checking organizations actively worked to debunk false claims on social media. In this paper, we present a BERT-based classification model to predict both the domain and class of misinformation tweets collected during the 2022 Russo–Ukrainian conflict. Additionally, we incorporate supplementary data from fact-checked articles to enhance model performance. The proposed classification model achieved a best weighted average accuracy of 82% which outperformed the baseline model by 84%.

Keywords

Misinformation Detection, Classification Model, Social Media, Russo-Ukrainian Conflict

1. Introduction

Data journalism is a new journalistic discipline that focuses mainly on data-driven research and presentation formats. However, a fundamental problem of data journalism and classical journalism is that much data of journalistic interest is only available in the unstructured form: as texts, tables and graphics in documents of various types (Word, PDF, e-mail, etc.) or on websites. Also, there is a lack of a centralized data hub for gathering information from different sources.

There is an increasing amount of fake news in the media, social media, and other web sources. Much research has been done for fake news detection and debunking of fake news [1]. In the last two decades, there is a tremendous increase in the spread of misinformation, which is also reflected by the number of fact-checking websites [1]. Fact-checking websites can help to investigate claims and assist citizens in determining whether the information used in an article is true or not. More than 213 fact-checking websites are working in 40+ languages across 100+ countries [2, 3, 4].

There is no standard protocol for fact-checking services across different fact-checkers, and they do not publish their proofed articles in a standard format, which leads to several conflicts. Shahi et al. [3] discuss the need to detect news articles potentially containing false information.

Crises are complex, involve different stakeholders, and evolve over time. As the crises change, so does crisis communication. One option to consider the dynamic course of crisis communication in the analysis is to separate the communication into several time frames [5]. Previous research categorized crises into multiple stages [6, 7], which captures the different behaviors and emotions that individuals might exhibit and experience depending on the current development of the crisis. For example, before (or at the beginning of) a crisis, there might be higher uncertainty, perceived risk, and associated information-seeking behavior than in later crisis stages. Furthermore, as the COVID-19 pandemic evolved differently in different countries and local authorities responded differently to the threat, an investigation of the connection between the development of the pandemic and public communication must consider geographical factors [8]. Therefore, similar to Park, Park, and Chong [9], who analyzed the COVID-19 conversation in Korea, our research focuses on German-language communication.

Through analysis of the data generated by social media, we want to deliver a useful insight into crisis communication, which can support the management of future crisis situations more effectively [10]. Additionally, by analyzing how information is distributed on social media, we can obtain information

Forum for Information Retrieval Evaluation, December 17–20, 2025, India

st192042@stud.uni-stuttgart.de (S. Kumari)

that might be useful in guiding health professionals, researchers, and policymakers to better influence public health, policy decisions, and quality health information [11].

Crisis management could greatly benefit from a stronger use of AI, both for information extraction and for information dissemination [12]. During COVID-19, a massive amount of data was generated on social media, which is a good infrastructure and facility to process it. The use of Artificial Intelligence in crisis management has already been tested in data analysis for disaster [13], public health [14].

In this paper, we discussed the method used for the fake news detection for the shared task at PROMID [15]. The task uses the dataset from 2022 Russo–Ukrainian conflict which consists of tweets collected during the one year of the crises.

The remainder of the paper is organized as follows: the related work describes the past work, the task description gives an overview of the task, the Experiment section emphasizes the experiment details, the Conclusion, and the Future Work focuses on the conclusion and future aspects of the work.

2. Related Work

Fact-checking is a damage control method that is both essential and not scalable. It might be hard to take the human component out of the picture any time soon. But still, automatic fake news classification could help reduce the workload for the fact-checkers. The fake content is spread in multiple formats; many of them are repurposed by changing the text, location, etc. Fake news detection is a complex problem. Research has been done on fake news detection using social media data like tweets [2], YouTube videos [3], but less research has focused on the news articles. One of the primary reasons is the lack of a corpus of news articles. Fake news is spread in several domains, such as crime, elections, and the economy.

There is a lack of corpus to train Machine Learning based on fact-checking. In the last few years, a collection of small datasets related to fact-checking has been published. These datasets are a mixture of fake news topics, for instance, the US election 2016[16]. Still, fact-checking is dominant in the English language, and a few sources like Snopes, Politifact, etc.

Multi-FC corpus describes the different kinds of fact-checking datasets available and their limitations[17]. The authors have come up with a multi-domain, evidence-based fact-checking dataset. They have also described the metadata of Fact-check articles. [18] describes the task of fact-checking and the construction of a dataset using the process used by journalists. Several methods are published for automatic detection of fake news, Pérez-Rosas et al. discusses the automatic detection of fake news and linguistic differences in false and legitimate content. [20] analyzes, compares, and summarizes several different methods available for fake news detection. The authors give an overview of methods that have been tested for fake news detection. Research has been done to look inside the feature of a news story in the modern diaspora, along with different kinds of story and their impact on people[21].

