# Team enolp musk@SMM4H'22 : Leveraging Pre-trained Language Models for Stance And Premise Classification

**Anonymous ACL submission**

## Abstract

This paper covers our approaches for the Social Media Mining for Health (SMM4H) Shared Tasks 2a and 2b. Apart from the baseline architectures, we experiment with Parts of Speech (PoS), dependency parsing, and Tf-Idf features. Additionally, we perform contrastive pretraining on our best models using a supervised contrastive loss function. In both the tasks, we outperformed the mean and median scores and ranked first on the validation set. For stance classification, we achieved an F1-score of **0.636** using the CovidTwitterBERT model, while for premise classification, we achieved an F1-score of **0.664** using BART-base model on test dataset.

## 1 Introduction

With the growth of social media platforms such as Twitter and Facebook, people can express their views on any topic. This lets them reach a vast audience, as these platforms have millions of users, and helps them get recognition for their issues online. In the case of trending events on social media, many people post their opinions around the same time, thus rapidly shaping the public opinion on those events. Misinformation may be spread unknowingly or knowingly by malicious parties, creating the wrong public opinion among social media users.

Generally, when governments introduce new schemes or policies, they need to know the people's reviews, suggestions, and complaints. Social Media sites like Twitter are essential in this aspect as they let users post microblog tweets of up to 280 characters and allow them to use hashtags to link the tweet to other tweets of the same topic that use this hashtag. During the COVID crisis, many public mandates were imposed regarding the closure of public places, lockdown in many places, and social distancing norms. This unusual change made many people post their views on these topics. To derive the public opinion on these topics, we must know the text's claim i.e. towards what entity is this argument targeted, the user's stance associated with this claim i.e. whether the argument is in favor or against the claim, and the premise associated with this claim i.e whether the text has a convincing argument or not.

Similar to the work carried out by (Glandt et al., 2021), we perform stance classification for health mandate-related tweets by employing pre-trained transformer models and then trying to improve classification results by using additional features as described in the section 3.2.

## 2 Task & Dataset Description

Task 2 of the shared tasks for Social Media Mining for Health has the motive for stance and premise classification of tweets related to the health mandates - "Stay at Home Orders", "Face Masks" and "School Closures". The two subtasks are:

- Task 2a: Classification of stance in tweets about health mandates related to COVID-19 (in English)
- Task 2b: Classification of premise in tweets about health mandates related to COVID-19 (in English)

The dataset consists of views of the general public on various issues and government mandates for COVID-19 in the form of Twitter Comments. It is manually labeled for claims, stance, and premise. The claims are of three categories "Stay at Home Orders", "Face Masks" and "School Closures". Each of the annotated entries in the dataset consists of stance which is of one of the three classes- "FAVOR", "AGAINST" or "NEITHER", and the premise which contains the binary annotations 0 (if the tweet does not have any premise) or 1 (if the tweet contains a premise). The training, validation, and test sets are of size 3556, 600, and 10000 tweets, respectively.

## 3 Implementation Details

Section 3.1 describes the baseline architectures used for our experiments, Section 3.2 and Section 3.3 describe the techniques tried to improve the models using additional features and contrastive learning respectively. We discuss the details of our experimental setting in Section 3.4.

### 3.1 Baseline Architectures

We adopt pre-trained transformer models (Vaswani et al., 2017) from HuggingFace and add linear layers over the 786 or 1024 dimensional sentence representation from the models (Figure 1) and finetune them for our use case. We use monolingual pretrained language models as the data corpus consists of English comments only.

The details of exact architectural details is discussed in this section.

