






Semi-Supervised Learning Classifier for Misinformation Related to Earthquakes Prediction on Social Media

Or Elroy^{1,2}  and Abraham Yosipof^{1,2}  

¹ Faculty of Information Systems and Computer Science, College of Law and Business,
Ramat-Gan, Israel
aviyo@clb.ac.il

² International Institute for Applied Systems Analysis, Laxenburg, Austria

Abstract. Social media is a fertile ground for the growth and distribution of misinformation. The belief in misinformation can have devastating consequences, and may lead to unnecessary loss of life. Properly identifying and countering misinformation on social media is therefore necessary for the fight against misinformation. In this research, we developed an Adjusted Semi-Supervised Learning for Social Media (ASSLSM) method to classify and analyze tweets regarding misinformation related to earthquakes prediction. The ASSLSM method adjusts the pseudo-labeling constraints based on assumptions related to metadata of the tweets and users, with the goal of providing better information to the underlying models. We collected a dataset of 82,129 tweets related to the subject of earthquakes prediction. Expert seismologists manually labeled 4,157 tweets. We evaluated and compared the performance of ASSLSM, supervised learning, and semi-supervised learning (SSL) methods on the dataset. We found that the ASSLSM methodology provides better and more consistent performance in comparison to supervised learning and SSL. Finally, we used an ASSLSM classifier to classify the full dataset and analyzed the classified dataset.

Keywords: Semi-Supervised Learning · Misinformation · RoBERTa · NLP · Earthquakes · Social Media

1 Introduction

Social media has a key role in the expression and distribution of authoritative as well as speculative information on different subjects in recent times, primarily because of its massive adoption, audience, and accessibility.

Social media platforms reach different sectors of the population and are often accessed multiple times a day, or even continuously throughout the day, for recreational purposes as well as for receiving important information. Considering its reachability and instantaneous nature, social media inevitably became a viable channel of communication for information, such as warnings about upcoming and ongoing emergencies and disasters.

During the COVID-19 pandemic, social media served as a communication channel for news and updates about the spread of the virus throughout the world, as well as medical recommendations [1].

It is important to acknowledge that social media is also a ground for the growth and dissemination of misinformation. Misinformation is false or inaccurate information according to the best factual evidence that is available at a given point in time, regardless of an intention to mislead or deceive [2].

Misinformation and the belief therein are driven by a natural need to rationalize unexplained or unexpected emergencies and disasters. The lack of authoritative sources with reliable information regarding emergencies and disasters, such as an outbreak of a virus or an earthquake, combined with circumstantial evidence, promotes misinformation [3, 4]. The belief in misinformation can have devastating consequences, and even lead to unnecessary loss of life.

Earthquakes prediction is one of the topics of misinformation that is being discussed on social media. According to the current state of research, earthquakes cannot be predicted. The exact location, time and magnitude of future events cannot be specified [5, 6]. However, misinformation regarding earthquakes prediction or advance warnings is constantly spread on social media [5, 7]. Earthquakes contribute to anxiety, shock, and panic of the population, which consequently make the population more vulnerable to misinformation [8]. A population that frequently encounters misinformation regarding earthquake predictions may hesitate to take necessary actions to protect itself when a real earthquake alert is issued. Misinformation regarding earthquake predictions causes a variety of symptoms, such as confusion, anxiety, or misguided beliefs that further lead to unnecessary actions like evacuation. Misinformation on social media is a commonly researched topic in general and on Twitter specifically [1, 9, 10].

In this research, we developed a new semi-supervised classifier to classify and analyze tweets regarding misinformation related to earthquake predictions. This methodology presents several challenges, including the collection of enough relevant data, labeling the data as misinformation or not-misinformation, and the development of a classifier to detect tweets that spread misinformation. Finally, the classified tweets are analyzed to gain knowledge and insights on how to support the fight against misinformation regarding earthquakes prediction.

