A Weakly-Supervised Iterative Graph-Based Approach to Retrieve COVID-19 Misinformation Topics

Harry Wang, Sharath Chandra Guntuku
University of Pennsylvania,
Department of Computer and Information Science,
Philadelphia, PA 19104, USA
harrywan@seas.upenn.edu, sharathg@seas.upenn.edu

Abstract
The COVID-19 pandemic has been accompanied by an ‘infodemic’ – of accurate and inaccurate health information across social media. Detecting misinformation amidst dynamically changing information landscape is challenging; identifying relevant keywords and posts is arduous due to the large amount of human effort required to inspect the content and sources of posts. We aim to reduce the resource cost of this process by introducing a weakly-supervised iterative graph-based approach to detect keywords, topics, and themes related to misinformation, with a focus on COVID-19. Our approach can successfully detect specific topics from general misinformation-related seed words in a few seed texts. Our approach utilizes the BERT-based Word Graph Search (BWGS) algorithm that builds on context-based neural network embeddings for retrieving misinformation-related posts. We utilize Latent Dirichlet Allocation (LDA) topic modeling for obtaining misinformation-related themes from the texts returned by BWGS. Furthermore, we propose the BERT-based Multi-directional Word Graph Search (BMDWGS) algorithm that utilizes greater starting context information for misinformation extraction. In addition to a qualitative analysis of our approach, our quantitative analyses show that BWGS and BMDWGS are effective in extracting misinformation-related content compared to common baselines in low data resource settings. Extracting such content is useful for uncovering prevalent misconceptions and concerns and for facilitating precision public health messaging campaigns to improve health behaviors.

Introduction
Since the COVID-19 pandemic hit the world, there have been multiple myths and misconceptions regarding the virus and associated health behaviors (Singh et al. 2020). Many of the sources and contents of this misinformation (ranging from false cures to government conspiracies) have been disseminated online in various forms (Cinelli, Quattrociocchi, and Galeazzi 2020; Kouzy et al. 2020; Siwakoti et al. 2021). Misinformation and myths, especially during a public health crisis, are harmful and can lead to increased death along with poor health outcomes. (Dharawat et al. 2020; Memon and Carley 2020; Ojo et al. 2021). Gauging public perceptions aids in identifying specific sources of misinformation and concerns that need to be addressed in a tailored manner.

Identifying and correcting misinformation is a challenging task given the different shapes and sizes it is propagated in (Koltai, Moran, and Grasso 2022). Given the large volume of social media and news article posts available, prior works have developed labeled COVID-19 datasets (Agarwal 2020; Memon and Carley 2020). These datasets label text based on whether misinformation is present and sometimes the type of misinformation that is present. Many approaches leverage posters’ profiles and lists of credible sources to label posts as either misinformation or non-misinformation (Vosoughi, Roy, and Aral 2018; Shahi, Dirkson, and Majchrzak 2020). Others rely on labeling either through professionals or crowd-sourcing (Vraga and Bode 2017; Meer and
Jin 2020; Mujumdar and Kumar 2021). We reviewed several existing approaches to analyzing misinformation and summarize them in Table 1. In this paper, we present a low-cost approach for retrieving COVID-19 misinformation related topics and themes with the goal of contributing to developing precision public health messaging campaigns to counter misinformation.

One of the challenges in developing misinformation detection models is sifting through large amounts of unlabeled misinformation related sources to find relevant data to train models on. Our approach provides a complementary solution to this problem by generating a list of keywords, topics, and themes related to COVID-19 based on given seed words. The outputs of our approach can inform public health experts and model developers during the data collection process by providing keywords and topics to look for. Social media posts containing the keywords and topics generated from our approach can then be included in training data and/or further processed/curated. We also demonstrate the application of our approach on a down-stream misinformation detection task to provide a quantitative analysis of its performance.

Another challenge of building misinformation detection models for a public health crisis during its early stages, such as the COVID-19 pandemic, is the lack of labels on pandemic-related misinformation. Fact-checking is a challenging task and can require large amounts of time and resources. While utilizing manual classification and/or supervision may achieve greater accuracy on a small amount of data, these approaches do not scale in terms of data volume and resources such as time and number of workers, especially in dynamically changing public health environments. Thus, it is important to develop a method to detect misinformation even in the presence of few labeled data points.

