Linguist Geeks on WNUT-2020 Task 2: COVID-19 Informative Tweet Identification using Progressive Trained Language Models and Data Augmentation

Vasudev Awatramani[1] **Anupam Kumar**[1]
[1]Maharaja Agrasen Institute of Technology
vasudev.w13@gmail.com, anupamkumar@mait.ac.in

Abstract

Since the outbreak of COVID-19, there has been a surge of digital content on social media. The content ranges from news articles, academic reports, tweets, videos, and even memes. Among such an overabundance of data, it is crucial to distinguish which information is actually informative or merely sensational, redundant or false. This work focuses on developing such a language system that can differentiate between Informative or Uninformative tweets associated with COVID-19 for WNUT-2020 Shared Task 2. For this purpose, we employ deep transfer learning models such as BERT along other techniques such as Noisy Data Augmentation and Progress Training. The approach achieves a competitive F1-score of 0.8715 on the final testing dataset.

1 Introduction

The aim of WNUT 2020 Task 2 (Nguyen et. al., 2020), is to produce methods that automatically classify whether an English Tweet associated with the novel coronavirus or COVID-19 is informative or not. An informative tweet may report information regarding recovered, suspected, confirmed and death cases or may include the knowledge of location or travel history of such occurrences. To accomplish such a system, we are provided with a dataset of 10,000 tweets, consisting of 7000 tweets for training, 1000 tweets for a validation set and 2000 tweets for the evaluation phase.

Our solution employed an ensemble of pre-trained models such as BERT (Devlin et. al., 2019), fine-tuned earlier on the task dataset, see Section 2. We also investigated text augmentation techniques such as replacing tokens with synonyms, random removal and swapping of tokens, see Section 3. The work also explored Progressive Training of a given model see Section 4. Certain aspects of these methodologies seemed to perform well and are discussed in detail.

2 Related Work

Sequence Labelling or Text Classification is one of the primary tasks in Computational Linguists. In general, the task contains different levels of scope such as Document Level, Paragraph Level, Sentence Level and Sub-sentence Level (words or groups of words). Typical pipeline for Sequence Labelling consists of feature extraction, followed by dimensionality reduction and a classification technique. Initial approaches for feature extraction involved techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) (Salton et. al., 1998), Word2Vec (Goldberg et. al., 2014) or Global Vectors for Word Representation (GloVe) (Pennington et. al., 2014). Dimensionality reduction can help in reducing time and memory complexity if datasets contain a large vocabulary of unique words. Therefore, this step is sometimes left out nonetheless, prevalent methods for feature extraction include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) or t-distributed stochastic neighbor embeddings (t-SNE). Early models for classification such as Naive Bayes, and Support Vector Machines have now been superseded by deep learning models. Such models incorporate sequential processing ability through architectures such as Recurrent Neural Networks which can be worked with a varied length of text sequences as well.

Lately, transfer learning has taken over these pipelines, producing high-quality textual representation by employing enormous corpora used in pre-training. Deep Learning methods have produced dominant performances in many

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 414–418
Online, Nov 19, 2020.

domains delivering state-of-the-art results. Transfer Learning is one such technique that has dominated this trend with models such as BERT (Devlin et. al., 2019) and RoBERTa (Liu et al., 2019) producing high performance on NLP benchmarks such as GLUE ().

3 Transfer Learning Ensemble

In our approach, we apply models like BERT that have already been pre-trained on data related to COVID-19. For this purpose, we used huggingface's transformers package (Wolf et. al.,2019).

Model	Acc.	F1-Score
CT-BERT	90.2	0.893
mrm8488/bioclinicalBERT-finetuned-covid-papers[1]	88.4	0.872
deepset/covid_bert_base[2]	87.0	0.860

Table 1. Various Models and their performance

Majority of the models that were tested are some variations of BERT. They vary in terms of data that have been fine-tuned on or number of parameters such as BERT-base or BERT-large (model architecture variants for BERT). For instance, deepset/covid_bert_base[2] is fine-tuned on CORD-19 dataset whereas CT-BERT (Müller et al.,2020) is trained on Crowbreaks Dataset(Müller et al.,2019).

