Classification of Tweets Self-reporting Adverse Pregnancy Outcomes and Potential COVID-19 Cases Using RoBERTa Transformers

Lung-Hao Lee, Man-Chen Hung, Chien-Huan Lu,
Chang-Hao Chen, Po-Lei Lee, and Kuo-Kai Shyu
Department of Electrical Engineering, National Central University, Taiwan
Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan

Abstract

This study describes our proposed model design for SMM4H 2021 shared tasks. We fine-tune the language model of RoBERTa transformers and their connecting classifier to complete the classification tasks of tweets for adverse pregnancy outcomes (Task 4) and potential COVID-19 cases (Task 5). The evaluation metric is F1-score of the positive class for both tasks. For Task 4, our best score of 0.93 exceeded the median score of 0.925. For Task 5, our best of 0.75 exceeded the median score of 0.745.

1 Introduction

The Social Media Mining for Health Application (SMM4H) shared tasks involve natural language processing challenges using social media data for health research. We participated in the SMM4H 2021 Task 4 (Klein et al., 2020a; 2020b), focusing on automatically distinguishing tweets that self-report a personal experience of an adverse pregnancy including miscarriage, stillbirth, preterm birth, low birthweight, and neonatal intensive care (annotated as "1") from those that do not (annotated as "0"). This task is a follow-up to SMM4H 2020 Task 5, which involves three classes of tweets that mention birth defects.

We also participated in SMM4H 2021 Task 5. This new binary classification task involves automatically distinguishing tweets that self-report potential cases of COVID-19 (annotated as "1") from those that do not (annotated as "0"). Potential cases includes those tweets indicate the user or a member of the user's household was denied testing for, was symptomatic of, was directly exposed to presumptive or confirmed COVID-19 cases, or had experiences that pose a higher risk of exposure to COVID-19. Other tweets related to COVID-19 may discuss topics such as testing, symptom, traveling, or social distancing, but do not indicate someone may be infected.

This paper describes the **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical **E**ngineering, **N**atural **L**anguage **P**rocessing Lab) system for the SMM4H 2021 Task 4 and Task 5. Our solution explores how to use the RoBERTa transformers (Liu et al., 2019) with involved language models and classifier fine-tuning to predict tweet classes. The evaluation metrics of both tasks are F1-score for the positive class (i.e., tweets annotated as "1"). For Task 4, our best score of 0.93 exceeded the median score of 0.925. For Task 5, we achieved a best score of 0.75 exceeding the median score of 0.745.

The rest of this paper is organized as follows. Section 2 investigates the related studies. Section 3 describes the NCUEE-NLP system for the tweet classification tasks. Section 4 presents the evaluation results and performance comparisons. Conclusions are finally drawn in Section 5.

2 Related Work

Our participated SMM4H 2021 Task 4 is a follow-up to SMM4H 2020 Task 5, which focused on detecting tweets that mention birth defects. A hard-voting ensemble of nine BioBERT-based models was used to achieve a higher macro-averaging recall (Bai and Zhou, 2020). The ELMo word embeddings and data-specific resources were adopted to achieve a higher macro-averaging precision (Bagherzadeh and Bergler, 2020). Ensemble BERT flavors were studied to detect tweets that mention birth defects (Dima et al., 2020). Two-views based CNN-BiGRU networks

