NARNIA at NLP4IF-2021: Identification of Misinformation in COVID-19 Tweets Using BERTweet

Ankit Kumar*, Naman Jhunjhunwala*, Raksha Agarwal, Niladri Chatterjee
Indian Institute of Technology Delhi
Hauz Khas, Delhi-110016, India
{mt1170727, mt1170737, maz178296, niladri}@maths.iitd.ac.in

Abstract

The spread of COVID-19 has been accompanied with widespread misinformation on social media. In particular, Twitterverse has seen a huge increase in dissemination of distorted facts and figures. The present work aims at identifying tweets regarding COVID-19 which contains harmful and false information. We have experimented with a number of Deep Learning based models, including different word embeddings, such as Glove, ELMo, among others. BERTweet model achieved the best overall F1-score of 0.881 and secured the third rank on the above task.

1 Introduction

Rapid propagation of social media has revolutionized the way information is consumed by general public. The ability of web platforms, such as Twitter, Instagram and Facebook, to quickly and broadly disseminate huge volumes of information has encouraged any user to be a (super) conduit of information. This can be helpful for problem solving in stressful and uncertain circumstances. However, this has also raised serious concerns about the disability of naive internet users in distinguishing truth from widespread misinformation.

As the world reacts to the COVID-19 pandemic, we are confronted with an overabundance of virus-related material. Some of this knowledge may be misleading and dangerous. The wildfire of Fake News in the times of COVID-19 has been popularly referred to as an 'infodemic' by the WHO chief. Also, in literature, we see terms such as 'pandemic populism' and 'covidiocy' (Hartley and Vu, 2020). Distorted facts and figures formed by drawing false equivalence between scientific evidence and uninformed opinions and doctored videos of public figures have flooded the online space since the onset of COVID. In order to ensure safety and well being of online information consumers, it is crucial to identify and curb the spread of false information. Twitter should mark content that is demonstrably inaccurate or misleading and poses a serious risk of damage (such as increased virus transmission or negative impacts on public health systems). Hence, developing and improving classification methods for tweets is need of the hour.

In the present work, Fighting with Covid19 infodemic dataset (Shaar et al., 2021) comprising English tweets about COVID-19 has been utilised for identifying false tweets. Many Deep Learning models have been trained to predict several properties of a tweet as described in Section 3.

The rest of the paper is organized as follows. Section 2 discusses related research work. Section 3 describes the dataset and Section 4 describes the language models we have used for our predictions. Sections 5 and 6 report the results of the experiments we conducted for the different language models and the error analysis respectively. Finally, in Section 7, we discuss future work that can be done in this area and conclude our paper.

2 Related Work

Classification of tweets has been studied widely by many researchers. Most of the methods use traditional Machine Learning classifiers on the features extracted from individual tweets, such as POS, unigrams, bigrams. Gamallo and Garcia (2014) built a Naive Bayes classifier for detecting sentiment of tweets. They considered Lemmas, Polarity Lexicons, and Multiword from different sources and Valence Shifters as input features to the classifier.

In recent times, the advancement of deep learning approaches (e.g., neural networks and transformer-based pre-trained language models like BERT and GPT) have taken precedence over feature-based classifiers (e.g., Naive-Bayes, SVM, among others). Classification problems have primarily been tackled in two ways - Feature

[*] Joint First Author

based and by Fine-tuning of parameters. Feature based approaches use word-embeddings, such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), ELMO (Peters et al., 2018), and feed them into some Deep Learning model to perform downstream task. On the other hand, parameter fine-tuning based approach fine tunes all the pre-trained parameters on downstream tasks. We have experimented with both these approaches.

Recently, language models such as BERT (Devlin et al., 2018), pre-trained on large amount of unlabelled data and fine tuned on downstream task, have given state-of-the-art results in numerous NLP tasks. BERTweet (Nguyen et al., 2020) is one such model which is pre-trained on English tweets. It has been found that BERTweet outperforms other state-of-the art language models, e.g RoBERTa, XLM-R (Conneau et al., 2019) with respect to several NLP tasks, viz. text classification, NER etc. This motivates us to use BERTweet based approach for this task.

