Fighting against Fake News on Newly-Emerging Crisis: A Case Study of COVID-19

Migyeong Yang*
Sungkyunkwan University
Seoul, Republic of Korea
mgyang@g.skku.edu

Chaewon Park*
Sungkyunkwan University
Seoul, Republic of Korea
chaewonpark@g.skku.edu

Jiwon Kang
Sungkyunkwan University
Seoul, Republic of Korea
jiwonkang@g.skku.edu

Daeun Lee
Sungkyunkwan University
Seoul, Republic of Korea
delee12@skku.edu

Daejin Choi†
Incheon National University
Incheon, Republic of Korea
djchoi@inu.ac.kr

Jinyoung Han†
Sungkyunkwan University
Seoul, Republic of Korea
jinyounghan@skku.edu

ABSTRACT

As social media users can easily access, generate, and spread information regardless of its authenticity, the proliferation of fake news related to public health has become a serious problem. Since these rumors have caused severe social issues, detecting them in the early stage is imminent. Therefore, in this paper, we propose a deep learning model that can debunk fake news on COVID-19, as a case study, at the initial stage of emergence. The evaluation with a newly-collected dataset consisting of both the COVID-19 and Non-COVID-19 fake news claims demonstrates that the proposed model achieves high performance, indicating that the model can identify fake news on COVID-19 in the early stage with a small amount of data. We believe that our methodology and findings can be applied to detect fake news on newly-emerging and critical topics, which should be performed with insufficient resources.

CCS CONCEPTS

• **Security and privacy** → *Social aspects of security and privacy*; • **Computing methodologies** → **Natural language processing**; *Knowledge representation and reasoning*.

KEYWORDS

Fake News, COVID-19, Early detection

ACM Reference Format:
Migyeong Yang, Chaewon Park, Jiwon Kang, Daeun Lee, Daejin Choi, and Jinyoung Han. 2024. Fighting against Fake News on Newly-Emerging Crisis: A Case Study of COVID-19. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651506

*Both authors contributed equally to this research.
†Corresponding author.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651506

1 INTRODUCTION

Social media has become an integral part of our daily lives [6], allowing people to share common interests [16]. As a result, the use of social media has skyrocketed [6]. Unfortunately, the abundance of data and information on these platforms has led to a higher risk of data leakage, which in turn has caused various cybercrimes or social threats [9]. It is easy for social media users to generate and spread information at a low cost, regardless of its authenticity [12].

The situation becomes more serious, especially when the information is related to public health. A recent representative example would be the coronavirus disease, COVID-19. The infodemic by COVID-19 is reported as more critical than the fake news for other topics (e.g., politics) since it impedes people with insufficient knowledge about COVID-19 from adopting healthy practices, and spreads misperception of the government policy [17]. To make matters worse, the limited ability to fact-checking and monitoring mechanisms on social media platforms has made people more susceptible to encountering and believing in fake news [14], which can cause severe social problems such as animosity and hatred between various ethnic groups [18]. Therefore, it is essential to identify COVID-19 fake news in the early stage for maintaining social stability, and effectively preventing the spread of disease [1].

To prevent the dissemination of misinformation and its potential negative effect, diverse computational techniques and deep learning methods have been applied to detect COVID-19 fake news [3, 4]. Unfortunately, despite the valuable contribution to the identification of COVID-19 fake news, the models proposed by prior work were trained with a substantial amount of COVID-19 fake news data *after* a considerable period of time has elapsed since the onset of the COVID-19 pandemic. In this way, applying a fact-checking model on COVID-19 in the early phase of a COVID-19 outbreak is not easy; it is challenging to debunk COVID-19 fake news with the lack of relevant resources.

Therefore, in this paper, we propose a model that can debunk fake news on COVID-19 at its initial stage, where the resources related to COVID-19 are limited. In particular, we seek to answer the following research questions: (i) *How can we use models and data based on existing non-COVID-19 fake news to detect fake news on a newly emerging disease, COVID-19* and (ii) *How can the COVID-19 fake news data that are newly or additionally collected be used for accurate prediction at its early stage?*

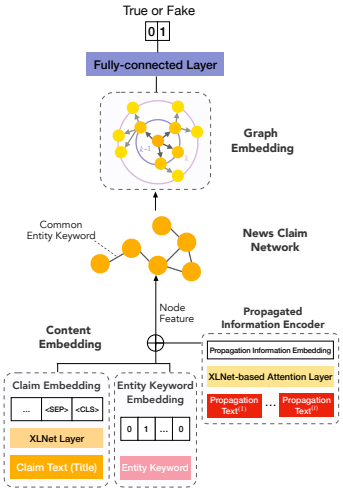

Figure 1: Overall architecture of the proposed model.

