Golden_Duck at #SMM4H 2024: A Transformer-based Approach to Social Media Text Classification

Md. Ayon Mia, Mahshar Yahan, Hasan Murad, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1804{128, 007}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd, muhammad_ikhan@cuet.ac.bd

Abstract

In this paper, we have addressed Task 3 on social anxiety disorder identification and Task 5 on mental illness recognition organized by the SMM4H 2024 workshop. In Task 3, a multi-classification problem has been presented to classify Reddit posts about outdoor spaces into four categories: Positive, Neutral, Negative, or Unrelated. Using the pre-trained RoBERTa-base model along with techniques like Mean pooling, CLS, and Attention Head, we have scored an F1-Score of 0.596 on the test dataset for Task 3. Task 5 aims to classify tweets into two categories: those describing a child with conditions like ADHD, ASD, delayed speech, or asthma (class 1), and those merely mentioning a disorder (class 0). Using the pre-trained RoBERTa-large model, incorporating a weighted ensemble of the last 4 hidden layers through concatenation and mean pooling, we achieved an F1 Score of 0.928 on the test data for Task 5.

1 Introduction

In the past few years, social media has become the primary source of communication. Unfortunately, millions of social media users suffer from mental illnesses such as social anxiety, attention-deficit or hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech or asthma. The World Health Organization (WHO) estimates that around 450 million individuals are impacted by these conditions. According to WHO, 1 in 7 adolescents from age 10 to 19 experience mental health issues, many of which go unnoticed and untreated.[1].

In this research work, we have presented our findings for the two shared tasks on social anxiety disorder identification (Task 3) and mental illness recognition (Task 5) organized under SMM4H 2024 workshop (Xu et al., 2024). By employing transformer-based pre-trained models alongside different techniques, we obtained F1 scores of 0.596 for Task 3 and 0.928 for Task 4.

By examining user-generated content from social media platforms like Reddit and Twitter, this research aims to shed light on the potential therapeutic benefits of outdoor environments and enhance our understanding of childhood developmental disorders.

2 Related Work

Automatic mental illness detection using social media data has been a key research area in literature. The key focus of this study (Klein et al., 2022) is to detect and identify conditions such as ADHD, ASD, delayed speech, and asthma in children using data from Twitter. It suggests Twitter data could offer a high-performance method for this epidemiological research. They have achieved notable performance using several models: SVM with an F1 score of 0.74, BERT-Base-Uncased achieving an F1 score of 0.85, and BERTweet-Large scoring an F1 score of 0.92. Mukherjee et al. (2023) discuss a system using traditional machine learning models for the early detection of Autism Spectrum Disorder (ASD). They have utilized various machine learning techniques, including SVM (accuracy of 0.71), Logistic Regression (accuracy of 0.71), K-nearest neighbor (accuracy: 0.62), and Random Forest (accuracy: 0.69). Additionally, they have explored the stacked deep learning models like CNN+SVM (accuracy of 0.916) and RNN+SVM (accuracy of 0.866). Ta et al. (2024) focus on classifying text data related to social disorders in children and adolescents using encoder-decoder models and data augmentation methods, achieving the best F1 scores of 0.627 in Task 3 and 0.841 in Task 5. The research highlights the importance of employing sophisticated deep learning networks to identify social disorders from social media data. According

[1] https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 28–31
August 15, 2024 ©2024 Association for Computational Linguistics

to Ameer et al. (2022), the paper explores the application of social media for mental health communication. It emphasizes the significance of automating the detection of mental illness using RoBERTa (achieving an accuracy and F1 score of 0.83) and BERT (with accuracy and F1 score of 0.78 and 0.80 respectively) models for classifying mental health disorders based on Reddit posts.

3 Task and Dataset Description

These shared tasks have been organized by the SMM4H 2024 workshop (Xu et al., 2024), including two tasks related to identifying and classifying medical information from social media data.