3 Dataset Description

The dataset used in this task contains English tweets, and the corresponding labels (which are mainly "yes"/"no"), that are the answers to the following questions:

1. Does the tweet contain a verifiable claim?
2. Does the tweet appear to contain any false information?
3. Will the tweet be of any interest to the public?
4. Can the claim made be harmful to society?
5. Does the claim need any verification?
6. Is the tweet harmful or misleading the society?
7. Should the govt pay any attention to the tweet?

As per the dataset specifications, Q2 to Q5 will be NaN if and only if Q1 is "no". Further, Q1, Q6 and Q7 are never supposed to be NaN. If there are some instances where this condition is violated, we have dropped the corresponding tweets (independently for all the questions) during training or validation. Finally, for the final predictions, we first obtain the predictions for Q1, and the tweets are checked for the labels Q2 to Q5 only when Q1 is "yes".

The given dataset has 869 tweets in the train dataset. We randomly split the dataset for training and in-sample validation purposes, with the splits having 695 and 174 tweets respectively ($80 - 20$ split). For validation, we are given a dev dataset with 53 tweets. The test dataset on which we submit our final predictions contains 418 tweets.

4 Model Description

A vast number of Language Models have been developed in the last decade. We used a number of them to solve the given problem, and they are described in the following subsections.

4.1 Pre-trained Embeddings

BERT and its variants have successfully produced state-of-the-art performance results for various NLP tasks. BERTweet is one such variant, which has been pre-trained for English tweets. It has three variants, that differ on the data they are trained on:

1. **Base:** This model has been trained on 845M (cased) English tweets along with 5M COVID-19 tweets.
2. **Cased:** It has been trained on additional 23M COVID-19 (cased) English Tweets
3. **Uncased:** It has been trained on additional 23M COVID-19 (uncased) English Tweets

However, using the pre-trained embeddings provided by BERTweet may not give the best results since they have been trained for a different dataset. So, to fine-tune the model for our task, we plug the BERTweet model to a fully connected neural network. We vary the number of hidden layers, optimization function (Adam and AdaFactor), learning rate and the number of epochs. Thus, for each label, we try all three of the BERTweet variants, and choose the best one depending upon the F1-score obtained.

Additionally, we have experimented with GloVe and ELMo embeddings. We have used the GloVe Twitter embeddings, which have been pre-trained on 2B tweets. To obtain the embeddings for the entire tweet from GloVe, we have taken average of the embeddings of the words present in the tweet. The pre-trained ELMo model available on the Tensorflow hub has also been utilised to obtain tweet embeddings. This model, however, was not trained on a tweet dataset. After obtaining the embeddings, the subsequent model used is the same as that for BERTweet.

4.2 SVM

In this method, we first trained our BERTweet based model (Section 4.1) and stored the output of the last fully connected layer for each dataset

Models	Q1	Q2	Q3	Q4	Q5	Q6	Q7
BERTweet	**0.94**	**0.84**	**0.92**	**0.96**	**0.76**	**0.84**	**0.76**
GloVe	0.83	0.23	0.90	0.53	0.50	0.17	0.15
ELMo	0.76	0.35	0.83	0.49	0.63	0.54	0.52
SVM	0.90	0.78	0.88	0.85	0.71	0.34	0.74
3-BERT Ensemble	0.91	0.84	0.85	0.84	0.73	0.75	0.71
5-BERT Ensemble	0.87	0.84	0.81	0.82	0.72	0.69	0.62

Table 1: Comparison of F1-score of different models

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Overall
0.831	0.925	0.976	0.822	0.854	0.909	0.849	0.881

Table 2: F1-score on official test dataset

(training, validation and test). We used these stored values as input features for training and testing SVM for each label separately.

