TATL at WNUT-2020 Task 2: A Transformer-based Baseline System for Identification of Informative COVID-19 English Tweets

Anh Tuan Nguyen
NVIDIA, Santa Clara, USA
tuananhn@nvidia.com

Abstract

As the COVID-19 outbreak continues to spread throughout the world, more and more information about the pandemic has been shared publicly on social media. For example, there are a huge number of COVID-19 English Tweets daily on Twitter. However, the majority of those Tweets are uninformative, and hence it is important to be able to automatically select only the informative ones for downstream applications. In this short paper, we present our participation in the W-NUT 2020 Shared Task 2: Identification of Informative COVID-19 English Tweets. Inspired by the recent advances in pretrained Transformer language models, we propose a simple yet effective baseline for the task. Despite its simplicity, our proposed approach shows very competitive results in the leaderboard as we ranked 8 over 55 teams participated in total.

1 Introduction

The COVID-19 pandemic has been spreading rapidly across the globe and has infected more than 20 millions men and women. As a result, more and more people have been sharing a wide variety of information related to COVID-19 publicly on social media. For example, there are a huge number of COVID-19 English Tweets daily on Twitter. However, the majority of those Tweets are uninformative and do not contain useful information, therefore, systems which can automatically filter out uninformative tweets are needed by the community. Tweets are generally different from traditional written-text such as Wikipedia or news articles due to its short length and informal use of words and grammars (e.g abbreviations, hashtags, marker). These special characteristics of Tweets may pose a challenge for many NLP techniques that focus solely on formally written texts.

In this paper, we present our participation in the W-NUT 2020 Shared Task 2: Identification of

category	#training	#valid	#test
informative	3303	472	944
uninformative	3697	528	1056

Table 1: Statistics of Shared task 2 dataset. "#training", "#valid" and "#test" denote the size of the training, validation and test sets, listed by categories, respectively.

Informative COVID-19 English Tweets (Nguyen et al., 2020b). Inspired by the recent success of Transformer-based pre-trained language models in many NLP tasks (Devlin et al., 2019; Lai et al., 2019; Chen et al., 2019; Nguyen and Nguyen, 2020; Lai et al., 2020), we propose a simple yet effective baseline for the task. Despite its simplicity, our proposed approach shows very competitive results.

In the following sections, we first describe the task definitions in Section 2 and proposed methods in Section 3. We then describe the experiments and their results in Section 4. Finally, in Section 5, we conclude this work and discuss potential future research directions.

2 Task Definitions

The goal of Shared task 2 is to identify whether a COVID 19 English Tweet is informative or not. Such informative Tweet provides information about recovered, suspected, confirmed and death cases as well as location and history of each case. The dataset introduced in this Shared task consists of 10K COVID 19 English Tweets. Dataset statistics can be found in Table 1

3 Method

3.1 Baseline Model

The task is formulated as a binary classification of Tweets into informative or uninformative classes. Figure 1 gives a high-level overview of our proposed approach. Given a Tweet consisting of n
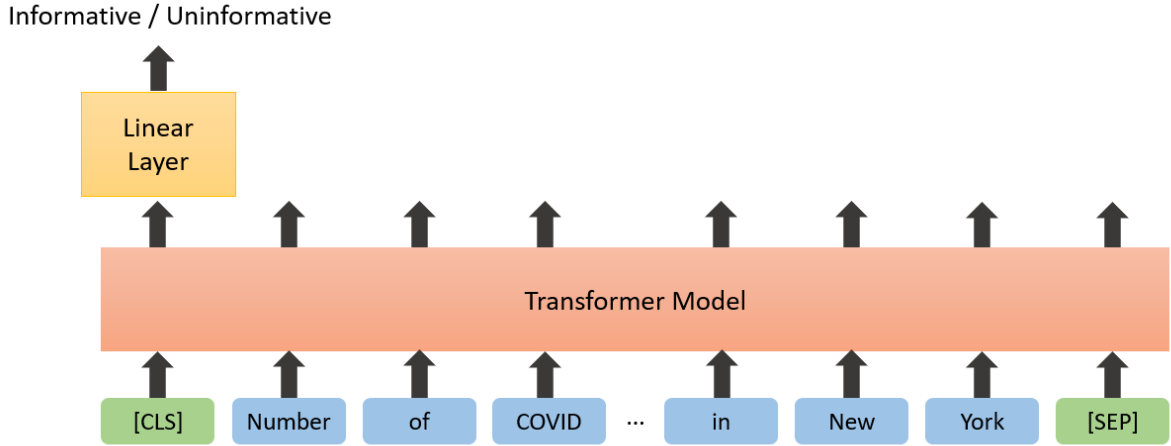
Figure 1: A high level overview of our proposed model for the task.

