CS230

Cross Domain Training for COVID Misinformation Detection

Cong Kevin Chen
Dept. of Computer Science
Stanford University
ckchen95@stanford.edu

Mariana Frangos
Dept. of Computer Science
Stanford University
mfrangos@stanford.edu

Nick Walker
Dept. of Computer Science
Stanford University
nhwalk13@stanford.edu

1 Introduction

Throughout the last year, highly-connected networks and quickly-diffusing contagions have dictated public health and well-being. This can be observed not only in the physical spread of COVID-19, but also in the virtual diffusion of misinformation on Twitter. Such misinformation is harmful when remedying a public health crisis that requires both cooperative and well-informed citizens. In terms of the latter, we see it as our responsibility as computer scientists to assist in limiting the spread of misinformation online, which is our motivation for identifying duplicitous COVID-19-related tweets in this project.

While COVID-19 is a relatively novel topic (about which there remains much to be discovered), the battle against misinformation and fake news is not. However, if deep learning is able to recognize patterns common to misinformation concerning a variety of seemingly otherwise unrelated topics, it can considerably hamper malicious attempts to deceive the general public.

The input of our algorithm is text from tweets. We are fine-tuning existing models (see below), training on data from Codalab and CoAID that contain tweets labeled "real" or "fake". The output of our algorithm is binary: either the tweet contains misinformation or it does not.

2 Challenges

Any attempt to identify misinformation comes with inherent bias: who decides what is misinformation? Additionally, the quickly-evolving nature of the pandemic poses a challenge. A tweet containing what may be misinformation one day may be true in the future as more is learned about the virus. We can't entirely overcome these issues, but to remove our team's biases and to establish a stable and unchanging source of "truth", we have decided to use external data sets labeled by previous researchers. We recognize that this could make our algorithm ignorant of new developments that make once false claims true. However, we contend that most tweets fall outside of this category.

3 Related work

Barbieri et al. [1] fine-tuned roBERTa-base [2] on a 60 million tweets to tackle a variety of social media specific tasks such as sentiment analysis, hate speech detection, irony detection, and stance analysis. For nearly every task, they determined that this method outperformed pre-training roBERTa-base directly on the tweets. One experiment we plan to conduct is fine-tuning roBERTa-base on our target dataset.

BERTweet [3] follows the same architecture as BERT-base [4], but unlike BERT-base, has been pre-trained on a corpus of 868 million tweets, including 23 million tweets specifically related to

,

COVID-19. The model was fine-tuned on a variety of tasks such named entity recognition, sentiment analysis, and irony detection, and outperformed default roBERTa-base on all three tasks. [3] uses a mix of COVID-19 related twitter data and general twitter data. This has inspired our decision to use the CREDBANK-data dataset — consisting of tweets related to general events — in conjunction with COVID-19-specific data.

Hamid et al. [5] have analyzed tweets to detect COVID-19 and 5G misinformation. Using binary classification for fake-news detection, they obtained F1-scores of 0.666 and 0.693 for BoW and BERT based solutions, respectively.

They used both binary and ternary classification for their text-based fake news detection: in their binary classification task, they obtained average F1 scores of 0.666 and 0.693 for BoW and BERT based solutions, respectively; for ternary classification, the average F1 scores for BoW and BERT were 0.606 and 0.566, respectively. They also applied structure-based fake news detection, for which they relied on Graph Neural Networks and generated an average ROC of 0.95.

4 Datasets

The target dataset was curated for the shared task contest of the first annual Constraint workshop, aimed at combating online hostility and misinformation and held in conjunction with the AAAI conference that took place in February 2021. [6] contains 10,700 total tweets, each labelled "real" or "fake", and split into train/val/test by a ratio of 60:20:20. This is the dataset on which our models will be fine-tuned and tested. The preprocessing steps include removal of all emojis and URLs, and replacement of user mentions and hashtags with generic tokens.

We plan to fine tune a BERTweet architecture on two different datasets. The first is the Co-AID dataset [7]. It contains approximately 20,000 tweets and replies related to the 232 fake news articles, as well as approximately 280,000 tweets and replies related to 4,019 real news articles, dated May to November 2020. Per Twitter's Terms of Service, only the Tweet IDs are posted publicly, but the original text of most tweets has been retrieved using the tweepy API. The items are all dated from May 1 to November 1, 2020 and separated by whether they refer to real or fake sources.

