UACH at SMM4H: a BERT based approach for classification of COVID-19 Twitter posts

Alberto Valdés Chávez
Jesús Roberto López Santillán
Facultad de Ingeniería
Universidad Autónoma de Chihuahua
Chihuahua, Chih., Mexico
valdeschaveza@gmail.com
jrlopez@uach.mx

Manuel Montes-y-Gómez
Department of Computational Sciences
INAOE
Sta. Ma. Tonantzintla, Puebla, Mexico
mmontesg@inaoep.mx

Abstract

This work describes the participation of the Universidad Autónoma de Chihuahua team at the Social Media Mining for Health Applications (SMM4H) 2021 shared task. Our team participated in Tasks 5 and 6, both focused on the automatic classification of tweets related to COVID-19. Task 5 considered a binary classification problem, aiming to identify self-reporting tweets of potential cases of COVID-19. On the other hand, Task 6 goal was to classify tweets containing COVID-19 symptoms. For both tasks we used models based on bidirectional encoder representations from transformers (BERT). Our objective was to determine whether a model trained on a corpus from the domain of interest could outperformed one trained on a much larger general domain corpus. Our F1 results were encouraging, 0.77 and 0.95 for Tasks 5 and 6 respectively, having achieved the highest score among all the participants in the latter.

1 Introduction

The Social Media Mining for Health Applications (SMM4H) 2021 shared task aimed to address the challenges presented in Natural Language Processing (NLP) applied to text obtained from social networks, specifically Twitter, to gain medical insights (Magge et al., 2021). This year's SMM4H proposed 8 different problems that involved classification and Named Entity Recognition (NER) tasks. Our team focused on Tasks 5 and 6, both dealing with *classification of COVID-19 related tweets* in different situations. We decided to approach this problem with transformer-based models (Vaswani et al., 2017), since they are considered state-of-the-art in many NLP applications. Also, we hypothesized that a model trained on domain specific texts could performed better than one trained in a much larger but general-domain corpus. To test our hypothesis, we implemented two models that share the same architecture but were trained on different data sets; on the one hand, the large uncased version of BERT (BERT-Large) (Devlin et al., 2018), which is a model pretrained on a very large corpus (Wikipedia), and, on the other hand, CT-BERT that is a model based on BERT-Large but pretrained on a smaller corpus of COVID-19 related tweets (Müller et al., 2020).

2 Tasks Description

2.1 Task 5: Classification of tweets self-reporting potential cases of COVID-19

This is a binary classification task that involves distinguishing tweets of potential cases of COVID-19 (including situations that pose high risk of contagion) annotated as "1", from those that do not represent danger (annotated as "0"). Next, we show a tweet example for each class (Klein et al., 2021).

I think I have the coronavirus I've been coughing nonstop all day and I feel really warm **Label: "1"**

With coronavirus we certainly need more doctors and surgeons and nurses and sonographs and radiologists, let them in, quick! **Label: "0"**

Table 1 shows the labels distribution over the training, validation and test sets, as given by the organizers. NA denotes that the corresponding number is currently unknown.

Dataset	"1"	"0"	#
Training	1,026	5,439	6,465
Validation	122	594	716
Test	NA	NA	10,000

Table 1: Distribution of labels over the training, validation and test sets for Task 5.

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 65–68
June 10, 2021.

2.2 Task 6: Classification of COVID-19 tweets containing symptoms

This task is a three-way classification problem where the target classes are self_reports, non-personal_reports and literature/news_mentions. Self reports are personal mentions where the user describes his/her own symptoms. Nonpersonal reports are tweets where the user describes symptoms that other people experience. In addition, literature/news mentions are tweets coming from news articles or other sources that describe medical symptoms. Next, we show a tweet example for each class; then, Table 2 shows the labels distribution over the dataset.