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 98–101

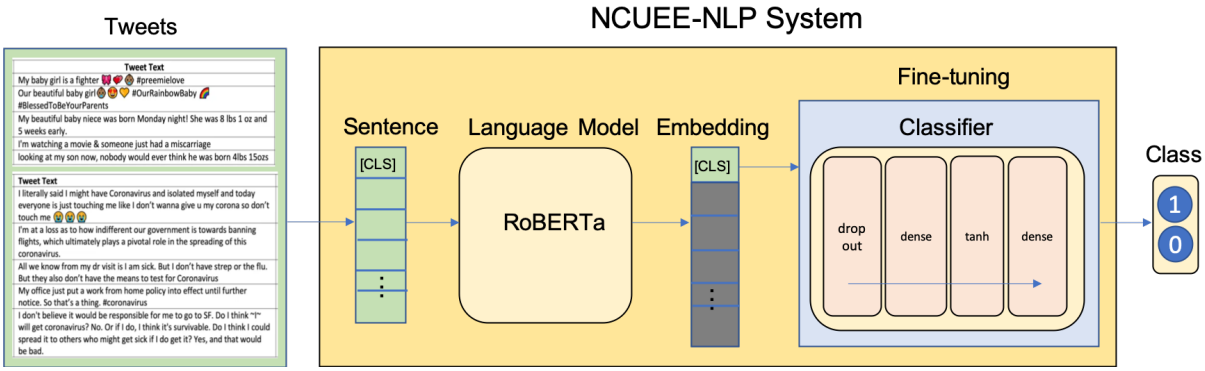

Figure 1: Our NCUEE-NLP system architecture for the SMM4H 2021 Task 4 and Task 5.

RoBERTa Transformers	Fine-Tuning		Validation Set			Test Set		
	LM	Classifier	Precision	Recall	F1-score	Precision	Recall	F1-score
	No	Yes	0.9253	0.9338	0.9296	0.9235	0.9248	0.92
	Yes	Yes	0.9141	0.9475	**0.9305**	0.9130	0.9480	**0.93**

Table 2: Submission results on the SMM4H 2021 Task 4 validation and test datasets.

RoBERTa Transformers	Fine-Tuning		Validation Set			Test Set		
	LM	Classifier	Precision	Recall	F1-score	Precision	Recall	F1-score
	No	Yes	0.7407	0.8197	0.7782	0.6750	0.7890	0.73
	Yes	Yes	0.7907	0.8361	**0.8128**	0.7452	0.7597	**0.75**

Table 2: Submission results on the SMM4H 2021 Task 5 validation and test datasets.

were also proposed to address this multi-class classification task (Reddy, 2020).

Our participated SMM4H 2021 Task 5 is new binary classification task, which aims at distinguishing tweets that self-report potential cases of COVID-19 from those that do not. COVID-19 Twitter Monitor was presented to show interactive visualizations of the analysis results on tweets related to the COVID-19 pandemic (Cornelius et al., 2020). An iterative graph-based approach was proposed to detect COVID-19 emerging symptoms using context-based twitter embeddings (Santosh et al., 2020). A large twitter dataset of COVID-19 chatter was used to identify discourse around drug mentions (Tekumalla and Banda, 2020).

3 The NCUEE-NLP System

Figure 1 shows our NCUEE-NLP system architecture for the SMM4H 2021 shared tasks. Specially, our system is composed of two main parts: RoBERTa transformers and fine-tuning. RoBERTa (a Robust optimized BERT pretraining approach) (Liu et al., 2019) is a replication study of BERT pretraining (Devlin et al., 2018) that carefully measures the impact of key parameters and training data size. We observe that RoBERTa transformers have usually performed well for many SMM4H 2020 tweet classification tasks (Klein et al., 2020c). Hence, we explore the usage of RoBERTa transformers and fine-tune the downstream tasks.

For Task 4, we use training, validation, and test datasets provided by task organizers to fine-tune the language model to improve the embedding representation. Then, the tweets with class labels in the training dataset were used to fine-tune the classifier.

For Task 5, because COVID-19 related tweets are relatively rare for fine-tuning the language model, we use the original training, validation, and test datasets from the Task 5 along with those tweets from Task 6 involving a three-class classification of COVID-19 tweets containing symptoms. To fine-tune the classifier, we only use the Task 5 training set that contains tweets with corresponding labels.

4 Evaluation

The experimental datasets were mainly provided by task organizers (Arjun et al., 2021). For Task 4, we have a total of 5,514 tweets in the training set, including 2,484 positive tweets and 3,030 negative tweets. The validation set contains 973 tweets (438 positive and 535 negative). Finally, there are a total of 10,000 tweets in the test set.

