yiriyou@SMM4H'22: Stance and Premise Classification in Domain Specific Tweets with Dual-View Attention Neural Networks

Huabin Yang[1], Zhongjian Zhang[1], Yanru Zhang[1,2]*

[1]University of Electronic Science and Technology of China, Chengdu
[2]Shenzhen Institute for Advanced Study, UESTC
`202121080420@std.uestc.edu.cn`
`zhongjianzhang6972@gmail.com`
`yanruzhang@uestc.edu.cn`

Abstract

The paper introduces the methodology proposed for the shared Task 2 of the Social Media Mining for Health Application (SMM4H) in 2022. Task 2 consists of two subtasks: Stance Detection and Premise Classification, named Subtask 2a and Subtask 2b, respectively. Our proposed system is based on dual-view attention neural networks and achieves an F1 score of 0.618 for Subtask 2a (0.068 more than the median) and an F1 score of 0.630 for Subtask 2b (0.017 less than the median). Further experiments show that the domain-specific pretrained model, cross-validation, and pseudo-label techniques contribute to the improvement of system performance.

1 Introduction

Since the outbreak of the COVID-19 pandemic, there has been a large amount of research on COVID-19 in the natural language processing (NLP) arena. Meanwhile, social media platforms such as Twitter are widely used to share and spread users' views on various issues, so the Social Media Mining for Health Application (SMM4H) Shared Task in 2022 (Weissenbacher et al., 2022) focuses on leveraging COVID-related tweets for health research. In this paper, we describe our methodology for the shared Task 2: Classification of stance and premise in tweets about health mandates related to COVID-19 (in English). We are provided with labeled training set containing texts from Twitter about three health mandates (claims) related to COVID-19 pandemic: *school closures*, *stay at home orders* and *face masks*[1]. Subtask 2a aims at detecting the stance of the text's author concerning the given claim (e.g., *school closures*), while Subtask 2b aims at identifying whether there are premises in an argument.

---

*Corresponding author

[1]see Davydova and Tutubalina (2022) for more details

2 Methodology

2.1 Data

We use the dataset provided by the organizers of Task 2. For Subtask 2a, the tweets in the training and validation sets are annotated for stance towards the given claim according to three categories: FAVOR, AGAINST, and NONE. For Subtask 2b, the tweets in the training and validation sets are annotated as 1 for containing a premise (argument) and 0 otherwise. In addition, we note that the data in both the training and validation sets are in English. And the emoji in the tweets have been replaced with the corresponding textual representations (e.g., :face_with_medical_mask). However, we find that the language of tweets in the test set is variant. Most tweets are in English, but there are tweets in other languages, such as Hindi or Urdu. Even for tweets in the same language, some of them contain emoticons that will affect the overall meaning, indicating that data preprocessing on the test set is essential before predicting the results. Table 1 gives the statistics of the Task 2 dataset.

Training	Validation	Test
3556	600	2000(10000)

Table 1: Statistics of the Task 2 dataset. Note that only 2000 tweets in the test set were included in the metrics computing, other 8000 tweets were added to avoid manual annotation.

2.2 Preprocessing

In order to reduce the noise of tweets in the test set, we preprocess the test set as follows:

- For those tweets are not in English, we use Google Translation to translate them into English, considering that multilingual text can affect the final results. However, we note that Google Translation does not always translate the sentence as its intended meaning, so this

kind of operation may result in unexpected effects.

- To unify the impact of emoticons in the test set, we perform emoji codification to convert the emoticons into textual representations using the python emoji package [2].