Information and facts on pandemics can evolve over its course (Brennen et al. 2020; Siwakoti et al. 2021). While manual labeling and annotation may provide a one-time solution to detecting misinformation, topics and sources of misinformation may change. New emergencies may also arise, rendering old labels obsolete. Thus, it is valuable to develop an automated, iterative method that requires significantly less human labor and manual inspection (Wang et al. 2018, 2020). This can be achieved by decreasing the amount of supervision necessary for extracting misinformation topics and can make the process of detecting misinformation more repeatable at a lower cost. We achieve this by building upon BERT-Based Word Graph Search (BWGS) (Santosh et al. 2020b) to propose Bert-Based Multi-directional Word Graph Search (BMDWGS) in conjunction with Latent Dirichlet allocation (LDA) topic modeling. These results can help medical policy makers and public health practitioners gauge public perception towards the pandemic situation, thus creating more targeted campaigns to inform health behaviors and correct misinformation. Complementary to previous works, our approach does not require scraping a list of pre-defined reliable/non-reliable social-media/news arti-

| Study | Annotation | Dataset | Domain | Platforms |
|-------|------------|---------|--------|-----------|
| (Wang et al. 2018) | Manual | Twitter: 7898 (fake), 6026 (real); Weibo: 4749 (fake), 4779 (real) | Fake News | Twitter, Weibo |
| (Wang et al. 2020) | Weakly Supervised | Labeled: 2090 (fake), 2090 (real); Unlabeled: 22981 | Fake News | WeChat |
| (Vosoughi, Roy, and Aral 2018) | URL matching | 126,285 rumor cascades corresponding to 2,448 rumors | Rumors | Twitter |
| (Dharawat et al. 2020) | Manual | 57,981 (real), 439 (possibly severe misinformation), 568 (highly severe), 1,851 (other), 447 (refutes/rebuts) | COVID-19 | Twitter |
| (Micallef et al. 2020) | Manual | IFCN: 6,840 (false); annotated 2,400 tweets each for 5G conspiracy theories and fake cures | COVID-19 | Twitter |
| (Memon and Carley 2020) | Manual | 4573 tweets annotated into 17 categories | COVID-19 | Twitter |
| (Cinelli, Quattricozzi, and Galeazzi 2020) | URL matching | Gab: 3,778; Reddit: 10,084, YouTube: 351,786, Instagram: 1328, Twitter: 356,448 | COVID-19 | Twitter, Instagram, YouTube, Reddit, Gab |
| (Shahi, Dirkson, and Majchrzak 2020) | Manual | 1274 (false), 226 (partially false) | COVID-19 | Twitter |
| (Sharma et al. 2020) | URL matching | 150.8K Misinformation Source tweets | COVID-19 | Twitter |
| (Brennen et al. 2020) | Manual | 225 misinformation posts | COVID-19 | Twitter, Youtube, Facebook |
| Our Approach | Weakly Supervised | COVID-19 Keywords Tweets (26M) (Santosh et al. 2020a); COVID19FN: 1596 (real), 1225 (fake) (Agarwal 2020); CM-COVID-19 (Memon and Carley 2020) | COVID-19 | Twitter, Articles |

Table 1: Summary of related works by Analysis Type, including whether or not manual inspection of tweets or tweet sources was done, scale of analysis, whether the analysis was done on a COVID-19 dataset, and the social media platforms analyzed.
Figure 3: Example run of Bert-Based Word Graph Search and sample results
and BMWG against two baselines: a BERT Text Classification model, and a Logistic Regression model that been trained on a Bag-of-Words (BOW) feature vector. For the BERT Text Classification and Logistic Regression baselines, we up-sample from the minority-class of misinformation-related tweets to address label imbalance. In later sections, we show that our approach outperforms the baselines in the low linguistic resource settings. The source code used in this paper is available at https://github.com/harryw1248/COVID19_Misinformation_Weakly_Supervised_BWGS

