

Innovators@SMM4H’22: An Ensembles Approach for Stance and Premise Classification of COVID-19 Health Mandates Tweets

Vatsal Savaliya^ϕ, Aakash Bhatnagar[§], Nidhir Bhavsar[&], Muskaan Singh[#], Petr Motlicek[#]

^ϕIndian Institute of Information Technology Lucknow, Indian

[§]Boston University, Boston, Massachusetts

[&]University of Potsdam, Germany

[#] Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland

lci2020013@iitl.ac.in, aakash07@bu.edu, nidbhavsar989@gmail.com,

(msingh, petr.motlicek)@idiap.ch

Abstract

This paper presents our submission for the Shared Task-2 of classification of stance and premise in tweets about health mandates related to COVID-19 at the Social Media Mining for Health 2022. There have been a plethora of tweets about people expressing their opinions on the COVID-19 epidemic since it first emerged. The shared task emphasizes finding the level of cooperation within the mandates for their stance towards the health orders of the pandemic. Overall the shared subjects the participants to propose system’s that can efficiently perform 1) Stance Detection, which focuses on determining the author’s point of view in the text. 2) Premise Classification, which indicates whether or not the text has arguments. Through this paper we propose an orchestration of multiple transformer based encoders to derive the output for stance and premise classification. Our best model achieves a F1 score of 0.771 for Premise Classification and an aggregate macro-F1 score of 0.661 for Stance Detection. We have made our code public [here](#)

1 Introduction

Users actively share their opinions on numerous problems on social media. These concerns are now frequently linked to the COVID-19 pandemic. Users, for example, share their feelings about quarantines and wearing masks in public places. Some comments are supported by arguments, while others are simply emotional claims.

Using Twitter tweets, automated algorithms for assessing people’s attitudes regarding COVID-19 health orders can assist in determining the extent of cooperation with the mandates.

In this paper, we try to provide a feasible solution towards the task of Stance-Detection (Glandt et al., 2021). We are extracting arguments from COVID-related tweets. According to argumentation theory, an argument must include a claim containing a stance towards some topic or object, and at least

Model	Precision	Recall	F1
BERT	0.672	0.667	0.659
RoBERTa	0.671	0.616	0.599
PubMedBERT	0.643	0.637	0.639
SciBERT	0.663	0.66	0.661
SPECTER	0.596	0.566	0.572
Ensemble	0.717	0.724	0.716

Table 1: Shows the evaluation scores obtained on the test-set of the Stance Detection dataset measured across the metrics precision, recall, and f1. We represent each individual score obtained over fine-tuning each of the previously mentioned pre-trained language models.

one premise supporting this stance. So basically, we are finding whether the given tweet contains a stance (favour, against, or none) and also if it contains a premise (reason to support the stance).

The first subtask of the shared task is to design a system that can determine the point of view (stance) of the text’s author in relation to the given claim (e.g., wearing a face mask). Given a tweet, there can be 3 different values for stance *FAVOR* – positive stance *AGAINST* – negative stance *NEITHER* – neutral/unclear/irrelevant stance

The second subtask is to predict whether at least one premise or argument is mentioned in the text. A given tweet is considered to have a premise if it contains a statement that can be used as an argument in a discussion. There are 2 values for premise. **1** – tweet contains a premise (argument) **0** – tweet doesn’t contain a premise (argument).

1.1 Motivation

Millions of users share and read tweets actively. During the COVID-19 pandemic, many users shared information about health mandates such as face masks, stay-at-home orders, and school closures. So basically, every tweet claims some particular topic.

The first subtask is to determine the point of view (stance) of the text’s author in relation to the given claim. This means that stance detection is very important for a tweet, especially when the