tokens $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, we first form a contextualized representation for each token using a Transformer-based encoder such as BERT (Devlin et al., 2019). Following common conventions, we append special tokens to the beginning and end of the input Tweet before feeding it to the Transformer model. For example, if we use BERT, x_1 will be the special [CLS] token and x_n will be the special [SEP] token. Let $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n\}$ denote the contextualized representations produced by the Transformer model. We then use $\mathbf{h}_1$ as an aggregate representation of the original input and feed it to a linear layer to calculate the final output:

$$y = \sigma(\mathbf{W}\mathbf{h}_1 + \mathbf{b}) \in \mathbb{R} \qquad (1)$$

where the transformation matrix $\mathbf{W}$ and the bias term $\mathbf{b}$ are model parameters. σ denotes the sigmoid function. It squashes the score to a probability between 0 and 1. y is the predicted probability of the input Tweet being informative.

In this work, we experiment with various state-of-the-art Transformer models including BERTweet (Nguyen et al., 2020a), XLM-RoBERTa (Conneau et al., 2020), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020). In the following subsections, we will briefly describe these Transformer models.

3.1.1 RoBERTa

RoBERTa (Liu et al., 2019) improved over BERT (Devlin et al., 2019) by leveraging different training objectives which leads to more robust optimization i.e removing next sentence prediction and using dynamic masking for masked language modelling.

Liu et al. (2019) also shows that training the language model longer and with more data hugely benefits the performance on downstream tasks.

3.1.2 XLM-RoBERTa

Inspired by the success of multilingual language model (Devlin et al., 2019; Lample and Conneau, 2019), XLM-RoBERTa (Conneau et al., 2020) significantly scaled up the amount of multilingual training data used in unsupervised MLM pretraining compares to previous work (Lample and Conneau, 2019) and achieved state-of-the-art performance in both monolingual and cross-lingual benchmarks.

3.1.3 BERTweet

BERTweet (Nguyen et al., 2020a) is a domain-specific language model pre-trained on a large corpus of English Tweets. Similar to the success of BioBERT (Lee et al., 2019) in BioNLP domain and the success of SciBERT (Beltagy et al., 2019) in ScientificNLP domain, BERTweet achieved state-of-the-art performance across many TweetNLP tasks, outperformed its counterparts RoBERTa (Liu et al., 2019) and XLM-RoBERTa (Conneau et al., 2020).

3.1.4 ELECTRA

ELECTRA (Clark et al., 2020) proposed a new pre-training objective which is different from Masked Language Modelling (Devlin et al., 2019; Liu et al., 2019). Instead of masking input tokens, ELECTRA corrupts the tokens using a small generator network to produces distribution over tokens, while the discriminator tries to guess which tokens are

actually corrupted by the generator. ELECTRA achieved state-of-the-art results across many tasks in the GLUE benchmark (Wang et al., 2019) while using much less compute resources compared to other pre-training methods (Devlin et al., 2019; Liu et al., 2019).

3.2 Ensemble Learning

To further boost the performance of our baseline models, we leverage ensemble learning technique. We performed ensemble learning over all of the Transformer models mentioned in the previous section and employed two different ensemble schemes, namely Unweighted Averaging and Majority Voting.

3.2.1 Unweighted Averaging

In this approach, the final prediction is estimated from the unweighted average of the posterior probability from all of our models. Thus, the final prediction is given by:

$$p = \arg \max_c \frac{1}{M} \sum_{n=1}^{M} p_i, \quad p_i \in \mathbb{R}^{\mathbb{C}} \quad (2)$$

where C is the number of classed, M is the number of models, and p_i is the probability vector computed using the softmax function of model i.