4.3 Ensemble

Finally, we created two ensemble models with BERTweet. Among the different models we obtained by fine-tuning BERTweet, we chose the best 3 and best 5 models for the ensembles.

5 Performance Evaluation

This section describes the evaluation scheme, followed by the results obtained for the different models.

5.1 Evaluation Scheme

We have used F1-score as the main evaluation scheme. Apart from Q2 to Q5, we have assumed the labels to be independent of each other (because Q2 to Q5 only need to be checked when Q1 is "yes"). Thus, we first train a model for Q1 and obtain the predictions on the dev/test dataset. Then, we pick the tweets for which Q1 is "yes", and assign Q2 to Q5 to be NaN for the rest of the tweets. Subsequently, we have treated all the models for all the questions to be independent of each other. Due to this, it may be possible that while some model performs extremely well on one label, its performance may not be that good for some other label(s). Thus, we can have different models for different labels. So, we calculate label-wise F1-score to compare different models, and choose the best one.

5.2 Evaluation of Different Models

Performance of different systems for the present task are described in the following subsections.

5.2.1 BERTweet

As was expected, in all our experiments, models based on BERTweet outperform all the other models that we described in Section 4. Detailed results (F1-score) for all the labels (along with results for all the different models) are given in Table 1.

5.2.2 GloVe

Although the dataset used in Glove Twitter is bigger than the one over which BERTweet was trained (2B vs 850M tweets), the GloVe vectors are "fixed", and unlike BERTweet, no Transfer Learning was involved for GloVe. As a result, GloVe performed much worse compared to BERTweet for most of the labels. The closest performance obtained is in Q3, when the GloVe based model was simply predicting all 1s (for Q3, the number of 1s is $> 90\%$ in the dataset (excluding NaNs)).

5.2.3 ELMo

Since we did not use an ELMo model pre-trained on the twitter dataset, it did not perform as well as BERTweet. But, as it was possible to use transfer learning here to fine-tune the weights of ELMo, it was mostly performing better than GloVe. On an average, the difference between the F1-scores of BERTweet and GloVe is 0.39, while for ELMo it is 0.28. Further, ELMo performs better than GloVe for four labels, namely, Q2, Q5, Q6 and Q7. Even on the labels when the F1-score of ELMo is lesser than that of GloVe (Q1, Q3 and Q4), the difference between their scores is low (average difference of 0.06), but that is not true for the labels when ELMo beats GloVe (average difference of 0.2475).

5.2.4 SVM

Since this method takes last fully connected layer output of our BERTweet based model as input features, it performes better than Glove and ELMo for almost all questions (except for ELMO for Q6).

5.2.5 Ensemble

For all the labels, the 3-BERTweet Ensemble (3BE) is atleast as good as 5-BERTweet Ensemble (5BE). Further, BERTweet is atleast as good as 3BE. In fact, BERTweet is better than 3BE, which is better than 5BE, for all labels other than Q2. For Q2, all the three models have the same F1-score: 0.84.

5.2.6 Best Model

In view of the results described in Table 1, we decided to use BERTweet for our final predictions. We combined the train + in-sample + dev splits to obtain a dataset with 912 tweets. Early stopping callback has been used with 10% validation split. Testing was done for the best five models we had for each label. We submitted two models (see Table 3). Their average F1-scores over the (new) validation dataset are 0.813 and 0.827, respectively. Even though Model 1 has a lesser F1-score on validation than Model 2, it has the final score of 0.881 (2), beating the latter (0.856).

