Yet@SMM4H'22: Improved BERT-based classification models with Rdrop and PolyLoss

Yan Zhuang
University Of Electronic Science And Technology Of China
delecisz@gmail.com

Yanru Zhang*
University Of Electronic Science And Technology Of China
Shenzhen Institute for Advanced Study, UESTC
yanruzhang@uestc.edu.cn

Abstract

This paper describes our approach for 11 classification tasks (Task1a, Task2a, Task2b, Task3a, Task3b, Task4, Task5, Task6, Task7, Task8 and Task9) from Social Media Mining for Health (SMM4H) 2022 Shared Tasks. We developed a classification model that incorporated rdrop to augment data and avoid overfitting, poly loss and focal loss to alleviate sample imbalance, and pseudo labels to improve model performance. The results of our submissions are over or equal to the median scores in almost all tasks. In addition, our model achieved the highest score in Task4, with a higher 7.8% and 5.3% F1-score than the median scores in Task2b and Task3a respectively.

1 Introduction

Nowadays, people are more and more willing to share their life status on social networks, and analyze and discuss their illness and medication. These texts can be useful for analyzing an individual's health, but they are rare among the vast number of social tweets and particularly difficult to collect. The Social Media Mining for Health Applications (SMM4H) 2022 workshop aims to bring together researchers interested in automatic methods for the collection, extraction, representation, analysis, and validation of social media data (e.g., Twitter, Facebook) for health informatics (Weissenbacher et al., 2022). There are total 14 tasks in SMM4H 2022 shared tasks, and we participated in all classification tasks, including Task1a: Classification of Adverse Drug Events (ADEs) in English tweets (Magge et al., 2021); Task2a & Task2b: Classification of stance and premise in tweets about health mandates (COVID-19) (Davydova and Tutubalina, 2022); Task3a & Task3b: Classification of changes in medication treatments in tweets and WebMD reviews; Task4: Classification of Tweets Self-Reporting Exact Age (in English)(Klein et al., 2022); Task5: Classification of tweets of self-reported COVID-19 symptoms in Spanish; Task6: Classification of tweets indicating self-reported COVID-19 vaccination status; Task7: Classification of self-reported intimate partner violence on twitter(Al-Garadi et al., 2022); Task8: Classification of self-reported chronic stress on Twitter(Yuan-Chi et al., 2022); Task9: Classification of social media forum posts self-reporting exact age.

To solve the above classification challenges, we developed an improved BERT-based classification model. It incorporates rdrop for data augmentation, poly loss and focal loss for balancing label distributions, and pseudo label for boosting the model performance. Our model scored above the median score in all tasks, and finally, our model achieved the highest score in Task4, with a higher 7.8% and 5.3% F1-score than the median scores in Task2b and Task3a respectively.

2 Data Analysis

Of all the 11 tasks we participated, only Task2a and Task5 are three-way classification tasks, the others are all binary-classification tasks, and all tasks use F1-score of certain categories as the evaluation metric. Each of the eleven classification tasks has a different classification objective and the data takes different forms. For example, Task2a is a sentence pair matching task, while Task3b contains several meta-information items such as 'Effectiveness', 'Satisfaction' and so on. However, all of these tasks can be transformed into either a single text classification or a sentence pair matching task, both of which can be trained and tested with the same model structure. Therefore, we focus more on designing generic models to solve as many problems as possible.

After analysing the training and test data, we found that the distribution of labels for each task was consistent between the training and valid sets, but the ratio between the categories within the task

*Corresponding author

Proceedings of the 29th International Conference on Computational Linguistic, pages 98–102
October 12–17, 2022.