- BERT[1](Devlin et al., 2018) is a transformers model pretrained on a large corpus of English data in a self-supervised fashion with the Masked Language Modelling objective. We use the uncased version of the model for our experiments.
- RoBERTa[2] (Liu et al., 2019), pretrained on a large corpus of English data in a semi-supervised setting with dynamic masked language modelling objective.
- DeBERTa-V3[3] (He et al., 2021), is an improvement over the original DeBERTa model that uses ELECTRA-style pre-training with Gradient Disentangled Embedding Sharing. DeBERTa improves the BERT and RoBERTa models using disentangled attention and enhanced mask decoder.
- BART[4] (Lewis et al., 2019) is a encoder-decoder transformer model with a BERT like bidirectional encoder and auto-regressive decoder like GPT (Radford and Narasimhan, 2018).
- CovidTwitterBERT-V2[5](Müller et al., 2020), is a BERT$_{large}$ model pretrained on large corpus of Tweets on COVID-19.

[1]huggingface.co/bert-base-uncased
[2]huggingface.co/roberta-large
[3]huggingface.co/microsoft/deberta-v3-large
[4]huggingface.co/facebook/bart-large
[5] huggingface.co/digitalepidemiologylab/covid-twitter-bert-v2

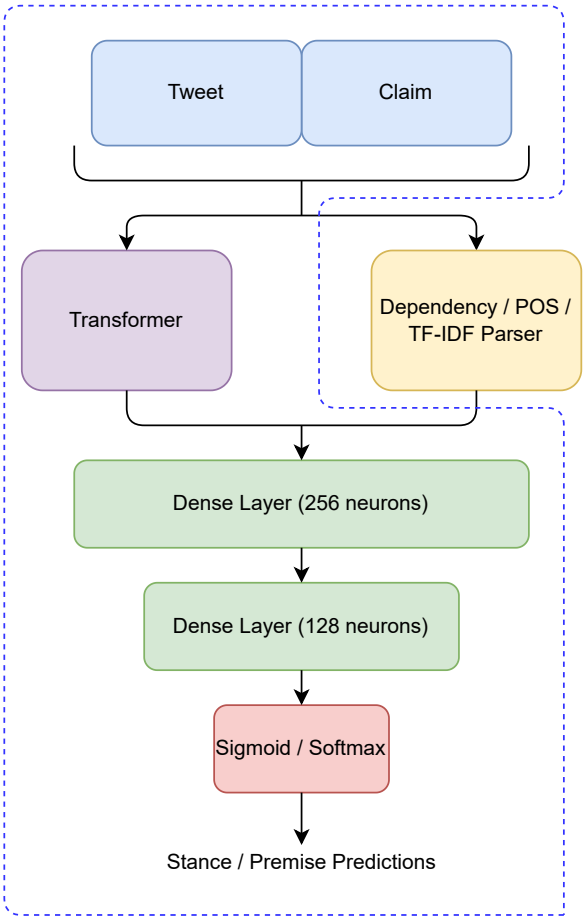


Figure 1: Model Architectures
The part of the figure inside the blue dotted line is the baseline architecture.

### 3.2 Additional Features

As suggested in (Kapadnis et al., 2021), we pass additional information to the model in the form of Parts of Speech features(POS), Dependency Parsing features and term frequency-inverse document frequency(Tf-Idf) features. The claims corresponding to the tweets are appended to the text before computing the feature vectors.

We find the POS label for each lexicon and then label encode it according to the descending order of occurrences. This feature vector is then merged with the pooled layer output and passed to the next set of dense layers.
Similarly, we do it for dependency parsing features and Tf-idf bi-gram features generation.

### 3.3 Contrastive Pretraining

We try to improve the representations in the embedding space of our best model (CovidTwitterBERT & BART-base) by using a supervised contrastive loss function (Khosla et al., 2020), instead of the

usual cross-entropy loss. This can leverage label information better and morph the embedding space by pulling data points from the same class closer and pushing apart data points from other classes. The loss is defined as follows,

$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{exp\,(z_i z_p / \tau)}{\sum_{a \in A(i)} exp\,(z_i z_a / \tau)} \tag{1}$$

We try three different strategies, finetuning with contrastive loss only, pretraining with contrastive loss then finetuning with cross-entropy loss and finetuning with a weighted loss with 0.7 weight given to cross-entropy and 0.3 to contrastive loss. The weights were chosen considering the performance when finetuning with cross-entropy and contrastive loss.