The second is the CREDBANK-data dataset, specifically the "Credibility Annotation File." [8] This file contains tweets about 1,300 events, each ranked on a "credibility" scale ranging from -2 to 2. This is not a COVID-19 specific dataset, and includes tweets of various topics collected between 2014 and 2015. The topics are grouped by event (which will be one of the 1,300 in the "Credibility Annotation File"), and we define a "real" tweet as one that corresponds to an event with a positive credibility rating, a "fake" tweet as one that corresponds to an event with a negative credibility rating. One topic covered extensively by this dataset is the Ebola epidemic, so we hypothesize that this prior knowledge may be particularly transferable to our model's inferences about COVID-19.

5 Baseline Results

We used pre-trained Twitter-roBERTa-base and BERTweet models from the HuggingFace library to predict on the Constraint validation set. We obtained accuracies of 0.523 for Twitter-roBERTa-base and 0.551 for BERTweet. In addition, we will use Hamid et al.'s 0.693 binary F1 score using a BERT network [5] as an additional measuring stick. Since the target dataset is not very large and easy to overfit to, this may be one reason the baseline accuracies are low.

6 Methods

Due to the intensive computing costs of training any variant of BERT from scratch, most of our experiments will involve fine-tuning - for example, supervised fine-tuning of the last few layers of Twitter-roBERTa-base and BERTweet on the target dataset. We would also like to determine whether a network trained on a domain specific - in this case, myths and facts about COVID-19 - knowledge base could perform better on a task specific to that domain. Therefore, we chose to fine tune a pre-trained BERTweet separately on the Co-AID and Constraint datasets. This will likely involve some form of weighted sampling due to the imbalanced real/fake distribution in the Co-AID dataset; out of approximately 17,000 tweets, 15,000 were labelled "real" and the rest "fake". Certain

tweets in this dataset appeared multiple times based on the number of times they were replied to or retweeted, but due to time constraints, we were unable to parse out the duplicates, which do not all fit into patterns that can be detected by a regex script.

The concept of fine-tuning on a specific domain of interest was explored by Chen, Ramjee, and Wang [9] who determined that fine tuning an adaptive cross-domain variant of BERT-base on a dataset of abusive/hateful tweets increased the F1 score on a cyberbullying detection task by nearly 15%. This has inspired us to explore whether patterns affecting COVID-19-related tweets can be inferred from a richer pool of training data consisting of general event-related tweets. This will ultimately point us towards the feasibility of using one level of misinformation to recognize another level. The key to this may lie in general stylistic patterns (such as deliberately sensationalist vocabulary) that distinguish real info from fake.

7 Evaluation

In evaluating our model's performance, we will be using binary classification to predict whether or not a given tweet contains misinformation. We are measuring our model's precision, recall, balanced accuracy, and F1-score. We are recognizing "real" tweets to be positives, and "fake" tweets to be negatives. Considering the damage that tweets with misinformation can cause, we are focusing specifically on capturing true negatives. In other words, our main focus for evaluation is obtaining the highest Precision score possible, while also making sure our Recall is sufficient. This will ensure that we have the smallest amount of fake tweets labeled as real tweets, false positives, possible. In addition, we picked balanced accuracy, or the arithmetic mean of specificity and recall, instead of standard accuracy in order to account for the our data being imbalanced heavily in favor of examples labelled "real".

Since we are using binary data, our loss function will be Binary Cross-Entropy (BCE), or

$$\mathcal{J} = -\sum_{i}^{M} y_i \log \hat{y_i} + (1 - y_i) \log(1 - \hat{y_i}) \tag{1}$$

where y_i corresponds to the true label for training example i and $\hat{y_i}$ is the value predicted by the model.

8 Results and Setbacks

We fine tuned BERTweet for 5 epochs each on the target dataset [6], the CoAID dataset, and a concatenation of the two, using the Adam optimizer with weight decay (an alternative to L2 regularization) [10] and a learning rate of 1e-3. If the validation metrics plateaued, we enforced early stopping after the third epoch. Our results on the Constraint validation and test sets ($n = 2140$ each) are shown in Tables 1 and 2.