Figure 1 shows examples of data preprocessing. (a) shows a good example of preprocessing while (b) proves that Google Translation doesn't always work well.
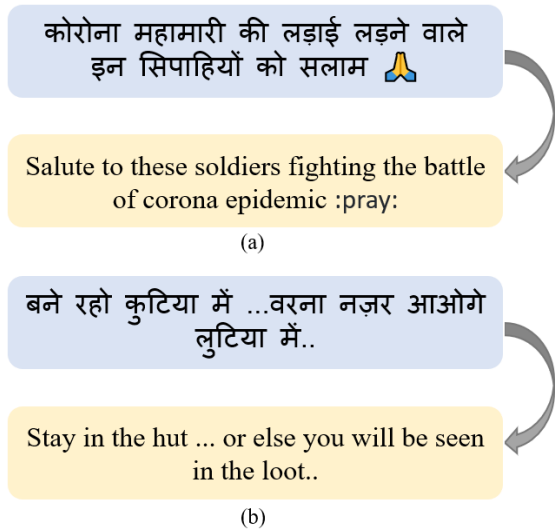

Figure 1: Examples of preprocessing. (a) Good preprocessing result. (b) Bad preprocessing result.

2.3 Modeling

Dual-view Attention Neural Networks (Xu et al., 2020) learn the representations of subjective and objective features of texts. Motivated by (Xu et al., 2020; Glandt et al., 2021), we focus on exploring the impact of subjective and objective information from tweets on stance detection and premise classification. Figure 2 shows the architecture of our proposed system. We use BERT-based models to extract subjective and objective features from texts, then we concatenate the two [CLS] embeddings (i.e., subjective and objective features, denoted by $\boldsymbol{f}_{subj}$ and $\boldsymbol{f}_{obj}$ respectively) and use the result as input to feed into a fully connected layer, followed by a sigmoid activation to produce a fusion vector:

$$\boldsymbol{g} = sigmoid(\boldsymbol{W}_u[\boldsymbol{f}_{subj} \oplus \boldsymbol{f}_{obj}] + \boldsymbol{b}_u),$$

where $\oplus$ denotes vector concatenation and $\boldsymbol{W}_u, \boldsymbol{b}_u$ are trainable parameters (Xu et al., 2020). We aim

[2] https://pypi.org/project/emoji/

to attain an optimal combination between $\boldsymbol{f}_{subj}$ and $\boldsymbol{f}_{obj}$, so vector $\boldsymbol{g}$ serves as a weight vector as follows:

$$\boldsymbol{f}_{dual} = \boldsymbol{g} \odot \boldsymbol{f}_{subj} + (\boldsymbol{1} - \boldsymbol{g}) \odot \boldsymbol{f}_{obj},$$

where $\odot$ denotes the element-wise product (Xu et al., 2020; Glandt et al., 2021). Vector $\boldsymbol{f}_{dual}$ stands for dual-view features of texts and is used to train our classifier, which consists of 2 linear layers with 1024 hidden units respectively and a ReLU activation function. For each subtask of Task 2, the model was trained for 50 epochs with a batch size of 16, a dropout probability of 0.15, and a learning rate of 0.00003.

2.4 Cross-validation and Pseudo-label

Cross-validation is a data resampling method used to evaluate models and prevent overfitting. The basic form of cross-validation is k-fold cross-validation (Refaeilzadeh et al., 2009). Pseudo-label (Lee et al., 2013) uses semi-supervised learning to increase the training data to improve the robustness of the model, which can be regarded as a method to increase the fitting ability of the model by expanding the decision boundary. This technique trains a model by labeled data and uses the trained model to predict labels for unlabeled data, then adds these newly obtained pseudo-labeled data to the training set and retrains the model.

3 Results and Discussion

	F1 score	
	Subtask 2a	Subtask 2b
Model$_{roberta}$	0.750	0.636
Model$_{v2}$	0.801	0.747
Model$_{v2} + cv$	0.802	0.750
Model$_{v2} + cv + pl$	**0.805**	**0.754**

Table 2: Performance of different approaches on the SMM4H 2022 Task 2 **Validation Set**. *roberta* denotes RoBERTa-large pre-trained model, *v2* denotes COVID-Twitter-BERT-v2 pre-trained model, *cv* denotes cross-validation, and *pl* denotes pseudo-label.

