Transformers at #SMM4H 2024: Identification of Tweets Reporting Children's Medical Disorders And Effects of Outdoor Spaces on Social Anxiety Symptoms on Reddit Using RoBERTa

Kriti Singhal, Jatin Bedi
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
kritisinghal711@gmail.com, jatin.bedi@thapar.edu

Abstract

With the widespread increase in the use of social media platforms such as Twitter, Instagram, and Reddit, people are sharing their views on various topics. They have become more vocal on these platforms about their views and opinions on the medical challenges they are facing. This data is a valuable asset of medical insights in the study and research of healthcare. This paper describes our adoption of transformer-based approaches for tasks 3 and 5. For both tasks, we fine-tuned large RoBERTa, a BERT-based architecture, and achieved an F1 score of 0.413 and 0.900 in tasks 3 and 5, respectively.

1 Introduction

The past few years have witnessed an exponential rise in the use of social media. People can voice their views and opinions on social media platforms such as Facebook, Reddit, and Twitter. Twitter and Reddit have become major platforms for people seeking help and sharing their medical problems. They also take to these platforms to share their health regimen and medical concerns. Thus, Reddit and Twitter are indispensable resources that aid in better comprehension of health services and exploration of avenues for improvement in health services.

The recent developments in the field of Natural Language Processing (NLP) have garnered great interest from the healthcare research community. Some of the major breakthroughs include Long Short Term Memory (Hochreiter and Schmidhuber, 1997), and Gated Recurrent Units (Chung et al., 2014). However, the advent of transformers (Vaswani et al., 2017) led to a significant improvement in the performance.

The Social Media Mining for Health Applications (SMM4H) (Xu et al., 2024) unites researchers from across the globe with the objective of developing and sharing NLP methods for mining, representation, and analysis of health-related data. This year, SMM4H hosted seven shared tasks involving extraction, classification, and Large Language Model (LLM) identification. Our team participated in two of the classification tasks, namely, Task 3 and Task 5.

Task 3 is a multi-class classification task aimed to qualitatively evaluate the impact of outdoor spaces on Social Anxiety Disorder (SAD). Around one-third of the people with SAD report showing symptoms for ten years before seeking medical help. However, people do share their symptoms on platforms such as Reddit to discuss and share their symptoms and seek help to alleviate those symptoms. Task 3 focuses on classifying such posts on Reddit into one of the four categories, 'positive effect', 'neutral or no effect', 'negative effect', and 'unrelated'.

Task 5 aims at distinguishing tweets that mention a disorder such as delayed speech from the tweets whose users reported their pregnancy on Twitter and also reported having a child with attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech, or asthma. This task facilitates the use of Twitter as a medium not just for epidemiological studies but also to investigate the parents' experiences and directly target support interventions.

2 Methodology

Transformers have shown great potential in various NLP classification tasks (Khatri et al., 2022). For the purpose of performing classification in both the tasks, various transformers were tested. However, RoBERTa showed the best performance for both the multi-class and binary classification tasks. The RoBERTa transformer model was first introduced by Facebook in 2019 (Liu et al., 2019). RoBERTa has been trained on 160GB of uncompressed text leading to improved performance on classification tasks. In this work, we present our approach to fine-tuning RoBERTa for the SMM4H classification

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 67–70
August 15, 2024 ©2024 Association for Computational Linguistics

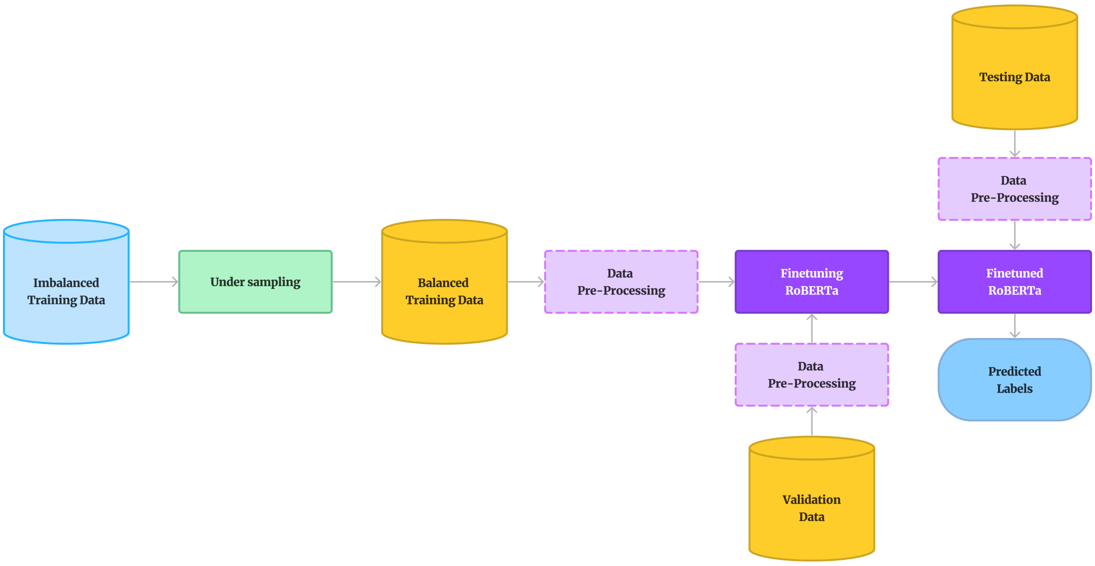

Figure 1: Proposed Methodology

Table 1: Data Distribution for Task 3

Dataset	Label		Total
	0	1	
Training	5118	2280	7398
Validation	254	135	389

Table 2: Data Distribution for Task 5

Dataset	Label				Total
	0	1	2	3	
Training	1131	395	160	114	1800
Validation	377	131	54	38	600

tasks. A pre-trained model, RoBERTa, was trained in a self-supervised manner, i.e., only raw texts were used to train it without the involvement of human intervention for labeling.