We could reasonably expect a fine tuned model to perform better than simply plugging the dataset into a model straight out of the box, but not necessarily by this much after only 5 epochs. The relatively small size of the datasets makes it easy to overfit, but this is unlikely to be the biggest concern since the test accuracy and F1 score were not significantly worse than that achieved on the validation set. However, when using CoAID only, the validation accuracy after three epochs was worse than that after one epoch, and the test accuracy was worse by about the same amount. The results improve significantly when the model is trained on both the CoAID and Constraint datasets, but overall the incorporation of CoAID still has a negative impact on overall performance.

A possible explanation for this is that the CoAID data is quite imbalanced compared to the Constraint dataset we're trying to predict. The large number of duplicates also calls into question how much actual information does the CoAID dataset provide for the model. In addition, many of the tweets are written in languages other than English, are unrelated to COVID-19 (such as the passing of Eddie Van Halen), or contain only a COVID-19 related question followed by an URL, where the contents of the URL determine whether the annotator will deem it real or fake. This is not the case for most of the Constraint tweets, which are also mostly complete sentences. While the authors of [9] demonstrated that cross-platform data sources can complement each other in the process of adaptive learning,

	Constraint	CoAID	Both
Balanced Accuracy	**0.9754**	0.6359	0.9318
Precision	**0.9757**	0.6523	0.9164
Recall	**0.9757**	0.6523	0.9696
F1	**0.9757**	0.6523	0.9304

Table 1: Fine tune performance on Constraint validation set

	Constraint	CoAID	Both
Balanced Accuracy	**0.9679**	0.5432	0.9392
Precision	**0.9687**	0.5645	0.9276
Recall	**0.9687**	0.5645	0.9607
F1	**0.9687**	0.5645	0.9402

Table 2: Fine tune performance on Constraint test set

this experiment does not appear to have produced that, as fine-tuning on noisy data worsened the classifier's performance.

Our team was delayed by getting approval to use the Twitter API in order to extract additional data from tweets, and also by the API's rate limits. This prevented us from collecting enough CREDBANK tweets to run any meaningful experiments with. To make matters worse, a significant portion of the CREDBANK and CoAID tweets came from deleted or suspended accounts. Generally we expect a majority of the tweets affected by this to have been labelled "fake", which would further skew the data distribution.

8.1 Qualitative Analysis

For our qualitative analysis, we looked at both fake tweets misclassified as real and real tweets misclassified as fake. We focused on analyzing mislabeled examples in order to identify causes of misclassification.

From the table below, one of the tweets that was fake mislabeled as real reads "I'm about to deliver remarks on the coronavirus pandemic. Tune in to watch live: <URL>". We believe that this tweet may have been misclassified because the tweet alone does not provide enough context to have it labeled as fake; even though it is a fake tweet, possibly due to the author impersonating a credible source, it does not appear to be fake when it is taken out of context.

Another fake tweet mislabeled as real is the following: "10 Million People contracted Tuberculosis last year. 1.5 Million People DIED. Did you even know? Were you scared for your life? Did we wear masks, close the economy, cancel schools, and ruin small businesses? No. Why? Because the media didn't tell you to be AFRAID!" This tweet may have been classified as real because of its use of statistics and medical terms. In many real COVID-19 tweets, we see the word "mask" or quantitative data about the spread of the virus. Since this tweet incorporates those both of those things, our model may have interpreted its message to be factual.

Overall, we noticed more fake tweets being misclassified as real than real tweets misclassified as fake, and this could be due to the skewed distribution of our training data, where there were many more real tweets than fake tweets. If someone wanted to spread fake information, a common way to do it would be to use authoritative vocabulary and statistics that a neural network would typically recognize as real; conversely, a source of real information would be unlikely to use the provocative tone and language that would lead to it getting classified as fake, but there are always exceptions.

