Exploring New Methods for Identifying False Information and the Intent Behind It on Social Media: COVID-19 Tweets

Lin Ai, Run Chen, Ziwei Gong, Julia Guo, Shayan Hooshmand, Zixiaofan Yang, Julia Hirschberg

[1]Columbia University
New York NY 10027
lin.ai@cs.columbia.edu

Introduction

The detection of false information is an important task today, because its spread erodes the trust of people with their government and each other and leads to atmospheres of suspicion and growing political divides. Social media, though it has many benefits, such as helping friends stay connected, has contributed to the spread of false information because of its accessible, free, and highly connected nature. Over the past year, false information on social media has played a particularly large role in the perpetuation of false information about COVID-19: where it started, how serious it is, what cures are effective, and how to avoid infection. While there has been much research on how to identify false information, very little work has focused on the **intent** behind such falsification. In this paper we present our ongoing work on identifying not only false information about COVID-19 but the **intent** behind its production: is this false information created for purposes of malice or for other purposes? What are the different types of malicious and non-malicious purpose? Can we identify these automatically?

Our COVID-19 Corpus and ML Models

In our current work on detecting false information on COVID-19 in social media, we have collected a very large corpus of COVID-19-related Twitter data, containing tweets from January 2020 to December 2020, using 2 publicly available datasets (2020; 2020). We cleaned the corpus further by filtering by language (English-only), embedded URLs, and keywords extracted from COVID-19 debunked rumor statements. In this way, we produced a new corpus of COVID-19 tweets potentially balanced between rumor tweets and debunking tweets. From the 480M tweets we initially collected, we produced a cleaned corpus of 35M.

Using a semi-supervised approach, we weak-labelled a balanced subset of our cleaned corpus, where each tweet is labeled as ”trustworthy” or ”untrustworthy” based on the credibility of the URLs shared in the tweet. To verify the quality of the weak-labelling, we hand annotated 270 tweets, and the weak-labels achieves an overall accuracy of 0.71, with 0.61 trustworthy F1 and 0.77 untrustworthy F1.

We then built an end-to-end trustworthiness model using our weak-labeled dataset. Our model contains 2 main components: a graph-based model representing a user graph linked by retweet interactions, and a textual based model capturing textual features of the tweets. The graph-based model is based on the GraphSAGE model (2017), and user features are extracted using the official Twitter API and the Botometer API (2020). The textual based model is an RCNN model (2015), and the tweets are embedded using COVID-19 Tweets fine-tuned BERTweet embeddings (2020). The tweet representations learned from these 2 components go into an attention mechanism to generate a weighted representations, which is then passed through a feed-forward layer to perform class prediction. Trained and tested on our weak-labeled dataset, this model achieved an overall accuracy of 0.892. Currently, we are also collecting human annotations using Amazon Mechanical Turk to produce gold labels for our test dataset.

Intent Detection

We have also explored identifying different types of **intent** behind these the falsified information in our corpus: are the tweeters intending malice or something else? When people distribute false information, their purposes might be to: 1) persuade people to support individuals, groups, ideas, or actions; 2) persuade people to oppose individuals, groups, ideas or actions; 3) produce emotional reactions toward individuals, groups, ideas or actions; 4) educate on ideologies; 5) prevent an embarrassing or criminal act from being believed; 6) exaggerate the seriousness of something said or done; 7) create confusion over past incidents and activities; 8) demonstrate the importance of detecting false information to public (e.g., Elizabeth Warren and Mark Zuckerberg dispute); 9) convey sarcasm or humor (e.g., The Onion).

To better categorize malicious vs. non-malicious intents, we look at categories proposed by (2017): 1) Misinformation: information that is false, but the intent behind it is not harmful; 2) Disinformation: information that is false and deliberately created, where the intent behind it might be harmful to individuals or groups; 3) Malinformation: information that is sometimes based on reality, but is intended for causing harm. Example scenarios are listed in Table 1.