For Task 5, we have 6,465 tweets (1,026 positive and 5,439 negative) in the training set. The validation set contains 716 tweets (122 positive and 594 negative). Finally, there are 10,000 tweets in the test set. We also have a total of 16,067 tweets from Task 6 for fine-tuning the language model.

All tweets were pre-processed to convert emojis into the corresponding codes defined by the unicode consortium. The pre-trained RoBERTa-Large model was downloaded from HuggingFace (Wolf et al., 2019). The hyper-parameters used for both tasks are as follows: training batch size 64, learning rate 4e-5, and maximum sequence length 128.

Tables 1 and 2 respectively summarize the results for Tasks 4 and 5. The evaluation metric is the F1-score of the positive class for both tasks. It's obvious that we have consistent results for both tasks, with a performance boost coming from fine-tuning the language model. Our best results for both tasks slightly exceeded than the respective median scores of all submissions by 0.005.

5 Conclusions

This study describes the NCUEE-NLP system participating in SMM4H 2021 Task 4 for adverse pregnancy outcome and Task 5 for potential COVID-19 cases, including system design, implementation and evaluation. For Task 4, our best F1-score of 0.93 exceeded the median score of 0.925. For Task 5, our best F1-score of 0.73 exceeded the median score of 0.725.

Acknowledgments

This study is partially supported by the Ministry of Science and Technology, under the grant MOST 108-2218-E-008-017-MY3 and MOST 108-2634-F-008-003- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan.

References

Ari Z. Klein, and Graciela Gonzalez-Hernandez. 2020a. An annotated data set for identifying women reporting adverse pregnancy outcomes on twitter. *Data in Brief*, 32(2020): 106249. https://doi.org/10.1016/j.dib.2020.106249

Ari Z. Klein, Haitao Cai, Davy Weissenbacher, Lisa D. Levine, Graciela Gonzalez-Hernandez. 2020b. A natural language processing pipeline to advance the use of twitter data for digital epidemiology of adverse pregnancy outcomes. *Journal of Biomedical Informatics: X*, 8(2020):100076. https://doi.org/10.1016/j.yjbinx.2020.100076

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020c. Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, pages 27-36.

Arjun Magge, Ari Z. Klein, Ivan Flores, Ilsear Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinova, Eulalia Farre-Maduell, Salvador Lima Lopez, 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 application (#SMM4H) shared tasks at NAACL 2021. *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics.

George-Andrei Dima, Andrei-Marius Avram, and Dumitru-Clementin Cercel. 2020. Approaching SMM4H 2020 with ensembles of BERT flavours. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, pages 153-157.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, https://arxiv.org/abs/1810.04805

Joseph Cornelius, Tilia Ellendorff, Lenz Furrer, and Fabio Rinaldi. 2020. COVID-19 twitter monitor: aggregating and visualizing COVID-19 related trends in social media. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, pages 1-10.

Parsa Bagherzadeh, and Sabine Bergler. 2020. CLaC at SMM4H 2020: birth defect mention detection. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*.

Association for Computational Linguistics, pages 168-170.

Ramya Tekumalla, and Juan M Banda. 2020. Characterizing drug mentions in COVID-19 twitter chatter. In *Proceedings of the 1st Workshop on NLP for COVID-19* (*Part 2*). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.nlpcovid19-2.25

Roshan Santosh, H. Schwartz, Johannes Eichstaedt, Lyle Ungar, and Sharath Chandra Guntuku. 2020. Detecting emerging symptoms of COVID-19 using context-based twitter embeddings. In *Proceedings of the 1st Workshop on NLP for COVID-19* (*Part 2*). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.nlpcovid19-2.35

Saichethan Miriyala Reddy. 2020. Detecting tweets reporting birth defect pregnancy outcome using two-view CNN RNN based architecture. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, pages 125-127.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Alex- ander M. Rush. 2019. HuggingFace's transformers: state-of-the-art natural language processing. *arXiv preprint*. https://arxiv.org/abs/1910.03771

Yang Bai, and Xiaobing Zhou. 2020. Automatic detecting for health-related twitter data with BioBERT. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. pages 63-69.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2019. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint*, https://arxiv.org/abs/1907.11692