We address these challenges and provide the following contributions. We collected 82,129 tweets according to a specific search query that expert seismologists curated. A key task in analyzing tweets related to misinformation is to label and classify the tweets [1]. Collecting a large amount of data is often efficient and fast, whereas labeling the data can be a lengthy, costly, and complicated process. Therefore, a relatively small amount of labeled data is often used to classify a relatively large amount of unlabeled data. In this work, expert seismologists labeled 4,157 tweets. The labeled dataset constitutes about five percent of the dataset.

Training a classifier where only five percent of data is labeled can provide a good model for the training set. However, applying the model on the rest of the data may be problematic regardless of how representative the training set and classifier are [11]. Semi-Supervised Learning (SSL) methods address this concern by making assumptions about the actual labels of the unlabeled data based on the confidence levels of the predictions.

In this work, we eased the confidence level assumption and introduced new assumptions that are more robust when labeling data from social media.

Classification models depend on features that represent samples in the dataset. For textual data, the features are often derived from Natural Language Processing (NLP) word embedding algorithms. The outcome is that two semantically-similar texts are likely to be classified as having the same label.

However, classification models based on word embedding features do not account for valuable metadata of social media posts, such as the classification of other posts of the same users, the number of followers a user has, or the time the user has existed on the platform. While semantically similar tweets share similar embeddings, the actual classification of other tweets posted by the same user can be even more useful than the classification of more semantically similar tweets from other users.

In this work, we present the Adjusted Semi Supervised Learning for Social Media (ASSLSM) method. ASSLSM takes into consideration different metadata of the users who posted the tweets, as well as metadata of the tweets. This approach is more robust and adapted to the environment of social media in general, and to misinformation regarding earthquake predictions specifically. To the best of our knowledge, this work is the first to consider features derived from social media metadata in the process of SSL.

2 Related Work

Twitter is a valuable and frequently used source of information for research regarding misinformation on social media. Erokhin et al. [9] analyzed the behavior of different conspiracy theories related to the COVID-19 pandemic. Batzdorfer et al. [12] investigated the dynamics of tweets that discuss COVID-19 conspiracy theories, by comparing tweets from a group of users that talked about conspiracy theories and a group of users that participated in the general discussion on the virus. Darwish et al. [13] created a fake news detection system and built a dataset of fake and real tweets about the Russian-Ukrainian conflict using deep-learning and machine learning methods.

NLP methodologies such as Bidirectional Encoder Representations from Transformers (BERT) [14], provide superior results for different NLP tasks, including word embedding [15, 16]. Micallef et al. [17] used BERT embeddings to investigate and counter misinformation in tweets related to COVID-19 over a period of five months. Elroy and Yosipof [1] transformed BERT word embeddings to sentence embedding using Sentence-BERT [18] to train a classifier and classify a dataset of over 300K tweets related to the COVID-19 5G conspiracy theory.

RoBERTa is a Robustly Optimized BERT Pretraining Approach model based on BERT, that was pretrained with different design decisions, leading to improved performance and state of the art results [19–21].

Sentence-BERT is a modification of the pre-trained BERT network that uses Siamese and triplet network structures on top of the BERT model and fine-tuned based on high quality sentence interface data to learn more sentence level information [18]. Sentence-BERT can also be applied on RoBERTa's embeddings to transform the word embeddings into a single sentence embedding, resulting in 768 features per tweet when used with RoBERTa-base.

In addition to embeddings, training a classifier for misinformation on social media also requires a labeled set. The labeled set is traditionally gathered through a lengthy process of manual labeling and often results in a relatively small number of labeled samples out of a relatively large amount of data. SSL addresses this issue by enhancing the labeled dataset with pseudo-labels based on assumptions regarding the appropriate labels for some of the unlabeled dataset.

Multiple SSL approaches exist, such as consistency training, proxy-label methods, generative methods, and graph-based methods [22]. SSL models operate under certain assumptions, such as that two samples that are close enough to each other in terms of distance should share the same labels [22]. Another possible common assumption is that two samples in the same cluster share the same label [22]. Depending on the task, these assumptions can be more or less strict, which is reflected in the size and quality of the pseudo-labeled dataset.