Task	Task2a	Task2b	Task3b	Task4	Task8	Task9
Training	38:37:25	63:37	55:45	68:32	63:37	69:31
Validation	41:33:26	63:37	57:43	68:32	63:37	69:31
Task	**Task1a**	**Task3a**	**Task5**	**Task6**	**Task7**	
Training	93:7	91:9	60:24:16	89:11	89:11	
Validation	90:10	91:9	60:24:16	89:11	90:10	

Table 1: Label distribution in each task. Among all tasks, only task2a and task5 are three-way classification tasks, the others are all binary classification tasks. The ratios in the table are sorted in descending order by the number of samples in each category in each task.

Task	Task2a	Task2b	Task3b	Task8	Task9
BioBERT	98.73	**98.62**	86.43	80.48	94.20
pubmedBERT	97.41	98.57	85.12	81.19	93.8
DeBERTa	97.55	98.09	**88.67**	84.05	94.5
BERTweet	98.85	98.57	87.51	84.52	94.5
Best+RD+PLoss	**98.88**	98.45	87.58	**85.03**	**94.80**

Table 2: Model performance in Task2a, Task2b, Task3b, Task8 and Task9 valid sets. To simplify the model, we here have uniformly used micro F1 as the evaluation metric rather than the metrics set for each task. So there may be deviations in the values. RD denotes the models using rdrop. PLoss represents the model with Taylor expansion of focal loss as the new loss function. Best means the model with the best performance in each task. The best performance in each task is highlighted in boldface, and the second highest F1 score is highlighted by underline.

was not consistent, as shown in the Table 1. The distribution of labels in these tasks, especially in Task1a, Task3a, Task6 and Task7, is quite unbalanced, reaching 93:7 in the worst cases, which requires a method to alleviate such situation.

3 Method

The whole procedure of our system model follows the BERT-style training and inference frame. The difference between the different models lies in the process of data pre-processing, encoders and tricks.

3.1 Data Pre-Processing

Since the data format of each task is different, we convert them all into the single text and text pairs for classification. Firstly, we normalized the data by changing the split words start with @ into @USER and removing the urls, @USER strings, non-English, non-numeric, and non-punctuation characters. Then we modified the task data for inputting into the classification model. For Task2a and Task2b, we selected the 'Tweet' and 'Claim' columns as the text pairs, and took 'Stance' and 'Premise' as the label respectively. While in Task3b, we transformed the number in "Ease of Use, Effectiveness, Satisfaction" columns into text, taking "Ease of Use" as an example, from 1 to 5 were transformed into "not easy, hardly easy, easy, quite easy, and extremely easy". Then the transformed data was appended together with the columns "DRUG" before the report text, and the whole data were then put into the model for training and inference. In other tasks, we just took the text and label columns for training and testing.

3.2 Encoders

Different tasks may involve different languages. For example, Task5 aims to classify the self-reported COVID-19 symptoms in Spanish tweets while the others using English tweets. So we chose various encoders, including the ones pretrained on general corpus and domain-specific corpus, to accommodate this situation.

BERTweet(Nguyen et al., 2020) is the fisrt large-scale pretrained model for English tweets. It was trained on 850M English tweets using RoBERTa(Liu et al., 2019) pre-training procedure, and showed good performance on tweet-related NLP tasks.

BioBERT(Lee et al., 2020) is short for "Bidirectional Encoder Representations from Transformers for Biomedical Text Mining". It is a domain-specific large-scale pret-rained model, and tries to handle word distribution shift from general domain corpora to biomedical corpora.

pubmedBERT (Gu et al., 2021) is pre-trained

Task	Task1a	Task3a	Task5	Task6	Task7	Task4
BioBERT	93.00	92.88	86.40	95.59	94.94	99.52
pubmedBERT	93.00	93.13	84.89	95.11	94.01	99.01
DeBERTa	92.90	94.02	86.35	95.55	94.76	99.73
BERTIN	-	-	84.93	-	-	-
BERTweet	94.97	93.51	-	94.78	94.76	99.73
Best_performing+RD	96.00	93.51	86.35	95.77	95.13	99.57
Best_performing+PLoss	96.28	94.21	86.50	**95.78**	95.32	99.57
Best_performing+RD+PLoss	**96.61**	**94.51**	**86.54**	95.77	**95.73**	99.58
Best_performing+RD+PLoss+PL	-	-	-	-	-	**99.90**