We highlight the main contributions of this paper as follows.

- We propose a fake news detection model that can effectively reveal falsehood from short claim text *in an early stage* by considering the relation among claims and incorporating additional features such as entity and propagated information.
- We provide empirical evidence of the effect of utilizing existing data to combat fake news related to newly arisen issues along with only a small amount of relative context where not enough data is available.
- We build a dataset of 7,133 COVID-19 and non-COVID-19 claims, which contains corresponding entity keywords and propagated information[1].

2 THE MODEL

We propose a deep learning model to debunk the claims of news related to COVID-19 at its early stage, considering not only the text information of the given news claim itself, but also the information of its related news articles and initial propagation in social media. To this end, we design three embedding layers: (i) Content Embedding (CE), (ii) Propagated Information Encoder (PIE), and (iii) Graph Embedding (GE). Figure 1 illustrates the overall architecture of the proposed model.

2.1 Content Embedding

We aim to extract the linguistic and semantic information of the given claim of a news article. To this end, we use a generalized autoregressive pretraining method (XLNet) [21], a popular model that can compute the vector representation of the content features for the given text. In particular, we put the claim text to the pre-trained XLNet model to obtain the feature vector of the given claim C, which is formally defined as follows:

$$C_{xlnet} = XLNet([w^0_{c_i}, w^1_{c_i}, ..., w^n_{c_i}]) \in \mathbb{R}^d \quad (1)$$

where w and n denote the tokens in the claim and the number of tokens, respectively. Note that d is the dimension size of the output embedding of XLNet.

[1]The data and code are available at https://github.com/DSAIL-SKKU/Fighting_Against_FakeNews_on_Emerging_Crisis-WWW24

In addition to the content features, we also encourage the proposed model to explicitly learn the information from the important keywords mentioned in news claims, by computing and concatenating entity keyword embeddings to content features. To this end, we first extract the entity names of all the claims using BERT-NER and count the appearance frequency of each keyword across all the claims. We then use the top 100 entity keywords in terms of the frequency of the appearance as a keyword set. Note that we excluded the name of news media, such as PesaCheck or PolitiFact, to consider only the topical information of the news claims. For a given claim C, we finally compute an entity keyword embedding K by representing the existence of the top 100 keywords as a 100-dimension vector.

2.2 Propagated Information Encoder

It is reported that initial propagation information of the fake news can be a key factor in identifying fake news [13]. Inspired by the prior work, we design the Propagated Information Encoder, which considers the propagated information of fake news on social media at an early stage. In particular, the Propagated Information Encoder first finds a set of relevant YouTube videos S by keyword searching with the extracted keywords by BERT-NER[2] and KeyBERT[3]. For the retrieved videos, we extract all the text information (e.g., title and description) and refine the texts by conducting special-characters removal, conversion to lowercase, and word unification. Since the videos in S may have different importance in debunking the given claim, we apply the attention mechanism that can give more weight to the important videos in the final decision. To this end, we first calculate the text features of each video s_i using pre-trained XLNet (similar to claim embedding), then concatenate the features of all the videos in S, which is formally defined as follows.

$$s^i_{xlnet} = XLNet([w^0_{s_i}, w^1_{s_i}, ..., w^n_{s_i}]) \quad (2)$$

$$S_{xlnet} = s^0_{xlnet} \oplus s^1_{xlnet} \oplus ... \oplus s^t_{xlnet}, \quad (3)$$

where $s_{xlnet} \in \mathbb{R}^{n \times d}$, $S_{xlnet} \in \mathbb{R}^{t \times n \times d}$, and t is the number of videos.