Both the COVID-19 pandemic and the infodemic spread in parallel. Mesquita et al. propose a framework to fight against fake medical news because fake news can intensify the effect of the COVID-19 pandemic [22]. All health workers, scientists, the government are trying to fight against fake medical news. In March 2020, several cases were discovered where people were consuming partially true information about COVID-19 on Facebook and WhatsApp without using the proper medical terms, which created panic about medicine and prevention for COVID-19 [23]. By analysing the search behaviour of people in Italy from January to March 2020 and found that a ”large number of infodemic monikers were observed across Italy” [24].

During the time of the pandemic, fake news was spread all over the world in different languages. The fake news is covered in different domains like origin and spread, conspiracy theory, etc. There is a lack of resources that are multilingual and cross-domain and have been collected from multiple sources.

3. Task Description

The objective of this task is to automatically classify tweets related to the Russo-Ukrainian conflict into two categories: misinformation and non-misinformation. The dataset consists of manually annotated tweets collected via the Twitter API during the first year of the 2022 Russo–Ukrainian conflict [25] using AMUSED framework [26]. The dataset consists of 36,174 non-misinformation tweets and 778 misinformation tweets, divided into training and the test set as shown in Table 1. A detailed description of the dataset and the task are given in the overview paper of PROMID [27, 15].

Table 1
Class Distribution of the PROMID Dataset

Class Label	Number of Tweets	Percentage (%)
Non-Misinformation	36,174	97.89
Misinformation	778	2.11
Total	**36,952**	**100.00**

4. Experiment & Results

Transfer learning is a technique where a deep learning model trained on a large dataset performs similar tasks on another dataset. We call such a deep learning model a pre-trained model. The most renowned examples of pre-trained models are the computer vision deep learning models trained on the ImageNet dataset. So, it is better to use a pre-trained model as a starting point to solve a problem rather than building a model from scratch.

To classify fake news articles, we used the state-of-the-art neural network language model BERT, which has been pre-trained on a large corpus to solve language processing tasks [28]. An essential advantage of BERT is that it can be fine-tuned for task-specific datasets and allows high text classification accuracy even for smaller datasets. In the context of fake news classification, BERT has already been applied for multiclass classification tasks, for example, on the Chinese social media platform Weibo, where it achieved considerable accuracy [29]. The randomisation of the data prevents seasonal patterns from being learned by the model. The randomisation of the data prevents seasonal patterns from being learned by the model. For BERT fine-tuning, the models for the videos were trained on for four epochs, and the models for the comments were trained on for three epochs with a learning rate of 2e-5. We set the hidden units to 300 and the training epoch to 200. Each training process continues until the restriction or validation loss is continued. The batch size is set to 10, and the learning rate is 0.001. After the individual prediction on the two test datasets, we evaluated the accuracy of the two models using the weighted F1 score. For classification, the model gave a macro F1 score of 82 % with a precision of 82 % and 84 %. A description of classification model is given in Table 2. The obtained result is also get compared with the baseline models as SVM, LSTM and CNN as shown in Table 3

Table 2
BERT Model Configuration for Fake News Classification

Parameter	Value
Pre-trained Model	BERT
Video Training Epochs	4
Comment Training Epochs	3
Fine-tuning Learning Rate	2e-5
Hidden Units	300
Maximum Training Epochs	200
Batch Size	10
Optimizer Learning Rate	0.001

Table 3
Performance Comparison of Different Models

Model	Macro F1 (%)	Precision (%)	Recall (%)
SVM	70	71	72
CNN	68	69	70
LSTM	72	72	74
BERT	**82**	**82**	**84**

5. Conclusion & Future Work

In this paper, we have presented a classification model for detecting fake news and its domain. Using the BERT model, the model performed well compared to the traditional machine learning technique. With the proposed model, the problem of fake news classification is addressed. The experimental results demonstrate that the proposed BERT-based model significantly outperforms traditional deep learning approaches such as CNN and LSTM, as well as the state-of-the-art baseline model. In particular, the proposed model achieves a Macro F1 score of 82%, showing an improvement of approximately 10% over the state-of-the-art baseline. This performance gain highlights the effectiveness of transformer-based contextual language representations for fake news classification tasks.

As a limitation of the work, the dataset size was highly imbalanced, and building a generalizable model was difficult. Getting a fake news corpus is a challenging task, and with the limited dataset, it is hard to enhance the machine learning model's performance. With the external dataset, the performance of the classifier is increased. We also observed that different fact-checking websites are checking several duplicates of claims. Several old claims are repurposed as fake news.

So, in future work, detecting similar claims would help find fake news classes using text matching. Incorporating semantic similarity techniques and transformer-based embedding models could improve the detection of paraphrased or reworded misinformation. Additionally, addressing class imbalance through data augmentation or synthetic sample generation may further enhance model robustness. Expanding the dataset across multiple domains and languages would also improve generalization capability. Integrating metadata such as temporal patterns and source credibility could strengthen detection performance. Finally, deploying the model in a real-time large-scale monitoring system could support proactive misinformation identification.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools in the preparation of this manuscript.

References

[1] L. Graves, F. Cherubini, The rise of fact-checking sites in europe (2016).

[2] G. K. Shahi, A. Dirkson, T. A. Majchrzak, An exploratory study of covid-19 misinformation on twitter, Online Social Networks and Media 22 (2021) 100104. URL: https://www.sciencedirect.com/science/article/pii/S2468696420300458. doi:10.1016/j.osnem.2020.100104.

