Stemming the Tide of Fake News about the COVID-19 Pandemic

Chih-Yuan Li[1], Soon Ae Chun[2], James Geller[3]

[1,3]New Jersey Institute of Technology, [2]College of Staten Island, City University of New York
{cl524, james.geller}@njit.edu, soonaechun@gmail.com

Abstract

While the world has been combating COVID, there has also been an ongoing "Infodemic," caused by the spread of fake news about the pandemic. Due to the rapid data sharing on social media, the impact of fake news can be quite damaging. Citizens might mistake fakes news for real news. Human lives have been lost due to fake information about COVID. Our goal is to identify fake news on social media and help stem the spread by deep learning approaches. To understand the different characteristics in fake and real news, we conducted behavioral and sentiment analyses between fake and real news regarding the COVID pandemic. We then further built detection models based on feature elimination, and we identified differences of model robustness based on selected features.

Introduction

Social media users are exposed to large amounts of information. In recent years, it has become harder to verify its authenticity. Some users distribute fake news due to evil intentions, ignorance, or for personal entertainment. While the COVID pandemic has led to unprecedented damage, fake news has also become an issue. For example, some users have accepted wrong reports that garlic or alcohol could prevent COVID. This in turn might have led them to ignore warnings about wearing masks and staying socially distanced, potentially leading to hospitalizations or even death. Fake news has also claimed that COVID is a hoax and vaccinations are ineffective or even dangerous. This misinformation has been termed an "Infodemic" (Rothkopf 2013).

In this paper, our goal is to use machine learning (ML) to find remedies for the Infodemic. The objectives of this paper are: (1) We aimed to provide deep learning-based detection models for differentiating fake news (=FaN) from real news (=ReN). (2) We identified feature differences between FaN and ReN with respect to lengths, expressed sentiments, and the use of hashtags and mentions. (3) We built models based on feature elimination of social media posts on COVID, to identify the feature influence on model robustness.

To help curb this Infodemic, we present a deep learning (DL) approach to distinguish between FaN and ReN. We compared BERT (Devlin et al., 2018), LSTM (Hochreiter & Schmidhuber, 1997), and DistilBERT (Sanh et al., 2019).

In order to further understand the features of FaN about COVID, we applied Natural Language Processing (NLP) techniques, and implemented statistical analyses. Several features were analyzed and compared between FaN and ReN, including sentiments, "concern indices," and the use of hashtags (e.g., #COVID) and mentions (e.g., @WHO). Then, to investigate how influential the features are for FaN detection, we built models based on feature elimination.

We are raising four research questions: Q1: Is there a difference of the expressed sentiments between FaN and ReN in social media posts regarding COVID? Q2: If yes, is the difference statistically significant? Q3: What are the distinguishing features between the uses of hashtags in FaN vs. ReN in social media posts regarding COVID? Q4: What are the distinguishing features between the uses of mentions in FaN vs. ReN in social media posts regarding COVID?

Related Work

Allcott et al. (2017) defined FaN as "news articles that are intentionally and verifiably false." Lazer et al. (2018) defined it as "fabricated information that mimics news media content." Trusting FaN can cost lives. In March 2020 in Iran, nearly 300 people died after ingesting methanol because of a FaN message "alcohol can wash and sanitize the digestive system" (Karimi and Gambrell 2020). FaN also caused riots in Novi Sanzhary, Ukraine (Korybko 2020).

Bojjireddy et al. (2021) presented an ML approach to recognizing misleading information. Ali et al. (2021) investigated the robustness of different DL architectural choices, Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and a recently proposed Hybrid CNN-RNN. Their experiments on (Kaggle fake news, ISOT, and LIAR) datasets suggest that RNNs are robust, compared to other architectures. Kaliyar et al. (2021) proposed a combined approach of different parallel blocks of single-layer deep CNNs with different kernel

sizes and filters and BERT used to handle ambiguity. FaN will likely cause further confusion of citizens and conflicts in society (Boyd et al., 2018; Ng 2018).

Dataset and Methods

We used the dataset by Patwa et al. (2021), which contains 4,480 ReN and 4,080 FaN about COVID. FaN items were collected from Facebook and Instagram posts, tweets, public statements, and press releases. They were verified as FaN by various fact-checking sites (Politifact 2020; Newschecker 2021; Boom Live 2021), and by tools such as Google fact-check-explorer, and (International Fact-Checking Network 2021). These sites present determinations about COVID and other topics, whether the items are fake or real. The ReN items were from Twitter using verified Twitter handles, including WHO (World Health Organization), CDC (Centers for Disease Control and Prevention), ICMR (Indian Council of Medical Research), etc. Each tweet was read by a human, and marked as ReN if it contained useful information on COVID. FaN and ReN examples are shown in Table 1.

