Binary classification of tweets reporting children's medical disorders

Ekaterina Valverde Bilenko
100432515@alumnos.uc3m.es
Computer Science and Engineering Department, Universidad Carlos III de Madrid
Leganés, Madrid, Spain

ABSTRACT

In recent years, health-care systems have been experiencing a change in the way problems are solved. The use of machine learning models is nowadays key to analyse data, not only in formal medical environments but also in more relaxed social interactions. This work highlights the importance of working with data provided by social platforms outside classical health domains such as hospitals, particularly classifying tweets related to children's disorders. Combining heterogeneous sources related to health could provide more powerful data to improve decision-making processes. In this work we will explore the use of large language models (LLM) in classification tasks with a special focus in small models that consume less energy because of the current context of ecological efficiency in Artificial Intelligence (AI) development.These smaller LLMs are compared with traditional machine learning models to evaluate their performance in enhancing accuracy. Initial results demonstrate that LLMs outperform traditional methods, thus confirming their potential for real-world applications in pediatric health monitoring.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; • **Machine learning**; • **Large Language Models**;

KEYWORDS

natural language processing, text classification, social media

1 INTRODUCTION

Social networks such as Twitter or Reddit could be a valuable source of information related to health. People use these platforms to explain health related issues as a mean of sharing problems with patients under similar circumstances. Processing user-generated content has become a challenge as the practice of discussing healthcare issues with family, friends, and even strangers in the setting of social networks has grown in recent years. Analysing user posts could help in online crowd surveillance for various applications, for instance, pharmacovigilance to detect inadequate use of drugs as well as adverse drug reactions and filtering health-related content in blogs, among others. Many efforts have been devoted to analyse user posts in different social networks, some of then in classification to filter posts related to different topics, others in named entity recognition (identity relevant concepts such as diseases, symptoms, drugs,...), see [4](Weissenbacher et al., 2023) and [3](González et al, 2022) for approaches to analyze social media in the medical domain.

Shared tasks organized in SEMEVAL [1] (International Workshop on Semantic Evaluation), CLEF [2](Conference and Labs of the Evaluation Forum), SMM4H [3](Social Media Mining for Health Applications Workshop) and many others propose challenges devoted to analyze unstructured data (such as texts) and extract relevant information to be used in decision-making processes.

2 RESEARCH PROBLEM

Many children suffer from disorders that can impact their life. Taking as example numbers from the USA there is 17% of children diagnosed with a developmental disability, and 8% with asthma. There is still little to no data for studying the relation between these diseases and pregnancy exposure. The objective of this research is to develop a binary classifier to detect tweets reporting children's medical disorders. The work is done in the framework of SMM4H 2024 Task 5 - Binary classification of tweets reporting children's medical disorders [4]. This binary classification task involves automatically distinguishing tweets, posted by users who had reported their pregnancy on Twitter, that report having a child with attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), delayed speech, or asthma (annotated as "1"), from tweets that merely mention a disorder (annotated as "0"). As a secondary objective, the differences are analyzed by applying the LLMs in their smaller versions and comparing them with classical machine learning modelsin the classification task.

3 PROPOSED SYSTEM

The data was provided by the organizers of the SMM4H 2024[5] competition and it consisted of a training and validation sets. These contain two columns, the first one being the actual tweet and the second one the label provided by the organizers (with 0,1 values). A testing set with just tweets was provided as well for making predictions.

The metrics for evaluating the systems were based on the F_1 score for the class 1 which denoted tweets reporting a child having a medical disorder. This F_1 score is based on Precision and Recall measures, where Precision = True Positives/(True Positives + False Positives), and Recall = True Positives/(True Positives + False Negatives).

We developed a baseline system using a traditional machine learning approach. The first thing that needed to be done was to clean the data and give it the appropriate format. The cleaning involved removing hastaghs and user tags, stopwords and leaving just

[1] https://semeval.github.io/
[2] https://www.clef-initiative.eu/
[3] https://healthlanguageprocessing.org/smm4h-2023/
[4] https://codalab.lisn.upsaclay.fr/competitions/17310
[5] https://healthlanguageprocessing.org/smm4h-2024/

Ekaterina Valverde Bilenko
100432515@alumnos.uc3m.es

the most relevant words. Each word left is a unit called token.For this we used the library provided by Spacy.io [6]. Once cleaned, the tokenized words were vectorized using Term Frequency-Inverse Document Frequency (TF-IDF). This needs to be done because the program does not understand text. This method calculates a word's relevance to a specific text.This relevance is directly proportional to the word's frequency within that text – the more often a word appears, the more important it likely is. However, TF-IDF also considers the word's frequency across the entire dataset. This Inverse Document Frequency (IDF) component helps to downweight common words and emphasize terms unique to a particular text, ensuring the most meaningful words are highlighted. There are other methods for this procedure but this gave the best results.