In a study done in Milan, Italy, 402 Covid-19 patients were surveyed after being discharged. 28% showed symptoms of PTSD, 31% suffered from depression, 40% had insomnia, amp; 42% had anxiety. Overall, 56% of participants manifested at least one mental disorder following the disease. **Label: "Lit-News_mentions"**

@mention My wife takes 3 subway and a bus one way to reach her downtown office. She started having erratic fever and slight cough in past 3 days. She also travelled from India on 18th January via Germany and London. Does is qualify for a covid-19 test? **Label: "Nonpersonal_reports"**

Agreed! My covid19 was considered mid-level I wouldn't wish what I went through on my worst enemy. 1st symptoms March 19th - STILL RECOVERING!!! #longtailcovid **Label: "Self_reports"**

Dataset	LitNews	NonP	Self R	#
Training	4,277	3,442	1,348	9,067
Validation	247	180	73	500
Test	NA	NA	NA	6,500

Table 2: Distribution of labels over the training, validation and test sets for Task 6.

3 Our approach

For both tasks we implemented models based on the deep neural *transformer* architecture (Vaswani et al., 2017), since they have achieved state-of-the-art (SOTA) results in several NLP tasks. We studied the impact of fine-tuned BERT Large and CT-BERT pretrained models (Devlin et al., 2018)(Müller et al., 2020). For both models we employed the Pytorch implementation available from the HuggingFace library (Wolf et al., 2020).

3.1 Model Architecture

Figure 1 shows the general architecture shared by the two used models. It follows a standard design for sentence classification tasks using BERT, which considers the hidden state h of the final layer over the special token [CLS] as the full representation of the input sequences, and on top of this a classifier. The classifier head consists of a fully-connected layer with dropout probability of 0.1, 1024 input units, with 2 output units for Task 5 and 3 units for Task 6, followed by a softmax activation function to predict the class probability given the hidden state representation.

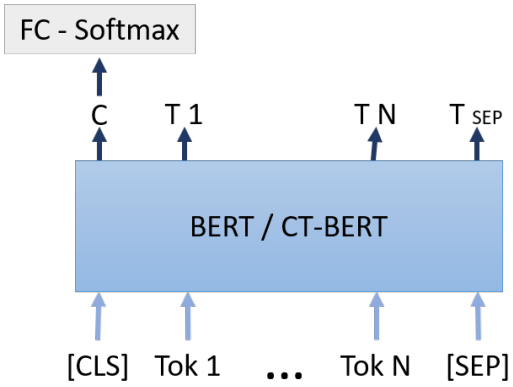

Figure 1: General model architecture

4 Experimental Setup

In this section we describe the training process for the two models in both tasks.

4.1 Data Preprocessing

All tweets in both tasks were preprocessed with the following operations[1]:

- Replace @usernames with a "user" token.
- Replace multiple occurrences of the "user" token with "n user", where n denotes the number of times the "user" token appears in the tweet.

[1]https://github.com/digitalepidemiologylab/covid-twitter-bert/tree/master/utils

- Replace URLs with a "url" token.
- Replace multiple occurrences of the "url" token with "n url", where n denotes the number of times the "url" token appears in the tweet.
- Convert emojis to their text aliases using the emoji library[2].
- Standardize text to ASCII representation. Using the Unidecode library[3], we removed all unicode symbols, punctuation and accented characters.
- Lowercase all the text.

4.2 Fine-tuning of models

In both tasks the models were fine-tuned with the Optuna framework (Akiba et al., 2019) using a random search approach by trying 10 different combinations for each model in each task. The search space described in Table 3 was defined so that the range of values stay close to the recommended values for BERT[4].

Weight Decay	LR	Epochs
0 - 0.3	1e-5 - 5e-5	1 - 4

Table 3: Hyper parameter search space

For Task 5, the best performing hyper parameter set according to the F1 score in the validation set was used in both models: 0.1328 *weight decay*, 3.154e-5 *learning rate* and 3 *epochs*. For Task 6 the values selected were 0.1423 *weight decay*, 3.278e-5 *learning rate* and 3 *epochs* for both models. Furthermore, the models for both tasks were trained over the concatenation of the training and validation sets for the final submission.

4.3 Results

Performance was measured using precision, recall and macro F1 metrics. Tables 4 and 5 show the performance attained in the test and validation sets of Task 5 and 6 respectively. Results in the validation and test sets of Task 5 were better for the CT-BERT model by a large margin, whilst for Task 6 both models achieved a high score with a small difference between them. CT-BERT outperformed both the BERT and the median submission score of all participants in the test set in both tasks achieving the highest score of all systems in Task 6.

