LT4SG@SMM4H'24: Tweets Classification for Digital Epidemiology of Childhood Health Outcomes Using Pre-Trained Language Models

Dasun Athukoralage
NirvanaClouds
dasun@nirvanaclouds.com

Thushari Atapattu
University of Adelaide
thushari.atapattu@adelaide.edu.au

Menasha Thilakaratne
University of Adelaide
menasha.thilakaratne@adelaide.edu.au

Katrina Falkner
University of Adelaide
katrina.falkner@adelaide.edu.au

Abstract

This paper presents our approaches for the SMM4H'24 Shared Task 5 on the binary classification of English tweets reporting children's medical disorders. Our first approach involves fine-tuning a single RoBERTa-large model, while the second approach entails ensembling the results of three fine-tuned BERTweet-large models. We demonstrate that although both approaches exhibit identical performance on validation data, the BERTweet-large ensemble excels on test data. Our best-performing system achieves an F1-score of 0.938 on test data, outperforming the benchmark classifier by 1.18%. Our code is available on Github[1].

1 Introduction & Motivation

Chronic childhood disorders like attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech, and asthma significantly impact a child's development and well-being, often extending into adulthood. Approximately 1 in 6 (17%) children aged 3-17 years in the United States experience a developmental disability, with ADHD, ASD, and others contributing to this statistic (Zablotsky et al., 2019). In previous studies (Guntuku et al., 2019; Hswen et al., 2019; Edo-Osagie et al., 2019), Twitter data have been utilized to identify self-reports of the aforementioned disorders; however, the identification of reports concerning these disorders in users' children has not been explored. It may be of interest to explore Twitter's potential in continuing to collect users' tweets postpartum, enabling the detection of outcomes in childhood.

2 Task and Data Description

2.1 Task

The SMM4H-2024 workshop and shared tasks have a special focus on Large Language Models (LLMs) and generalizability for natural language processing (NLP) in social media. We participated in Task 5, which is 'Binary classification of English tweets reporting children's medical disorders'. The objective is to automatically differentiate tweets from users who have disclosed their pregnancy on Twitter and mention having a child with ADHD, ASD, delayed speech, or asthma (annotated as "1"), from tweets that merely refer to a disorder (annotated as "0").

2.2 Data

There were three different datasets provided: training, validation, and test datasets. The training and validation datasets were labeled while the test dataset was not. All datasets are composed entirely of tweets posted by users who had reported their pregnancy on Twitter, that report having a child with a disorder and tweets that merely mention a disorder. The training, validation, and test sets contain 7398 tweets, 389 tweets, and 1947 tweets, respectively.

3 Methodology

3.1 Baseline

A benchmark classifier, based on a RoBERTa-large model (Liu et al., 2019), has achieved an F1-score of 0.927 for the 'positive' class (i.e., tweets that report having a child with a disorder) on the test data for Task 5 (Klein et al., 2024).

3.2 Models Used

We investigated three Transformer based models which are BioLinkBERT-large (Yasunaga et al., 2022), RoBERTa-large and BERTweet-large (Nguyen et al., 2020). BioLinkBERT was selected for its specialized understanding of biomedical NLP tasks, RoBERTa for its domain-independent NLP capabilities, and BERTweet for its superior performance in Tweet-specific NLP tasks. We fine-

[1] To access the code, please visit: [GitHub Source Code]

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 38–41
August 15, 2024 ©2024 Association for Computational Linguistics

tuned each model with the training dataset and evaluated its performance using the validation dataset.

3.3 Training Regime

Experiments were conducted using Google Colab Pro+ equipped with an NVIDIA A100 Tensor Core GPU boasting 40 gigabytes of available GPU RAM. The Hugging Face Transformers Python library (Wolf et al., 2019) and its Trainer API facilitated training procedures. Each model was trained on the training datasets for 3 iterations and 10 epochs per iteration. We used HuggingFace's Trainer Class's default 'AdamW' and 'linear warmup with cosine decay' as the optimizer and scheduler respectively. The maximum sequence length for all models was set to 512. FP-16 mixed precision training was employed to enable larger batch sizes and expedited training. Primary hyperparameters including learning rate, weight decay, and batch size were determined as described in the subsequent section.

