# DILAB at #SMM4H 2024: RoBERTa Ensemble for Identifying Children's Medical Disorders in English Tweets

**Azmine Toushik Wasi**
Shahjalal University of Science and Technology
Sylhet, Bangladesh
azmine32@student.sust.edu

**Sheikh Ayatur Rahman**
BRAC University
Dhaka, Bangladesh
sheikh.ayatur.rahman@g.bracu.ac.bd

## Abstract

This paper details our system developed for the 9th Social Media Mining for Health Research and Applications Workshop (SMM4H 2024), addressing Task 5 focused on binary classification of English tweets reporting children's medical disorders. Our objective was to enhance the detection of tweets related to children's medical issues. To do this, we use various pre-trained language models, like RoBERTa and BERT. We fine-tuned these models on the task-specific dataset, adjusting model layers and hyperparameters in an attempt to optimize performance. As we observe unstable fluctuations in performance metrics during training, we implement an ensemble approach that combines predictions from different learning epochs. Our model achieves promising results, with the best-performing configuration achieving F1 score of 93.8% on the validation set and 89.8% on the test set.

## 1 Introduction

Health informatics research often involves analyzing social media data from platforms like Twitter, Facebook, and Reddit to understand public sentiment on health-related topics. The 9th edition of the Social Media Mining for Health Research and Applications Workshop (SMM4H 2024) (Xu et al., 2024) is dedicated to advancing this area. Researchers at SMM4H 2024 aim to contribute by exploring topics like deriving health trends from social media, classifying health-related messages, identifying health-related or medical terms and monitoring diseases using social media content.

We decided to engage in Task 5 of this workshop, focusing on the binary classification of English tweets reporting children's medical disorders. This task aims to refine methods for accurately identifying and categorizing tweets related to pediatric health conditions. It aligns with broader efforts in health informatics research to understand public sentiment on pediatric health topics through social media analysis, inspired by Klein et al.'s work on using longitudinal Twitter data for digital epidemiology of childhood health outcomes.

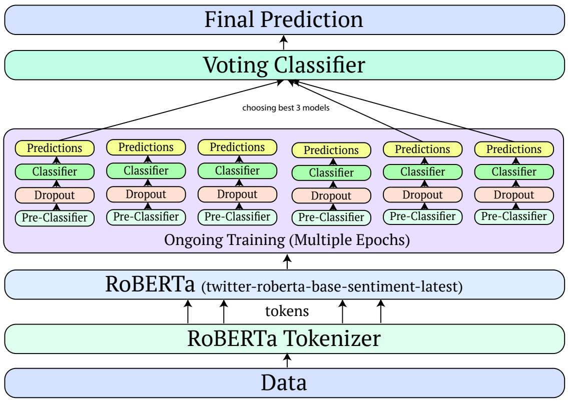

Figure 1: Model architecture, containing tokenizer, pre-trained RoBERTa, classifier and other components

## 2 System Description

### 2.1 Problem and Dataset

This binary classification task involves automatically distinguishing tweets indicating a user reported pregnancy on Twitter and subsequently mentioned having a child with specific disorders (ADHD, ASD, delayed speech, or asthma) labeled as "1", from tweets that simply mention a disorder (labeled as "0"). This task enables large-scale use of Twitter for epidemiologic studies and to understand parents' experiences with targeted support interventions (Klein et al., 2024). The dataset includes 7398 training tweets, 389 validation tweets, and 1947 test tweets.

### 2.2 RoBERTa Model

Our work uses a pre-trained RoBERTa model, namely "*cardiffnlp/twitter-roberta-base-sentiment-latest*" (R-TRBSL, in short) from HuggingFace[1], originally developed by Loureiro et al.

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

*Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks*, pages 10–12
August 15, 2024 ©2024 Association for Computational Linguistics

| Model | MLP | Dropout | T F1 | T Precision | T Recall | V F1 | V Precision | V Recall |
|---|---|---|---|---|---|---|---|---|
| BERT-base | 786, 786 | 0.3 | 88.92 | 87.32 | 89.12 | 86.67 | 82.55 | 89.31 |
| BERT-base | 786, 786 | 0.5 | 89.42 | 87.84 | 89.58 | 86.97 | 83.12 | 89.64 |
| RoBERTa-base | 786, 786 | 0.3 | 92.42 | 92.72 | 93.32 | 90.34 | 86.53 | 92.42 |
| RoBERTa-base | 786, 786, 512 | 0.3 3 | 92.22 | 92.27 | 92.78 | 90.02 | 87.46 | 92.24 |
| RoBERTa-base | 786, 786 | 0.5 | 92.12 | 92.22 | **93.46** | 90.13 | 87.55 | 92.23 |
| RoBERTa-base | 786, 786, 512 | 0.5 | 92.36 | 92.24 | 93.02 | 91.24 | 88.32 | 91.72 |
| RoBERTa-large | 786, 786 | 0.3 | 91.23 | 91.23 | 91.36 | 91.44 | 91.43 | 91.34 |
| RoBERTa-large | 786, 786 | 0.5 | 91.34 | 91.55 | 92.33 | 91.34 | 88.34 | 92.26 |
| R-TRBSL | 786, 786 | 0.3 | **92.81** | **92.89** | 92.69 | 93.42 | 93.43 | 93.27 |
| R-TRBSL | 786, 786 | 0.5 | 92.23 | 92.34 | 92.33 | 92.82 | 92.76 | 92.81 |
| R-TRBSL Top3 En.$^t$ | 786, 786 | 0.3 | 92.75 | 93.06 | 92.71 6 | 93.21 | 93.26 | 93.16 |
| R-TRBSL Best S.$^t$ | 786, 786 | 0.3 | 92.93 | 93.03 | 92.83 | **93.52** | **93.41** | **93.31** |