SSL proxy-label methods leverage a model that was trained on the labeled dataset to label samples of the unlabeled dataset using heuristic approaches. A common requirement that the prediction meets a certain threshold of confidence level is typical for proxy-label methods [22]. In this case, a label is considered a proxy label if the prediction probability is greater than a certain threshold.

Metadata and characteristics of social media posts, as well as their authors, were proven to be useful for classification tasks of tweets and were used to enhance classification models in previous works [1, 23–25]. These include the number of users who follow the author and the number of users the author follows as an indication of the author being a robot [23], or URLs, mentions, retweets, and tweet length as indicators for credibility [24, 25]. Balaanand et al. used tweets metadata in graph-based semi-supervised learning to detect fake users on Twitter [26]. Jan et al. used tweets metadata as features for the underlying classifier in a SSL methodology [27].

3 Workflow

To achieve the research objective, we developed the following workflow. Figure 1 describes the workflow used in this work. The workflow consists of four stages, namely data collection and preprocessing, models evaluation, results, and analysis.

The data collection and preprocessing phase involves the collection of tweets related to the discussion of earthquakes prediction misinformation on Twitter, the computation of the embedding for each tweet, and the hand-labeling of tweets.

The evaluation process of the models consists of testing and comparing the performance of different models using different techniques, namely supervised learning, SSL, and ASSLSM.

Following the evaluation, the complete dataset is classified using the model that provides the best performance, and the classified dataset is analyzed.

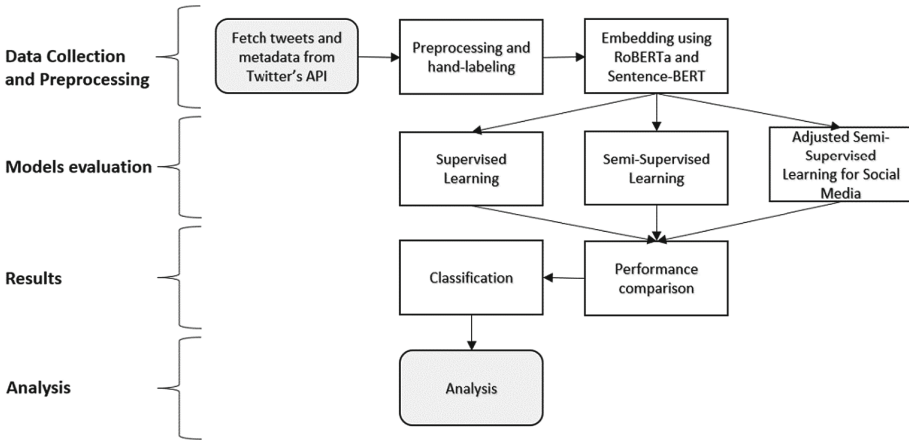


Fig. 1. Research workflow.

4 Dataset

We collected 82,129 tweets related to the subject of earthquakes prediction, forecasts, and notifications, and metadata of the users that posted the tweets, over a period of about two years, from March 1, 2020, to March 31, 2022. The data was collected using a Twitter API that is limited to academic research and provides access to Twitter's full archive. The search query used in this study was defined by expert seismologists as `[[predict OR forecast OR warn OR updates OR alert] AND [earthquake OR quake OR [seismic AND event] OR seismicity OR shaking OR EQ]]`.

The data includes the tweets themselves as well as metadata such as the number of retweets, likes, and replies a tweet received. Metadata of the users who posted the tweets was also collected, such as the total number of tweets the user posted on the platform, the number of followers they have, and the number of other users they follow.