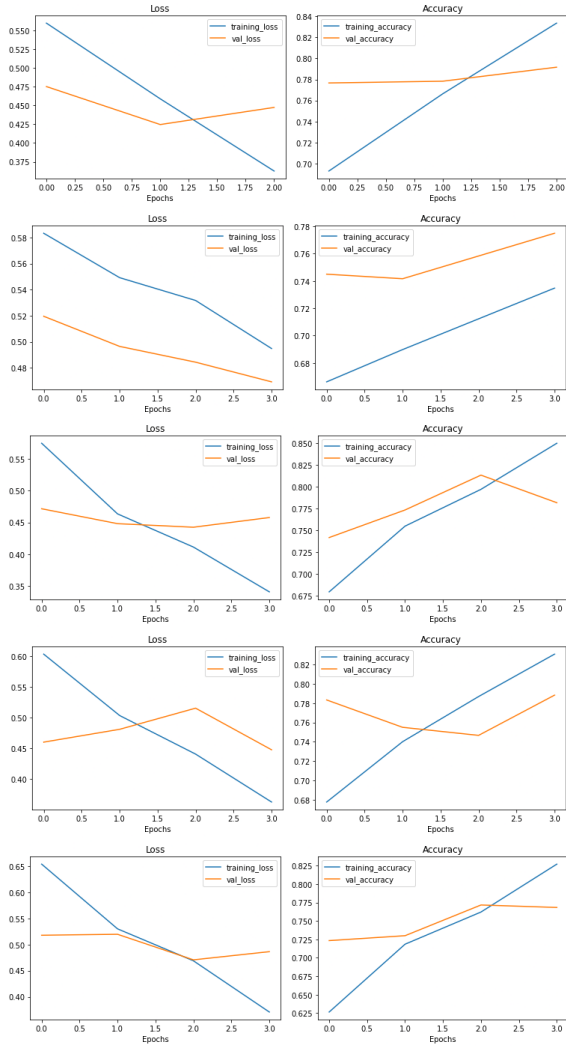


Figure 1: Graphical presentation for the loss and accuracy plot for (a)BERT (b)RoBERTa (c) PubMedBERT (d) SciBERT (e)SPECTER

tweet is done by a renowned personality. The second subtask is to predict whether at least one premise/argument is mentioned in the tweet. It's also important to find out whether the tweet is supported by an argument or not. This way, we can show tweets to users which are actually supported by an argument.

2 System Architecture

The main task was divided into two subtasks. Subtasks (a) STANCE DETECTION and (b) PREMISE CLASSIFICATION, respectively. Both subtasks have classification problems, with subtask (a) having 3 classes (favour, against, or none) and subtask (b) having 2 classes (1-tweet contains a premise, 2-tweet does not contain a premise).

We try out five different transformer models:

BERT, PubMedBERT, ROBERTS, SciBERT, and SPECTER. Finally, we combine all five models' predictions with a voting classifier.

In ensemble modeling, multiple diverse models are created to predict an outcome. In our case, we use a voting classifier to ensemble the predictions of all five models (BERT, RoBERTa, PubMedBERT, SciBERT and SPECTER). As a result, the overall accuracy and f1 scores increased for stance detection and premise classification.

- BERT(Devlin et al., 2018) is pretrained language model trained specifically for the task of masked language modelling (MLM) and next sentence prediction (NSP).
- RoBERTa(Liu et al., 2019) is a transformers based pretrained language model trained on a large corpus of English data in a self-supervised fashion. The model follows the pre-training architecture proposed by Devlin et al. (2018) tries to overcome the limitation relative to Next Sentence Prediction (NSP). It also states the impact of hyperparameter tuning and training data size.
- PubMedBERT(Gu et al., 2020) is a pretrained biomedical domain-specific model which follows the BERT architecture. It is trained on the publicly available PubMed and full-text articles from PubMedCentral.
- SciBERT(Beltagy et al., 2019) follows the BERT architecture and is pretrained on scientific training corpus of papers from the Semantic Scholar and constitutes of both abstracts and full-texts. The language model follows a different wordpiece vocabulary "scivocab" built specifically to match the training data.
- SPECTER(Cohan et al., 2020) is a pre-trained model to generate document-level embeddings of scientific texts. It is pretrained on a powerful signal of document-level relatedness also refereed as citation graphs. Furthermore, the language model incorporates SciBERT, and can be applied easily to downstream scientific domain specific tasks.

3 Results

We present our results in Table: 1 and 2 across five different models. BERT, PubMedBERT, RoBERTa, SciBERT, SPECTER which were trained and evaluated on validation and testing data for both the

Model	Precision	Recall	F1	Accuracy
BERT	0.7191	0.7681	0.7428	0.805
RoBERTa	0.6313	0.8409	0.7212	0.7616
PubMedBERT0.75	0.709	0.709	0.7289	0.8066
SciBERT	0.7759	0.6454	0.7047	0.8016
SPECTER	0.6762	0.75	0.7112	0.7766
ensemble	0.761	0.7818	0.7713	0.83

Table 2: Shows the evaluation scores obtained on the validation-set of the Premise Classification dataset measured across the metrics precision, recall, f1, and Accuracy. We represent each individual score obtained over fine-tuning each of the previously mentioned pre-trained language models.

	F1 Stance Detection	F1 Premise Classification
Ours	0.4816236592	0.64073476
Mean	0.4913350946	0.5739699138
Median	0.5501048571	0.6472029954

Table 3: Compares the F1 scores obtained by the best performing system at the SMMH’22 Task 2. It also provides and analysis over the Mean and Median of submission under the same task

subtasks. Our proposed approach of ensembled voting classifier, performs best (in terms of F1-scores of **0.716** for stance detection and **0.83** for premise classification).

4 Analysis

The overall dataset contains 6269 tweets, 3669 in training, 600 in validation and 2000 in testing. The tweets were further divided into three categories based on their claim, (i) face masks (ii) stay at home orders (iii) school closures.