We then apply the multi-head attention for S_{xlnet}, followed by residual connection and layer normalization, which are calculated as follows.

$$M = LN(S_{xlnet} + MultiHead(S_{xlnet})) \quad (4)$$

$$T = LN(M + FFN(M)), \quad (5)$$

where $MultiHead$, FFN, and LN denote multi-head self-attention layer, feed forward network, and layer normalization, respectively.

From the computed representation T, we finally compute the propagated information feature $\widetilde{T}$ by concatenating the [CLS] token of the propagated information features, and apply a fully-connected layer to have the same dimension with the claim representation C_{xlnet} as follows.

$$\widetilde{T} = [T^0_{CLS}, T^2_{CLS}, ..., T^t_{CLS}] \in \mathbb{R}^{td} \quad (6)$$

$$S_{attn} = P \cdot \widetilde{T} \in \mathbb{R}^d, \quad (7)$$

where P denotes trainable parameters.

[2]https://github.com/kamalkraj/BERT-NER
[3]https://github.com/MaartenGr/KeyBERT

2.3 Graph Embedding

The proposed model not only extracts the information of the given claim, but also considers the information of other news claims that are topically related to the given claim, which enables the proposed model to use more information related to the given claim. An example is to debunk a claim "5G spreads coronavirus", which is one of the well-known COVID-19-related fake news claims with a relevant claim, "Telecommunications worker exposed circuit boards being installed in 5G towers bearing markings that read COVID-19." In this case, the proposed model can learn not only from the claim text but also from the other false claim, including more specific information.

To model the (topical) relations among the claims, we construct a News Claim graph $G = (V, E)$, where V is the set of true and fake news claims and E is the set of edges, which represents whether two nodes (claims) share the same topic or not. In particular, the adjacency matrix A of graph G is defined as follows.

$$A_{i,j} = \begin{Bmatrix} 1 & K^i \cap K^j \geq 1 \\ 0 & \text{otherwise} \end{Bmatrix} \tag{8}$$

where K^i and K^j denote the set of entity keywords of i-th and j-th node, respectively.

Based on the News Claim graph, we compute the node embedding of a claim by adopting a message-passing mechanism that uses the information of the neighbors, who can receive the messages from the given node, to represent the features of the given node. In particular, for a node $v \in V$, the proposed model first samples a fixed-size of neighbor nodes, denoted as $N(v)$. The features of the sampled nodes in $N(v)$ are then put into the LSTM [8]. The last output of the LSTM layer is concatenated with the feature vector of the given node for the final node representation. Note that the proposed model can consider the information of the multi-hop neighbors (e.g., neighbors of a neighbor) in an inductive manner by iteratively performing the process (sampling and aggregating). Formally, the process of sampling and aggregating at k-th iteration can be defined as follows.

$$h^k_{N(v)} \leftarrow LSTM[h^{(k-1)}_u, \forall u \in N(v)] \tag{9}$$

$$h^k_v \leftarrow \sigma\left(P^k \cdot [h^{k-1}_v \oplus h^k_{N(v)}]\right), \tag{10}$$

where P^k is trainable parameters and σ is a nonlinear activation function. Note that h^0_v is initialized to the node feature X_v, which is the concatenation of the content embedding and the embedding calculated from the initial propagation encoder. The final node representation $h \in \mathbb{R}^g$ for V in graph G is obtained, where g is the hidden dimension size of Graph Embedding layer. Note that this approach is similar to GraphSAGE [7] with an LSTM as an aggregator, a popular graph neural network model to represent node features.

2.4 Classification

The model finally predicts the falsehood of a given claim by feeding h into a fully-connected layer followed by the softmax function as follows:

$$\hat{Y} = softmax\left(\mathcal{F}(h)\right), \tag{11}$$

where $\mathcal{F}$ is the fully-connected layer. To train the proposed model, we use the cross-entropy loss for detecting a fake news claim as follows:

$$\mathcal{L}(\hat{Y}) = -\sum_{i \in V} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), \tag{12}$$

where y_i and $\hat{y}_i$ denote the ground-truth label and the predicted probability of node i, respectively.

		Exp. 1		Exp. 2	
		Non-COVID	COVID	Non-COVID	COVID
Train	True	2,245	-	2,245	396
	Fake	2,245	-	2,245	391
	Total	4,490		5,277	
Test	True	944			
	Fake	912			
	Total	1,856			

Table 1: Description of fake news data for Exp. 1 and Exp. 2.

