NutCracker at WNUT-2020 Task 2: Robustly Identifying Informative COVID-19 Tweets using Ensembling and Adversarial Training

Priyanshu Kumar and Aadarsh Singh

Indian Institute of Technology (Indian School of Mines) Dhanbad, India
{kpriyanshu256, aadarshsingh191198 }@gmail.com

Abstract

We experiment with COVID-Twitter-BERT and RoBERTa models to identify informative COVID-19 tweets. We further experiment with adversarial training to make our models robust. The ensemble of COVID-Twitter-BERT and RoBERTa obtains a F1-score of 0.9096 (on the positive class) on the test data of WNUT-2020 Task 2 and ranks 1st on the leaderboard. The ensemble of the models trained using adversarial training also produces similar result.

1 Introduction

Since 2006, Twitter has been a popular social network where people express their thoughts and opinions on various topics. The enforcement of lockdown in various parts of the world, due to COVID-19, led to an increased usage of social media. Thus, the world has witnessed a plethora of tweets since the beginning of COVID-19. People have tweeted about many issues regarding the pandemic, mostly about the increasing infection rate, carelessness and incapability of governance and authorities to handle the increasing rate of cases. In addition, various preventive measures have also been conveyed through tweets.

Although there have been more than 600 million English tweets on COVID-19 (Lamsal, 2020) , only a few of them are informative enough to be used by various monitoring systems to update their databases. Manual identification of these informative tweets can be tedious and erroneous. Hence, there is a dire need to develop systems in the form of machine learning models that can help us in filtering informative tweets.

In this paper, we present our approaches for the shared task "Identification of informative COVID-19 English Tweets" organized under the Workshop on Noisy User-generated Text (W-NUT). Our method makes use of ensembles consisting of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) pretrained on COVID-19 tweets (Müller et al., 2020) and Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019). We also experiment with adversarial training so as to create models that generalise well and are robust.

The rest of the paper is organized as follows: Related work has been discussed in Section 2, followed by a brief description of the data used in Section 3. The proposed methods and experimental settings [1] have been elaborated in Section 4, 5 and 6. Section 7 and 8 contains the results and error analysis respectively. Section 9 concludes the paper and also includes possible future work.

2 Related Work

There has been much research to identify informative tweets during times of emergency and disaster. Neppalli et al. (2018) explored the performance of traditional machine learning algorithms and deep learning approaches for identifying informative tweets during disaster. They created manual features using the content of tweets for machine learning approaches and also tried out features obtained from Convolutional and Recurrent networks for deep learning approaches. A multi-modal approach for classifying informative tweets during disaster was proposed by Madichetty and Sridevi (2020). They combined the features obtained from text by a Convolutional Neural Network and the VGG-16 features obtained from the image accompanying the tweet, using late fusion for better performance than models using only text or only image.

[1] Source code available at https://github.com/kpriyanshu256/WNUT-2020-Task-2

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 404–408
Online, Nov 19, 2020.

Roy et al. (2020) proposed a classification and summarisation approach to identify informative tweets during the Fani cyclone (which occurred in 2019, affecting large parts of South-East Asia). They trained a Support Vector Machine (SVM) on linguistic features and Parts of Speech (POS) tags from tweets to identify informative tweets. To summarise the informative tweets, they experimented with Latent Semantic Analysis (LSA) and Luhn summarisation techniques. Zahera et al. (2019) experimented with the transformer based architecture BERT to classify disaster related tweets into multi-label information types. They preprocessed the tweets of TREC-IS dataset before feeding them into BERT to produce significantly better results than the median score. In addition, they also experimented with Focal loss instead of binary cross-entropy loss.

A new dataset for identifying informative tweet was released by Aggarwal (2019). The results of various machine learning algorithms using GloVe (Pennington et al., 2014) word embeddings, syntactic information in the form of tf-idf vectors and BERT embeddings were also presented in the work.

Due to the abundance of tweets related to COVID-19, many works have been done to analyze the content, intent and effect of such tweets. Singh et al. (2020) presented an analysis of COVID-19 tweets on the grounds of location, content and misinformation spread. A comprehensive study about the content of misinformative COVID-19 tweets and other aspects associated with them have been done by Shahi et al. (2020).

3 Dataset

The dataset (Nguyen et al., 2020) provided to the participants of the shared task contains 10,000 English COVID-19 tweets, out of which 4719 are labeled as INFORMATIVE and 5281 are labeled as UNINFORMATIVE. The tweets were annotated by 3 independent annotators and an inter-annotator agreement score of Fleiss' Kappa at 0.818 was obtained. The dataset contains the tweet ID, the tweet and the corresponding label.

