CogAI@SMM4H 2024: Leveraging BERT-based Ensemble Models for Classifying Tweets on Developmental Disorders

Liza Dahiya and **Rachit Bagga**
Computer Science & Engineering
Indian Institute of Technology, Bombay
{lizadahiya, rachitbagga}@cse.iitb.ac.in

Abstract

This paper presents our work for the Task 5 of the Social Media Mining for Health Applications 2024 Shared Task - Binary classification of English tweets reporting children's medical disorders. In this paper, we present and compare multiple approaches for automatically classifying tweets from parents based on whether they mention having a child with attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech, or asthma. We use ensemble of various BERT-based models trained on provided dataset that yields an F1 score of **0.901** on the test data.

1 Introduction

The prevalence of child developmental disorders, including attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech, and asthma, presents significant challenges for families and healthcare systems worldwide. Understanding the experiences and needs of parents raising children with these conditions is crucial for informing support interventions and public health policies. In response to this challenge, social media data can be leveraged (Kim et al., 2020) after classifying tweets from parents mentioning if their child has one of these developmental disorders. This task holds significant promise for advancing our understanding of the prevalence and correlates of these conditions, particularly in relation to pregnancy exposures. Some studies have been done previously to assess the relationship between prenatal exposure to various substances during pregnancy to the risk of developing these disorders but many had low statistical power (Castro et al., 2016; Linnet et al., 2003). In this paper, we present a comprehensive analysis of multiple approaches for classifying tweets related to parental experiences with ADHD, ASD, delayed speech, and asthma. Our study focuses on a binary classification task, aiming to discern whether tweets mention these developmental disorders or not. Through this work, we aim to contribute to the advancement of epidemiologic research (Cortese and Coghill, 2018; Rowland et al., 2002) and provide valuable insights into parental experiences with developmental disorders.

In this paper, we provide a detailed explanation of our system architecture, describe our experiments and share results obtained during this task.

2 Methods

2.1 Dataset

The dataset provided was divided into three partitions: training, validation, and test sets consisting of 7398, 389, and 1947 tweets respectively. Tweets that describe a child with ADHD, ASD, delayed speech, or asthma are annotated **1**, and those that simply mention a disorder are annotated **0**.

Text	Label
Finally a dr has diagnosed my 3.5yr old with asthma. Now he will be on chronic medicine and we can hopefully keep him healthy and thriving.	1
Can u give any tips to "live with it" please. I think my son has ADD. Trying to help him	0
Flying tomorrow...during a pandemic with a nonverbal 3 year old. We could use some prayers, please.ðŸ~ðŸ¥´	1

Table 1: Examples of Tweets with Annotations

2.2 Pre-processing

In the preprocessing of the data, two steps were undertaken to ensure the quality & balance of the dataset.

2.2.1 Data Cleaning

Since the twitter data contains large amounts of noise, we applied several data cleaning procedures. This included the removal of URLs, mentions (ex: @USER), and hashtags (ex: #funny). Additionally, handling of emojis was crucial. Emojis were replaced with their corresponding textual representations to remove potentially distracting elements and simplify the text for further analysis.

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 114–116

Model	F1			Recall			Precision		
	M1	M2	M3	M1	M2	M3	M1	M2	M3
MentalBERT	0.885	0.897	**0.917**	0.854	0.868	0.896	0.896	0.905	0.923
PsychBERT	0.843	0.855	**0.869**	0.821	0.835	0.854	0.866	0.874	0.879
TwitterBERT	0.885	**0.902**	0.897	0.854	0.868	0.896	0.896	0.905	0.923
DistilBERT	0.877	0.885	**0.887**	0.848	0.853	0.864	0.863	0.875	0.882

Table 2: Scores on the validation set of tweets, where *M1* is trained without preprocessing, *M2* is trained with augmented data, and *M3* is trained with cleaned and augmented data.

2.2.2 Data Augmentation

The technique used for data augmentation was "gender-based" text augmentation. Within the texts labelled as **1**, all instances of the term "son" (and corresponding pronouns "he," "his," and "him") were systematically substituted with "daughter" (and corresponding pronouns "she" and "her").The distribution of true labels increased from 30.82% before augmentation to **42.37%.** after augmentation improving the class imbalance situation.