3.2.2 Majority Voting

Majority Voting counts the votes of all the models and select the class with most votes as prediction. Formally, the final prediction is given by:

$$v_c = \sum_{n=1}^{M} F_i(c), \quad p = \arg \max_c v_c \quad (3)$$

where v_c denotes the votes of class c from all different models, F_i is the binary decision of model i, which is either 0 or 1.

4 Experiments

4.1 Finetuning

To fine-tune our baseline models, we employ `transformers` library (Wolf et al., 2019). We use AdamW optimizer (Loshchilov and Hutter, 2019) with a fixed batch size of 32 and learning rates in the set $\{1e-5, 2e-5, 5e-5\}$. We fine-tune the models for 30 epochs and select the best checkpoint based on performance of the model on the validation set.

Model	Dev F1
XLM-RoBERTa (base)	0.905
XLM-RoBERTa (large)	0.906
RoBERTa (base)	0.911
RoBERTa (large)	**0.918**
BERTweet	0.909
ELECTRA (base)	0.907
ELECTRA (large)	0.914
Ensemble (averaging)	**0.927**
Ensemble (voting)	0.922

Table 2: Performance of individual models as well as ensemble models on the validation set.

Model	Test F1
Ensemble (averaging)	0.8988
Ensemble (voting)	**0.9008**

Table 3: Performance of our system on the test set.

4.2 Performance of our baselines

Table 2 shows the overall results on the validation set. The large version of RoBERTa achieves the highest F1 score on the validation set (compared to other individual models). To our surprise, we find that BERTweet does not outperform the base version of RoBERTa on the validation set, even though BERTweet was trained on English Tweets using the same training procedure of RoBERTa. Finally, XLM-RoBERTa achieves lower F1 score than both RoBERTa and ELECTRA, suggesting that using a multilingual pretrained language models may not improve the performance since the shared task is mainly about English Tweets. We also evaluate the performance of our ensemble models. The results show that ensemble learning improves the F1 score compare to each individual model and Unweighted Averaging perform better than Majority Voting on the validation set. We also submitted the predictions of both ensemble scheme to the competition and final results on the leaderboard are shown in table 3. We notice that Majority Voting slightly performs better than Unweighted Averaging on the hidden test set.

5 Conclusion

In this paper, we introduce a simple but effective approach for identifying informative COVID-19 English Tweets. Despite the simplicity of our approach, it achieves very competitive results in the leaderboard as we ranked 8 over 56 teams partici-

pated in total. In future work, we will conduct thorough error analysis and apply visualization techniques to gain more understandings of our models (Murugesan et al., 2019). Furthermore, we will also extend our approach to other languages. Finally, we will investigate the use of advanced techniques such as transfer learning, few-shot learning, and self-training to improve the performance of our system further (Pan et al., 2017; Huang et al., 2018; Lai et al., 2018; Yoon et al., 2019; Xie et al., 2020).

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Qian Chen, Zhu Zhuo, and W. Wang. 2019. Bert for joint intent classification and slot filling. *ArXiv*, abs/1902.10909.

K. Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia. Association for Computational Linguistics.

Tuan Lai, Trung Bui, Nedim Lipka, and Sheng Li. 2018. Supervised transfer learning for product information question answering. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1109–1114. IEEE.

Tuan Lai, Quan Hung Tran, Trung Bui, and Daisuke Kihara. 2019. A gated self-attention memory network for answer selection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5953–5959, Hong Kong, China. Association for Computational Linguistics.

Tuan Manh Lai, Quan Hung Tran, Trung Bui, and Daisuke Kihara. 2020. A simple but effective bert model for dialog state tracking on resource-limited systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8034–8038. IEEE.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

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. 2019. Decoupled Weight Decay Regularization. In *Proceedings of ICLR*.

Sugeerth Murugesan, Sana Malik, Fan Du, Eunyee Koh, and Tuan Manh Lai. 2019. Deepcompare: Visual and interactive comparison of deep learning model performance. *IEEE computer graphics and applications*, 39(5):47–59.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. *Findings of EMNLP*.

Dat Quoc Nguyen, Thanh Vu, and A. Nguyen. 2020a. Bertweet: A pre-trained language model for english tweets. *EMNLP 2020*.

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

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint*, arXiv:1910.03771.

Qizhe Xie, E. Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2019. A compare-aggregate model with latent clustering for answer selection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2093–2096.