[2]https://pypi.org/project/emoji/
[3]https://pypi.org/project/Unidecode/
[4]https://github.com/google-research/bert

System	Data	F1	P	R
BERT	Val	0.82	0.83	0.81
CT-BERT	Val	**0.89**	**0.89**	**0.90**
BERT	Test	0.68	0.71	0.65
CT-BERT	Test	**0.77**	**0.76**	**0.77**
Median	Test	0.74	0.73	0.74

Table 4: Results on test and validation data for Task 5

System	Data	F1	P	R
BERT	Val	**0.99**	**0.99**	**0.99**
CT-BERT	Val	0.98	0.98	0.98
BERT	Test	0.94	0.93	0.93
CT-BERT	Test	**0.95**	**0.94**	**0.94**
Median	Test	0.93	0.93	0.93

Table 5: Results on test and validation data for Task 6

5 Discussion

The training set for Task 5 was unbalanced, approximately at a ratio of 1:5 for classes "1" and "0". During experimentation an attempt was made to balance the classes to see if this would improve the results, but this approach was abandoned as the metrics dropped considerably. Also, it was observed that data preprocessing had a greater impact than hyper parameter search on the models for both tasks.

We analyzed the errors on the validation set in Task 5 for both models. The BERT model mislabeled 60 tweets vs 44 for CT-BERT, with 27 errors in common. Next we show two tweets that both models wrongly predicted:

- *i cough once and people think i have the coronavirus*. **Predicted = 1, True label = 0**
- *I legit feel super sick to my stomach and really weak hopefully I'm not dying from coronavirus*. **Predicted = 0, True label = 1**

Data in Task 5 showed that tweets labeled as "1" contain more mentions of the words "i", "got", "cough" and other symptoms compared to the tweets labeled as "0". While analyzing the common errors in both BERT and CT-BERT on the validation set, we discovered that "0" tweets that included these words were often misclassified as "1". On the other hand, in the validation set of Task 6 the BERT model mislabeled 4 tweets vs 9 tweets for CT-BERT, with only 3 errors in common. The following are two examples of errors made by our model.

- *@user Hi @user. The symptoms of Covid-19 are similar to that of a common cold or flu. These symptoms are: fatigue, fever, coughing, stuffy nose, sore throat or diarrhea. Seek medical attention if you, your child or family member show any of these signs. url.* **Predicted = Lit-News, True label = Nonpersonal Reports**

- *@user1 @user2 @user3 @user4 @user5 @user6 @user7 @user8 @user9 Man life is full of tortures. Has everyone with covid19 shown excessive damage in India? It is the opposite. 80% are asymptomatic. 10% have fever and 5% require medical supervision and rest need oxygen support. No need to panic..* **Predicted = Nonpersonal Reports, True label = Lit-News**

All errors in the validation set were misclassifications between the Lit-News and Nonpersonal Reports labels, where the model struggles to differentiate between the size of the audience of the tweet. In addition, tweets written in an impersonal style tend to be classified as News, whereas tweets written in first person tend to be classified as Nonpersonal or Self Reports.

6 Conclusions and future work

Transfer learning has shown to achieve above average results for various NLP tasks, where domain specific models can attain better results even if they were trained with less data than a general domain model. Based on the results obtained, we conclude that a model trained with quality domain-specific data (CT-BERT) can outperform a model trained with a much larger amount of general domain data (BERT).

In the experimental stage we also considered the BioBERT model, which was trained on a medical corpus (Lee et al., 2019), but it was not possible due to time constraints. Thus, we envision a potential future work to further compare the reach of several domain-specific models in the prediction of social media posts that could embed Covid-19 infection risk information.

Based on the error analysis, we plan to further improve our models' performance by considering *wide & deep* learning techniques (Cheng et al., 2016), which help to enhance their generalization ability by adding other types of features through the wide branch.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide deep learning for recommender systems.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ari Z Klein, Arjun Magge, Karen O'Connor, Jesus Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez Hernandez. 2021. Toward using twitter for tracking covid-19: A natural language processing pipeline and exploratory data set. *J Med Internet Res*, 23(1):e25314.

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*, 36(4):1234–1240.

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*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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. Huggingface's transformers: State-of-the-art natural language processing.