Table 3: Model performance in Task1a, Task3a, Task5, Task6, Task7 and Task4 validation sets. To simplify the model, we also here uniformly used micro F1 as the evaluation metric rather than the metrics set for each task. RD and PL denote the models using rdrop and pseudo label respectively. PLoss represents the model with Taylor expansion of focal loss as the new loss function. Best_performing means the model used the encoder with the best performance in each task. '-' means we did not implement the model on that task. The best performance in each task is highlighted in boldface, and the second highest F1 score is highlighted by underline.

on the PubMed abstracts, including 14 million abstracts and 3.2 billion words.

DeBERTa(He et al., 2020) is short for Decoding-enhanced BERT with disentangled attention. It improves the BERT(Kenton and Toutanova, 2019) and RoBERTa by introducing disentangled attention mechanism and an enhanced mask decoder. Given the same pretraining corpus, DeBERTa shows better performance in many downstream tasks.

BERTIN (De la Rosa et al., 2022) is pre-trained on Spanish corpus following the roberta-style pretraining procedure. It uses a data-centric technique, which is called 'perplexity sampling', to train the model with roughly half the amount of the steps and one fifth of the data.

3.3 Tricks

In order to alleviate the unbalanced label distribution and overfitting, we took the following tricks to increase the model performance.

Rdrop(Wu et al., 2021) is a contrastive learning technique. It generates positive samples through putting the input in two sequential dropout layers, and computes the Kullback-Leibler (KL) divergence between the two outputs. It can be used for data augmentation and alleviating overfitting problems.

FocalLoss(Lin et al., 2017) is a new loss function which focuses more on the hard-to-classify samples during training by reducing the weights of the easy-to-classify samples. Thus, it can be used to alleviate the class imbalance through reshaping the standard cross entropy.

PolyLoss(Leng et al., 2022) approximates the loss functions via Taylor expansion, and designs the loss functions as a linear combination of polynomial functions. Thus it can be combined with focal loss and cross entropy functions.

Pseudo Label(Lee et al., 2013) can be seen as the semi-supervised learning for it predicts the labels of the unannotated samples, and re-inputs the samples with the predicted labels into training, thus can improve the model performance in some cases.

4 Experiments and Analysis

All the model we used shared a fixed training config. They were all trained for 5 epochs with learning rate $4e-5$. Besides, the max length of the input was 128, the weight decay rate was 0.01 and the Adam parameter was 1e-8. Table 2 and Table 3 show the performance of different models in each task. We implemented four different encoders to do the classification in each task and the micro F_1-score was chosen as the evaluation metric on our validation process. BioBERT, pubmedBERT, DeBERTa and BERTweet were first applied in the tasks shown in Table 2, then we modified the loss function in the best performing model using rdrop, poly loss tricks to increase the performance.

Results show that the introduction of the tricks help the model get the highest performance in Task2a, Task8 and Task9. While in Task2b, BioBERT achieved the best performance, and the tricks reduced the F_1 by 0.17. The reason for this decline is likely to be the way we processed the data. The purpose of Task2b is to predict whether at least one premise/argument is mentioned in the text, but we transformed this into a text pair match-

	Ours			Median		
Task	**Precision**	**Recall**	**F_1-Score**	**Precision**	**Recall**	**F_1-Score**
Task1a	76.5	58.4	66.2	-	-	-
Task2a	-	-	**57.1(+2.1)**	-	-	55.0
Task2b	-	-	**65.3(+7.9)**	-	-	57.4
Task3a	68.3	60.1	**63.9(+5.3)**	61.7	55.8	58.6
Task3b	85.7	85.4	**85.6(+1.3)**	84.4	86.5	84.3
Task4	93.3	90.8	**92.0(+5.1)**	86.9	88.9	86.9
Task5	85.0	85.0	**85.0(+1.0)**	84.0	84.0	84.0
Task6	90.0	74.0	**81.0(+4.0)**	90.0	68.0	77.0
Task7	90.3	70.5	**79.1(+2.8)**	79.0	71.6	76.3
Task8	79.7	76.9	**78.3(+3.3)**	72.0	76.0	75.0
Task9	95.7	91.9	**93.8(+4.7)**	89.6	91.9	89.1

