# Fine-tuning BERT to Classify COVID19 Tweets Containing Symptoms

**Rajarshi Roychoudhury**
Dept of Computer Science and Engg.
Jadavpur University, India
rroychoudhury2@gmail.com

**Sudip Kumar Naskar**
Dept of Computer Science and Engg.
Jadavpur University, India
sudip.naskar@gmail.com

## Abstract

Twitter provides a source of patient-generated data that has been used in various population health studies. The first step in many of these studies is to identify and capture Twitter messages (tweets) containing medication mentions. Identifying personal mentions of COVID19 symptoms requires distinguishing personal mentions from other mentions such as symptoms reported by others and references to news articles or other sources. In this article, we describe our submission to Task 6 of the Social Media Mining for Health Applications (SMM4H) Shared Task 2021. This task challenged participants to classify tweets where the target classes are - (1) self-reports, (2) non-personal reports, and (3) literature/news mentions. Our system uses a handcrafted preprocessing and word embeddings from BERT encoder model. We achieve. F1 score of 93%.

## 1 Introduction

The classification of medical symptoms from COVID-19 Twitter posts presents two key issues. Firstly, there is plenty of discourse around news and scientific articles that describe medical symptoms. While this discourse is not related to any user in particular, it enhances the difficulty of identifying valuable user-reported information. Secondly, many users describe symptoms that other people experience, instead of their own, as they are usually caregivers or relatives of people presenting the symptoms. This makes the task of separating what the user is self-reporting particularly tricky, as the discourse is not only around personal experiences. Moreover, detecting tweets containing health-related words such as diseases, treatments and medications is a fundamental yet difficult step. These difficulties are exacerbated by the short length and informal nature of tweets, which often contain non-standard grammar, frequent misspellings, many contractions, extensive slang, and combined symbols (emojis/emoticons) to express emotion (Dang et al., 2020).

From the types of class-labels that are to be predicted, it is clear that contextual representations play an important role beside semantics. Recurrent models are typically used for this task which computes along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states $h_t$, as a function of the previous hidden state $h_{t-1}$ and the input for position $t$. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples. Earlier in context-representation there were two strategies for applying pre-trained language representations to downstream tasks: feature-based and fine-tuning. However, both are limited to the fact that they are unidirectional language models and are unable to learn general language representations. The latest advances in Bidirectional Encoder Representations from Transformers (BERT) address both of these issues, as it is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al., 2019). We have used small-BERT preprocessing and encoding to get vector representation of sentences, and finetuned BERT for the ternary classification task.

## 2 Data

Task 6 of SMM4H Shared Task 2021(Magge et al., 2021) challenged participants to develop an automatic classification system to identify tweets mentioning (1) self-reports, (2) non-personal reports, and (3) literature/news mentions. The task was formulated as a multi-class classification task, in which given a set of tweets a system should predict the label for each tweet. Table 1 gives the statistics of the dataset.

*Proceedings of the Sixth Social Media Mining for Health Workshop 2021*, pages 138–140
June 10, 2021. ©2021 Association for Computational Linguistics

| Dataset | LN | NP | Self | total |
|---|---|---|---|---|
| Train | 4277 | 3442 | 1248 | 9067 |
| Validation | 247 | 180 | 73 | 500 |
| Test | - | - | - | 6500 |

Table 1: Statistics of the dataset. LN: Lit-News, NP: Non-personal

## 3 System Description

### 3.1 Data Preprocessing

All apostrophe containing words were expanded. Characters like : , & ! ? were removed . Words were lower-cased to avoid capitalized version of the same word being treated as a different word. The emojis were removed using Python "emoji" library. Hashtags, mentions (words beginning with @) and urls were also removed.

### 3.2 Model

We used Small_BERT (Tsai et al., 2019) encoder and preprocessing models to extract features from the sentence and used the pooled outputs from the encoder and fed it into a fully connected dense layer, a dropout layer (dropout rate=0.1) and a final dense layer with softmax activation. We used the learning rate of 3e-5 and the adam optimizer. We tested it for 5-10 epochs and obtained the best result after training the model for 9 epochs.

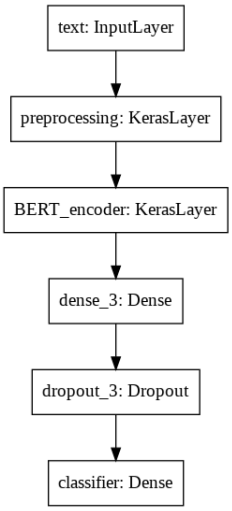

Figure 1: Model

## 4 Results and Analysis

We obtained an F1 score of 0.968 on the validation set and 0.9325 on the test set (cf. Table 2).

On analysing the wrongly classified tweets in the validation set, we observed some interesting patterns. The sentence “Me and my girl swear we have already had COVID-19. We were sick for nearly a month, fever, cough, sore throat, the doctors told me I had the flu combined with bronchitis because some days I felt like I was drowning in chest mucus.” was classified as self-report, while it is an ambiguous case of self-report and non-personal review. The misclasssification of the tweet “I had crippling body aches, fatigue and couldn't concentrate' - was @tomhanksanother COVID19 long-hauler? Sounds v. familiar! LongCovid @HadleyFreeman @guardian” was due to shortcoming of the preprocessing. This tweet is originally labelled as a Lit-News, though it was classified as self report. The main reason is that after preprocesing all the hashtags and the mentions were removed; therefore the overall context the model understood was in first-person and hence it classified the tweet as a self report.

| Dataset | F1 | Precision | Recall |
|---|---|---|---|
| Validation | .968 | .968 | .968 |
| Test | .9325 | .9325 | .9300 |

Table 2: Results

## 5 Conclusion

We present a Small BERT based model with custom preprocessing to classify tweets containing COVID-19 symptoms. We tested the model by adjusting various hyperparameters and presented the best result that we obtained using this model. We achieved F1 score of 93%. We observed that the preprocessing needed to include the mentions for some tweets for proper classification, though it was necessary to remove the mentions for the overall increase in the performance of the system.

## References

Huong Dang, Kahyun Lee, Sam Henry, and Özlem Uzuner. 2020. Ensemble BERT for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 37–41, Barcelona, Spain (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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical bert models for sequence labeling. *arXiv preprint arXiv:1909.00100*.