4 Data Preprocessing

Twitter data contains a lot of noise. Therefore, preprocessing on Twitter data will help the pretrained models in better performance. We perform the following data preprocessing steps, most of which have been inspired from Müller et al. (2020) :

1. Unescape HTML tags
2. Remove unnecessary spaces, tabs and newlines
3. Replacing the the mentioned hyperlinks in the tweets (depicted as HTTPURL), with URL. A simple explanation for this could be that "URL" is a more commonly used expression of hyperlinks than HTTPURL.
4. Using the Python emoji [2] library to demojise the emojis i.e. replace them with a short textual description.

The user handles were already replaced by @USER in the tweets, hence no processing was required.

5 Models

We experiment with the following models and techniques:

1. **COVID-Twitter-BERT**: BERT is based on the Transformer architecture (Vaswani et al., 2017). It consists of multi-attention heads which apply a sequence-to-sequence transformation on the input text sequence. For its training, BERT makes use of the following objectives: (a) learn to predict a masked token using the left and right context of the text sequence (Masked Language Model) (b) learn to predict whether two sentences occur in continuation or not (Next Sentence Prediction)

   Müller et al. (2020) pretrain the large version of BERT (BERT-Large) on COVID-19 related tweets posted by users between January 12 and April 16, 2020. This version of BERT has a better understanding of the given data as compared to BERT-large, which is pretrained on texts from Wikipedia. Hence, COVID-Twitter-BERT will be even more beneficial for the task when fine tuned.

2. **RoBERTa Large**: RoBERTa has the same architecture as BERT, but is different from BERT on the grounds of pretraining, which helps in better optimisation and performance. RoBERTa is pretrained on a larger dataset as compared to BERT, uses a larger batch size and replaces the Next Sentence Prediction objective. It also uses dynamic masking pattern

[2] https://pypi.org/project/emoji/

as a better alternative to the static masking pattern used in BERT, i.e. RoBERTa duplicates the data and masks those differently each time, whereas BERT will mask the data only once.

3. **Adversarial Training**: With time, adversarial training is gaining popularity in Natural Language Processing (NLP) as well. In the field of Computer Vision, adversarial training is done by perturbing the input images slightly and minimising the adversarial loss. In NLP, the nature of input being discrete, small perturbations are done on the word embeddings. Adversarial training not only increases the robustness of models but also helps in better generalisation. Both properties are beneficial and desirable for identification of informative tweets.

   Although many approaches for adversarial training in NLP have been developed, we experiment with the approach proposed by Miyato et al. (2016) with a slight modification. In their approach, first the word embeddings are normalized. The gradients are then computed using the data and the required perturbations are created using the obtained gradients.

   Let the sequence of (normalized) word embedding vectors of a text be $\boldsymbol{t}$. The model parameters are represented by $\boldsymbol{\theta}$. The probability of the text belonging to class y is given by $p(y|\boldsymbol{t};\boldsymbol{\theta})$. The adversarial perturbations $\boldsymbol{z}_{adv}$ are computed as follows:

$$\boldsymbol{g} = \nabla_t \, log \, p(y|\boldsymbol{t};\boldsymbol{\theta})$$

$$\boldsymbol{z}_{adv} = -\epsilon \boldsymbol{g} / \parallel \boldsymbol{g} \parallel_2$$

   where ϵ is a hyper-parameter controlling the size of the perturbations. The adversarial loss is defined as :

$$\boldsymbol{L}_{adv}(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^{N} log \, p(y_n|\boldsymbol{t}_n + \boldsymbol{z}_{adv,n};\boldsymbol{\theta})$$

   By using the gradients calculated from the above loss, the weights of the model are updated (the non-perturbed word embeddings of the model are updated). The slight modification in our experiments is that we do not normalize our pretrained word embedding of the model, since it might change the semantic meaning of the pretrained word embeddings. We perform adversarial training on both COVID-Twitter-BERT and RoBERTa Large models using $\epsilon = 1$.

4. **Ensembling**: The ensembling of the predictions for our submissions is done at two levels-

   (a) Fold level i.e. the predictions obtained by the models trained using the different folds (during cross validation) are averaged.

   (b) Model level i.e. the fold level averaged predictions of the two different models are ensembled using averaging.

6 Experimental Settings

We concatenate the training and the validation data and perform a 5-fold stratified cross validation to train our models. Each fold is trained for 5 epochs using early stopping with patience of 3 and tolerance of 1e-3. The models are optimised using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 2e-5 and a batch size of 16. The models have been implemented using Pytorch (Paszke et al., 2019) and Huggingface's Transformers (Wolf et al., 2019) library.

7 Results

We evaluate the performance of the models and their ensemble using cross-validation (CV). We also tabulate the models' performance on the test set as evaluated by the F1 score on the positive class.