Expert seismologists manually labeled 4,157 tweets into three categories in accordance with the Communication Guide [6]. The three categories are misinformation, not-misinformation, and irrelevant tweets (see Table 1). Tweets claiming to be able to predict future earthquakes were labeled as misinformation (835 tweets, Table 1). Tweets notifying about current earthquakes, rejecting others' ability to predict future earthquakes, or explaining how certain services work, were labeled as not-misinformation (1,416 tweets, Table 1). Other tweets that are not directly related to earthquakes, such as secondary hazards, were labeled as irrelevant (1,906 tweets, Table 1).

We used RoBERTa-base to calculate the word embeddings of each tweet in the dataset, and transformed the word embeddings of each tweet to a sentence embedding using Sentence-BERT [18], yielding a vector of 768 features per tweet.

Table 1. Categories of the manually labeled dataset with examples, number of tweets and proportion of each category.

Category	Example	# of Tweets	% of Tweets
Misinformation	“24 h WARNING: 5.5 + earthquake is likely in the Mammoth Lakes - Bridgeport area and 5.0 + earthquake is likely within 50 miles of Santa Clarita - NW of Los Angeles during the next 24 h.”	835	20.09
Not misinformation	“No one can accurately predict earthquakes. The USGS issues long term earthquake forecasts for certain areas.”	1,416	34.06
Irrelevant	“Could end in 5 billion gallons of lava or nothing will happen. Hard to say. <i>[link]</i> ”	1,906	45.85

5 ASSLSM: Adjusted Semi-Supervised Learning for Social Media

SSL based on proxy-method uses an underlying supervised learning model to predict the labels for the unlabeled data. Predictions that meet certain criteria are assigned as pseudo-labels to the labeled dataset for the purpose of training a model. A common assumption in SSL is that predictions with a confidence level above a certain threshold are correct.

ASSLSM implements additional constraints for the pseudo-labels, based on features of the metadata of the tweets and the users who posted them.

The ASSLSM methodology provides better information to the models by adjusting the constraints used in SSL, to achieve more consistent performance across different underlying models. The constraints used in the ASSLSM methodology require all of the following:

- (A) The prediction matches most of the existing labels for that user in the labeled dataset.
- (B) The user who posted the tweet has ≥ 100 tweets in the dataset.
- (C) The confidence level of the prediction is above threshold T .

Constraint A is based on the presumption that a new tweet by a user who mostly posted tweets belonging to a certain category is very likely to also belong to the same category. Constraint B reduces the number of exceptions to the previous presumption by ignoring users without enough samples in the dataset. Constraint C uses a lower threshold than basic SSL methods, to compensate for the lower number of samples due to constraints A and B.

6 Results

According to the workflow previously described, we evaluated three methodologies, namely supervised learning, semi-supervised learning, and ASSLSM with two different confidence thresholds. Each evaluation phase tested the performance of five different

machine learning models, namely k-NN with $k = 3$, Random Forest with 100 trees, Gaussian Naïve-Bayes, Logistic Regression, and a Voting Ensemble classifier of all previous models with soft voting. We used 5-fold cross-validation for each model. Table 2 presents the weighted F1, precision, and recall scores of the models using supervised learning, SSL, and ASSLSM.

First, we tested traditional supervised learning using the labeled dataset. The results of the supervised learning performance are presented in Table 2, Supervised Learning.

Second, we tested a proxy-method based SSL model by calculating the prediction probability for the sentence embedding of each unlabeled tweet using each model at a time. Predictions with a confidence level over a threshold of 0.9 were added to the labeled dataset as pseudo-labels (Table 2, Semi-Supervised Learning). For the SSL method, we used a single constraint that requires the confidence level of the prediction to be greater than a fixed threshold of 0.9.

Finally, we tested the ASSLSM method (Table 2, ASSLSM) using a confidence level threshold of >0.7 , and a confidence level threshold of >0.8 .

The supervised learning models provided an average F1 score of 0.752 with a standard deviation of 0.04, using a dataset of only 4,157 labeled samples.

