CHAAI@SMM4H'24: Enhancing Social Media Health Prediction Certainty by Integrating Large Language Models with Transformer Classifiers

Sedigh Khademi, Christopher Palmer, Muhammad Javed,
Jim Buttery, Gerardo Luis Dimaguila
Murdoch Children's Research Institute
{sedigh.khademi, chris.palmer, muhammad.javed,
jim.buttery, gerardoluis.dimaguil}@mcri.edu.au

Abstract

This paper presents our approach for SMM4H 2024 Task 5, focusing on identifying tweets where users discuss their child's health conditions of ADHD, ASD, delayed speech, or asthma. Our approach uses a pipeline that combines transformer-based classifiers and GPT-4 large language models (LLMs). We first address data imbalance in the training set using topic modelling and under-sampling. Next, we train RoBERTa-based classifiers on the adjusted data. Finally, GPT-4 refines the classifier's predictions for uncertain cases (confidence below 0.9). This strategy achieved significant improvement over the baseline RoBERTa models. Our work demonstrates the effectiveness of combining transformer classifiers and LLMs for extracting health insights from social media conversations.

1 Introduction

The SMM4H Shared Tasks focus on using machine learning and natural language processing to solve the challenges associated with extracting health insight from social media (Xu et al., 2024). This paper presents our work in SMM4H 2024 Task 5. The focus of task 5 is to identify tweets where users talk about their own child having attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech, or asthma, and differentiate them from tweets that simply discuss these conditions in general. For a description of this dataset, see (Klein et al., 2024).

One of the main challenges of processing social media text is dealing with the lay language that often lacks proper grammar and structure, containing misspellings, abbreviations, and unconventional phrasing (Gonzalez-Hernandez et al., 2017). Additionally, texts from advertising, news, social bots, and other non-personal posting modalities need to be excluded when identifying personal conversations (Javed et al., 2023). This requires robust algorithms capable of understanding colloquial language while maintaining accuracy in analysis and interpretation (Khademi Habibabadi et al., 2022). In this paper, we present a pipeline that combines transformer-based classifiers with the GPT-4 LLM to classify the tweets containing the health conditions of interest.

1.1 Task description

The dataset of Task 5 contained texts from Twitter. The training data consisted of an imbalanced set of 7,398 tweets (2,280 positive and 5,118 negative). A separate validation set of 389 tweets (135 positive and 254 negative) was also provided. An unlabelled test dataset of 10,000 records was supplied after the model building and validation stages of the competition, predictions over this test data were the competition submission.

To prepare the text data, we used the clean-text and html python libraries to fix Unicode errors and convert character entities, transliterate to ASCII, replace user mentions and URLs with generic placeholders, and remove line breaks. We observed that in the validation data emojis were replaced with double question marks, so we adopted the same strategy.

2 Method

Our approach consisted of three main steps. First, for data preparation we used a topic-modelling based under-sampling method to address data imbalance issues. Secondly, we trained a RoBERTa-based classifier on this adjusted training data. Finally, we employed GPT-based LLMs to refine the classifiers' predictions where the classifier's confidence in its predictions dropped below a threshold of 0.9.

2.1 Data preparation

To select a more balanced set our strategy was to retain all the positive labels and under-sample neg-

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 71–73
August 15, 2024 ©2024 Association for Computational Linguistics

ative labels. Using BERTopic, we performed topic modelling on all the data. We discovered one topic with a repetitive subject and content, for which we retained only a few records that represented the subject. For the other topics, we sampled 60% of the negative labels per topic, selecting them based on decreasing proximity from the topic's cluster centroid, as determined by text similarity. This method (Khademi et al., 2023) ensured a diverse selection of negative examples. We identified 10 records in the negative training data that appeared more likely positive, so these were relabeled. The dataset then consisted of 2,290 positive and 3,345 negative records. Subsequently, we divided this into training and validation datasets of 5,345 and 500 respectively, preserving the competition's supplied 389 records validation dataset as a hold-out test set.

2.2 Classifier training

We trained BERTweet-Large (Nguyen et al., 2020) and Twitter-RoBERTa (Loureiro et al., 2023) classifiers for 6 epochs with a batch size of 16 for training and validation. The AdamW optimizer was used; the evaluation strategy was "steps", the learning rate was 2e-5, weight decay was 0.01, with no warmup steps.

Checkpoints were saved every 10 steps and the best checkpoints were subsequently identified based on a balance of loss, ROC-AUC, and F1-score on validation data. Their F1 scores were evaluated against the hold-out test set. The top-performing BERTweet models achieved a rounded F1 score of 0.912, while the best Twitter-RoBERTa model reached 0.938.

2.3 LLM prompt engineering

We used two state-of-the-art language models: GPT-4-Turbo-2024-04-09 ("GPT4") with a 128K token context window and GPT-3.5-Turbo-16K-0613 ("GPT3") with a 16K token context window. To ensure consistent predictions, we set the model temperature to zero for all experiments. Default context length settings for each model were used. Our evaluation involved a comparison of zero-shot (where no training data is provided) and few-shot (with limited training data) prompting strategies combined with standard and chain of thoughts (CoT) prompting approaches. Based on its superior performance in our evaluations, GPT-4's zero-shot CoT prompting was chosen. Appendix A provides the prompt used for this task.