[3] G. K. Shahi, D. Röchert, S. Stieglitz, Covid ct: Analysis and detection of different conspiracy theories on youtube in the context of covid-19 (2020).

[4] G. K. Shahi, T. A. Majchrzak, Exploring the Spread of COVID-19 Misinformation on Twitter, Technical Report, EasyChair, 2021.

[5] S. Stieglitz, D. Bunker, M. Mirbabaie, C. Ehnis, Sense-making in social media during extreme events, Journal of Contingencies and Crisis Management 26 (2018) 4–15.

[6] S. Fink, A. M. Association, et al., Crisis management: Planning for the inevitable, Amacom„ 1986.

[7] S. Youngblood, Ongoing crisis communication: Planning, managing, and responding, 2nd edition (coombs, w. t.) and handbook of risk and crisis communication (heath, r. l. and o'hair, h. d., eds.)

[book reviews, IEEE Transactions on Professional Communication 53 (2010). doi:10.1109/TPC.2010.2046099.
[8] S. Bhochhibhoya, A. E. Burnette, D. D. Cali, A. Davisson, D. R. Dewberry, M. Eisenstadt, R. L. Fox, M. Gibbons, M. Horning, J. W. Kirk, et al., Public Communication in the Time of COVID-19: Perspectives from the Communication Discipline on the Pandemic, Rowman & Littlefield, 2022.
[9] H. W. Park, S. Park, M. Chong, Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea, Journal of Medical Internet Research 22 (2020). doi:10.2196/18897.
[10] C. Ehnis, M. Mirbabaie, D. Bunker, S. Stieglitz, The role of social media network participants in extreme events, in: Proceedings of the 25th Australasian Conference on Information Systems, 2014.
[11] C. Chew, G. Eysenbach, Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak, PloS one 5 (2010) e14118.
[12] W. Sun, P. Bocchini, B. D. Davison, Applications of artificial intelligence for disaster management, Natural Hazards 103 (2020) 2631–2689.
[13] L. Cao, Ai and data science for smart emergency, crisis and disaster resilience, International journal of data science and analytics 15 (2023) 231–246.
[14] A. P. Adekugbe, C. V. Ibeh, Harnessing data insights for crisis management in us public health: lessons learned and future directions, International Medical Science Research Journal 4 (2024) 391–405.
[15] A. Hegde, G. K. Shahi, S. Satapara, P. Mehta, S. Modha, D. Ganguly, D. Nandini, H. L. Shasirekha, A. K. Jaiswal, G. Pasi, T. Mandl, Prompt recovery for misinformation detection at fire 2025, in: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '25, Association for Computing Machinery, 2025.
[16] W. Y. Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).
[17] I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. Hansen, C. Hansen, J. G. Simonsen, Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims, arXiv preprint arXiv:1909.03242 (2019).
[18] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, 2014, pp. 18–22.
[19] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th international conference on computational linguistics, 2018, pp. 3391–3401.
[20] X. Zhou, R. Zafarani, Fake news: A survey of research, detection methods, and opportunities, arXiv preprint arXiv:1812.00315 2 (2018).
[21] S. B. Parikh, P. K. Atrey, Media-rich fake news detection: A survey, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2018, pp. 436–441.
[22] C. T. Mesquita, A. Oliveira, F. L. Seixas, A. Paes, Infodemia, fake news and medicine: Science and the quest for truth, International Journal of Cardiovascular Sciences (2020).
[23] D. Orso, N. Federici, R. Copetti, L. Vetrugno, T. Bove, Infodemic and the spread of fake news in the covid-19-era, European Journal of Emergency Medicine (2020).
[24] A. Rovetta, A. S. Bhagavathula, Covid-19-related web search behaviors and infodemic attitudes in italy: Infodemiological study, JMIR Public Health and Surveillance 6 (2020) e19374.
[25] G. K. Shahi, Y. Mejova, Too little, too late: Moderation of misinformation around the russo-ukrainian conflict, Websci '25, 2025. doi:10.1145/3717867.3717876.
[26] G. K. Shahi, T. A. Majchrzak, Amused: An annotation framework of multimodal social media data, in: F. Sanfilippo, O.-C. Granmo, S. Y. Yayilgan, I. S. Bajwa (Eds.), Intelligent Technologies and Applications, Springer International Publishing, Cham, 2022, pp. 287–299.
[27] G. K. Shahi, A. Hegde, S. Satapara, P. Mehta, S. Modha, D. Ganguly, D. Nandini, H. L. Shasirekha, A. K. Jaiswal, G. Pasi, T. Mandl, Overview of the first shared task on prompt recovery for misinformation detection (promid 2025), in: K. Ghosh, T. Mandl, S. Pal, S. Majumdar, A. Chakraborty

(Eds.), Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, Varanasi, India. December 17-20, 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[29] T. Wang, K. Lu, K. P. Chow, Q. Zhu, Covid-19 sensing: negative sentiment analysis on social media in china via bert model, Ieee Access 8 (2020) 138162–138169.