3 EXPERIMENT

3.1 Dataset

To evaluate the performance our model, we construct a dataset that consists of both COVID-19 and non-COVID-19 news articles. We use two publicly available datasets on COVID-19-related news claims, FakeCovid [15] and CoAID [2]. We also crawled the non-COVID-19 news claims, which were published before 21 January 2020, the official date of the outbreak of COVID-19, from the two popular fact-checking services, Snopes and Politifact. In addition, we collected the title and description of relevant YouTube videos, uploaded before or after two weeks of the published date based on each claim, to be used in Propagated Information Encoder. The final dataset consists of 7,133 claims, which contain corresponding entity keywords and propagated information.

3.2 Experiment Setup

Experiemental Scenarios: We evaluate the proposed model in two different scenarios designed from a practical standpoint. First, we assume the situation that the critical event (i.e., COVID-19) has just happened and no data for the event is available. For this scenario, we use only claim texts of the non-COVID-19 news claims and the news graph built by the claims, denoted as *Exp. 1*. Second, we additionally use the news claims on COVID-19, assuming a few news claims with the context are available, denoted as *Exp. 2*.

Data Split: The description of the dataset is summarized in Table 1. For training in *Exp. 1*, we solely use the collected non-COVID-19 news claims. We added COVID-19 news claims published before 1st April in 2020 into the training set for *Exp. 2*. The COVID-19 news claims published after 1st April in 2020 are used as a test set, which is the same in both experiments. Note that 1st April in 2020 is the date after about three months from the COVID-19 outbreak, which assumes an early stage of the pandemic with the lack of the dataset to learn. While training, we also restrict the model to learning only the resources published before 1st April in 2020, to avoid the model using future resources.

4 RESULTS

Exp. 1 and Exp. 2. Table 2 summarizes the performance results in both scenarios, *Exp. 1* and *Exp. 2*, which assume the situation that no COVID-19-related news is available and a few news related to COVID-19 can be used, respectively. In *Exp. 1*, the proposed method outperforms all the other baseline models in terms of f1-score, which demonstrates that the proposed model can identify

Models		Exp. 1				Exp. 2			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
Text-based	SVM-LIWC	0.411	0.408	0.445	0.428	0.504	0.496	0.560	0.526
	SVM-TFIDF	0.574	0.551	0.725	0.626	0.726	0.741	0.679	0.709
	HAN [20]	0.600	0.565	0.805	0.664	0.732	0.717	0.751	0.733
	TextCNN [10]	0.636	0.595	0.816	0.687	0.754	0.738	0.778	0.756
	BERT [5]	0.585	0.568	0.647	0.605	0.817	0.846	0.769	0.805
	XLNet [21]	0.683	0.686	0.681	0.680	0.900	0.889	0.910	0.900
Graph-based	GCN [11]	0.519	0.509	0.626	0.561	0.785	0.742	0.862	0.798
	GAT [19]	0.534	0.514	**0.924**	0.661	0.782	0.725	0.898	0.802
Prior Work	EM [3]	0.677	0.683	0.679	0.676	0.920	0.923	0.921	0.920
	MiCNA [4]	0.647	0.667	0.643	0.632	0.900	0.903	0.900	0.900
Proposed Model		**0.711**	**0.711**	0.710	**0.710**	**0.962**	**0.963**	**0.961**	**0.962**

Table 2: Performance results in Exp. 1 and Exp. 2.

	Acc.	Pre.	Rec.	F1
CE Only	0.920	0.935	0.899	0.917
PIE Only	0.491	0.491	1.000	0.659
CE + PIE	0.928	0.914	0.942	0.928
CE + GE	0.946	0.934	0.957	0.945
CE + PIE + GE	0.962	0.963	0.961	0.962

Table 3: Performance in the ablation experiments.

fake news dealing with COVID-19 accurately even in a situation where no COVID-19 resource is available. Moreover, we find that the text-based models show similar or higher performance than graph-based models, which implies that retrieving information from non-COVID-19 news claims is less helpful in debunking COVID-19-related news claims or can even make the models inaccurate.