In our system, we fine-tuned these models to the data in Shared Task 2 of WNUT 2020. Table 1 describes their performance in terms of accuracy and f1-score. There are slight variations in the training of these networks such as the optimiser used or number of epochs, however for the major part of the system we employed the following:

1. Weighted Adam or AdamW (Loshchilov et al.,2017) as the optimizer as opposed to Adam. The models seemed to converge with the learning rate of 3×10^{-5} and 1×10^{-8} as the epsilon value. We tried LAMB (You et. al.,2019) as well. Both of these optimisers, had similar effect on performance.
2. Dropout Layers with 0.5 dropout probability to counter overfitting.
3. The epochs varied between 3 to 5, corresponding to the model.
4. Some of the models were trained on TPU and rest of GPUs due to computational constraints. Batch size of 128 on TPU proved more benefitting in terms of performance over smaller batch-sizes for the same model.

Ensemble inference is a very popular trick in machine learning competitions. We employed an averaging ensemble of the models, to get an improved f1-score and accuracy on the validation set of 0.897 and 90.6 respectively. Though marginal, we employed the ensemble strategy for our final submission as well.

4 Data Augmentation

Another typical trick to enhance the performance of neural networks is to employ more training data. To this purpose, we applied Data Augmentation techniques that are loosely inspired by those used in computer vision. Moreover, we wanted to add some noise to the input text, in order to produce more robust models, particular because the tweets are ordinarily noisy due to use of emoticons, social media lingos, short forms and symbols such as #, @ or URLs in them. For a given sentence in the training set, performed one of the following procedures with respect to a random probability:

1. **Replacing tokens with Synonym:** Randomly choose n words from the sentence and replace each of these words with one of its WordNet synonyms.
   Original Tweet:
   This week in podcast heaven: Roman Mars singing that one obscure song Reply All "covered" in a recent Super Tech Support ***segment*** *to time his hand washing This* ***week*** *in podcast* ***hell****: Ira Glass* ***self*** *quarantining because he shook hands with someone who* ***tested*** *positive for COVID-****19***
   Augmented Tweet:
   This week in podcast heaven: Roman Mars singing that one obscure song Reply All " covered " in a recent Super Tech Support ***section*** *to time his hand washing*

[1] https://huggingface.co/mrm8488/bioclinicalBERT-finetuned-covid-papers.
[2] https://huggingface.co/deepset/covid_bert_base
[3] https://github.com/makcedward/nlpaug.

*This **calendar week** in podcast **netherworld**: Ira Glass **ego** quarantining because he shook hands with someone who **essay** positive for COVID – **xix***

2. **Random Swapping:**
   Find a pair of random words in the sentence and rearrange their positions in the sentence.
   Original Tweet:
   *President Trump revealed a grim projection in the **coronavirus pandemic** on Tuesday: Even with the **social distancing** the US is doing now, 100k to 200,000 Americans will likely die as a result of **the ongoing** outbreak. "When you see 100**,000 people**, that's a minimum number" Trump said*
   Augmented Text:
   *President Trump revealed a grim projection in the **pandemic coronavirus** on Tuesday Even: with the **distancing social** the US is doing now, 100k to 200, 000 will Americans likely die as a result of **ongoing the** outbreak. you "When see 100, **people 000**, that' s a minimum number" Trump said*

3. **Removing Tokens Randomly:**
   Randomly removing words in the sentence.
   Original Tweet:
   *Hawaii **has** its first case of #Coronavirus. @USER has been devoting more than an hour. Hawaii currently has capacity to only test 1600. **Officials** trying to retrace steps of 2 **tourists** from cruise ship who disembarked in Hilo &amp; tested positive for virus. We are now affected too.*
   Augmented Text:
   Hawaii its first case. @USER has been devoting more than an hour. Hawaii currently has capacity to test 1600. trying retrace steps of 2 from cruise ship who disembarked in Hilo &amp; tested positive for virus. now. We are now affected too.

For implementing such augmentations, we used Edward Ma's nlpaug package[3], however, we note that such techniques have been studied most recently in the paper EDA (Jason et. al.,2019).

Model	Acc.	F1-Score
Synonym Insertion	90.4	0.901
Random Removal	89.1	0.887
Random Swapping	90.1	0.893
Combination of all 3 Augmentations	90.7	0.902

Table 2. Effect of Augmentations

Table 2. compares the performance of CT-BERT (Müller et al.,2020) trained using augmentation training data, over original validation set.

Other augmentation techniques that we tried but did not use were Antonym Replacement and Sentence Augmentation with models such as GPT-2 (Radford et al.,2019). For instance, replacing with antonyms seemed to distort the meaning of the text unless, the random selection of words somehow resulted in some form of double negation.