The SSL models with a confidence threshold of >0.9 presented a much higher average F1 score of 0.938 and a slightly higher standard deviation than the supervised learning models. These results represent an increase of almost 25% in the average F1 over the average F1 of the supervised learning models. The SSL approach significantly increased the number of samples in the labeled dataset by hundreds of percent with pseudo-labeled samples.

ASSLSM provided even better results with average F1, precision, and recall scores of 0.961, 0.971, and 0.958, respectively, using a confidence level threshold of >0.7 ; and 0.956, 0.969, and 0.953, respectively, using a confidence level threshold of >0.8 . ASSLSM also provided a significantly lower standard deviation. The additional constraints introduced in the ASSLSM method enable the use of a lower threshold for the confidence level of the predictions while achieving better F1 scores than those of the SSL method. The ASSLSM methodology performed better on average than the SSL methods, using a lower average number of samples because of the tighter constraints in ASSLSM.

The SSL methodology provided an average of 42,598 labeled and pseudo-labeled samples. ASSLSM provided a lower number of labeled and pseudo-labeled samples, with an average of 27,819 and 29,990, using a threshold of >0.7 and >0.8 , respectively. The standard deviation of the number of labeled and pseudo-labeled samples in the different models is also significantly larger in SSL (21,890) compared to ASSLSM (3,468 and 1,967). The results show that more samples do not necessarily imply better performance of the models. For example, the SSL Naïve Bayes model (80,423 labeled samples, mean F1 of 0.853) and the SSL k-NN model (49,600 labeled samples, mean F1 of 0.967) with a relatively larger number of samples, performed worse than the SSL logistic regression (36,552 labeled samples, mean F1 of 0.979) and the SSL voting ensemble (31,191 labeled samples, mean F1 of 0.968) models.

Table 2. Supervised, Semi-Supervised, and ASSLSM classification performance metrics.

Model	Labeled	F1	Precision	Recall
Supervised Learning				
k-NN	4,157	0.742 ± 0.053	0.756 ± 0.047	0.748 ± 0.050
Random Forest	4,157	0.763 ± 0.075	0.812 ± 0.047	0.780 ± 0.065
Naïve Bayes	4,157	0.681 ± 0.100	0.683 ± 0.098	0.688 ± 0.100
Logistic Regression	4,157	0.799 ± 0.056	0.812 ± 0.046	0.804 ± 0.051
Voting Ensemble	4,157	0.773 ± 0.066	0.785 ± 0.056	0.781 ± 0.060
<i>Average ± Std</i>	<i>4,157 ± 0</i>	<i>0.752 ± 0.04</i>	<i>0.770 ± 0.048</i>	<i>0.760 ± 0.04</i>
Semi-Supervised Learning with threshold >0.9				
k-NN	49,600	0.967 ± 0.024	0.967 ± 0.024	0.967 ± 0.024
Random Forest	15,222	0.923 ± 0.072	0.950 ± 0.037	0.922 ± 0.083
Naïve Bayes	80,423	0.853 ± 0.054	0.875 ± 0.047	0.843 ± 0.057
Logistic Regression	36,552	0.979 ± 0.037	0.980 ± 0.035	0.978 ± 0.038
Voting Ensemble	31,191	0.968 ± 0.043	0.971 ± 0.040	0.969 ± 0.043
<i>Average ± Std</i>	<i>42,598 ± 21,890</i>	<i>0.938 ± 0.047</i>	<i>0.949 ± 0.038</i>	<i>0.936 ± 0.05</i>
ASSLSM with threshold >0.7				
k-NN	28,151	0.959 ± 0.036	0.966 ± 0.025	0.958 ± 0.041
Random Forest	27,126	0.961 ± 0.044	0.976 ± 0.021	0.958 ± 0.052
Naïve Bayes	32,010	0.947 ± 0.045	0.960 ± 0.025	0.943 ± 0.053
Logistic Regression	31,546	0.968 ± 0.038	0.976 ± 0.024	0.965 ± 0.045
Voting Ensemble	31,116	0.969 ± 0.040	0.979 ± 0.021	0.966 ± 0.047
<i>Average ± Std</i>	<i>29,990 ± 1967</i>	<i>0.961 ± 0.008</i>	<i>0.971 ± 0.007</i>	<i>0.958 ± 0.008</i>
ASSLSM with threshold >0.8				
k-NN	27,793	0.958 ± 0.036	0.966 ± 0.025	0.957 ± 0.041
Random Forest	21,473	0.945 ± 0.057	0.968 ± 0.028	0.944 ± 0.063
Naïve Bayes	31,996	0.948 ± 0.045	0.961 ± 0.025	0.943 ± 0.053
Logistic Regression	28,931	0.964 ± 0.044	0.974 ± 0.026	0.961 ± 0.051
Voting Ensemble	28,902	0.966 ± 0.042	0.977 ± 0.022	0.962 ± 0.050
<i>Average ± Std</i>	<i>27,819 ± 3468</i>	<i>0.956 ± 0.008</i>	<i>0.969 ± 0.006</i>	<i>0.953 ± 0.008</i>