We next analyze the performance result in *Exp. 2*. Overall, the performance metrics of all the models are improved in *Exp. 2*, indicating that using COVID-19 resources significantly helps to debunk COVID-19-related news claims. Notably, the performance of the proposed model is significantly improved, showing that the proposed model can identify COVID-19-related fake news with high accuracy. The proposed model shows much higher performance than graph-based baselines, implying that the proposed model explores the deeper structure of the fake news claim network and captures more effective features from the graph. Additionally, the proposed model outperforms all the models in prior work, which implies that the proposed model is useful in accurately finding fake news despite the lack of relative resources, whereas prior models are not effective in a newly emerging issue.

Ablation Study. We conduct an ablation study to understand how each component of the proposed model plays a role in identifying COVID-19 fake news. In particular, we additionally evaluate the four different variants of the proposed model: (i) Content Embedding Only (CE Only), (ii) Propagated Information Encoder Only (PIE Only), (iii) CE + PIE, and (iv) CE + Graph Embedding (CE + GE). Note that we only consider the *Exp. 2* scenario here.

Table 3 describes the results of the ablation study. The models with a single component, such as CE Only and PIE Only, show lower performance than the models with combined components like CE + PIE and CE + GE. These results imply that the components used in the proposed model contribute to fake news identification in a complementary way.

5 CONCLUSION

We proposed a COVID-19 fake news detection model that uses (i) existing non-COVID-19 fake news as well as a few newly added COVID-19 related fake news, (ii) the propagated information of the claim on social media by its early state, and (iii) related fake news claims in terms of topic similarity. The evaluation demonstrated that the proposed model can debunk COVID-19-related fake news with high performance. We believe that the proposed method gives insight into improving the model performance in detecting fake news, especially on newly emerging issues, such as COVID-19.

ACKNOWLEDGMENTS

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2023-00230337, Advanced and Proative AI Platform Research and Development Against Malicious Deepfakes) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2023R1A2C2007625).

REFERENCES

[1] Danroujing Chen, et al. 2023. CNFRD: A Few-Shot Rumor Detection Framework via Capsule Network for COVID-19. *International Journal of Intelligent Systems* 2023 (2023).
[2] Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885* (2020).
[3] Sourya Dipta Das, et al. 2021. A heuristic-driven ensemble framework for COVID-19 fake news detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*. Springer, 164–176.
[4] Arkadipta De and Maunendra Sankar Desarkar. 2022. Multi-Context Based Neural Approach for COVID-19 Fake-News Detection. In *Companion Proceedings of the Web Conference 2022*. 852–859.
[5] Jacob Devlin, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the NACCL: Human Language Technologies, Volume 1*. 4171–4186.
[6] Maeve Duggan, et al. 2015. Social media update 2014. *Pew research center* 19 (2015), 1–2.
[7] William L Hamilton, et al. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on NeurIPS*. 1025–1035.
[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[9] Ankit Kumar Jain, et al. 2021. Online social networks security and privacy: comprehensive review and analysis. *Complex & Intelligent Systems* 7, 5 (2021), 2157–2177.
[10] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181
[11] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[12] Shudong Li, et al. 2022. False alert detection based on deep learning and machine learning. *IJSWIS* 18, 1 (2022), 1–21.
[13] Federico Monti, et al. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673* (2019).
[14] Sonia Mukhtar. 2021. Psychology and politics of COVID-19 misinfodemics: Why and how do people believe in misinfodemics? *International Sociology* 36, 1 (2021), 111–123.
[15] Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid–A Multilingual Cross-domain Fact Check News Dataset for COVID-19. *arXiv preprint arXiv:2006.11343* (2020).
[16] Santhoshkumar Srinivasan and Dhinesh Babu LD. 2021. A social immunity based approach to suppress rumors in online social networks. *International Journal of Machine Learning and Cybernetics* 12 (2021), 1281–1296.
[17] Samia Tasnim, et al. 2020. Impact of rumors and misinformation on COVID-19 in social media. *Journal of preventive medicine and public health* 53, 3 (2020), 171–174.
[18] Nicolas Velasquez, et al. 2021. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. *Scientific reports* 11, 1 (2021), 11549.
[19] Petar Veličković, et al. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
[20] Zichao Yang, et al. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the NACCL: human language technologies*. 1480–1489.
[21] Zhilin Yang, et al. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in NeurIPS* 32 (2019).