3.4 Hyperparameter Optimization

We conducted hyperparameter optimization that relied on HuggingFace's Trainer API with the Ray Tune backend (Liaw et al., 2018). We utilized Ray Tune's built-in "BasicVariantGenerator" algorithm[2] for hyperparameter search, paired with the First-In-First-Out (FIFO) scheduler. Since BasicVariantGenerator has the ability to dynamically generate hyperparameter configurations based on predefined search algorithms (e.g., random search, Bayesian optimization), it enables more efficient exploration of the search space. The hyperparameters optimized using BasicVariantGenerator are presented in Table 1.

Model	Learning Rate	Weight Decay	Batch Size
BioLinkBERT-large	6.10552e-06	0.00762736	16
RoBERTa-large	7.21422e-06	0.00694763	8
BERTweet-large	1.17754e-05	0.01976150	8

Table 1: Hyperparameters optimized via BasicVariantGenerator.

4 Preliminary Experiments

Each selected model was trained for 3 iterations, with 10 epochs per iteration. At the end of each epoch, its F1-score was recorded. The F1-score for each model was determined based on its performance with the validation dataset. We saved the best-performing epoch (i.e., the best F1-score for the positive class) for each model in each iteration. The results are shown in Table 2.

Model	1st run	2nd run	3rd run	Mean F1	SD
BioLinkBERT-large	0.855019	**0.875969**	0.863159	0.864716	0.010561
RoBERTa-large	0.931408	**0.945055**	0.931408	0.935957	0.007879
BERTweet-large	**0.940741**	0.934307	0.933824	**0.936291**	**0.003862**

Table 2: The F1 scores of the BioLinkBERT-large, RoBERTa-large, and BERTweet-large classifiers on the validation data. The mean F1 score and standard deviation are also provided.

As shown in Table 2, RoBERTa-large and BERTweet-large perform similarly on the validation dataset, and considerably better than BioLinkBERT-large, even though it has been pretrained on a large corpus of biomedical data. Therefore, we decided to remove BioLinkBERT-large to carry out further experiments for this task (Guo et al., 2020).

4.1 Ensembling Strategy

The issue arises when fine-tuning large-transformer models on small datasets: the classification performance varies significantly with slightly different training data and random seed values, even when using the same hyperparameter values (Dodge et al., 2020). To overcome this high variance and provide more robust predictions, we propose ensembles of multiple fine-tuned RoBERTa-large models and BERTweet-large models separately. We create two separate ensemble models using the best models corresponding to three iterations for each RoBERTa-large and BERTweet-large. All three iterations use the same hyperparameters, and only differ in the initial random seed. A hard majority voting mechanism combines the predictions of these models:

$$\hat{y} = \arg\max_{c} \sum_{i=1}^{n} \mathbf{1}(\hat{y}_i = c) \tag{1}$$

where $\mathbf{1}(\cdot)$ represents the indicator function, which returns either '1' or '0' for the class label c predicted by the i-th classifier.

Classifier	F1-score	Precision	Recall
RoBERTa-large Ensemble	0.934783	0.914894	**0.955556**
BERTweet-large Ensemble	**0.945055**	**0.934783**	**0.955556**

Table 3: Performance results for ensemble classifiers on validation data.

As shown in Table 3, the BERTweet-large ensemble performs better than the RoBERTa-large

[2] https://docs.ray.io/en/latest/tune/api/doc/ray.tune.search.basic_variant.BasicVariantGenerator.html

ensemble. This is fundamentally because its performance variation is less for three iterations, as indicated in Table 2. Another noteworthy observation is that the performance of the BERTweet-large ensemble is identical to that of the best iteration (Table 2, 2nd run) of RoBERTa-large as shown in Table 4. Figure 1 shows the corresponding confusion matrices for both classifiers which are also identical.

Classifier	F1-score	Precision	Recall
RoBERTa-large best-run	**0.945055**	**0.934783**	**0.955556**
BERTweet-large Ensemble	**0.945055**	**0.934783**	**0.955556**

Table 4: Performance comparison of the RoBERTa-large best-run vs the BERTweet-large ensemble on validation data.