2.1 Data Pre-processing

For the tasks, the data was imbalanced in nature, as can be seen from the data distribution in Table 1 and Table 2 for Task 3 and Task 5, respectively. Due to this, there is a possibility that the results can be skewed with a preference towards the majority class. To address this issue, undersampling was performed. In this, random sampling was performed on all classes such that the number of instances for all the classes become equal to the number of instances in the minority class.

The data provided for both the tasks has been sourced from social media platforms. As a consequence of this, there was extensive use of various emojis, numbers and other special characters. In the pre-processing step, all the characters are first converted to lowercase. Then, emojis, numbers, and special characters are removed from the text.

2.2 Transformer Fine-tuning

The procedure adopted to fine-tune the model has been shown in the Figure 1. The RoBERTa transformer was fine-tuned in two different ways. The difference between the two approaches was that pre-processing was performed in one and it was not performed in the other. The results for both these approaches have been detailed in Table 3 and 4 for Task 3 and Task 5, respectively.

In Task 3, to train the transformer, the learning rate was set to 1e-6, and the weighted Adam optimizer was used. The Cross-Entropy loss function was utilized to penalize the mistakes made by the model during the training process. The model was fine-tuned till 45 epochs when data pre-processing was performed. And when data pre-processing was not performed the model was trained for 42 epochs.

In Task 5, to fine-tune the RoBERTa, the learning rate used was 1e-5. The cross-entropy function and weighted Adam optimizer were used as the loss function and optimizer, respectively. The model was fine-tuned for 10 epochs when pre-processing

Table 3: Model Performance for Task 3

Description	F1 Score	Precision	Recall	Accuracy
RoBERTa with Pre-Processing	0.383	0.413	0.482	0.378
RoBERTa without Pre-Processing	0.413	0.431	0.52	0.411
Mean	0.5186	0.5649	0.5379	0.5746
Median	0.5795	0.63	0.5885	0.627

Table 4: Model Performance for Task 5

Description	F1 Score	Precision	Recall
RoBERTa with Pre-Processing	0.900	0.854	0.950
RoBERTa without Pre-Processing	0.870	0.807	0.944
Mean	0.822	0.818	0.838
Median	0.901	0.885	0.917

was performed, and when no pre-processing was performed, the model was fine-tuned for 11 epochs.

To determine the number of iterations for which the model should be trained, the early stopping was used. If there was no significant improvement in the performance of the validation data for 5 consecutive epochs during training, then the process was not carried further and was halted at that point.

3 Results and Discussion

An in-depth analysis was performed on the performance of the large RoBERTa transformer model in the work. The performance was analyzed in different scenarios, both with and without pre-processing. The results obtained on the test data, along with the mean and median of the overall performance of all teams, have been summarised in Table 3 and Table 4 for Task 3 and Task 5, respectively.

To fine-tune the transformers, we first perform under-sampling on the data to avoid bias in the model. Then, we use two approaches for training using the balanced data. In the first approach, we use the text as is, without any pre-processing, and in the second approach, we perform pre-processing as described in Section 2.1. The early stopping approach was used to determine the number of epochs.

In Task 3, large RoBERTa without pre-processing performed better and achieved an F1-score of 0.413 than large RoBERTa with pre-processing, which achieved an F1-score of 0.383.

However, in Task 5, large RoBERTa with pre-processing achieved an F1-score of 0.900, whereas large RoBERTa without pre-processing achieved only 0.870.

4 Conclusion and Future Work

In this work, we present our adoption of the large RoBERTa transformer model to perform classification on social media data sourced from Reddit and Twitter. In Task 3, we use the model to perform multi-class classification on the effects of outdoor spaces on social anxiety using Reddit posts. In Task 5, we use the model to perform binary classification of English tweets to classify whether or not they report medical disorders in children. We also analyze the performance of the transformer, both with and without pre-processing.

In the future, an ensembling approach can be implemented. This approach can help amalgamate the results of different transformers, which may lead to improved results (Dima et al., 2020; Montañés-Salas et al., 2022; Lin et al., 2022).

References

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

George-Andrei Dima, Andrei-Marius Avram, and Dumitru-Clementin Cercel. 2020. Approaching SMM4H 2020 with ensembles of BERT flavours. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 153–157, Barcelona, Spain (Online). Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Roshan Khatri, Sougata Saha, Souvik Das, and Rohini Srihari. 2022. UB health miners@SMM4H'22: Exploring pre-processing techniques to classify tweets using transformer based pipelines. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 114–117, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tzu-Mi Lin, Chao-Yi Chen, Yu-Wen Tzeng, and Lung-Hao Lee. 2022. NCUEE-NLP@SMM4H'22: Classification of self-reported chronic stress on Twitter using ensemble pre-trained transformer models. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 62–64, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

Rosa Montañés-Salas, Irene López-Bosque, Luis García-Garcés, and Rafael del Hoyo-Alonso. 2022. ITAINNOVA at SocialDisNER: A transformers cocktail for disease identification in social media in Spanish. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 71–74, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.