Original Tweet:

*Frey says Minneapolis has 131 **known** positive cases of Covid19. So far the city has declined to give **regular** updates on this beyond Friday meetings. Would be nice to get some ward/neighborhood breakdowns, even if known positives don't paint the whole picture.*

Augmented Text with Antonym Replacement:

*Frey says Minneapolis has 131 **ignored** positive cases of Covid19. So far the city has accepted to give **irregular** updates on this beyond Friday meetings. Would be nice to get some ward/ neighborhood breakdowns, even if known positives don' t paint the whole picture.*

Similarly, with GPT-2 augmentations produced were distorting in cases we observed, such as unwarranted repetition of input sentence in the augmented text.

5 Progressive Training of Model

Progressive Training is one of the highly recommended techniques among deep learning practitioners, especially for its practical utility. This may involve training the model first on a smaller segment of the problem such that it training on a smaller sample of data, or priorly training on a smaller number of classes as opposed to all classes. A popular variation in Image Classification has been popularised by Jeremy Howard's FastAi lectures and library (Howard et al.,2020), that involves first training an image classification network on smaller sized images and then gradually increasing the dimensions. Similar intuition was extended for training of Generative Adversarial Networks by the ProGANS (Karras et al.,2017) study.

Inspired by the concept, we developed a similar method by performing the following steps:

1. Assume, the model consisting of two parts:
    a. Transformer Head: Consisting of a BERT architecture
    b. Classifier Network: Usually consists of Dense or Linear Layers.
2. Initially, the entire model is trained on smaller sized encoded vectors of the tweets, having length 128.
3. This is followed by fitting the same transformer head over encoded vectors having length 192, with an uninitialized classifier network
4. The above steps can be iterated up till the encode input length is 256.

Tokenized Vector Length	Acc.	F1-Score
128	90.2	0.893
192	91.13	0.905
256	91.63	0.913

Table 3. Comparison of Progressive Training

In our work, we followed the above described method gaining a marginal improvement as shown in Table 3.

6 Results

We made 2 submissions on the test set involving the following:

a. Ensemble of various BERT models trained over original inputs.
b. Progressively trained Digital Epidemiology Lab’s CT-BERT model over augmented inputs.

From the results on the leaderboard, submission b. scored better with an f1-score of 0.8715.

7 Conclusion

We have outlined the motivation, design, and results of the WNUT Shared task 2 on detection of informative tweets pertaining to COVID-19 pandemic. We employed techniques such as Transfer Learning, Ensemble Learning, Data Augmentation and Progressive Training. However, we hope to study the task and come up with new findings in the future. One of the major aspects, we would like to analyze is the timestamp of the tweets. COVID-19 has had a very dynamic impact on social media, with varying trends concerning the number of cases, information regarding vaccine trials or new developments in social distancing guidelines. Therefore, in further work, we would like to include a temporal factor associated with these tweets for more reliable prediction on their informativeness. Furthermore, cross-lingual research and applications such as XLM-Roberta (Conneau et. al, 2020) have been rising in NLP studies. Therefore, data augmentations from multi-lingual or code-mixed tweets and microblogs sources about COVID-19 can comprehensively contribute to robust of the system such as ours. Lastly, to further increase the robustness of the model, we would like to explore techniques such as Knowledge Distillation.

References

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for language understanding. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 4171–4186.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019c). Roberta: A robustly optimized bert pretraining approach.

Gerard Salton and Chris Buckley. (1998). Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 1988, 24, 513–523.

Yoav Goldberg. and Omer Levy. (2014). “word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method.”

Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Volume 14, pp. 1532–1543

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface’s transformers: State-of-the-art natural language processing.

Ilya Loshchilov and Frank Hutter (2017), Decoupled Weight Decay Regularization.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer and Cho-Jui Hsieh (2019), Large Batch Optimization for Deep Learning: Training BERT in 76 minutes

Jason Wei and Kai Zou (2019), EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)

Martin Müller, Marcel Salathé and Per E Kummervold (2020), COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter

Radford, Alec and Wu, Jeff and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya (2019). Language Models are Unsupervised Multitask Learners

Martin M Müller and Marcel Salathé (2019). Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. Frontiers in public health, 7, 2019.

Jeremy Howard and Sylvain Gugger (2020). fastai: A Layered API for Deep Learning

Tero Karras, Timo Aila, Samuli Laine and Jaakko Lehtinen (2017) Progressive Growing of GANs for Improved Quality, Stability, and Variation

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov (2020). Unsupervised Cross-lingual Representation Learning at Scale