Table 1: Examples of FaN and ReN.

Text	*Label*
Politically Correct Woman (Almost) Uses Pandemic as Excuse Not to Reuse Plastic Bag #coronavirus #nashville	*Fake*
Covid Act Now found "on average each person in Illinois with COVID-19 is infecting 1.11 other people. Data shows that the infection growth rate has declined over time this factors in the stay-at-home order and other restrictions put in place."	*Real*

Deep Learning Models

For FaN detection, we built BERT (Devlin et al., 2018), LSTM (Hochreiter & Schmidhuber, 1997) and DistilBERT (Sanh et al., 2019) models. BERT is a powerful DL system for language modelling, and is the first deeply bidirectional model. It uses bidirectional transformers, such that a transformer is used for converting a sequence using an encoder and a decoder into another sequence. LSTM (long short term memory) is a specific recurrent neural network (RNN) that can handle long term dependencies and in turn solve the problem of vanishing gradients. A common LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell. DistilBERT (Sanh et al., 2019) is a pre-trained version of BERT. It leverages knowledge distillation during a pretraining phase. Thus, it has fewer parameters than a corresponding BERT model (bert-base-uncased) (40%), while it retains 97% of its language understanding capabilities and runs 60% faster. The token-type embeddings and the poolers are removed from BERT, and the number of layers is reduced by a factor of 2. For each model we performed 5-fold cross validation.

Data pre-processing: We used all the words occurring in a post including words in hashtags and mentions. We use regular expressions to capture any word starting with a "#" or a "@". There were hashtags expressing the same meaning but in different representations, such as "Covid_19" and "covid19." We lowercased each hashtag and removed the punctuations to unify such hashtags.

Experimental Results

Detection Models: The performances of our DL models are in Table 3. FaN detection in BERT outperformed the other models. Compared with previous approaches, our BERT model achieved a higher accuracy than the models proposed by (Patwa et al., 2021) (see Table 4).

Table 2: Performance of our DL models.

Model	LSTM	BERT	DistilBERT
Accuracy	87.21%	***95.61%***	75.37%

Table 3: Accuracies of our models and previous approaches.

Model	**Accuracy**
Decision Tree (Patwa et al., 2021)	85.23%
Gradient Boost (Patwa et al., 2021)	86.82%
Our proposed LSTM	**87.21%**
Logistic Regression (Patwa et al., 2021)	92.76%
Support Vector Machine (Patwa et al., 2021)	93.46%
Our proposed BERT	**95.61%**

Behavioral and Sentiment Analysis

We also analyzed the characteristics of FaN and ReN through sentiment analysis, length of posts, hashtag and mentions.

Concern Index: We measured the sentiments of the news items to determine their emotional impact, using the Stanford NLP library (Manning et al., 2014). A post is labeled either as "Very Negative," "Negative," "Neutral," "Positive," or "Very Positive" (Table 2). To monitor the concerns expressed by FaN and ReN, we computed a "*concern index*" (CI) modifying the CI of (Ji et al., 2013). The higher the CI is, the bigger the negative sentiment that is expressed by the news item.

Definition 1. Concern index (CI)

$$CI = \frac{N}{N+P+1} \qquad (1)$$

N is the count of items with Negative and Very Negative sentiments. P is the count of items with Positive and Very Positive sentiments. We are purposefully not using the Neutral items for CI. Table 5 is used to derive the statistical significance of this sentiment analysis. The FaN items result in a higher CI than ReN items by 10% (72% vs. 62%). This 10% difference in CI is statistically significant, based on Z-score calculation. We obtained a Z-score of 9.3. A lookup (Social Science Statistics (n.d.)) of a two-tailed p-value from

the Z-score identified a p-value < 0.00001. Thus, the difference of CIs between FaN and ReN is highly significant.

Table 4: Five sentiment classes expressed by FaN texts.

Text	Sentiment
COVID-19 is no worse than other outbreaks that have occurred in every election year suggesting that the new coronavirus is being hyped to hurt President Donald Trump.	Very Negative
Italy has surrendered to the coronavirus pandemic as all the measures to control COVID-19 have been exhausted.	Negative
Did You Already Have Coronavirus?	Neutral
Native Americans in North Dakota will be the first subjects to receive a novel coronavirus vaccine	Positive
foundation is truly one of the most inspirational forces of social change	Very Positive

Table 5: Statistical Result of Sentiment Analysis

	FaN	ReN
Very Negative (VN)	2,512	1,794
Negative (Neg)	247	553
Neutral (Neu)	240	677
Positive (Pos)	503	716
Very Positive (VP)	578	740
Concern index (CI)	0.72	0.62

Length of Posting: ReN items are on average 40% longer than FaN items (175 vs. 125 in chars; 29.89 vs. 20.7 in words). We hypothesize that to defend the facts, scientists/authorities might need to draft longer paragraphs of precise content, e.g., to combat the claim "Garlic can cure COVID," it may be necessary to cite published studies.