Task 3, entitled "Classification of reported effects of outdoor spaces on social anxiety symptoms", entails a multi-class classification challenge. Its objective is to sort Reddit posts containing specific keywords related to outdoor environments into four distinct categories: (i) Positive effect (0), (ii) Neutral or no effect (1), (iii) Negative effect (2), and (iv) Unrelated (keywords not pertaining to actual outdoor spaces or activities) (3).

Task 5, named "Binary classification of tweets reporting children's medical disorders", introduces a binary classification challenge. The goal of this task is to automatically identify tweets from users who disclosed their pregnancy on Twitter and report a child with ADHD, ASD, delayed speech, or asthma (labeled as "1"), versus tweets that just mention these disorders (labeled as "0"). In Task 3, the data is divided into 1800 training samples, 600 validation samples, and 1200 test samples, with the test data not publicly provided. Similarly, in Task 5, there are 7398 training samples, 389 validation samples, and 10,000 test samples, with the test data also not publicly provided.

Sets	**Classes**				**Total**
	0	1	2	3	
Train	1131	160	395	114	1800
Valid	377	54	131	38	600

Table 1: Dataset statistics of Task 3.

Sets	**Classes**		**Total**
	0	1	
Train	5118	2280	7398
Valid	254	135	389

Table 2: Dataset statistics of Task 5.

4 Methodology

4.1 Preprocessing

We have applied similar preprocessing steps for both tasks: removed URLs, hashtags, and user mentions from text, removed emojis, lowered all English letters, and removed duplicate punctuations. For Task 5, we have conducted additional preprocessing by expanding contractions like "wouldn't" to "would not".

4.2 Augmentation

In Task 3, as the classes are imbalanced, we have employed random oversampling using the scikit-learn library. For minority classes 1, 2, and 3, instances have been resampled with replacements to a target count of 350 each.

4.3 Fine-tuning Process

We have explored different transformer models for both shared tasks. For Task 3, we have fine-tuned the following transformer models:

- XLM-RoBERTa-base, followed by Mean pooling of the output
- RoBERTa-base with the concatenation of Mean pooling, CLS, and Attention Head
- RoBERTa-base with additional Attention Head on top of the model
- BERT-base-uncased, with weighted layer pooling from the last 4 layers

For Task 5, we have employed ensemble methods combining multiple output representations from large pre-trained language models. Specifically, we have fine-tuned following models:

- RoBERTa-large (Liu et al., 2019) utilizes a weighted ensemble approach that concatenates the last 4 hidden layers and employs mean pooling.
- RoBERTa-large uses a combination of weighted ensemble by concatenating the last 4 hidden layers, weighted pooling of these layers, and mean pooling.
- BERT-large uses a weighted ensemble by concatenating the last 4 hidden layers and utilizes mean pooling.

Model	F1 Score	Precision	Recall	Accuracy
XLM-RoBERTa-base (Mean pooling)	0.540	0.543	0.560	0.697
RoBERTa-base: Concat of Mean pooling, CLS, Attention Head	**0.562**	0.563	0.573	0.692
RoBERTa-base + Attention Head	0.333	0.533	0.332	0.67
BERT-base-uncased: WeightedLayer-Pooling (last 4 layers)	0.226	0.253	0.266	0.64

Table 3: Validation set outcomes for Task 3 for different models

Model	F1 Score	Precision	Recall	Accuracy
XLM-RoBERTa-base (Mean pooling)	0.59	0.586	0.612	0.623
RoBERTa-base: Concat of Mean pooling, CLS, Attention Head	**0.596**	0.603	0.601	0.618
RoBERTa-base + Attention Head	0.361	0.577	0.376	0.505
Mean	0.5186	0.5649	0.5379	0.5746
Median	0.5795	0.63	0.5885	0.627

Table 4: Results of various models in Task 3 on the official test set

Model	F1 Score	Precision	Recall
RoBERTa-large: Weighted Ensemble of LastConcat, LastPooled, Mean Pooling	0.912	0.899	0.926
RoBERTa-large: Weighted Ensemble of LastConcat and Mean Pooling	**0.923**	0.913	0.933
BERT-large: Weighted Ensemble of LastConcat, Mean Pooling	0.864	0.834	0.896