**BERT-based Word Graph Search (BWGS)**

We leverage BERT, a state-of-the-art Deep Bidirectional Transformers-based language model for constructing word-embeddings (Devlin et al. 2019). We fine-tune a bert-base-uncased model using a masked language modeling protocol with the following parameters: Num. Epochs: 5, Learning Rate: 1e-5, Batch Size: 8, Max Sequence Length: 128 (Santosh et al. 2020b). We obtained 768-dimensional word-embeddings from this model by summing the hidden states of the last 4 BERT layers. HuggingFace library is used for loading pre-trained model weights as well as Language Model training (Wolf et al. 2019).

Using the word-embeddings, we construct a word-graph where node locations are determined by the word-embeddings. Nodes in this graph represent words, and edge weights represent the degree of similarity between two words. A search is performed through this word-graph, where tokens have their embedding compared against a query embedding \( q \). If the cosine similarity between a token’s embedding and the query embedding \( q \) is below the minimum similarity threshold, \( \text{minSimThresh} \), the token is dropped and will not be explored further.

Previous work has found success in using a “Context Embedding,” \( \text{CEmb} \) to search through these word graphs. These Context Embeddings are constructed by using a seed word for initialization and integrating information from other words over iterations, following the formula:

\[
\text{CEmb} \leftarrow \frac{\text{CEmb} + \sum_{m=0}^{m} \text{Emb}(x)}{m+1}
\]

where \( m \) corresponds to the top \( m \) words at a certain depth in the search process. This creates

**Algorithm 1: BERT-Based Word Graph Search (BWGS) (Santosh et al. 2020b)**

**Require:** Initialized graph \( G \), initialized queue \( Q \), initialized Context Embedding \( \text{CEmb} \), minimum similarity threshold \( \text{minSimThresh} \), maximum depth \( \text{maxDepth} \), top k-words \( k \)

1. Pop next word from \( Q \), denoted by \( t \). Initialize a new node in \( G \) based on \( t \) and set the node embedding to \( \text{Emb}\{t\} \). Append \( t \) to \( \text{keywords} \).
2. Initialize the query embedding \( q \) as \( q \leftarrow k \times \text{CEmb} + (1-k) \times \text{Emb}\{t\} \).
3. Iterate through all tokens in the data. We compare the tokens’ embeddings against the query embedding \( q \). We drop all tokens with similarity less than the minimum similarity threshold \( \text{minSimThresh} \).
4. Select the top \( k \) words that are most similar to \( q \). Append these words to \( \text{keywords} \). Instantiate new nodes (if one does not already exist) for these words in \( G \) and add outgoing edges from \( t \) to these new nodes.
5. If all words for given depth are explored, the top \( k \) words corresponding to that depth are selected based on the query embedding \( q \). We drop all tokens with similarity less than the minimum similarity threshold \( \text{minSimThresh} \).
6. Terminate iterations when either \( Q \) is empty or when \( \text{depth} = \text{maxDepth} \). Otherwise, repeat from Step 1. If a stopping condition is met, return \( \text{keywords} \).

![Figure 4: Example run of Bert-Based Multi-directional Word Graph Search and sample results](image-url)
a more robust and comprehensive representation of the desired context. The hyperparameter, $k$, allows us to control the number of tokens that are considered at each depth in the search process. Thus, the full formula for the query embedding $q$ is $q = k \times C\text{Emb} + (1 - k) \times \text{Emb}(t)$ where $t$ is the current token being explored. A visualization for BWGS is shown in Figure 3 (Santosh et al. 2020b).

Algorithm 2: BERT-based Multi-directional Word Graph Search (BMDWGS)

Require: minimum similarity threshold $\text{minSimThresh}$, maximum depth $\text{maxDepth}$, top k-words $k$, list of seed words $\text{seedWords}$

Initialize empty list $\text{keywordsUnion}$

for seedWord in $\text{seedWords}$ do

1. Initialize $G$, queue $Q$, Context Embedding $C\text{Emb}$ using seedWord
2. Run BWGS with $G$, $Q$, $C\text{Emb}$, $\text{minSimThresh}$, $\text{maxDepth}$, $k$ to obtain keywords
3. $\text{keywordsUnion} \leftarrow \text{keywordsUnion} \cup \text{keywords}$

end for

return $\text{keywordsUnion}$

the Carnegie Mellon COVID-19 Tweet Dataset to validate our approach on a dataset with non-binary labels.