Hashtag and Mention Analysis: Figure 1 and Figure 2 show the top 20 hashtags used in FaN and ReN. In FaN, there are 2,021 hashtags, 794 of which are unique, while in ReN there are 4,743 hashtags, 386 of which are unique. Hashtags in FaN tend to include inspiring and admonishing messages, such as "*staysafe*," "*indiawillwin*," "*wearamask*," etc. Figure 3 and Figure 4 present the bar graphs of the top 20 mentions in FaN and in ReN. In FaN, there are 669 mentions, 486 of which are unique, while in ReN, there are 2,090 mentions, 568 of which are unique. In FaN, the top mentions are political: "*realDonalTrump*," "*narandramodi*," etc. In ReN, top mentions are related to public health: "*MoHFW_INDIA*" (Ministry of Health and Family Welfare of India), "*DrTedros*" (Director General of WHO), etc.

BERT Models with Feature Elimination

As the second experiment, we trained BERT models with feature elimination to see the effects of hashtags and mentions in predicting the FaN. We compared the accuracies, as follows: 1) remove hashtags, 2) remove mentions, and 3) remove both.

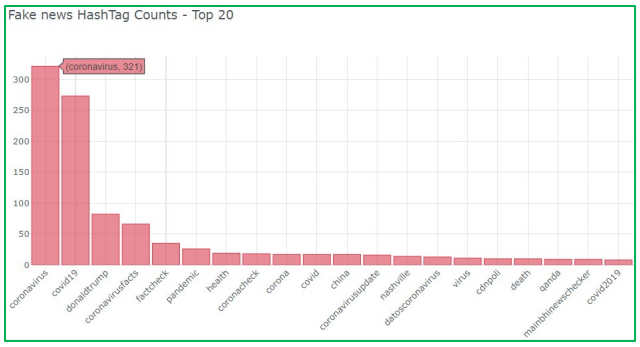

Figure 1: Bar graph showing top 20 hashtags used in FaN.

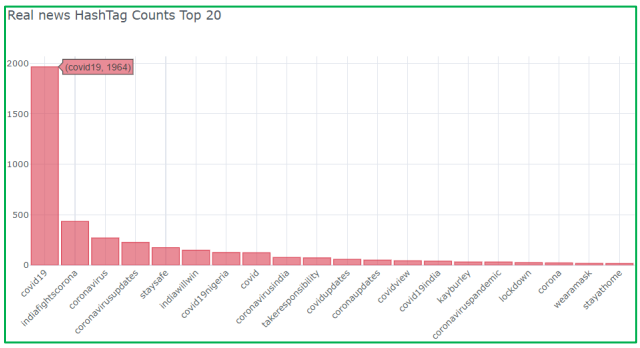

Figure 2: Bar graph showing top 20 hashtags used in ReN.

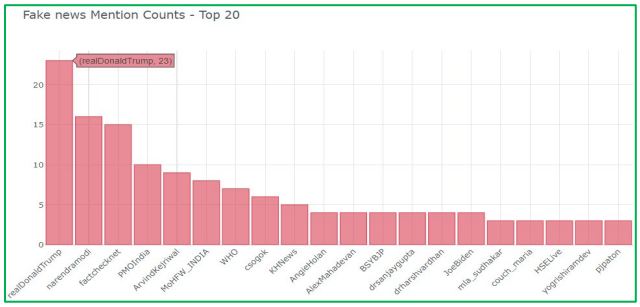

Figure 3: Bar graph showing top 20 mentions in FaN.

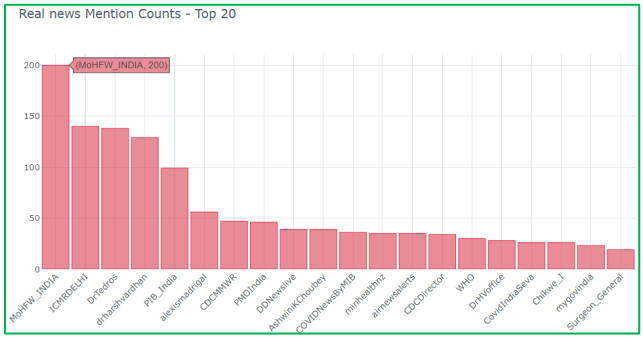

Figure 4: Bar graph showing top 20 mentions in ReN.

