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

Priyanshu Kumar(kpriyanshu256) and Aadarsh Singh(aadarshsingh191198) @gmail.com

Indian Institute of Technology (Indian School of Mines) Dhanbad, India

Introduction

- The world has witnessed a plethora of tweets since the beginning of COVID-19.
- However, only few of them are informative enough to be used by various monitoring systems to update their databases.
- Hence, there is a dire need to develop systems in the form of machine learning models that can help us in filtering informative tweets.
- Our method makes use of **ensembles** consisting of **Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) pretrained on COVID-19 tweets** [2] 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.

Data

- The dataset [3] 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 dataset contains the tweet ID, the tweet and the corresponding label.

Data Preprocessing

- We remove unnecessary spaces, tabs, newlines; unescape HTML tags and demojise emojis.
- User handles had already been replaced with @USER in the dataset.

Models

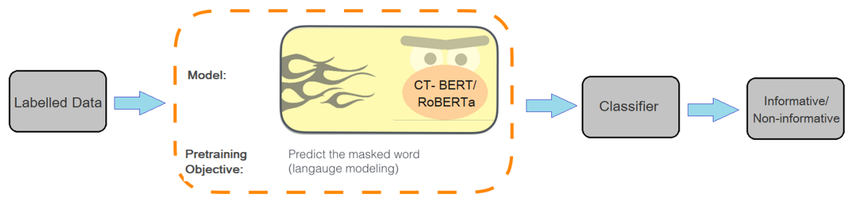

Figure 1:Model Architecture

Adversarial Training

We experiment with the technique as proposed in [1] with a modification that we do not normalise word embeddings. Let the sequence of word embedding vectors of a text be t. The model parameters are represented by θ. The probability of the text belonging to class y is given by $p(y|t;\theta)$. The adversarial perturbations z_{adv} are computed as follows:

$$g = \nabla_t \, log \, p(y|t;\theta) \qquad z_{adv} = -\epsilon g / \parallel g \parallel_2$$

$$L_{adv}(\theta) = -\frac{1}{N} \sum_{n=1}^{N} log \, p(y_n|t_n + z_{adv,n};\theta)$$

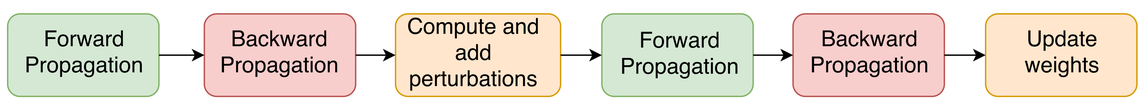

Figure 2:Adversarial Training Pipeline

Ensembling

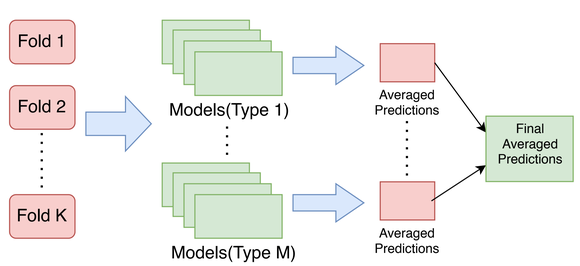

Figure 3:Ensembling Pipeline

Results

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.

Discussion

- The ensemble of COVID-Twitter-BERT and RoBERTa-Large achieves the state-of-the-art performance.
- Adversarial training is found to improve our model further.

References

[1] T. Miyato, A. M. Dai, and I. Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

[2] M. Müller, M. Salathé, and P. E. Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*, 2020.

[3] D. Q. Nguyen, T. Vu, A. Rahimi, M. H. Dao, L. T. Nguyen, and L. Doan. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*, 2020.