COVID-19 Keywords Tweet Dataset

Data and Model Hyperparameters

We utilize tweets collected in November 2020 related to COVID-19. This dataset was also used in other works related to studying emerging topics, symptoms, and varying health behaviors during COVID-19 (Van Loon et al. 2020; Zamani et al. 2020; Santosh et al. 2020a). We select a random subset of 10M tweets from over 26M total tweets. After filtering for English tweets, the dataset yielded over 5.5M tweets. These tweets serve as our training corpus for finding keywords in all of our experiments. For running BWGS, we used $\text{seedWord}$: “conspiracy”, $\text{similarityThreshold}$: 0.4, $\text{maxDepth}$: 2, and top $k$: 4. We then run LDA using the MALLET implementation in DLATK (Schwartz et al. 2017) with $\text{numTopics}$: 25 on these tweets to examine topics. We select “conspiracy” as the seed word because it is commonly associated with the context of misinformation. Additionally, many COVID-19 misinformation theories have been conspiracy-themed, such as relating COVID to government labs or certain political agendas. In subsequent analyses, we utilize other seed words, such as “claim” and “hoax,” as these words are commonly associated with misinformation and can be connected to a broader set of topics related to misinformation. Furthermore, during the early stages of a public health crisis, specific misinformation topics may be unknown. Thus, data collection could rely on general words related to the context of misinformation.

Findings

The COVID-19 Misinformation Keyword tokens returned by BWGS include: “fake”, “theories”, “racist”, “theory”, “fabricated”, “fraud”, “theorists”, “##cies”, “horrible”, “hatred”, “nationalist”, “disgusting”, “venture”, “conspiracy.”

According to our approach, 0.6717% of tweets (37026 out of the over 5.5M tweets) were labeled as COVID misinformation-related tweets. Due to the lack of labels for this dataset, we relied on the manual inspection of 100 tweets classified by our approach as misinformation-related and 100 tweets classified by our approach as non-
misinformation-related. Upon inspection, we found that 84 of the 100 tweets classified as misinformation-related were indeed true positives and that 91 of the 100 tweets classified as non-misinformation-related were indeed true negatives.

LDA topics associated with misinformation and myths were related to: Trump, politics, government conspiracies, 5G, China, vaccines, lockdowns, quarantine, COVID-19 affecting voter-fraud, racism, etc. We present the top words related to sample topics in Figure 5. While we started off with the general misinformation-related word “conspiracy,” our approach was able to retrieve a wide-range of specific misinformation-related topics due to shared context.

COVID19FN Dataset

Data and Model Hyperparameters We validate our approach using the COVID19FN dataset (Agarwal 2020). This dataset consists of labelled news articles of misinformation spread during the COVID-19 infodemic. It contains approximately 2800 news articles, real and fake, scraped from Poynter and other fact-checking sites. This dataset has binary labels that indicate whether an article is related to misinformation.

We experiment with different values for the hyperparameters in both BWGS and BMDWGS. For each set of hyperparameters, we measure precision, recall, and f1-score, three commonly used metrics in Natural Language Processing and Information Retrieval (Goutte and Gaussier 2005).