Table 2 shows the performance of the suggested approaches on the validation set for both subtasks. We can see that COVID-Twitter-BERT-v2 based model (Müller et al., 2020) performs better than RoBERTa-large based model (Liu et al., 2019). Additionally, we applied techniques including pseudo-label and 5-folds cross-validation with a majority
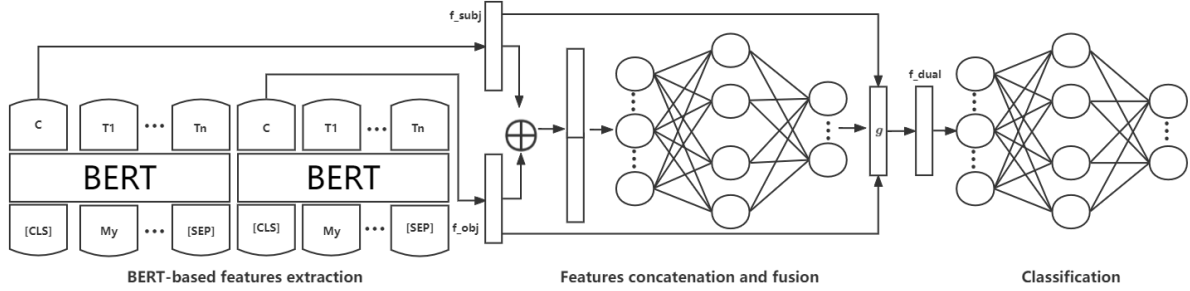
Figure 2: Architecture of our proposed system

voting strategy to our experiments, both methods proved to be useful in improving the performance of the system. We conducted an error analysis and found that, in the Stance Detection task, positive or negative words such as "support" or "don't" might lead to misclassification to some extent, especially when distinguishing NONE from the other two stances (e.g., "Schools are doing their best to ensure that education for children is not disrupted. I support my children's school and all self financed schools wholeheartedly" is predicted as FAVOR label while the truth label is NONE, we note that it is also difficult for humans to distinguish between FAVOR and NONE). As for the Premise Classification task, it also comes with challenges of how to identify a given statement that can be used as an argument in a discussion, it is also necessary to distinguish between sentiment polarity and argumentation.

In the evaluation phase, our submitted systems including $Model_{v2}$, $Model_{v2} + cv$, and $Model_{v2} + cv + pl$ ($v2$ denotes COVID-Twitter-BERT-v2 pretrained model, cv denotes 5-folds cross-validation, and pl denotes pseudo-label. As for the pseudo-label method, we selected 2000 English tweets in the test set to predict labels and added them to the training set to train our final submitted model), and the last one achieved the highest score for all of our submissions. Table 3 shows the results of the $Model_{v2} + cv + pl$ compared to the mean and median of all competing submissions.

4 Conclusion

Our proposed methods achieve an F1 score of 0.618 for Subtask 2a (the mean score is 0.491 and the median score is 0.550) and an F1 score of 0.630 for Subtask 2b (the mean score is 0.574 and the median score is 0.647) on the test set. We observe that the domain-specific pre-trained model can significantly

	F1 score	
	Subtask 2a	**Subtask 2b**
Ours	**0.618**	0.630
Median	0.550	**0.647**
Mean	0.491	0.574

Table 3: Comparison of our best performing system (i.e., $Model_{v2} + cv + pl$) with the mean and median of all competing systems in the evaluation phase.

outperform the general pre-trained model. Besides, we also notice that the system's performances on the test set are lower than on the validation set, which can be attributed to overfitting. In terms of future work, stochastic weight averaging, adversarial training, and exponential moving averaging can be investigated to improve the generalizability and robustness of our system.

Acknowledgements

We would like to thank the reviewers for their insightful suggestions.

References

Vera Davydova and Elena Tutubalina. 2022. SMM4H 2022 Task 2: Dataset for stance and premise detection in tweets about health mandates related to COVID-19. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages –.

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.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for

deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

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.

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*.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2009. Cross-validation. *Encyclopedia of database systems*, 5:532–538.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. 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*, pages –.

Chang Xu, Cécile Paris, Surya Nepal, Ross Sparks, Chong Long, and Yafang Wang. 2020. DAN: Dual-view representation learning for adapting stance classifiers to new domains. *arXiv preprint arXiv:2003.06514*.