NLP Based Detection of COVID-19 Misinformation

Sarthak Behl, Harshavardhan Kamarthi, Shlok Natarajan

Georgia Institute of Technology

The Problem

Misinformation or "Fake News" has become one of the leading causes of distrust on many popular social media sites. This was increasingly evident during the COVID-19 pandemic, where conflicting information on treatment and spread caused many to lose faith in media and even harmed patients. There have been a number of efforts to combat misinformation online using Natural Language Processing (NLP) and Deep Learning (DL) techniques. Prominent papers in this space focus on surveying misinformation detection techniques, applying them to the transcripts of videos on YouTube, and trying to classify individual tweets on Twitter as fake news using sentiment analysis and binary classification. We build on previous tweet classification research with the following key additions to previous literature:

- Use 3 novel COVID-19 specific datasets relatively unexplored by similar research and not been used in conjunction.
- Incorporation of sentiment analysis on tweets as well as its quote replies. This can be used as a measurement for emotional responses to the original tweet.
- Taking keywords from poster's bio related to occupation and account topic as we believe accounts dedicated to epidemiology/health are more likely to be factually correct

The Datasets

COVID Lies (CLDS)

- Contains 6591 tweets each labeled from a set of 62 misconception IDs which we converted into 2 binary classes.
- Labels were human-generated by researchers from the UCI School of Medicine

CoVaxxy

- 600k+ tweets curated based on COVID-19 keywords with a list of "High-Credibility" and "Low-Credibility" sources associated with the tweets
- We plan to use this data for BERT fine-tuning for potential improvement of model accuracy

Covid Cures Misinformation Dataset (CMDS)

- We extracted 2386 tweets (801 labeled misinformation)
- Contains tweets manually annotated for COVID-19 misinformation divided further into evidence-based misinformation and non-evidence-based misinformation which we combined into 2 classes (misinformation vs truth).

Tweet and User features

Engagement:

- Likes
- Retweets
- Replies
- Quote Tweets

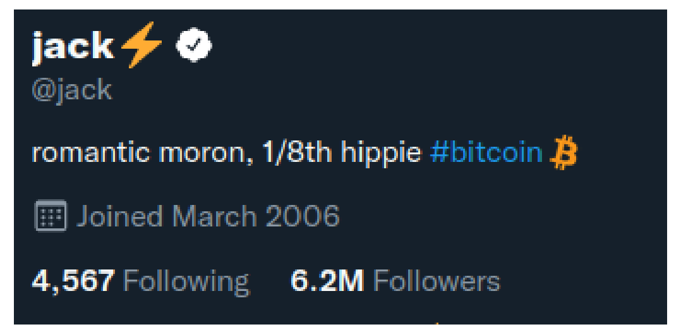
Tweet text:

- Text Embeddings
- Sentiments

(positive, negative, neither)

User engagement

- Followers
- Following
- Total Tweets
- Total tweets


jack⚡ ✓
@jack
romantic moron, 1/8th hippie #bitcoin ₿
Joined March 2006
4,567 Following 6.2M Followers

User description

- Text Embeddings
- Health-related
- Verified

Sentiment and public-health specific features

Sentiment of tweet

- 2 Models: Empath [2] and Transformer
- Used **Empath** to mine words related to positive and negative sentiment
- Transformers: used IMDB and Amazon reviews to train for sentiment.

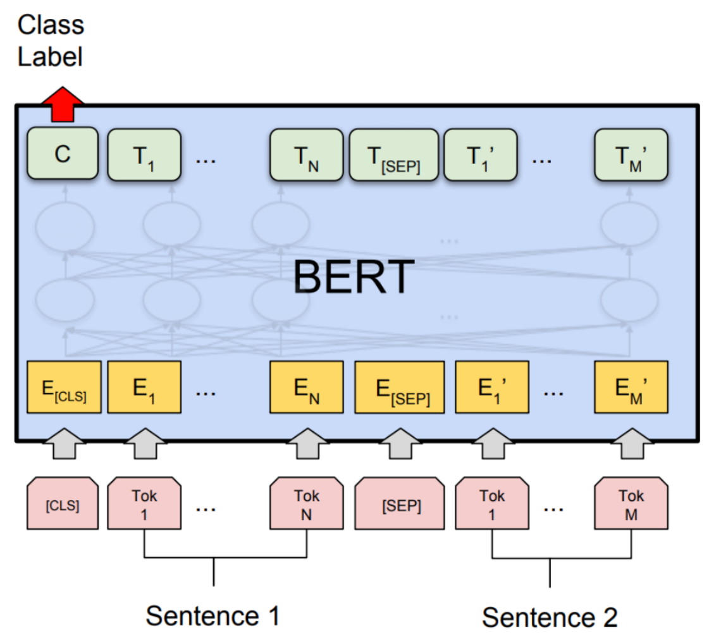
Health-related users

- Could relate to official sources
- **Empath** to mine words similar to "doctor", "physician", "hospital", "health", "disease"
- Classify user as health-related based on occurrence of these words

Embedding and Prediction Models

Text Embeddings

- Pre-trained Bert [1]
- Both tweet text and user description



Classification models applied and compared on the text embeddings + above features

- Logistic Regression
- Random Forest
- SVM
- Feed-forward Neural Network

Results

Features/Model	Logistic	Random Forest	SVM	FNN
Text Embeddings	0.66	0.68	0.7	0.7
+Engagement	0.68	0.7	0.66	0.7
+ User features	0.68	0.71	0.7	0.71
+Sentiment	0.71	**0.73**	0.68	0.7
+Health-related	0.71	**0.75**	0.69	0.71

- Evaluated on a 60%:40% train-test split
- Improved by 4% over similar past work [3]
- Additional Sentiment and Health-related features helped improve performance of Random Forest

Summary

We found that analysis on quote text replies, key words in the user's bio/description, in addition to commonly used features (original tweet embedding, like-count, etc.) modestly improved our model performance. We plan on using the CoVaxxy dataset to fine-tune BERT and incorporate additional sentiments extracted from quote text replies for further improvement.

References

[1] Devlin, Jacob et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL (2019).
[2] Fast, Ethan et al. "Empath: Understanding Topic Signals in Large-Scale Text." Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (2016): n. pag.
[3] Micallef, Nicholas et al. "The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic." 2020 IEEE International Conference on Big Data (Big Data) (2020): 748-757.