| Baseline          | Precision | Recall | F1-Score |
|-------------------|-----------|--------|----------|
| BERT Text Classification | 0.396     | 0.832  | 0.537    |
| Logistic Regression BOW | 0.171     | 0.0699 | 0.0992   |

| BWGS                |          |        |          |
|---------------------|----------|--------|----------|
| Seed Word(s)        | Similarity Threshold | Max Depth | Top k   |
| "conspiracy"        | 0.4      | 2      | 4        | 0.912  | 0.288 | 0.438 |
| "claim"             | 0.4      | 2      | 4        | 0.813  | 0.820 | 0.817 |
| "hoax"              | 0.4      | 2      | 4        | 0.936  | 0.323 | 0.480 |
| "false"             | 0.4      | 2      | 4        | 0.835  | 0.788 | 0.811 |
| "myth"              | 0.4      | 2      | 4        | 0.872  | 0.815 | 0.843 |
| "myth"              | 0.4      | 3      | 4        | 0.752  | 0.919 | 0.827 |
| "myth"              | 0.4      | 4      | 4        | 0.720  | 0.972 | 0.828 |
| "myth"              | 0.3      | 2      | 4        | 0.815  | 0.921 | 0.865 |
| "myth"              | 0.5      | 2      | 4        | 0.877  | 0.814 | 0.844 |
| "myth"              | 0.4      | 2      | 5        | 0.758  | 0.597 | 0.668 |
| "myth"              | 0.4      | 2      | 6        | 0.658  | 0.987 | 0.790 |

| BMDWGS              |          |        |          |
|---------------------|----------|--------|----------|
| ["conspiracy","claim"] | 0.4     | 2      | 4        | 0.837  | 0.822 | 0.829 |
| ["myth","claim"]  | 0.4      | 2      | 4        | 0.640  | 0.987 | 0.777 |
| ["myth","false"]  | 0.4      | 2      | 4        | 0.565  | 0.999 | 0.722 |
| ["conspiracy","myth"] | 0.4    | 2      | 4        | 0.834  | 0.795 | 0.814 |
| ["conspiracy","false"] | 0.4   | 2      | 4        | 0.814  | 0.897 | 0.853 |
| ["conspiracy","false"] | 0.4  | 3      | 4        | 0.768  | 0.918 | 0.836 |
| ["conspiracy","false"] | 0.4  | 4      | 4        | 0.709  | 0.960 | 0.816 |
| ["conspiracy","false"] | 0.3  | 2      | 4        | 0.815  | 0.896 | 0.854 |
| ["conspiracy","false"] | 0.5  | 2      | 4        | 0.824  | 0.879 | 0.850 |
| ["conspiracy","false"] | 0.4  | 2      | 5        | 0.565  | 0.998 | 0.722 |
| ["conspiracy","false"] | 0.4  | 2      | 6        | 0.565  | 1.00  | 0.722 |

Table 2: BWGS and BMDWGS Results on COVID19FN Dataset based on different hyperparameter sets; highest f1-scores are in bold

Findings We present our findings on this dataset in Table 2. We observe that most of the BWGS trials outperformed both the BERT Text Classification and Logistic Regression baselines. Of the 11 BWGS trials, 9 outperformed both of the baselines. BWGS with “myth” as the seed word, similarityThreshold = 0.3, maxDepth = 2, and top k = 4 yielded an f1-score of 0.865, the highest of all the BWGS trials on this dataset. It also yielded a recall score of 0.921 and a precision score of 0.815. It achieved 32.8% higher f1, 8.9% higher recall, and 41.9% higher precision compared to the BERT Text Classification baseline. The Logistic Regression Baseline did not achieve above 0.50 for any of the three metrics.

We observe that all of the BMDWGS trials outperformed both the BERT Text Classification and Logistic Regression baselines. BMDWGS with [“conspiracy”, “false”] as the seed word set, similarityThreshold = 0.3, maxDepth = 2, and top k = 4 yielded an f1-score of 0.854, the highest of all the BMDWGS trials on this dataset. It also yielded a recall score of 0.896 and a precision score of 0.815. It achieved 31.7% higher f1, 6.4% higher recall, and 41.9% higher precision compared to the BERT Text Classification baseline.