	Model Specs	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Model 1	BERTweet	Base	Cased	Base	Uncased	Uncased	Cased	Base
	Optimizer	AdaF	Adam	Adam	AdaF	Adam	Adam	AdaF
	Learning Rate	5e-5	1e-5	1e-5	1e-4	2e-5	1e-5	1e-4
	F1-Score	0.83	0.88	0.98	0.86	0.72	0.69	0.73
Model 2	BERTweet	Uncased	Cased	Base	Uncased	Uncased	Uncased	Base
	Optimizer	Adam	Adam	Adam	Adam	Adam	AdaF	AdaF
	Learning Rate	2e-5	1e-5	1e-5	1e-5	2e-5	5e-5	1e-4
	F1-Score	0.86	0.85	0.98	0.86	0.71	0.80	0.73

Table 3: Hyperparameters corresponding to the best models

Labels	Example 1	Example 2
Q1	(452) Instead of prioritizing regular Americans who need tests for #coronavirus, or paid sick leave because they live paycheck to paycheck, @realDonaldTrump wants to bail out oil billionaires. Thank goodness the House of Representatives, and not @POTUS, has the Power of the Purse. URL"	(490) We love this handwashing dance from Vietnamese dancer, Quang ng. Washing your hands with soap and water is one of the first steps to protect yourself from #coronavirus.
Q2	(491) Just like all the other fake stuff they do, the COVID-19 over-hype will backfire on the Democrats. The economy will come roaring backs with China's grip on trade weakened and Trump's high approval on handling the virus will only help.	(498) But, but...Trump didn't prepare for the coronavirus...his admin still doesn't have a clue...we are just not ready to combat a pandemic...Trump ignored the HHS, CDC? #FakeNews WATCH ?? #coronavirus #RepMarkGreen thank you! URL"
Q4	(461) The Italian COVID-19 outbreak has just been explosive... look at the numbers &amp; timeframe. Time is not a luxury we have! Feb 18: 3 cases Feb 21: 20 cases Feb 24: 231 cases Feb 27: 655 cases Mar 1: 1,694 cases Mar 4: 3,089 cases Mar 7: 5,883	(462) A Youtuber who recently made a racist remark regarding BTS by relating them to Corona virus will now be making a video about them where he roasts the band and our fandom I request ARMYs to pls block him and report his channel, Ducky Bhai on YouTube URL

Table 4: Example tweets (from dev data) on which the BERTweet model fails. For each tweet the preceding number in parenthesis denotes the tweet number in the database

6 Error Analysis

For Q1, only three tweets in the dev data (452, 490, 492: all having a verifiable claim) are predicted wrong by our model. Similarly, for Q2, three examples (491, 498, 500), which also have a positive label (denoting that the tweet appears to contain false information), have been predicted wrong while for Q4, four examples (461, 462, 484, 485), all having negative labels (denoting that the claim made in the tweet cannot be harmful to the society), are predicted wrong by our model. Some of these tweets (as described above) can be found in Table 4. Rest of the labels do not have such pattern.

7 Conclusion and Future Work

We implemented five models, described in section 4, and showed that the BERTweet based models outperforms the rest. However, apart from the dependence of Q2 to Q5 on Q1 (refer section 3), we have assumed all questions to be independent. But, by the definitions of questions (section 3), it is evident that Q4 & Q6 and Q5 & Q7 have some dependence on each other. This can be seen in the dataset labels as well, because Q4 & Q6 have the same label for 87.6% of the tweets. Similarly, Q5 and Q7 have the same label 83.3% of the times. Since correlation does not imply causation, this property can be further explored to see if there is some dependence between the labels, which can possibly be incorporated in the model to improve the predictions for Q4 to Q7. Moreover, in this work, we have not experimented with Multi-class classification techniques, which can be further explored for a possible improvement.

Acknowledgments

Raksha Agarwal acknowledges Council of Scientific and Industrial Research (CSIR), India for supporting the research under Grant no: SPM-06/086(0267)/2018-EMR-I.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pablo Gamallo and Marcos Garcia. 2014. Citius: A naivebayes strategy for sentiment analysis on english tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 171–175.

Kris Hartley and Minh Khuong Vu. 2020. Fighting fake news in the covid-19 era: policy insights from an equilibrium model. *Policy Sciences*, 53(4):735–758.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. 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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the NLP4IF-2021 shared task on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF@NAACL' 21, Online. Association for Computational Linguistics.