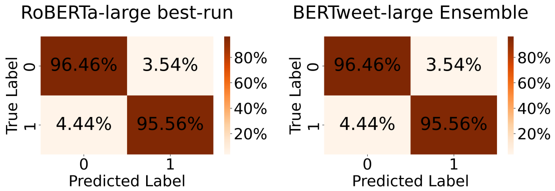

Figure 1: Confusion matrices of the RoBERTa-large best-run and BERTweet-large ensemble on the validation dataset.

5 Results and Conclusion

Since RoBERTa-large best-run and BERTweet-large ensemble are performing equally well on the validation data, we tested the performance of both classifiers on unseen, unlabeled test data. As shown in Table 5, the BERTweet-large ensemble classifier outperforms the mean and median performance on the test data among all teams' submissions by a considerable margin, as well as the benchmark classifier by 1.18%. Additionally, we can observe that even though both classifiers perform equally well on validation data, the BERTweet-large ensemble model performs significantly better on test data. One possible reason for this is that different runs of BERTweet-large might excel at capturing different aspects of the data or learning different patterns.

Previously, authors (Klein et al., 2024) have achieved an F1-score of 0.92 using a BERTweet-large classifier on the same dataset. We believe that our approach achieved better results for several fundamental reasons. First, we performed more thorough hyperparameter tuning compared to the previous authors. Better-optimized hyperparameters can significantly improve model performance. Second, by creating an ensemble of BERTweet-large models from the three best epochs of three runs, we captured more robust and generalized patterns in the data. Ensemble methods typically outperform individual models because they reduce variance and mitigate the risk of overfitting to specific subsets of data. Third, a likely reason is the diversity in model runs. Each run of our BERTweet-large model may have encountered slightly different initialization and training dynamics (e.g., random seed), resulting in diverse decision boundaries. Combining these diverse models helps in making more accurate predictions by leveraging the strengths of each individual model.

Model	F1-score	Precision	Recall
Baseline	0.927	0.923	0.940
Mean	0.822	0.818	0.838
Median	0.901	0.885	0.917
RoBERTa-large best-run	0.925	0.908	0.942
BERTweet-large Ensemble	**0.938**	**0.930**	**0.946**

Table 5: Results for our two proposed approaches on the test data, including the mean, median, and baseline scores.

When fine-tuning complex pre-trained language models, one issue on small datasets is the instability of the classification performance. To overcome this, we combined the predictions of multiple BERTweet-large models in an ensemble. By doing so, we achieved significantly better results in terms of F1-score for SMM4H'24 Task 5. For future work, it's interesting to investigate how the system performance varies when adding more BERTweet-large iterations (i.e., runs) to the ensemble.

References

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Oduwa Edo-Osagie, Gillian Smith, Iain Lake, Obaghe Edeghere, and Beatriz De La Iglesia. 2019. Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. *PLoS One*, 14(7):e0210689.

Sharath Chandra Guntuku, Jared R. Ramsay, Raina M. Merchant, and Lyle H. Ungar. 2019. Language of adhd in adults on social media. *Journal of Attention Disorders*, 23(12):1475–1485.

Yuting Guo, Xiangjue Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cécile Paris, and Diego Mollá-Aliod. 2020. Benchmarking of transformer-based pre-trained models on social media text classification datasets. In *Workshop of the Australasian Language Technology Association*, pages 86–91.

Yulin Hswen, Ashwin Gopaluni, John S. Brownstein, and Jared B. Hawkins. 2019. Using twitter to detect psychological characteristics of self-identified persons with autism spectrum disorder: A feasibility study. *JMIR mHealth and uHealth*, 7(2):e12264.

Ari Klein, José Gutiérrez Gómez, Lisa Levine, and Graciela Gonzalez-Hernandez. 2024. Using longitudinal twitter data for digital epidemiology of childhood health outcomes: An annotated data set and deep neural network classifiers. *J Med Internet Res*, 26:e50652.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. 2018. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv preprint arXiv:1807.05118*.

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*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Makoto Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016.

Benjamin Zablotsky, Lindsey I. Black, Matthew J. Maenner, Laura A. Schieve, Melanie L. Danielson, Rebecca H. Bitsko, et al. 2019. Prevalence and trends of developmental disabilities among children in the united states: 2009-2017. *Pediatrics*, 144(4):e20190811.