Out of 8,560 posts, there are 286 posts containing mentions, and 648 records containing hashtags. In addition, we found that there are no posts containing both a mention and a hashtag. We compared the accuracies among the models trained on the original text with all features retained, and text

with feature(s) eliminated (Table 6). The accuracies show small differences. The model trained on original text performed best, the model trained on the text with both features eliminated had the lowest accuracy. The other models were in between. The differences are too small to be significant. This might be due to the low number of data records containing mentions or hashtags.

Table 6: BERT Model with feature elimination.

Original Text	Eliminate Hashtag	Eliminate Mention	Eliminate both
95.61%	95.03%	95.41%	94.81%

Discussion, Conclusions and Future Work

FaN circulating on social media has created trust issues among citizens and discord in society. We built DL models for FaN detection based on a dataset regarding the COVID pandemic. Our BERT model achieved state-of-the-art results compared with previous studies. DL models with feature elimination show differences between detection models' robustness, though they are not significant.

Further analyses show that ReN posts are on average 40% longer than FaN. This implies that to recognize FaN, length can provide a hint. The CI of FaN is greater than that of ReN by 10%, which is statistically significant. This result answered Q1 and Q2. FaN contains more unique hashtags; ReN has more total hashtags. FaN and ReN prefer different hashtags. Hashtags in ReN include inspiring and admonishing messages, answering Q3. ReN contains more unique mentions and more total mentions than FaN. In FaN, top mentions are the handles of politicians and fact checking sites, while in ReN, top mentions are the handles of public health experts and institutes. The findings about mentions answered Q4. Currently, we are developing more powerful detection models by identifying data features and adjusting model variables that are necessary to achieve better transferability when working in different domains of FaN data. We are working on a platform where users can copy and paste news items and get immediate responses stating whether an item is likely to be real or fake. Social media operators could apply our research to build systems that block FaN posts, or show warning messages.

Acknowledgement

N. Kollapally has supported us with LSTM and DistilBERT.

References

Ali, H., Khan, M. S., AlGhadhban, A., Alazmi, M., Alzamil, A., AlUtaibi, K. and Qadir, J. (2021). All Your Fake Detector are Belong to Us: Evaluating Adversarial Robustness of Fake-News Detectors Under Black-Box Settings. IEEE Access 9, 81678–81692.

Allcott, H and Gentzkow, M. (2017). "Social Media and Fake News in the 2016 Election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236.

Bojjireddy, S., Chun, S. A. and Geller, J. (2021). Machine Learning Approach to Detect Fake News, Misinformation in COVID-19 Pandemic. The 22nd Annual International Conference on Digital Government Research (DG.O'21), 575–578.

Boom Live. (2021). https://twitter.com/boomlive_in

Boyd, R. L., Spangher, A., Fourney, A., Nushi, B., Ranade, G., Pennebaker, J and Horvitz, E. (2018). Characterizing the Internet Research Agency's Social Media Operations During the 2016 U.S. Presidential Election using Linguistic Analyses.

Devlin, J., Chang, M. -W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

International Fact-Checking Network. (2021). www.poynter.org/ifcn/

Ji, X., Chun, S. A., and Geller, J. (2013). "Monitoring Public Health Concerns Using Twitter Sentiment Classifications," 2013 IEEE International Conference on Healthcare Informatics, Philadelphia, PA, USA, 2013, pp. 335-344.

Kaliyar, R.K., Goswami, A. & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed Tools Appl 80, 11765–11788.

Karimi, N. & Gambrell, J. (2020). Hundreds die of poisoning in Iran as fake news times of Israel.

Korybko, A. (2020). Ukraine's anti-COVID-19 riot is due to fake news and media-driven fear. https://news.cgtn.com/news/2020-02-22/Ukraine-s-anti-COVID-19-riot-is-due-to-fake-news-and-media-driven-fear-Oi9eszIfzW/index.html

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, et al. (2018). "The Science of Fake News," Science, vol. 359, no. 6380, pp. 1094–1096.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60

Newschecker. (2021) from https://newschecker.in/

Ng, A. (2018). This was the most viewed Facebook ad bought by Russian trolls. https://www.cnet.com/news/this-was-the-most-viewed-facebook-ad-bought-by-russian-trolls/

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Shad Akhtar, Md., Ekbal, A., Das, A. & Chakraborty, T. (2021). Fighting an Infodemic: COVID-19 Fake News Dataset. Communications in Computer and Information Science, 21–29.

Politifact. (2020). https://www.politifact.com/

Rothkopf, D. J. (2003). When the Buzz Bites Back. www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd/

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv.

Social Science Statistics. (n.d.). P Value from Z Score Calculator. https://www.socscistatistics.com/pvalues/normaldistribution.aspx