BWGS To analyze the effect of varying hyperparameters, we experimented with five different seed words with fixed similarityThreshold, maxDepth, and top k values. We then selected the seed word that achieved the highest f1-score to observe the effects of varying
**similarityThreshold, maxDepth, and top k values. Further experimentation that involves trying every hyperparameter combination could have been done. However, we were primarily interested in examining the effects of varying similarityThreshold, maxDepth, and top k values with respect to a certain seed word.**

We select “myth” as the seed word for testing different similarityThreshold, maxDepth, and top k values as it yielded the highest f1-score (0.843) in our first 5 trials. We observe that increasing maxDepth produced increasing trends in recall and decreasing trends in precision. This is expected, due to the fact that more keywords are being returned, increasing the number of articles retrieved, thus leading to more true positives and false positives. A similar trend can be observed with increasing the top k value. We found no definitive pattern with increasing similarityThreshold. Similar to many other information retrieval studies, we observe a general trade-off between precision and recall. However, multiple sets of hyperparameters were able to achieve greater than 0.800 in precision, recall, and f1-score.

**BMDWGS** Our analyses with BMDWGS are similar to those with BWGS except that five different seed word sets are used instead.

We select [“conspiracy”, “false”] as the seed word set for testing different similarityThreshold, maxDepth, and top k values as it yielded the highest f1-score (0.853) in our first 5 trials. Interestingly, the two seed words from the first 5 BWGS trials (“myth” and “claim”) that yielded the highest f1-score did not form the seed word set that produced the highest f1-score using BMDWGS. For BMDWGS, we observe similar patterns in precision and recall varying with similarityThreshold, maxDepth, and top k values.

BMDWGS achieves 17.0% higher average recall (0.920 versus 0.750), but has 9.20% lower average precision (0.722 versus 0.814). This is likely due to the fact that the keywordUnion returned by BMDWGS has a greater cardinality than the keywords set returned by BWGS given a fixed hyperparameter set, leading to more texts being returned. We also observe that overall, BMDWGS had more robust performance on this dataset – the standard deviation of f1-scores and recall scores across BMDWGS hyperparameter sets was 9.20% lower (0.0520 versus 0.144) and 16.4% lower (0.0693 versus 0.233) respectively, while the standard deviation of precision scores was only 2.99% higher (0.111 versus 0.0811). This is likely due to the fact that utilizing multiple seed words allows BMDWGS to take more contextual information into account, leading to a more comprehensive language representation. BMDWGS is also less likely to be sensitive to the effects of picking one particular seed word.

**Carnegie Mellon COVID-19 Tweet Dataset**

**Data and Model Hyperparameters** We also validated our approach using the Carnegie Mellon COVID Tweet Dataset because it contained multiple classes of labels (beyond binary) and multiple labels per tweet (Memon and Carley 2020). The dataset contains 4573 annotated tweets and 17 classes of labels. Our experiments on the COVID-19 dataset suggest that our approach is successful in detecting conspiracy-based misinformation. Thus, we aimed to retrieve tweets labeled as “conspiracy” and “calling out or corrections,” as they provide insights into perceptions towards conspiracies related to COVID. We experimented with several sets of hyperparameters for both BWGS and BMDWGS. We found that BMDWGS outperformed BWGS in terms of average f1-score and found the following hyperparameter set to yield the highest f1-score of our experiments: seedWordSet : [“conspiracy”, “myth”], similarityThreshold : 0.45, maxDepth : 6, and top k : 6. We found the following hyperparameter set to yield the highest f1-score for BWGS: seedWord : “conspiracy”, similarityThreshold : 0.55, maxDepth : 4, and top k : 4.

**Findings** We present our findings on this dataset in Tables 3 (baseline performance), 4 (BWGS performance), and 5 (BMDWGS performance). BMDWGS achieved higher f1-score and recall than BWGS while BWGS achieved higher precision. This is consistent with the explanations and observations from the experiments run on the COVID19FN dataset.

We observe that our approach reveals certain limitations of keyword based retrieval methods. The fact that this dataset contained non-binary labels and multiple labels per
tions allows them to be easily automated, complementing methods only require raw text input without manual annotation to utilize greater starting context. The fact that our BWGS method by introducing a novel BMDWGS algorithm to utilize greater context may arise. Thus, an approach for detecting topics related to COVID-19 misinformation. Our approach was able to extract misinformation related texts. Expand- ing them to n-grams would allow them to go beyond this scope. Our current approach achieved relatively high performance on datasets with binary labels, but further investigation is needed to improve performance on more complex datasets with several classes. Further exploration could also be done on analyzing misinformation regarding events other than COVID-19. While this work specifically analyzes COVID-related misinformation given current events, its methodology is meant to be domain-agnostic.

