PheonixTrio918 at SMM4H 2024: 5 Fold Cross Validation for Classification of tweets reporting children's disorders

B Rahul Naik
Indian Institute of Technology Jodhpur
naik.9@iitj.ac.in

Oppangi Poojita
Indian Institute of Technology Jodhpur
poojita.1@iitj.ac.in

PothiReddy Kovidh Reddy
Indian Institute of Technology Jodhpur
reddy.19@iitj.ac.in

Lipika Dey
Ashoka University
lipika.dey@ashoka.edu.in

Abstract

This document describes our system used for the Social Media Mining for Health (SMM4H) 2024 Task 05. The objective of this task was to perform binary classification on the tweets provided in the dataset. The dataset contained two categories of tweets: those reporting medical disorders and those merely mentioning the disease. We tackled this problem using a 5-fold cross-validation approach. Our method utilizes the RoBERTa-Large model with 5-fold cross-validation. The evaluation results yielded an F1-score of 0.886 on the validation dataset and 0.823 on the test dataset.

1 Introduction

Recent advancements in Natural Language Processing have been groundbreaking, simplifying classification systems while making them more resilient to complex challenges. Integrating real-world data with NLP techniques has enhanced system efficiency, and for practical tasks such as Intent Recognition, Sentiment Analysis, or Sentence Classification, the RoBERTa model (Liu et al., 2019) and its variants like RoBERTa-small and RoBERTa-large have demonstrated outstanding performance. These models' comprehension mechanisms have consistently been the preferred choice for these applications.

Social Media platforms, such as Twitter, consistently produce vast quantities of data in various formats, including text. Extensively utilized for sharing and disseminating users' opinions on numerous topics, Twitter data has led to initiatives like the Social Media Mining for Health Application (SMM4H) Shared Task in 2024 (Xu et al., 2024). This paper outlines our approach for shared task 05: Binary classification of tweets that report children's medical disorders (in English).

2 Methodology

2.1 Data

We utilized the dataset made available by the organizers of Task 05. In this task, we were provided with two categories of tweets. Tweets that merely mention Child Disorder are labeled as 0, while tweets that aim to report the disorders are labeled as 1. The dataset is divided into three sections. The training set encompasses 7,398 tweets, with 5,118 tweets labeled as 0 and 2,280 tweets labeled as 1. The validation set includes 389 tweets, with 254 tweets labeled as 0 and 135 tweets labeled as 1. The test set consists of approximately 10,000 tweets without labels.

Label	Training	Validation
0	5118 (69.2%)	254 (65.3%)
1	2280 (30.8%)	135 (34.7%)

Table 1: Data distribution across Training, Validation sets for each label.

2.2 Preprocessing

Neattext is a simple NLP package designed for cleaning textual data and preprocessing text by efficiently removing various types of unnecessary text elements, such as usernames, hashtags, links, emojis, and dates.[1] In this dataset, Neattext played a crucial role in enhancing the preprocessing phase by using its library of functions to systematically remove these elements, thereby enriching the model's understanding. Initially, we utilized Neattext's functions to remove user handles, hashtags, emojis, and URLs, ensuring a basic level of preprocessing. However, we observed that Neattext's function for removing emojis was not exhaustive. Consequently, we created a custom Python function to

[1] https://pypi.org/project/neattext/

ensure the removal of all types of emojis, further refining the preprocessing.

The detailed preprocessing steps, including the removal of user handles, hashtags, emojis, and URLs, significantly improved the model's performance. By eliminating these extraneous elements, the model focused more on the core textual content, which enhanced its ability to detect relevant patterns and reduced noise. This preprocessing resulted in better feature extraction and improved the accuracy and robustness of the model's predictions.

After preprocessing, we merged the training and validation datasets to augment the data available to the model. This strategy aimed to enhance its ability to generalize and mitigate overfitting. The combined dataset totaled 7,787 tweets, with 5,372 labeled as 0 and 2,415 as 1.

2.3 Resampling

We noticed a clear imbalance in the dataset before and after concatenating, where tweets labeled as mere mentions labeled as 0 outnumbered those labeled with medical disorders labeled as 1. To address this, we employed the RandomUnderSampler technique (Lemaitre et al., 2017) to balance the dataset by reducing the number of examples in the Majority class. This approach equalized both classes to 2,415 samples each, resulting in a total dataset of 4,830 samples. This resampling technique helps mitigate the bias in the dataset and ensures that the model is trained more effectively on both classes, improving its ability to generalize across different categories.

	Label 0	**Label 1**
Before Resampling	5372	2415
After Resampling	2415	2415

Table 2: Data distribution of labels before and after applying the resampling.

2.4 Model

RoBERTa(Liu et al., 2019), an optimized variant of BERT, demonstrates superior performance in natural language understanding tasks. It utilizes larger training datasets, longer training times, and dynamic masking patterns during pretraining to effectively capture intricate linguistic nuances. RoBERTa was chosen for its proven efficacy in various NLP benchmarks, consistently outperforming baseline models like BERT due to refined training methodologies and architectural enhancements.

2.5 Cross validation

Cross-validation is a data resampling technique employed to evaluate models and is computationally efficient. The fundamental approach to cross-validation is k-fold cross-validation (Brownlee, 2023). In this study, we implemented 5-fold cross-validation to robustly assess the model's performance across different subsets of the dataset.

2.6 Model Parameters

In implementing the 5-fold cross-validation method, we trained our model using parameters set as follows: 15 epochs, a batch size of 6, and a maximum tokenization length of 280 characters. The criterion employed was Cross Entropy Loss, with optimization performed using the Adam Optimizer at a learning rate of 1.5e-8.

3 Results

After applying the model parameters and conducting 5-fold cross-validation on both the validation and test datasets, we achieved an f1 score of 0.886 with precision of 0.831 and recall of 0.948 on the validation dataset. On the test dataset, the model yielded an f1 score of 0.823, with precision and recall values of 0.721 and 0.959, respectively.

	F1 Score	**Precision**	**Recall**
Valid dataset	0.886	0.831	0.948
Test dataset	0.823	0.721	0.959

Table 3: Results of the model on Validation and Test Datasets.

4 Discussion

Despite achieving an average f1 score of 0.823, our methodology falls short of matching the baseline model performance. Several factors may contribute to this disparity. One significant factor is the suboptimal choice of the learning rate. In our methodology, the learning rate was not effectively optimized, potentially impeding the model's ability to converge to the global minimum of the loss function efficiently. Consequently, the model may have struggled to achieve optimal parameter settings.Furthermore, the decision not to implement text conversion to lowercase could have limited the standardization of the model's vocabulary and text, potentially hindering the training process. Additionally, not setting the tokenization length to

the maximum of 512 tokens, as recommended for RoBERTa, may have reduced the model's ability to capture the complete contextual meaning of tweets, thus possibly leading to suboptimal performance.Our methodology consistently achieves an impressive recall score across the training, validation, and test datasets. This can be attributed to the RoBERTa model's dynamic masking, which adeptly captures intricate linguistic nuances. Consequently, while our model excels in identifying relevant cases, it also tends to exhibit higher false positive rates, thereby impacting precision.

Acknowledgments

We would like to thank our supervisor, Lipika Dey, for her invaluable guidance, support and mentorship throughout the competition. In addition, we extend our gratitude to the reviewers for their insightful suggestions.

References

Jason Brownlee. 2023. A gentle introduction to k-fold cross-validation. *Machine Learning Mastery*.

Guillaume Lemaitre, Fernando Nogueira, and Christos K. Aridas. 2017. imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

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.

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.