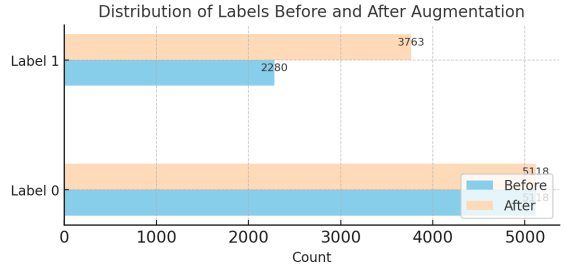

Figure 1: Data Distribution before and after Augmentation

This augmentation technique was straightforward to implement and has shown to improve performance by including additional context.

3 Experiments and Results

During our model exploration, we fine-tuned multiple pre-trained models including MentalBERT (Ji et al., 2022), PsychBERT (Vajre et al., 2021), TwitterBERT (Zhang et al., 2022), and DistilBERT (Sanh et al., 2019). We trained each of them for 200 epochs with a learning rate of 1e-5 and a batch size of 32. Each of these models offers unique advantages that could be beneficial for our purpose.

MentalBERT and PsychBERT are tailored for mental health and biomedical text datasets which could potentially capture features relevant to our task of identifying tweets related to "developmental disorders". TwitterBERT's training on Twitter data makes it adept at handling the informal language of tweets which can improve classification accuracy while DistilBERT's computational efficiency can make it suitable for scalable deployment, an advantage in handling large volumes of Twitter data.

We systematically evaluated the performance of each of these models by adopting an incremental approach. Initially, we trained the models using the raw data without any pre-processing (**M1**). Subsequently, we performed training after augmenting the data (**M2**). Finally, we conducted training using cleaned data with data augmentation (**M3**). This comprehensive approach allowed us to assess the impact of classification performance of each model, providing valuable insights into their robustness and effectiveness for our specific task.

Our final model then was an ensemble of all these methods assigned weights according to their "F1-scores" on validation dataset. We hardcoded weights of 0.6 score to MentalBERT, 0.2 score to TwitterBERT, 0.1 score to each PsychBERT and DistilBERT. This was decided based on their individual performance as can be seen in **Table 2**.

Statistic	F_1-score	Precision	Recall
Our	0.901	0.885	0.917
Mean	0.822	0.818	0.838
Median	0.901	0.885	0.917

Table 3: Performance on Test Data for Task 5

4 Conclusion

In this task, we experimented with various BERT models and presented the results in Table 2. Table 3 presents our results along with the mean and median results obtained during the challenge, depicting a final F1-score of **0.901**. In future, we would like to experiment with more ensemble modelling methods, use better data augmentation techniques and For greater scope of research, a potential direction could involve incorporating multimodal data sources, such as images or audio recordings, to enrich the understanding of parental experiences and enhance the accuracy of classification models.

References

VM Castro, SW Kong, CC Clements, R Brady, AJ Kaimal, AE Doyle, EB Robinson, SE Churchill, IS Kohane, and RH Perlis. 2016. Absence of evidence for increase in risk for autism or attention-deficit hyperactivity disorder following antidepressant exposure during pregnancy: a replication study. *Translational psychiatry*, 6(1):e708–e708.

Samuele Cortese and David Coghill. 2018. Twenty years of research on attention-deficit/hyperactivity disorder (adhd): looking back, looking forward. *BMJ Ment Health*, 21(4):173–176.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Eric S Kim, Peter James, Emily S Zevon, Claudia Trudel-Fitzgerald, Laura D Kubzansky, and Francine Grodstein. 2020. Social media as an emerging data resource for epidemiologic research: characteristics of regular and nonregular social media users in nurses' health study ii. *American Journal of Epidemiology*, 189(2):156–161.

Karen Markussen Linnet, Søren Dalsgaard, Carsten Obel, Kirsten Wisborg, Tine Brink Henriksen, Alina Rodriguez, Arto Kotimaa, Irma Moilanen, Per Hove Thomsen, Jørn Olsen, et al. 2003. Maternal lifestyle factors in pregnancy risk of attention deficit hyperactivity disorder and associated behaviors: review of the current evidence. *American Journal of Psychiatry*, 160(6):1028–1040.

Andrew S Rowland, Catherine A Lesesne, and Ann J Abramowitz. 2002. The epidemiology of attention-deficit/hyperactivity disorder (adhd): a public health view. *Mental retardation and developmental disabilities research reviews*, 8(3):162–170.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. 2021. Psychbert: a mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1077–1082. IEEE.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.