Table 5: Performance of different models in task 5 on the validation set where LastConcat, and LastPooled denote "Concatenation of last 4 hidden layers" and "Weighted pooling of last 4 layers" respectively

Model	F1 Score	Precision	Recall
RoBERTa-large: Weighted Ensemble of LastConcat, LastPooled, Mean Pooling	0.908	0.902	0.914
RoBERTa-large: Weighted Ensemble of LastConcat and Mean Pooling	**0.928**	0.919	0.937
Mean	0.822	0.818	0.838
Median	0.901	0.885	0.917

Table 6: Official results in Task 5 on the test set using different models

For the ensemble models, the weights have been treated as hyperparameters and tuned on the validation set to maximize the F1 score. First, we have evaluated each output representation independently on the validation data. Next, the top performing representations have been linearly combined using tuned weights, effectively creating an ensemble of multiple complementary representations.

The textual data has been preprocessed as per subsection 4.1 and tokenized with maximum lengths of 512 and 100 for Tasks 3 and 5, respectively, using a subword tokenizer.The hyperparameters used for both tasks include the AdamW optimizer, with learning rates set to 2e-5 for Task 3 and 1e-5 for Task 5. Batch sizes were 16 for Task 3 and 32 for Task 5, with training conducted over 10 epochs and early stopping determined by validation performance. A dropout rate of 0.3 was applied for regularization purposes. The whole experiment has been run on the Kaggle environment with the infrastructure including the NVIDIA Tesla P100 GPU with 16GB VRAM, 29GB RAM, and 4 CPU cores.

5 Results Analysis

We have evaluated the model's performance on the validation and test sets for both Task 3 and Task 5. Specifically, Table 3 and Table 4 present the results for Task 3 on the validation and test sets, respectively. On the validation set, the RoBERTa-base model with the concatenation of Mean pooling, CLS, and Attention Head has achieved the best F1 score of 0.562. This model with the same configuration has obtained the highest F1 score of 0.596, precision of 0.603, recall of 0.601, and accuracy of 0.618 on the test case. For Task 5, Table 5 and Table 6 present the validation and test set performance respectively. On the validation set, the RoBERTa-large model with a weighted ensemble of the last 4 hidden layers concatenation and Mean pooling has attained the best F1-score of 0.923. This model configuration has shown the best performance on the test dataset, with an F1-score of 0.928, precision of 0.919, and recall of 0.937.

6 Conclusion

In this paper, we outline our contributions to Task 3 and Task 5 for the 2024 SMM4H workshop on Social Media Mining for Health Applications. We utilized several transformer models to perform multiclass classification on Reddit posts and binary classification on tweets. Utilizing RoBERTa-base with Mean Pooling, CLS, and Attention Head concatenation, we have reached an F1 score of 0.596 for Task 3. For Task 5, we have employed RoBERTa-large with a weighted ensemble of the last 4 hidden layers concatenation and Mean Pooling, obtaining an F1-score of 0.928. Our experiments demonstrated our system's reliability and adaptability, and we are working on further improvements.

References

Iqra Ameer, Muhammad Arif, Grigori Sidorov, Helena Gòmez-Adorno, and Alexander Gelbukh. 2022. Mental illness classification on social media texts using deep learning and transfer learning. *arXiv preprint arXiv:2207.01012*.

Ari Z Klein, José Agustín Gutiérrez Gómez, Lisa D Levine, and Graciela Gonzalez-Hernandez. 2022. Automatically identifying childhood health outcomes on twitter for digital epidemiology in pregnancy. *medRxiv*, pages 2022–11.

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*.

Prasenjit Mukherjee, Sourav Sadhukhan, Manish Godse, and Baisakhi Chakraborty. 2023. Early detection of autism spectrum disorder (asd) using traditional machine learning models. *International Journal of Advanced Computer Science and Applications*, 14(6).

Hoang-Thang Ta, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and Alexander Gelbukh. 2024. Thangdlu at# smm4h 2024: Encoder-decoder models for classifying text data on social disorders in children and adolescents. *arXiv preprint arXiv:2404.19714*.

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.