The results show that applying the ASSLSM methodology optimized the data provided to the models, resulting in more consistent performance results between the different models.

The dataset contains 82,129 tweets posted by 34,219 unique users. Only 42 users had 100 or more tweets in the labeled and unlabeled dataset, hence meeting constraint B, according to which the user who posted the predicted tweet has ≥ 100 tweets in the labeled and unlabeled datasets. These 42 users are responsible for 33,084 tweets (about a third of the dataset). It may therefore be presumed that participants in the discussion of a certain domain of misinformation are likely to be repeating actors, and therefore more predictions could be pseudo-labeled.

7 Analysis

We applied the ASSLSM methodology with a confidence level threshold of >0.7, which provided the best performance, to train a Voting Ensemble model and classify the unlabeled dataset. Table 3 presents the distribution of tweets in each classification label. The results show that tweets in both the misinformation and not-misinformation groups are posted by a relatively small number of authors (2,644 and 2,760 users, respectively). On the other hand, authors who posted tweets that are classified as irrelevant, typically quit the discussion after posting a little more than a single tweet on average.

Table 3. Distribution of tweets in each label.

Label	# of Tweets	% of Tweets	# of Unique Authors	Tweets/Author
Misinformation	7,412	9.0	2,644	2.803
Not misinformation	32,539	39.6	2,760	11.789
Irrelevant	42,178	51.4	30,530	1.382

We analyzed the time series of the tweet frequency in both groups. Figure 2 presents the daily tweet frequency of the misinformation and not-misinformation groups. Certain peaks are immediately visible and can be attributed to actual earthquakes that happened at the time. For instance, the highest peaks on February 10, 2021 and March 4, 2021 (Fig. 2, annotations 1–2), correlate with a Mw 7.7 earthquake near Loyalty Islands and a Mw 8.1 earthquake near Keramedac Islands. The peak on August 14, 2021 (Fig. 2, annotation 3) correlates with a Mw 7.2 earthquake near Haiti.

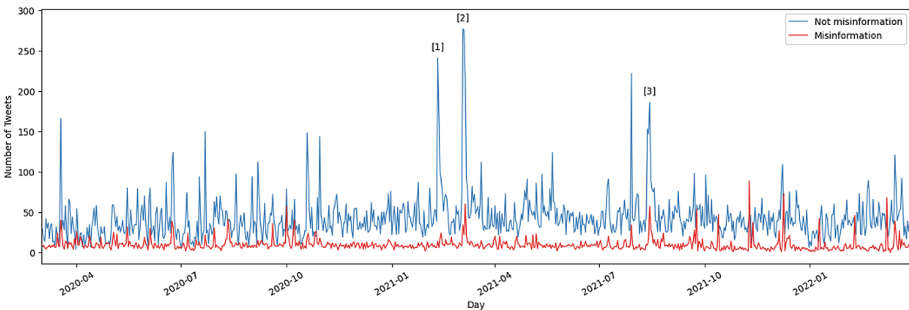


Fig. 2. Daily tweets frequency per category. The blue line represents the not-misinformation tweets, and the red line represents the misinformation tweets.

