PheonixTrio918 @ SMM4H2024: 5-Fold Cross Validation for Classification of Tweets reporting Children's Disorders

B Rahul Naik and **Oppangi Poojitha** and **PothiReddy Kovidh Reddy**
Indian Institute of Technology Jodhpur
{naik.9, poojita.1, reddy.19} @iitj.ac.in

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

Abstract

Our system for the SMM4H 2024 Task 05 performs binary classification on tweets to identify those reporting medical disorders versus those merely mentioning the disease. We employed a 5-fold cross-validation approach using the RoBERTa-Large model. The evaluation results yielded an F1 score of 0.886 on the validation dataset and 0.823 on the test dataset.

Introduction

Advancements in NLP have simplified classification systems and enhanced their resilience to complex challenges. Integrating real-world data with NLP techniques has improved system efficiency. Models like RoBERTa and its variants excel in tasks such as Intent Recognition, Sentiment Analysis, and Sentence Classification, making them the preferred choice for these applications. Social media platforms like Twitter generate vast amounts of data, which is extensively used to share user opinions. This data has inspired initiatives like the SMM4H 2024 Shared Task. Our paper presents an approach for Task 05: binary classification of tweets reporting children's medical disorders (in English).

Data

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.

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

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.

- **Removal of User Handles**: Neattext was initially used to remove user handles (@usernames) from the tweets.
- **Removal of Hashtags**: Neattext was used to remove hashtags (like #happy, #sad) from the tweets.
- **Removal of Emojis**: Initially, Neattext's function for removing emojis was used, but it was found to be incomplete. A custom Python function was then created to ensure comprehensive removal of all types of emojis.
- **Removal of URLs**: Neattext was used to remove URLs (web links) from the tweets.
- **Combining Training and Validation Datasets**: After preprocessing, the training and validation datasets were merged to increase the amount of data available to the model.

These preprocessing steps aimed to clean and standardize the textual data, removing unnecessary elements like usernames, hashtags, emojis, and URLs. This cleaning process helped improve the model's performance by allowing it to focus more effectively on the core textual content of the tweets.

Resampling

Random Under Sampling is a technique used to balance a dataset by reducing the number of examples in the majority class. This approach equalizes both classes, helping to mitigate bias in the dataset. It ensures that the model is trained more effectively on both classes, improving its ability to generalizability.

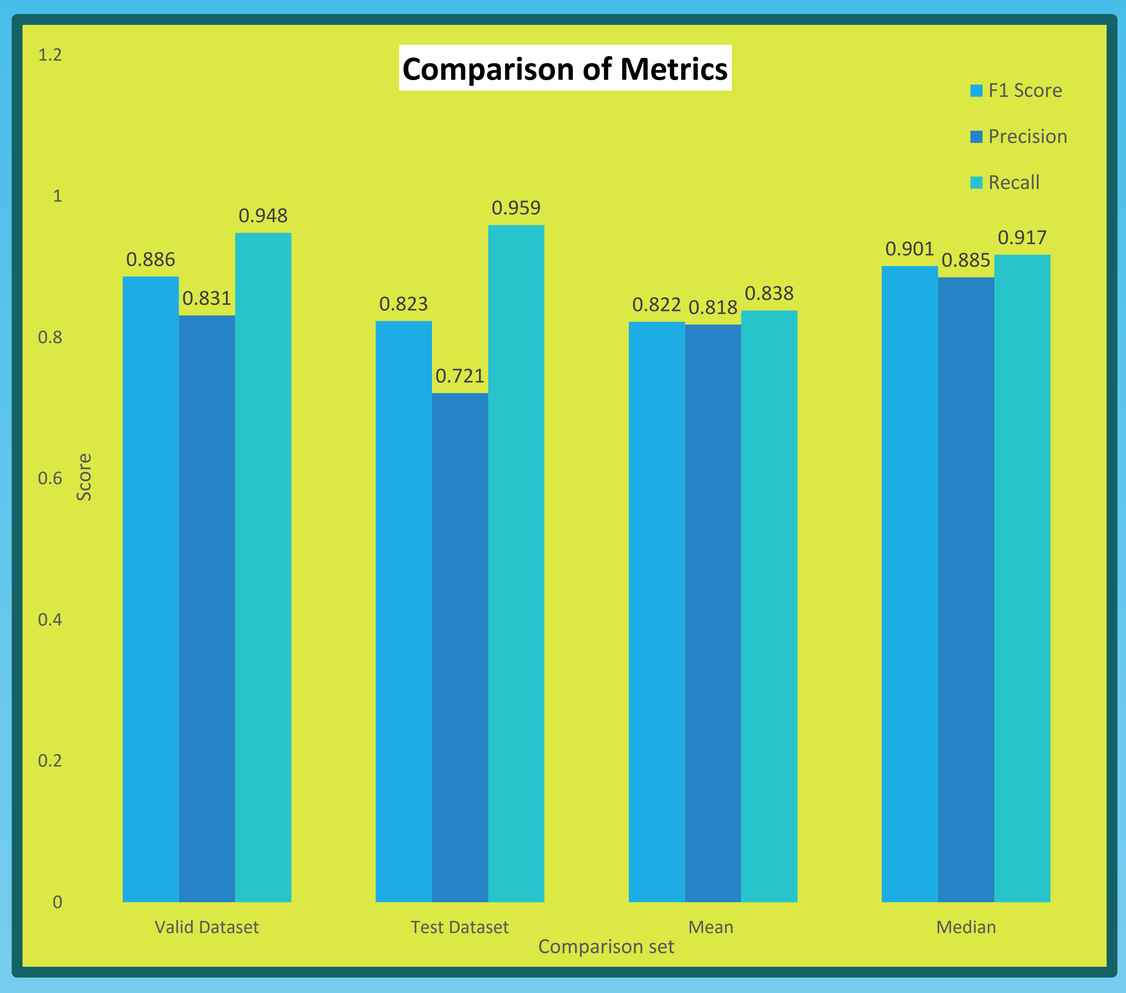
	Label 0	Label 1
Before Resampling	5372	2415
After Resampling	2415	2415

Training Regimen

Model:-RoBERTa-Large was preferred because of it's dynamic masking patterns to capture linguistic nuances.
Training Strategy:- 5-Fold-Cross Validation was preferred as it is a computationally efficient

Model parameters

- Epochs = 15
- Batch size = 6
- Tokenization length = 280
- Criterion = Cross Entropy Loss.
- Adam Optimizer
- Learning rate = 1.5e-8



Discussion

- **Suboptimal Learning Rate**: The learning rate was not effectively optimized, potentially impeding the model's ability to converge to the global minimum of the loss function efficiently.
- **No Text Conversion to Lowercase**: 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.
- **Tokenization Length**: 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.
- **Higher False Positive Rates**: While the model excels in identifying relevant cases, it tends to exhibit higher false positive rates, thereby impacting precision.

Acknowledgements

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

Future Work

This methodology can potentially be extended to Sentence Transformers, as they excel at embedding text based on context. Utilizing Sentence Transformers could simplify the development of a more effective machine learning model by providing richer contextual embeddings.

ACL 2024 Bangkok, Thailand
SMM4H