As previously indicated, our proposed system employs an ensemble voting classifier, which assigns distinct weights to each of the predictions from the 5 included models during inference and determines the output label appropriately. Table 1 provides separate results obtained by each of our fine-tuned models on the stance detection dataset. It is clearly justifiable that the ensemble approach easily outmatch the individual models. Next, the individual results from each of our trained models on the given premise classification dataset are shown in Table 2. Although the SciBERT model produces a higher precision-score of 0.7759, the total f1-score obtained by ensemble surpasses the same by a factor of 0.01 compared to the prior method, yielding better results overall.

Table 4 compares our models prediction against the ground truth labels. As it is clearly seen, our system can easily classify both stance and premise appropriately. For better clarification, we also provide the associated claims alongside the given

tweets. The proposed model may be partially treating all of the tweets that express negativity as the stance *AGAINST*, which is indirectly related to premise being mis-classified, and as a result, the models may occasionally discover the presence of premise for assertions like the stay at home orders.

Table 3 shows the complete task statistics for the test dataset for both stance detection and premise classification. Our premise classification proposal received a f1 score of 0.64, which is significantly higher than the typical submission for this subtask. However, for stance detection, our model underperformed the average submission, while the median shows that the majority of recursive submissions earned a f1 score of 0.51, showing that our suggested approach has to be improved.

5 Conclusions and Future work

We explored the use of a simple transformer based architecture for task 2 proposed by SMM4H’22. The five different transformer models (BERT, RoBERTa, PubMed-BERT, SciBERT, and SPECTER) were trained separately on both subtasks, and their results were then combined to create a voting classifier. As expected, the overall accuracy and f1 scores of the voting ensemble classifier were better as compared to individual model prediction for both the aforementioned subtasks. However, there are still greater possibilities towards improving the overall accuracy. We intent to mine crucial medical information from the text and train a separate classifier on them in this manner we’ll have a more precise and robust data to infer from. Also, due to time constraint we were restricted towards training a monolingual model, however the given data also entail written text in languages other than English which are ignored by our proposed system. Recently proposed Language-Models like Bloom¹ can help the model in dealing with the multilingual data.

6 Acknowledgements

This work was supported by the European Union’s Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROX-ANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

¹<https://huggingface.co/bigscience/bloom>

References

- Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. [Scibert: Pretrained contextualized embeddings for scientific text](#). *CoRR*, abs/1903.10676.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. [SPECTER: document-level representation learning using citation-informed transformers](#). *CoRR*, abs/2004.07180.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. [Stance detection in COVID-19 tweets](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. [Domain-specific language model pretraining for biomedical natural language processing](#). *CoRR*, abs/2007.15779.
- 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](#). *CoRR*, abs/1907.11692.

A Appendix

Tweet	Claim	Actual		Predicted	
		Stance	Premise	Stance	Premise
Ordered a mask that had a cute chain is attached so you don't lose it during the day: raising_hands_medium-light_skin_tone	face masks	Favour	0	Favour	0
How many of #US would've agreed to #shutdown had we known it would be for 4+ months!?!#California rules regarding #COVID19 #lockdown are completely ridiculous. Many family businesses & corporate #Restaurants will file #bankruptcy due to prolonged compliance burden.	stay at home orders	Against	1	Against	1
@ananavarro That representative from Texas said almost the same crap in his mind, refused to wear a mask when asked...Guess what, God doesn't appreciate that crap.. he's now dealing with Covid-19 wherever he is.	face masks	Favour	1	Favour	0
we gotta go back to school, i bet once they start pushing kids back to school, there will be a massive increase in cases. Kinda sad how they still don't realize we are in the middle of a pandemic where cases are still rising every day. #SchoolReopening	school closures	Favour	0	Favour	1
Ok, so we can't trust educators to safe distance or wear masks? Why in the hell are we letting them teach our children?	school closures	Against	1	Favour	1
@realDonaldTrump And now I bet they're wishing they had stayed in space #TrumpVirus #TrumpFailsAmerica	face masks	None	0	None	0
@garethicke In last week, 2 of my family have been made redundant and 1 has attempted suicide. #endthelockdownuk #EndTheNightmare	stay at home orders	Against	1	None	0
@BetsyDeVosED you are crazy if you think the NBA and all it's resources in a bubble aren't able to keep Covid out, but the extremely underfunded schools are going to be able to.	school closures	Favour	1	Favour	1
@ottawacity The TRUTH @ottawahealth is hiding from you: Once the temperature reaches above 74F degrees 23.3C the virus CANNOT survive!. Why the masks? (mind control) Why the increase in numbers? (more mind control) Why distancing? (mind control) Awake now? #OttawaMaskContest	face masks	Against	1	Against	1
What's more important? A students education? Or the wellness of the student and staff? Closing schools for a day may work for bad weather, but not for the coronavirus. Districts all around need to open their eyes and realize one day closures do nothing to help!	school closures	Favour	1	Favour	1

Table 4: Result Analysis for empirical