One real tweet that was misclassified as fake is "Brazil is a worrying combination of pandemic and pandemonium." One hypothesis as to why this tweet was incorrectly labeled as fake is its inclusion of

Fake misclassified as real	Real misclassified as fake
"I'm about to deliver remarks on the coronavirus pandemic. Tune in to watch live: <URL>"	"Brazil is a worrying combination of pandemic and pandemonium."
"10 Million People contracted Tuberculosis last year. 1.5 Million People DIED. Did you even know? Were you scared for your life? Did we wear masks, close the economy, cancel schools, and ruin small businesses? No. Why? Because the media didn't tell you to be AFRAID!"	"People are drinking sanitizer to get an alcohol high a dangerous trend."

Table 3: Common mistakes made by the model fine-tuned on CoAID + Constraint data

"panic" words, such as "worrying" and "pandemonium." In the context of COVID-19 related tweets, we have noticed that many fake tweets use language intended to cause worry or panic. Use of this language in this real tweet may be why our model labeled it as fake.

A second real tweet that was misclassified as fake is "People are drinking sanitizer to get an alcohol high a dangerous trend." This tweet may have been classified as fake because even though it is mentioning a real trend, the trend itself is problematic and not a safe idea. The phrase "drinking sanitizer" most likely had an impact on the mislabeling of this tweet; since drinking sanitizer for any reason is ill-advised, the model might have seen the phrase as an indication that the tweet contained misinformation. In reality, the tweet was talking about a trend that was most likely fueled by misinformation, but our model does not pick up on that subtlety. Also, the use of the panic word "dangerous", which is prominently used in fake COVID-19 tweets, could have also caused our model to identify this tweet as fake.

9 Future Work

Future work in this space could include the use of more than two classes for misinformation ratings (beyond the current binary classification). A scale rating of very true, partly true, neither true nor false, partly false, and very false could be used. Many datasets exist that use this rating scheme, so this would be a practical and useful extension of our current model.

Another way to extend our current work would be stance detection, which is used to predict the stance that is held by an individual regarding a particular topic. For example, in our project, stance detection could have been useful to apply to tweets like the hand sanitizer tweet in the qualitative analysis section. This tweet was real, but misclassified as fake, because it discussed a behavior associated with believing misinformation (drinking sanitizer to get drunk), yet addressed it in a true/honest way (stating that drinking sanitizer is dangerous). If a classifier could detect a speaker's stance on a particular issue, then we could tell what their true feelings are about it, which could help determine if an assertion is truly misinformation or just sarcasm/reprimanding misinformation.

However, this idea of stance identification can extend even further, while staying within our topic of interest of COVID-19. Perhaps one's stance on COVID-19 could be predicted based on a feed of tweets/statements they have made, or their political affiliation could even be predicted given a prediction of their stance on COVID-19.

10 Contributions

Kevin Chen: I decided on the model architecture based on work that I had done previously and found the main target data source that required relatively little cleaning. I also implemented and executed the fine-tuning code on AWS.

Mariana Frangos: I wrote the slides for the final video presentation, found the CREDBANK-data dataset, and wrote large portions of each of our papers. I also helped Nick debug the Python script and understand the Tweepy API.

Nick Walker: My contributions, beyond more standard contributions to paper-writing/meeting attendance, to the model were primarily data-related. I filled out the application for the Twitter Developer account and was accepted. After this, I used Tweepy and our access to the Twitter API, learning via reading through the documentation for each, to write the Python script that collected the CoAid data (the dataset only had files labeled fake or real that contained lines of tweet IDs, the script collected the actual text and formed a CSV with columns for tweet text and corresponding label of true or fake). Finally, I did further research into datasets because we still weren't convinced that we had enough data.

References

[1] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification, 2020.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[3] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] Abdullah Hamid, Nasrullah Shiekh, Naina Said, Kashif Ahmad, Asma Gul, Laiq Hassan, and Ala Al-Fuqaha. Fake news detection in social media using graph neural networks and nlp techniques: A covid-19 use-case, 2020.

[6] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset, 2021.

[7] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset, 2020.

[8] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations, 2015.

[9] Cong Chen, Sharan Ramjee, and Joseph Wang. Aspect-Target Sentiment Classification for Cyberbullying Detection. `http://http://web.stanford.edu/class/cs224n/reports/final_reports/report000.pdf/`, 2021. [Online; accessed 18-April-2021].

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.