The cross correlation between the daily tweet frequency of the misinformation and the not-misinformation groups examines how one group dynamics depending on the dynamics of the other group. The results show a positive correlation of $r = 0.36$ between the daily tweet frequency in the misinformation and not-misinformation groups at time

t. This finding means that the tweet frequency of either group is associated with the tweet frequency of the other group on the same day.

8 Conclusion

In this study, we introduced the Adjusted Semi-Supervised Learning for Social Media methodology for the classification of misinformation tweets related to earthquakes prediction. ASSLSM takes into consideration useful metadata from social media that is not directly related to the text of the posts. We compared the performance of ASSLSM to the performance of supervised learning and SSL. We found that ASSLSM achieves significantly better results than supervised learning, and a model that is much more fit with more consistent results than SSL.

The results show that the additional constraints introduced in ASSLSM helped achieve better performance on average, while using a lower average number of samples than SSL. The variance of the results also decreased significantly when using ASSLSM. This finding suggests that using constraints that are more relevant to the data improves the performance and consistency of the models, despite reducing the number of pseudo-labels.

We used ASSLSM to classify the complete dataset of tweets related to earthquakes prediction into three categories, namely misinformation, not-misinformation, and irrelevant, and analyzed the resulting labeled dataset.

The analysis of the classified dataset shows that relatively small groups of authors are responsible for most tweets in the misinformation and not-misinformation groups. In the group of irrelevant tweets, however, more authors participate in the discussion but typically quit the conversation after a little more than one post in average. We also found that the daily tweet frequencies of the misinformation and not-misinformation groups are positively correlated and peak during an earthquake.

Valid information regarding ongoing events and the effective spread thereof, especially in case of potentially hazardous events, is important for public safety [28]. At the same time, the prevention of misinformation is of similar importance. As such, it can be recommended to communicate authoritative and correct information in a timely manner as an effective measure against misinformation on social media.

Future works can extend the ASSLSM to classify misinformation of other disaster and emergencies events in social media, by fine-tuning the constraints to better fit other datasets, as well as generalize the constraints to fit a wider variety of datasets, either on different topics or from different social media platforms.

Acknowledgements. This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101021746, CORE (science and human factor for resilient society).

