NLP@NISER: Classification of COVID19 tweets containing symptoms

Deepak Kumar[1,*], **Nalin Kumar**[2,*], **Subhankar Mishra**[3]
School of Computer Sciences, NISER, Bhubaneswar- 752050
Homi Bhabha National Institute, Training School Complex, Anushakti Nagar, Mumbai-400094, India
`{deepak.kumar1, nalin.kumar2, smishra3}@niser.ac.in`

Abstract

In this paper, we describe our approaches for task six of Social Media Mining for Health Applications (SMM4H) shared task in 2021. The task is to classify twitter tweets containing COVID-19 symptoms in three classes (self-reports, non-personal reports & literature/news mentions). We implemented BERT and XLNet for this text classification task. Best result was achieved by XLNet approach, which is F1 score 0.94, precision 0.9448 and recall 0.94448. This is slightly better than the average score, i.e. F1 score 0.93, precision 0.93235 and recall 0.93235.

1 Introduction

In the beginning of the COVID-19 pandemic and even now, with variety of strains, caused great deal of information deficiency about symptoms as reported by people affected by it. As this disease is highly contagious and rapidly changing, one of the best sources for the live information is on the social media. There can be multiple sources of symptoms information on social media such as news/scientific articles (facts), other people's account (second or third person statements) and self report (first person statements). In this paper we will discuss our approach as a team participating in SMM4H (Magge et al., 2021) shared task 6 related to the classification of such information from social media platform like twitter. We will be looking at using pre-trained NLU models like BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019) for this task.

2 Task and Data Description

2.1 Task

The task 6 of SMM4H is to classify twitter tweet dataset containing COVID-19 symptoms into three classes:

- self-reports: mentioning ones own experience of COVID-19
- non-personal reports: mentioning other peoples account of COVID-19 experience
- literature/news mentions: mentioning scientific or news articles telling about COVID-19 symptoms

2.2 Data Description

The training dataset contained $9,000$ labeled tweets, validation dataset contained $5,76$ labeled tweets and test dataset contained $6,500$ unlabeled tweets.

3 Methodology

3.1 Pre-processing

- **Data Cleaning** : For cleaning we removed, part of sentences starting from "@" to first white space, links, numbers, "#" and extra white space.
- Further in pre-processing, we convert data into machine readable form. We change the labels into three discrete integers and tweet text text into tokens using tokenizer as mentioned in Model section.
- Then we truncate the tweet text to reduce the amount of padding. For BERT we truncate before applying tokenizer and making sentence length 65, while for XLNet we truncate after applying tokenizer to make sequence length 150.

3.2 Model

We explore both autoencoding as well as autoregessive models for the classification task from which BERT and XLNet are picked respectively. For our experiments, we use the pre-trained models provided by Huggingface (Wolf et al., 2020). For both of the models we are using cased versions which differentiates between upper and lower case.

[*] Equal contribution. Deepak implemented BERT while Nalin performed experiments on XLNet

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 102–104
June 10, 2021. ©2021 Association for Computational Linguistics

This is needed as upper case letters are important for identifying nouns and pronouns which in turn are important to identify first, second and third person in a sentence.

3.2.1 BERT

The first system we use for the task is BERT (Devlin et al., 2018). BERT, which stands for Bidirectional Encoder Representations from Transformers, learns by predicting the randomly masked token during pre-training using both left and right context of the masked word token. It also has the second objective of predicting if the given two sentences are consecutive. We use both base and large cased versions of this model for our experiments. The base version of BERT has 12 encoder layers, 768 hidden layer dimension and 12 attention heads with 109M parameters, while the large one has 24 encoder layers, 1024 hidden layer dimension and 16 attention heads with 335M parameters. We use their respective tokenizers.

3.2.2 XLNet

We use XLNet as our second system. XLNet (Yang et al., 2019) is an auto-regressive model, which, unlike BERT, uses autoregressive formulation to learn the bidirectional contexts. The word token output is calculated by taking into account the permutation of all word tokens in the sentence, in contrast to the traditional approaches, which used just left or right of the target token. We experiment with both base and large versions of this model. The base version has 12 layers, 768 hidden layer dimension and 12 attention heads with number of model parameters to be $110M$, whereas the large one has 24 layers, 1024 hidden layer dimension and 16 attention heads having $340M$ parameters. We use their respective tokenizers.

4 Experiments

We perform several experiments on cleaned and uncleaned data. We explore both base and large versions of BERT and XLNet along with different training methods, fine-tuning and retraining the whole model, scores of which are mentioned in Table 1. For both the NLU models, we use appropriate classification layer. For all these experiments, loss function is cross entropy loss and optimizer is adamW with learning rate $2e - 5$. We find BERT large version retrained on uncleaned data perform the best. For the XLNet version, we find that the large version of the XLNet tokenizer with base version of the model works the best among all. We get the best results on retraining the model over the uncleaned data. The best systems' scores on the test data for the shared task are given in Table 2. Our code is shared on Github [*].

Approach	**BERT**		**XLNet**	
	Base	Large	Base	Large
C_Retrain	0.976	0.978	0.974	0.968
C_Finetune	0.666	0.729	0.844	0.784
U_Retrain	0.98	0.998	0.984	0.984
U_Finetune	0.76	0.734	0.924	0.878

Table 1: All results are F1 scores of models trained over training set and calculated over validation set. All models are run for 6 epochs with batch size 16.

Model	**Precision**	**Recall**	**F1 Score**
XLNet	0.9448	0.94448	0.94
BERT	0.926	0.926	0.93
Median	0.93235	0.93235	0.93

Table 2: Performance of the best version of each model on the test set. Models are run for 6 epochs with batch size 16.

5 Results and Conclusion

We can have the following observations from Table 1 and 2:

- In comparison to training on cleaned data, the uncleaned versions show better results. We suspect that the information removed while data cleaning (such as "@", links, etc) are significant for predictions.
- We also observe that the retraining method performs significantly better than the fine-tuned one.
- BERT, the large version performs better than the base one, whereas in XLNet, the base version has better scores than the large one.
- Finally, the BERT performs better on validation set while the XLNet has better performance on the test set.

Between the given two systems, the XLNet performs the best with results shown in Table 2. The system performs slightly better than the median of all submissions made for the task. In the future work, one can look for new approach towards

[*] https://github.com/smlab-niser/2021smm4h

cleaning of tweets, as traditional way of cleaning tweets tend to decrease the F1-score (as shown by our results).

References

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*.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.