For the supervised methods we chose SVM (Support Vector Machine) algorithm [1] and RF (Random Forest). Both were trained and then tested with the validation set.

We wanted to investigate the power of current approaches based on Large Language Models (LLM) and compare the results with the ones obtained with the baseline code. To implement it, the preprocessing was different from the one above. LLM models work with big amounts of data and the training set that we had was too small, so we implemented a data augmentation code in order to create new instances. For this purpose a back translation method with Google Translate was used. It simply takes the texts and translates them twice, to a source language of choice and then back to the original language. This creates small variations that uplift the data set.

When choosing the pretrained LLM we filtered in the web HuggingFace[7] models that worked with tweets, for classification problems, in english and that had been trained with some sort of medical data. We found one called margotwagner/roberta-psychotherapy-eval that is based on the model type roBERTa (family of the BERT (Bidirectional Encoder Representations from Transformers) [2] models) .

In order to use this model, we had to load it and then fine-tune it with our data for it to learn about our specific problem. We modified some hyperparameters such as the number of epochs, the batch size and the learning rate to see what gave us the best results.

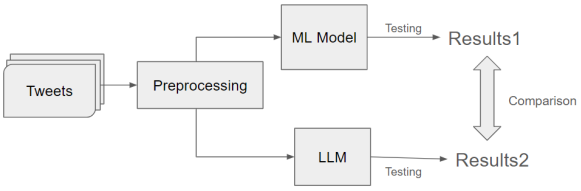

Figure 1: Proposed system

[6]https://spacy.io/
[7]https://huggingface.co/

Table 1: Results

Model	Precision	Recall	F1-score
SVM	0.67	0.51	0.58
RF	0.76	0.27	0.40
Roberta	0.91	0.90	0.91

4 RESULTS

As can be observed in Table1, for the ML approach a F1-score of 0.58 was obtained using SVM. With Random Forest the metric worsens giving just a 0.40 F1-score for the positive class. These poor results are attributed to suboptimal recall measurements (0.51 in SVM and 0.27 in RF) despite having a good precision in both cases (0.70 or above).

In stark contrast, the pretrained LLM outperformed these models significantly. With just 3 epochs, the LLM achieved a remarkable F1-score of 0.91, with both precision and recall at 0.90 and 0.91 respectively, which far surpasses the performance of the traditional ML models. This underscores the efficiency and potential of LLMs for real-world applications.

5 CONCLUSIONS

The study's objectives have been satisfactorily met. The application of small, efficient LLMs has proven beneficial for society, particularly in identifying tweets concerning medical disorders in children with high accuracy. This capability is critical for enhancing public health initiatives by enabling early detection, fostering research, and raising public awareness beyond conventional healthcare settings. Furthermore, these excellent results are achieved with a small LLM that optimizes resource use. Future work could explore fine-tuning the LLM on larger datasets and investigating its capacity to identify specific medical conditions to further enhance its utility.

6 ACKNOWLEDGEMENTS

As the author of this work, I would like to extend my heartfelt gratitude to Professor Paloma Martínez from the Computer Science and Engineering Department at Universidad Carlos III de Madrid. Her invaluable supervision and expert guidance have been essential to the success of this research.

REFERENCES

[1] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]
[3] Graciela Gonzalez and Davy Weissenbacher. 2022. Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task.*
[4] Davy Weissenbacher, Karen O'Connor, Siddharth Rawal, Yu Zhang, Richard Tzong-Han Tsai, Timothy Miller, Dongfang Xu, Carol Anderson, Bo Liu, Qing Han, et al. 2023. Automatic extraction of medication mentions from tweets—overview of the biocreative VII shared task 3 competition. *Database* 2023 (2023), baac108.