### 3.4 Experimentation Details

Preliminary analysis of the training set reveals that the dataset is balanced among the classes for stances. However, this is not the case for the premise prediction task. The most frequent tri-grams derived from the tweets that favor the claim are *'#closetheschools #closetheschools #closetheschools'*, *'wear damm mask'* & *'people wearing masks'*; the same for against *'#schoolreopenindia #schoolreopenindia #schoolreopenindia'*, *'open. don't think'* & *'don't wear mask'*. The none class has tri-grams from random topics like *'liveon #ourapponplaystore #websiteonbio'*.

We preprocess the dataset before passing it through the transformer models. The whitespaces, URLs, HTML tags, and emoji patterns are removed from the tweet string. The processed tweets and claims are passed to the model tokenizer, which returns the encoded input which consists of input ids and attention masks. The stance values are mapped to numerical values using a label binarizer.

All the models are trained with a batch size of 16, except CovidTwitterBERT which is trained with a batch size of 8 due to computational constraints. A very low learning rate of $10e^{-5}$ is used to finetune the transformer models for 10 epochs. Weight decay and learning rate scheduling is used to prevent overfitting. The input sequence length after tokenization is set at 512 to get the full advantage of these large language models.

| MODELS | base | large | large + noun | large + dependency | Contrastive |
|---|---|---|---|---|---|
| BERT | 0.606 | 0.602 | 0.475 | 0.537 | - |
| RoBERTa | 0.645 | 0.731 | 0.744 | 0.743 | - |
| DeBERTa-V3 | 0.630 | 0.773 | 0.784 | 0.784 | - |
| BART | 0.479 | 0.691 | 0.728 | 0.708 | - |
| CovidTwitterBERT | - | **0.855** | 0.843 | 0.827 | 0.835 |

Table 1: Performance on Stance Prediction

| MODELS | base | large | large + noun | large + dependency | Contrastive |
|---|---|---|---|---|---|
| BERT | 0.803 | 0.785 | 0.753 | 0.655 | - |
| RoBERTa | 0.823 | 0.817 | 0.741 | 0.675 | - |
| DeBERTa-V3 | 0.819 | 0.812 | 0.760 | 0.729 | - |
| BART | **0.857** | 0.803 | 0.715 | 0.738 | 0.704 |
| CovidTwitterBERT | - | 0.781 | 0.764 | 0.734 | - |

Table 2: Performance on Premise Prediction

## 4 Evaluation Metrics

The F1-scores for stance and premise are calculated as follows,

$$F_1 = 1/n \sum_{c \in C} F_{1rel,c}$$

where, C is the set of claims, n is the number of unique claims and $F_{1rel}$ is the macro $F_1$-score averaged over first two relevance classes (the class "neither" is excluded).

## 5 Summary of Results

We note the crucial observations for Stance & Premise prediction on the development dataset in Table 1 & Table 2[6] respectively. The first column mentions the model name, and the rest of the columns are the performance while using a specific model size or other improvement strategies described earlier.

We observe that, in general the transformer models pre-trained on in-domain datasets perform better than the base models. Furthermore, we also notice a jump in performance gained with CovidTwitterBERT by pretraining the baseline $\text{BERT}_{large}$ model with in-domain data.

We performed contrastive training only for the best model for each subtask. However, the performance didn't improve with this strategy. Passing more contextual information in the form of additional features gives a considerable boost to the performance of baseline models for stance prediction but fails for premise prediction.

Surprisingly, BART-base outperforms the large and in-domain pre-trained models for Task 2b.

[6]Note that the results for Task 2b are according to the evaluation metric described in the Task Description and doesn't match with Codalab 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*.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Manav Kapadnis, Sohan Patnaik, Siba Panigrahi, Varun Madhavan, and Abhilash Nandy. 2021. Team enigma at ArgMining-EMNLP 2021: Leveraging pre-trained language models for key point matching. In *Proceedings of the 8th Workshop on Argument Mining*, pages 200–205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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*.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.