Evaluating Model Generalizability for Mental Health Detection on Reddit datasets

First Author Matteo Gorni
Second Author Tomas Guarini
Third Author Nemanja Ilic
Fourth Author Arianna Morandi
Fifth Author Davide Nieto
Sixth Author Beatrice Spagnolo

Abstract

In this study we investigate the robustness of machine learning models in detecting mental health conditions using social media data. Recognizing the critical importance of early detection in improving patient outcomes, this research leverages natural language processing (NLP) to classify mental health statuses based on Reddit posts. A key focus is the generalizability of models trained on specific datasets to other, similar datasets on the same platform. We utilized two distinct Reddit datasets: one containing comments labeled as depressed or not depressed, and another comprising comments of users with self-reported depression diagnoses. Various machine learning algorithms, including Random Forest, Support Vector Machine (SVM), BERT, and MentalBERT, were employed to assess model performance. The results indicate that while these models perform well on their respective training datasets, their performance significantly declines when cross-tested on different datasets, underscoring generalizability challenges. Additionally, the study demonstrates that context-specific models like MentalBERT outperform general models, highlighting the critical role of contextually relevant training data in mental health detection tasks. This research underscores both the challenges and potential strategies for creating more robust and generalizable models for detecting mental health issues using social media data.

1. Introduction

Mental illnesses, also known as mental health disorders, are highly prevalent worldwide and have emerged as significant public health concerns. These conditions encompass a range of disorders, including depression, suicidal ideation, bipolar disorder, anxiety disorder, and schizophrenia. Each of these disorders can profoundly impact an individual's physical health and overall well-being. Recent statistics reveal that millions of people globally suffer from one or more mental disorders, creating a substantial social problem and straining healthcare systems worldwide. Early detection of mental illness has been conclusively linked to improved patient outcomes. In recent years, natural language processing (NLP) has played a pivotal role in analyzing and managing large-scale textual data. NLP facilitates tasks such as information extraction, sentiment analysis, emotion detection, and mental health surveillance. Detecting mental illness from text can be framed as a text classification or sentiment analysis task, where NLP techniques automatically identify early indicators of mental health issues. The ultimate goal of these efforts is threefold: to better personalize psychiatric care, to enable early intervention, and to monitor population-level health outcomes in real time [1]. Despite the increased number of studies on the topic, the field is still far from producing clinically viable and usable models. In a recent article, Wen-Jing Yan, Qian-Nan Ruan, and Ke Jiang highlighted some of the most significant issues for these tasks, one of them being the lack of well-documented, publicly available datasets.

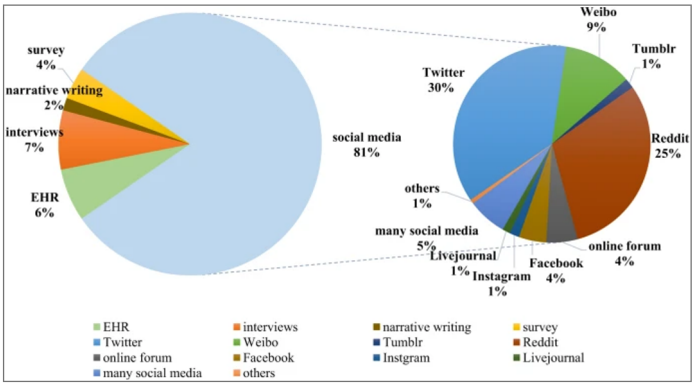

Figure 1.

A significant amount of potentially useful textual data could be gathered from sources such as medical records and interview transcripts. However, privacy-related issues complicate this procedure, making it difficult to create such datasets and even more challenging to share them with outside organizations [2]. This problem leads to the vast majority of research models being trained on data from social networks, this phenomenon has been quantified in a literature review written by Zhang, Annika et al in 2022 [1].

As we can observe in Figure 1, the vast majority of the models developed in the field's research use social media data, with most of it coming from just two social media platforms. This leads to an important question: how well do models of mental health conditions trained on automatically annotated social media data actually generalize to other and populations? According to Harrigian, Carlos, and Dredze, transfer learning performance is poor when models trained on data from a specific platform are tested on data from another one (in their case, Reddit and Twitter) [3]. In this work, we attempt to replicate a similar experiment but with a twist: we are going to compare two datasets that were taken from the same social media platform (Reddit) but were built using different approaches. We aim to see if even different data collection techniques on the same textual platform can lead to a significant drop in performance.

2. Datasets

The first dataset, collected by Kayalvizhi S. and Thenmozhi D., consists of 16,632 comments, each labeled with one of three categories: not depressed, moderately depressed, or severely depressed. Data was gathered from specific subreddits, including r/MentalHealth, r/depression, r/loneliness, r/stress, and r/anxiety. The inter-rater agreement (κ) for this dataset is 0.686 [4].
We performed two preprocessing operations on this data: first, we decided to merge the 'moderately depressed' and 'severely depressed' categories to transform the task from a multi-class classification problem to a binary one. According to the authors of the study, 'severely depressed' labels were assigned to comments expressing more than one disorder condition or mentioning past suicide attempts. However, this distinction does not align with the existing psychological literature, which classifies a depression diagnosis as 'severe' if the patient exhibits psychotic symptoms when depression peaks (e.g., derealization, visual or auditory hallucinations) without considering suicide attempts and the presence of other mental health issues. Secondly, since our data was unbalanced, we followed the dataset's authors' indications and implemented Synthetic Minority Over-sampling Technique (SMOTE) on it, SMOTE is a data augmentation technique which addresses class imbalances in classification problems by creating synthetic samples of the least popular class.