References

1. Elroy, O., Yosipof, A.: Analysis of COVID-19 5G conspiracy theory tweets using sentence-BERT embedding. In: Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, 6–9 September 2022, Proceedings, Part II, pp. 186–196 (2022)
2. Komendantova, N., et al.: A value-driven approach to addressing misinformation in social media. *Human. Soc. Sci. Commun.* **8**, 1–12 (2021)
3. Aschwanden, A., Demir, C., Hinselmann, R., Kasser, S., Rohrer, A.: Zika and travel: public health implications and communications for blood donors, sperm donors and pregnant women. *Travel Med. Infectious Disease* (2018)
4. Ortiz-Martínez, Y., García-Robledo, J.E., Vásquez-Castañeda, D.L., Bonilla-Aldana, D.K., Rodríguez-Morales, A.J.: Can Google® trends predict COVID-19 incidence and help preparedness? the situation in Colombia. *Travel Med. Infect. Dis.* **37**, 101703 (2020)
5. Fallou, L., Corradini, M., Bossu, R., Cheny, J.-M.: Preventing and debunking earthquake misinformation: insights into EMSC’s practices. *Front. Commun.* **7**, 287 (2022)
6. Fallou, L., Marti, M., Dallo, I., Corradini, M.: How to fight earthquake misinformation: a communication guide. *Seismol. Res. Lett.* **93**, 2418–2422 (2022)
7. Cochran, E.S., et al.: Research to improve ShakeAlert earthquake early warning products and their utility. US Geological Survey (2018)
8. Huang, Y.L., Starbird, K., Orand, M., Stanek, S.A., Pedersen, H.T.: Connected through crisis: Emotional proximity and the spread of misinformation online. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 969–980 (2015)
9. Erokhin, D., Yosipof, A., Komendantova, N.: COVID-19 conspiracy theories discussion on Twitter. *Social Media + Soc.* **8**, 20563051221126051 (2022)
10. Elroy, O., Erokhin, D., Komendantova, N., Yosipof, A.: Mining the discussion of monkeypox misinformation on Twitter using RoBERTa. In: IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 429–438 (2023)
11. Yosipof, A., Senderowitz, H.: Optimization of molecular representativeness. *J. Chem. Inf. Model.* **54**, 1567–1577 (2014)
12. Batzdorfer, V., Steinmetz, H., Biella, M., Alizadeh, M.: Conspiracy theories on Twitter: emerging motifs and temporal dynamics during the COVID-19 pandemic. *Int. J. Data Sci. Anal.* pp. 1–19 (2021)
13. Darwish, O., et al.: Identifying fake news in the russian-ukrainian conflict using machine learning. In: Barolli, L. (eds.) *Advanced Information Networking and Applications. AINA 2023. Lecture Notes in Networks and Systems*, vol. 655. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28694-0_51
14. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
15. Piskorski, J., Haneczok, J., Jacquet, G.: New benchmark corpus and models for fine-grained event classification: to BERT or not to BERT? In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6663–6678 (2020)
16. González-Carvajal, S., Garrido-Merchán, E.C.: Comparing BERT against traditional machine learning text classification. arXiv preprint [arXiv:2005.13012](https://arxiv.org/abs/2005.13012) (2020)
17. Micallef, N., He, B., Kumar, S., Ahamad, M., Memon, N.: The role of the crowd in countering misinformation: a case study of the COVID-19 infodemic. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 748–757 (2020)
18. Reimers, N., Gurevych, I.: Sentence-bert: sentence embeddings using siamese bert-networks. arXiv preprint [arXiv:1908.10084](https://arxiv.org/abs/1908.10084) (2019)

19. Adoma, A.F., Henry, N.-M., Chen, W.: Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 117–121 (2020)
20. Naseer, M., Asvial, M., Sari, R.F.: An empirical comparison of bert, roberta, and electra for fact verification. In: 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), pp. 241–246 (2021)
21. Tarunesh, I., Aditya, S., Choudhury, M.: Trusting roberta over bert: insights from checklisting the natural language inference task. arXiv preprint [arXiv:2107.07229](https://arxiv.org/abs/2107.07229) (2021)
22. Ouali, Y., Hudelot, C., Tami, M.: An overview of deep semi-supervised learning. arXiv preprint [arXiv:2006.05278](https://arxiv.org/abs/2006.05278) (2020)
23. Beskow, D.M., Carley, K.M.: Bot-hunter: a tiered approach to detecting and characterizing automated activity on twitter. In: Conference paper. SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, vol. 3, p. 3 (2018)
24. ODonovan, J., Kang, B., Meyer, G., Höllerer, T., Adalii, S.: Credibility in context: an analysis of feature distributions in Twitter. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 293–301 (2012)
25. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: Tweetcred: real-time credibility assessment of content on Twitter. In: International Conference on Social Informatics, pp. 228–243 (2014)
26. Balaanand, M., Karthikeyan, N., Karthik, S., Varatharajan, R., Manogaran, G., Sivaparhipan, C.: An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter. *J. Supercomput.* **75**, 6085–6105 (2019)
27. Jan, T.G., Khurana, S.S., Kumar, M.: Semi-supervised labeling: a proposed methodology for labeling the twitter datasets. *Multimedia Tools Appl.* **81**, 7669–7683 (2022)
28. Yosipof, A., Woo, G., Komendantova, N.: Persistence of risk awareness: manchester arena bombing on 22 May 2017. *Int. J. Disaster Risk Reduction* 103805 (2023)