Table 1: Evaluation results by our models on the training and validation set on different setups.
$^t$Test set submissions, T=Train, V=Validation, E=Epoch, En.=Ensemble, S.=Single

| Models Submitted | F1 | Precision | Recall |
|---|---|---|---|
| 1. R-TRBSL Top3 En. | 0.898 | 0.883 | 0.914 |
| 2. R-TRBSL Best S. | 0.892 | 0.866 | **0.921** |
| Task Mean | 0.822 | 0.818 | 0.838 |
| Task Median | **0.901** | **0.885** | 0.917 |

Table 2: Evaluation results by our models on the test set, together with the mean and median results of the task.

(TimeLMs) and fine-tuned by Camacho-collados et al. (TweetNLP) on TweetEval dataset, developed by Barbieri et al. (2020). We also use the original tokenizer (*RobertaTokenizer* class) used to get embeddings for our model, from HuggingFace transformers library (Wolf et al., 2020).

### 2.3 Implementation Details

The *RobertaTokenizer* uses a maximum length of 256 tokens with lowercase text processing. Both the pre-classifier and classifier are MLPs with 768 nodes, employing a Sigmoid activation function and ReLU between layers, with 0.3 dropout. Training batch size is 8, validation, and test batch sizes are 4. Learning rate is $1e-05$, using Binary Cross Entropy loss and Adam optimizer (Kingma and Ba, 2017). All the random seeds used are 101. In R-TRBSL, we train the model for 10 epochs and combines predictions from the best 3 epochs using a voting classifier as shown in Figure 1. Other models are also trained for 10 epochs with early stopping if over-fitted. No additional data is used.

## 3 Evaluation

In Table 1, we can see both train and test results of different model setups. Among BERT (Devlin et al., 2018), RoBERTa-base (Liu et al., 2019), RoBERTa-large (Conneau et al., 2019) and R-TRBSL (Camacho-collados et al., 2022), R-TRBSL works the best. MLP setup with (786,786) consistently performs the best, and a dropout probability of 0.3 generally outperforms 0.5 in most cases. Overall, the best single model score is 93.52 on F1-Macro score, with 92.82% accuracy in validation set. While Accuracy and F1 scores remain consistent, precision and recall exhibit variability. Certain models excel in precision but show lower recall, while others demonstrate strong recall but lower precision. Despite these variations, the combined Macro-F1 score remains stable across different configurations. Also, R-TRBSL performs better than other models, probably because it is previously trained on a twitter sentiment dataset.

The results on the test set are outlined in Table 2. The two models we sent for evaluation, the single best model (R-TRBSL Best S.+ MLP (786, 786) + Dropout p=0.3, 2 Epochs) and the ensemble of top 3 epochs obtained almost similar results, with a 89.8% F1 for the former and 89.2% for the later. This puts our solution 6.4% F1 above the mean task score. Though the ensemble model performs weakly in validation set than single model (see Table 1, 93.82 vs 93.16 on F1 score), it works better in the test data (89.8 vs 89.2); showing our approach to keep the model stable worked.

## 4 Conclusion

Studying social media data continues to be vital in health informatics research, providing valuable insights into public sentiment on health-related topics. Using pre-trained language models like RoBERTa and BERT, we obtain text embeddings and fine-tuned them using MLP classifiers on the task dataset with strategic adjustments to model layers and hyperparameters, enhancing performance. Additionally, we applied an ensemble approach on epochs to ensure performance stability. Our top models achieved 0.9316 and 0.9382 F1-macro on validation and 0.898 and 0.892 on the test set, demonstrating its effectiveness.

## References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Ari Z Klein, José Agustín Gutiérrez Gómez, Lisa D 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. *Journal of Medical Internet Research*, 26:e50652.

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.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Sai Tharuni Samineni, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of the 9th Social Media Mining for Health Applications Workshop and Shared Task*, Bangkok, Thailand. Association for Computational Linguistics.