Table 4: Our model performance and median scores in all tasks we participated. All results and evaluation metrics were provided by the official. '-' denotes officials did not provide that figure. The highest F_1-score in each task is highlighted in boldface. The numbers in brackets indicate how much higher our model's results are than the median score.

ing problem where the claim part may have been introduced as noise, and the tricks increased the noise.

Besides, without tricks, DeBERTa performed the best in Task3b, but with tricks, the scores dropped by 1.09 to 87.58. It may imply that there is a problem with the way we changed meta-information into text information and concatenate it.

Except for Task4, the label distribution of other tasks in Table 3 is very unbalanced. After the introduction of poly loss, the performance of all models has been improved, just as Table 3 shows. In addition, with the addition of rdrop, the model's performance is further improved, which help the model reach the highest F_1 score in Task1a and, Task3a, Task5 and Task7. Besides, the best score of Task1a and Task4 were 0.63 and 0.914 respectively (Magge et al., 2021; Klein et al., 2022), and our model achieved the state-of-the-art performance. In Task6, however, the best encoder using poly loss performed the best, with the additional rdrop dropping slightly.

However, in Task4, these tricks did not seem to improve the performance of the model, but decreased the score a lot, which may be caused by the relatively balanced label distribution of the task. So we took the pseudo label trick to augment the training data.

Specifically, we first used the three best-performing models to predict the samples in the validation set (here is BERTweet, DeBERTa and BioBERT, and we use the test set in the testing phase), and then processed the obtained three logits in the following way: if the two logits are greater than 0.8, it means that the sample is likely to belong to a certain category, and we would add the sample and the predicted label to the training set as a new training sample. Based on the augmented training dataset, we retrained BERTweet encoder with rdrop and poly loss, and got the highest score, which helped us win the first place in this task.

We submitted the highest scoring models incorporating tricks at each task, and the final scores are shown in the Table 4. The official did not provide the median score but provide the mean score of the Task1a, of which the Precison, Recall and F_1-Score are 64.6, 49.7 and 56.2 respectively. It can be seen that in all the tasks we participated in, our results exceeded the median scores. In addition, our model achieved the highest score in Task4, with a higher 7.8% and 5.3% F1-score than the median scores in Task2b and Task3a respectively.

5 Conclusion

In this work, we developed a classification model based on the BERT. It incorporated with rdrop to do the data augmentation, with poly loss to mitigate the label imbalance, and with pseudo label to boost the model performance. We participated 11 classification tasks in SMM4H 2022 shared tasks, and our model scored above the median score in all tasks. Finally, our model achieved the highest score in Task4, with a higher 7.8% and 5.3% F1-score than the median scores in Task2b and Task3a respectively.

References

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, page 100217.

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 –.

Javier De la Rosa, Eduardo G Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ari Z Klein, Arjun Magge, and Graciela Gonzalez-Hernandez. 2022. Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets. *PloS one*, 17(1):e0262087.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Xiaojie Shi, Shuyang Cheng, and Dragomir Anguelov. 2022. Polyloss: A polynomial expansion perspective of classification loss functions. *arXiv preprint arXiv:2204.12511*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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*.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Deepademiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

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 –.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Yang Yuan-Chi, Xie Angel, Kim Sangmi, Hair Jessica, Al-Garadi Mohammed Ali, and Sarker Abeed. 2022. Automatic detection of twitter users who express chronic stress experiences via supervised machine learning and natural language processing. *Computers, Informatics, Nursing*, pages –.