The ensembling done at two levels increases the robustness of the results. The out-of-folds predictions were used to find the optimal threshold of the submissions, 0.498 for the ensemble without adversarial training and 0.487 for the ensemble with adversarial training. We also compare our results with the fastText-baseline (Joulin et al., 2016) by the organizers. Table 1 shows the results of our experiments.

The ensemble of COVID-Twitter-BERT and RoBERTa Large performs the best on the test data and also obtains the 1st rank on the leaderboard. Owing to its pretraining, COVID-Twitter-BERT performs better than RoBERTa Large. The adversarial training of the models is also found to boost the scores. The ensemble of adversarial models produces results similar to the ensemble of models trained without adversarial training.

Model	CV	Test
Baseline - fastText	-	0.7503
COVID-Twitter-BERT	0.9622	-
RoBERTa Large	0.9560	-
COVID-Twitter-BERT Adv.	0.9632	-
RoBERTa Large Adv.	0.9578	-
COVID-Twitter-BERT + RoBERTa Large	0.9636	**0.9096**
COVID-Twitter-BERT Adv. + RoBERTa Large Adv.	**0.9655**	0.9082

Table 1: Comparison of results.

8 Analysis

We perform our analysis on the out-of-folds predictions from our two ensembles - without adversarial training and with adversarial training.

We calculate the binary cross entropy loss for all samples and then examine some of the top common mis-classifications by both the ensembles (Table 2).

We observe that there are instances where our models are mistaken. A possible reason might be the variance in gold-labels of samples from annotator-to-annotator. This acts as noise in the labels of the data which is used to train the normal ensemble models. On the other hand, adversarial training adds some noise to the samples. Hence, the models in the adversarial ensemble have been trained on data which has a combination of existing noise and externally added noise. We believe that because of this difference in noise, the models in the two ensembles must be behaving in different ways. To inspect this, we examine the number of samples which have been inferred incorrectly by one ensemble and correctly by the other.

We observe that the normal ensemble misclassified a total of 277 samples whereas, only 260 samples were misclassified by the adversarial ensemble. Moreover, out of the 277 examples that were misclassified by the normal ensemble, 102 were correctly predicted by the adversarial ensemble. On the other hand, only 85 out of the 260 samples misclassified by the adversarial ensemble, were correctly predicted by the normal ensemble. Thus, it is evident that the models in both the ensembles have learnt different patterns from the data.

9 Conclusion

We explored the performance of COVID-Twitter-BERT and RoBERTa-Large at identifying COVID-19 English tweets that are informative in nature. Their ensemble achieves the state-of-the-art performance. Adversarial training is found to improve our model further. For future work, we can pre-train other Transformer-based models on COVID-19 tweets. Data augmentation techniques can help us generate more data for training models. We can also experiment with combining models trained with and without adversarial training.

Tweet	Label
Election Judge Hospitalized After Primary Dies Of Coronavirus #RIP #SemperFi fought for his life while @USER got her hair done #LetThatSinkIn "face of Chicago" thinks she's more important than Chicagoans welfare #LightfootLiedPeopleDied HTTPURL	UNINFORMATIVE
Austin area nursing home residents who test positive for COVID-19 but do not need to be in the hospital will soon be moving to one of two new "isolation facilities," one in Travis County, one in Williamson County: HTTPURL @USER	INFORMATIVE
LOCAL NEWS SHOUTOUT: Have a family member or close friend with a Michigan connection who has died from COVID-19 and would like to share their story with @USER Please contact Georgea Kovanis at gkovanis@USER	INFORMATIVE
BREAKING: Southern AB #coronavirus case rumored to be Steve Busey, the younger brother of movie star @USER According to our sources, Steve works in the oil &amp; gas industry and is a huge @USER fan. #COVID19 😷🦠 HTTPURL	UNINFORMATIVE

Table 2: Some common highly misclassified samples.

Acknowledgments

We thank Google Colab for providing free GPU services for experimentation purposes.

References

Piush Aggarwal. 2019. Classification approaches to identify informative tweets. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 7–15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

R Lamsal. 2020. Corona virus (covid-19) tweets dataset. *IEEE Dataport*.

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

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Sreenivasulu Madichetty and M Sridevi. 2020. Classifying informative and non-informative tweets from the twitter by adapting image features during disaster. *Multimedia Tools and Applications*, pages 1–23.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. 2018. Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *ISCRAM*.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Sujoy Roy, Sumit Mishra, and Rakesh Matam. 2020. Classification and summarization for informative tweets. In *2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–4. IEEE.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2020. An exploratory study of covid-19 misinformation on twitter. *arXiv preprint arXiv:2005.05710*.

Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hamada M Zahera, Ibrahim A Elgendy, Rricha Jalota, and Mohamed Ahmed Sherif. 2019. Fine-tuned bert model for multi-label tweets classification. In *TREC*.