Model	Precision	Recall	F1
LLM+Clf	0.949	0.963	0.956
Clf	0.921	0.956	0.938
LLM	0.777	0.956	0.857

Table 1: Validation scores

Model	Precision	Recall	F1
LLM+Clf	0.907	0.949	0.927
Median	0.885	0.917	0.901
Mean	0.818	0.838	0.822

Table 2: Test scores

2.4 Predictions

On validation data, we compared the LLM's predictions against those of the Twitter-RoBERTa classifier's predictions, when the classifier's probabilities fell below 0.9, which we took as a threshold for a degree of certainty. The LLM's accuracy in correcting predictions in this cohort outweighed its errors, so by using its predictions for these records instead of the classifier's predictions, the F1 score increased by two percentage points. We employed this strategy for our entry with the competition's test data.

3 Results

The Twitter-RoBERTa classifier performed well independently, while the LLM showed comparatively lower performance when assessed against the validation dataset. However, by using the LLM for those texts where the classifier was less certain, we gained 2 percentage points over using the classifier only – this model is depicted as LLM+Clf in Table 1, with the other models below it in order of F1 score. Using the same strategy on the supplied test dataset gave us an F1 score of 0.927, which exceeded the median of all the competitors' results by almost 3 percentage points, as shown in Table 2.

4 Conclusions

While standard Transformer models excel at the classification task, using LLMs as arbiters on less interpretable texts can be beneficial. With good prompt design an LLM can be tuned to discern subtleties in texts that are difficult to teach a classifier without extensive training data. However, LLMs on their own may not be ideal for all classification tasks due to limitations in overall accuracy, efficiency, and resource requirements.

References

Graciela Gonzalez-Hernandez, Abeed Sarker, Karen O'Connor, and Guergana Savova. 2017. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(01):214–227.

Muhammad Javed, Gerardo Luis Dimaguila, Sedigh Khademi Habibabadi, Chris Palmer, and Jim Buttery. 2023. Learning from machines? social bots influence on covid-19 vaccination-related discussions: 2021 in review. In *Proceedings of the 2023 Australasian Computer Science Week*, pages 190–197.

Sedigh Khademi, Christopher Palmer, Muhammad Javed, Gerardo Luis Dimaguila, Jim P Buttery, and Jim Black. 2023. Detecting asthma presentations from emergency department notes: An active learning approach. In *Australasian Conference on Data Science and Machine Learning*, pages 284–298. Springer.

Sedigheh Khademi Habibabadi, Pari Delir Haghighi, Frada Burstein, and Jim Buttery. 2022. Vaccine adverse event mining of twitter conversations: 2-phase classification study. *JMIR Medical Informatics*, 10(6):e34305.

Ari Z Klein, José Agustín Gutiérrez Gómez, Lisa D Levine, and Graciela Gonzalez-Hernandez. 2024. Using longitudinal twitter data for digital epidemiology of childhood health outcomes: an annotated data set and deep neural network classifiers. *Journal of Medical Internet Research*, 26:e50652.

Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2023. Tweet insights: a visualization platform to extract temporal insights from twitter. *arXiv preprint arXiv:2308.02142*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

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.

A Prompts

Here is the zero-shot chain of thought prompt used in this task.

You are a highly intelligent and accurate tweet classifier with reasoning capabilities.

You will receive a tweet enclosed within triple quotes. Set the final Answer to Yes, and answer the following questions with Yes or No: Questions:

Q1: Does the tweet report a child having ADHD, ASD, autism, speech delay, asthma, or being nonverbal, or mention medications used for these conditions? Descriptions of assessment or testing of the child for these health conditions do not count as definite evidence of having them. Enquiries about the possibility of the child having these health conditions also do not count. If the answer to Q1 is yes, describe the health condition of the child and go to Q2. Otherwise, set the final answer to No and go to Instruction1.

Q2: Does the tweet describe a child with one of the health conditions—ADHD, ASD, autism, delayed speech, being nonverbal, or asthma—as the child of the author of the tweet, irrespective of any other relationships discussed? State the relationship of the writer of the tweet with the child who has one of these health conditions. If the answer to Q2 is yes, go to Q3; otherwise, set the final answer to No and go to Instruction1.

Q3: Is the text containing the health conditions of ADHD, ASD, autism, delayed speech, being nonverbal, or asthma an original statement, not a quote from someone else? If the answer to Q3 is yes, go to Instruction1; otherwise, set the final answer to No and go to Instruction1.

Intruction1: The final answer should be Yes if Q1, Q2, and Q3 are Yes; otherwise, the final answer is No. Your output should include answers to Q1, Q2, and Q3, followed by the final answer and reasoning sentences that show the proof step by step within 30 words.