We began our initial exploration of intent detection by focusing on intent category 9 above: conveying sarcasm or

Mis-information	Dis-information	Mal-information
False Connection	Fabricated Content	Hate Speech
Misleading Content	Impostor Content	Harassment
Sarcasm/Humor	Propaganda	Leaks

Table 1: Examples of Mis-, Dis- and Mal-information

humor. For sarcasm, we collected a dataset of 3.5K tweets posted between January and December 2020 by selecting the COVID-19 related tweets with hashtags #sarcasm, #sarcastic, #irony, #satire, #irony and #lol, and an additional filter based on sarcastic markers (2018). We fine-tuned a BERTweet model (2020) on the iSarcasm dataset of 4484 author-labeled tweets (2020). Our model reached 0.67 accuracy on a held-out test set. We have also begun examining multi-modal features in videos.

Due to the subtlety and culture-specific characteristics of humor, it is very difficult to obtain reliable humor labels from crowd-sourcing methods. Here we have used a novel approach to obtain semi-supervised humor labels utilizing a unique function of Facebook posts: viewers can react to posts with various emotions, including a laughing emoji called “haha”. Our humor score is determined by this “haha” reactions ratio with a post popularity multiplier. We collected 1K COVID-19 related humorous posts by this score and 1K neutral posts by the “like” (the most neutral reaction) reactions ratio. We fine-tuned a BERT-based model on this dataset which achieved 0.91 F1 on a held-out test set. We plan to collect more noisy-labeled Facebook posts and adapt the model to tweets using domain adaptation techniques.

We have also begun to collect human annotations on other intents, such as the ones listed in Table 1, including propaganda. With a high quality intent-annotated dataset, we will build models to detect additional intent categories. This will help us to build a more general model to identify malicious or harmless intents behind untrustworthy tweets. Our models will also be augmented with modules capturing useful features, such as our neural models for identifying sentiment, emotions, and deception (2018).

Time-Sensitive False Information and Intent Detection

Another challenge for false information detection, especially in a rapidly evolving situation such as the COVID-19 pandemic, is that some information may be considered true in an early stage, but later proven to be false. For example, if a tweet contains information that was not considered to be false at the time that it was posted, but was later debunked, should it be labeled as untrustworthy? Was it created deliberately to spread false information, or was it posted because no one had enough knowledge about the subject discussed?

To address this issue, we will collect set of true claims from fact-checking websites such as Snopes and PolitiFact along with the debunking dates. We will then use topic extraction (2010) or sentence similarity (2020) techniques to find tweets with false information associated with these claims. This will give us a novel dataset of true claims, and their related tweets before and after the debunking dates. With this dataset, we will be able to learn characteristics of early stage false information, and subsequently, build models to detect the spread of early stage false information.

Acknowledgments

This research was done with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Banda, J. M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; and Chowell, G. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research–an international collaboration. *arXiv:2004.03688* .

Chen, E.; Lerman, K.; and Ferrara, E. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance* 6(2): e19273.

Ghosh, D.; and Fabbri, Alexander R. and Muresan, S. 2018. Sarcasm Analysis Using Conversation Context. *Computational Linguistics* 44(4): 755–792.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *arXiv:1706.02216* .

Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In *AAAI 2015*, volume 29.

Levitan, S. I.; Maredia, A.; and Hirschberg, J. 2018. Linguistic cues to deception and perceived deception in interview dialogues. In *NAACL:HLT, Vol. 1 (Long Papers)*, 1941–1950.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. In *EMNLP 2020: System Demonstrations*, 9–14.

O’Connor, B.; Krieger, M.; and Ahn, D. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4.

Oprea, S.; and Magdy, W. 2020. iSarcasm: A Dataset of Intended Sarcasm. In *ACL 2020*, 1279–1289. Online.

Passaro, L.; Bondielli, A.; Lenci, A.; and Marcelloni, F. 2020. UNIPI-NLE at CheckThat! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers. *Cappellato et al.[10]* .

Sayyadiharikandeh, M.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2725–2732.

Wardle, C.; and Derakhshan, H. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Technical Report 27, Council of Europe Report. 1–107.