Leveraging Pre-trained Language Models for Stance And Premise Classification

Millon Madhur Das, Archit Mangrulkar, Ishan Manchanda, Manav Nitin Kapadnis, Sohan Patnaik

Indian Institute of Technology Kharagpur

Shared Task Description

The Social Media Mining for Health 2022 Shared Task 2: Classification of stance and premise in tweets about health mandates (COVID-19) involves 2 subtasks 2a Stance Detection and 2b Premise Detection.

In the first subtask, given a tweet and claim, participants must report the stance of the text's author in relation to the claim. The second subtask is to predict whether the tweet contains at least one argument/premise mentioned in its text.

Dataset Description

The given dataset [4] for the shared task consists of **6,156** tweets. Each tweet is manually labeled for claims ("Stay at Home Orders", "Fake Masks", and "School Closures"), stance ("FAVOR", "AGAINST", "NEITHER"), and premise (0 if the tweet does not have any premise and 1 if it does).

The training, validation, and test sets are of size 3556, 600, and 2000 tweets, respectively.

Tweet	Claim	Stance	Premise
The fact that anti-masking is a thing is a completely terrifying insight into the nature of some beings who look, walk and breathe just like us.	face masks	FAVOR	0
Masks help prevent the spread of the disease. Please, #WEARAMASK	face masks	FAVOR	1
Woah, so you're telling me that my five kids have to go to school while this coronavirus outbreak is happening. It was fair enough with all the strikes.	school closures	FAVOR	0
@GodFamilyJesus Masks Kill Our Immune Systems. They cover the mouth which needs to be seen when negotiating and conversing. Not seeing facial expressions leads to depression and misunderstanding. Fungal Respiratory Infections, Heat Stroke. Slave Muzzle, Panic Attacks Fear Mongering.	face masks	AGAINST	1

Contrastive Pretraining

- We try to improve the representations in the embedding space of our best model by using a supervised contrastive loss function [2], 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.

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

- Three different strategies were carried out, 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

Overview of the Approach

The approach to our system revolves around leveraging transfer learning by using pre-trained language models (especially transformers) along with finetuning. We use monolingual pretrained language models as the data corpus consists of English tweets only. The models themselves are a set of BERT and its variants including RoBERTa, DeBERTa-V3 [1], BART, and CovidTwitterBERT-V2 [3]. CovidTwitterBERT-V2 is a BERT large model that is pretrained on a large corpus of tweets related to the COVID-19 crises. We hypothesize that this model should perform better as the data domain closer.

In both the subtasks we use the transformer models to generate embeddings and then apply linear layers on top to get the predictions. We pass the preprocessed tweet along with the claim separated by specific token. We experiment with both the base and large variants of the mentioned 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

Fig. 1: Results on Stance Detection

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	-

Fig. 2: Results on Premise Detection

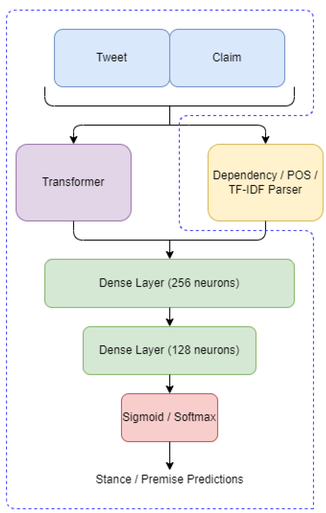

Fig. 3: Model Architecture

Extra features

As mentioned earlier, we experimented with three types of features i.e., Dependency Parsing features, Parts of Speech features and Tf-idf features.

1. **Dependency Parsing features**: In order to capture the syntactic structure of structure of the sentences, we added dependency parse tree of the sentences as an additional feature. The dependency features such as aux, amod, nsubj were label encoded according to the descending order of their occurances in the dataset. Finally, these encoded features were concatenated with the output of the transformer model and passed to the subsequent layers.
2. **Parts of Speech features**: With a similar approach to dependency parsed features, we obtained the label encoded Parts of Speech features according to descending order of their occurrences and concatenated them with the transformer output before feeding them into the fully-connected layers.
3. **Tf-idf features**: With an intuition to get better results by combining lexical overlap-based features with semantic features obtained from the transformer, we obtained Tf-idf vectors of the tweet and concatenated with the transformer output before feeding them into the subsequent layers. Tf is short for Term-frequency that gives the degree of presence of a word in a tweet, whereas idf is short for inverse-document-frequency that gives the measurement of the uniqueness of a word to a particular tweet with respect to the set of all tweets in the dataset. Tf-Idf can be understood by the word frequencies weighted by the words' uniqueness to the selected dataset. A word which can be seen almost everywhere in a target tweet and cannot be found in other tweets in the corpus has high degree of uniqueness on the target tweet and hence a high Tf-Idf weight on that triplet. This feature along with CovidTwitterBERT model helped us achieve the best F1 score.

References

[1] Pengcheng He, Jianfeng Gao, and Weizhu Chen. "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing". In: *arXiv preprint arXiv:2111.09543* (2021).

[2] Prannay Khosla et al. "Supervised contrastive learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 18661–18673.

[3] Martin Müller, Marcel Salathé, and Per E Kummervold. "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter". In: *arXiv preprint arXiv:2005.07503* (2020).

[4] Davy Weissenbacher et al. "Overview of the Seventh Social Media Mining for Health Applications SMM4H Shared Tasks at COLING 2022". In: *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop Shared Task*. 2022, pp. –.