Our approach also has applications in Information Retrieval and Recommender Systems, particularly with Query Expansion. Query expansion is the process of supplementing additional terms or phrases to the original query to improve the retrieval performance (Imran and Sharan 2020). Our approach can be used to automatically search for related terms to a query based on learned context and embeddings in low data resource settings, alleviating the data cold-start problem in certain scenarios. The ability to search for related terms is also applicable in psychology research, where word lists such as LIWC are useful for analyzing participants’ mental states when analyzing words used in text messages (Tausczik and Pennebaker 2010).

Conclusion

We introduce a weakly-supervised iterative graph-based approach for detecting topics related to COVID-19 misinformation. Our approach successfully extracted many key topics related to misinformation and produced lists of important keywords while using little training data. The novel BMDWGS algorithm we propose also shows greater performance in terms of f1-score, recall, and overall robustness to hyperparameter tuning. Further work could explore the effects of using our approach for different domains, as well as expanding upon its methodology. The work presented in this paper sets the groundwork for developing low data-resource, weakly-supervised, graph-based approaches for misinformation detection.

Acknowledgements

This project was supported by Penn Global and the India Research and Engagement Fund.

References

Agarwal, I. 2020. COVID19FN. Mendeley Data.
Brennen, J. S.; Simon, F.; Howard, D. P. N.; and Nielsen, R. K. 2020. Types, sources, and claims of COVID-19 misinformation. *Reuters Institute, University of Oxford.*

Cinelli, M.; Quattricocchi, W.; and Galeazzi, A. e. a. 2020. The COVID-19 social media infodemic. *Sci Rep.*

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL (North American Chapter of the Association for Computational Linguistics).*

Dharawat, A.; Lourentzou, I.; Morales, A.; and Zhai, C. 2020. Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformation. *Arxiv.*

Goutte, C.; and Gaussier, E. 2005. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *European Conference on Information Retrieval.*

Imran, H.; and Sharan, A. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media. *EMNLP.*

Koltai, K.; Moran, R. E.; and Grasso, I. 2022. Addressing the root of vaccine hesitancy during the COVID-19 pandemic. *XRDS: Crossroads, The ACM Magazine for Students.*

Kouzy, R.; Jaoude, J. A.; Kraitem, A.; Alam, M. B. E.; Karam, B.; Adib, E.; Zarka, J.; Traboulsi, C.; Akl, E. W.; ; and Baddour, K. 2020. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *US National Library of Medicine.*

Meer, T. G. L. A. V. D.; and Jin, Y. 2020. Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source. *Health Communication.*

Memon, S. A.; and Carley, K. M. 2020. Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset. *The 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN 2020).*

Micalef, N.; He, B.; Kumar, S.; and Ahamad, M. 2020. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. *PrePrint - IEEE Big-Data 2020.*

Mujumdar, R.; and Kumar, S. 2021. HawkEye: a robust reputation system for community-based counter-misinformation. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* 188–192.

Ojo, A.; Guntuku, S. C.; Zheng, M.; Beidas, R. S.; and Ranney, M. L. 2021. How Health Care Workers Wield Influence Through Twitter Hashtags: Retrospective Cross-sectional Study of the Gun Violence and COVID-19 Public Health Crises. *JMIR Public Health Surveill.*

Santosh, R.; Schwartz, H.; Eichstaedt, J.; Ungar, L.; and Guntuku, S. C. 2020a. Detecting Emerging Symptoms of COVID-19 using Context-based Twitter Embeddings. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020.*

Santosh, R.; Schwartz, H. A.; Eichstaedt, J. C.; Ungar, L. H.; and Guntuku, S. C. 2020b. Detecting Emerging Symptoms of COVID-19 using Context-based Twitter Embeddings. *EMNLP 2020 workshop on COVID-19.*