The second dataset, created by Yates, Cohan, and Goharian, is the Reddit Self-reported Depression Diagnosis (RSDD) dataset, consisting of 7,731 instances of posts from users identified as either depressed or not depressed [5]. It was created by labeling users from a publicly available Reddit dataset. Potentially depressed users were selected by identifying those who made posts between January 2006 and October 2016 containing specific sentences, such as "I was just diagnosed with depression." Three annotators reviewed each user's post history to determine if the user claimed to have been diagnosed with depression. Only users with at least two positive annotations were classified as depressed. On the other hand, the pool of potential control users was identified by selecting users who had never posted in a subreddit related to mental health and had never used terms related to depression or mental health. We performed no preprocessing operations on this dataset. From now on the two datasets will be called respectively: "MAIN" and "RSDD".

3. Methods

To test our hypothesis, we carried out a detailed procedure using both of our datasets with the following steps. First, we selected one of the datasets and performed a train-test split to evaluate the model's performance. We then retrained the model on the full dataset and assessed its performance on the other dataset. This entire procedure was repeated with the second dataset to ensure consistency and reliability. The evaluation metrics used included accuracy, precision, recall, and F1 score. We conducted our experiments with four different models, which comprised two classic machine learning algorithms and two deep learning models. The machine learning algorithms used were Random Forest and Support Vector Machine (SVM). These algorithms were selected based on their strong performance in tests conducted by Kayalvizhi and Thenmozhi. The deep learning models we used were BERT and MentalBERT. BERT is a transformer model pre-trained on BookCorpus, a dataset of 11,038 unpublished books, and English Wikipedia using a self-supervised approach. It uses two main pre training objectives: Masked Language Modeling (MLM), where 15% of the words in a sentence are randomly masked and predicted, enabling the model to learn bidirectional sentence representations, and Next Sentence Prediction (NSP), where the model predicts if two sentences were sequentially following each other in the original text. MentalBERT, a specialized variant of the BERT model, has been trained with an extensive dataset comprising mental health-related posts collected from Reddit. We selected this model because it aligns well with both the nature of our task and the characteristics of our dataset. Our goal is to evaluate whether this tailored approach can lead to significant improvements in the resulting metrics.

4. Experiments

4.1. Procedure

HYPERPARAMETERS	
Batch size	8 (default)
Maximum token length	256
Optimizer	AdamW (default)
Learning rate	4e-5
Epochs	10

Table 1.

The procedure for building our machine learning models began with data preprocessing using SpaCy. This involved tokenization, stop word removal, and lemmatization. We expanded contractions and converted all text to lowercase. The processed text was then used to build a continuous bag-of-words model.
After constructing our model, we utilized it to transform our text into embeddings. Each entry corresponded to the mean of the line's embedding vectors. This straightforward technique has been proven effective across various tasks and architectures [7]. For the pretrained transformer models, we used the SimpleTransformers library, which is built on top of the HuggingFace Transformers library. This facilitated the implementation and fine-tuning of BERT and MentalBERT on our dataset.

All experiments were done with a set random seed (42) to guarantee reproducibility. In Table 1, we present a list of some key model hyperparameters.

4.2. Results

In this section we present a summary table (Table 2) with all the methods used and the metrics results obtained.
The results highlighted in yellow represent models trained and tested on the same dataset, while those highlighted in green indicate models trained on one dataset and tested on another. For instance, the first row shows the random forest model trained on the MAIN dataset, with its performance evaluated on both the MAIN dataset (left side of the table) and the RSDD dataset (right side of the table). Notably, as expected, the accuracy score for the model tested on the same dataset (0.889) is significantly higher than when tested on the other dataset (0.501). In addition, accuracy scores and F1 scores in red highlight that MentalBERT model, with respect to BERT model, has higher performances in all the cases studied.

methods/metrics		test data: MAIN				test data: RSDD			
		accuracy	prec	rec	F1 score	accuracy	prec	rec	F1 score
train data: MAIN	random forest	0,889	0,915	0,855	0,884	0,501	0,235	0,003	0,006
	SVM	0,853	0,840	0,866	0,853	0,361	0,335	0,294	0,313
	BERT	0,856	0,830	0,894	0,860	0,474	0,438	0,216	0,290
	MentalBERT	0,858	0,860	0,864	0,862	0,542	0,597	0,236	0,338
train data: RSDD	random forest	0,498	0,167	0,001	0,002	0,899	0,935	0,855	0,909
	SVM	0,499	0,220	0,001	0,002	0,854	0,840	0,886	0,862
	BERT	0,526	0,514	0,960	0,660	0,980	0,988	0,971	0,980
	MentalBERT	0,531	0,517	0,954	0,671	0,984	0,985	0,983	0,984

Table 2.

5. Conclusions

Our results yield several important insights. Firstly, all the models we implemented achieved relatively high scores when we performed both training and testing on the same dataset. However, their performance deteriorated significantly during cross-testing, rendering them practically unusable in those scenarios. Interestingly, despite being considerably smaller than the MAIN dataset, the RSDD dataset consistently outperformed it. This discrepancy can be attributed to the differing comment collection strategies. As previously explained, the RSDD dataset is user-based, encompassing comments from a variety of subreddits, while the MAIN dataset is confined to specific subreddits focused on depression. This broader range of sources in the RSDD dataset likely contributes to its superior performance, as it benefits from a more diverse set of comments, aiding in more effective classification.

In addition, our findings provide further evidence that context is crucial. As expected, MentalBERT model consistently outperforms BERT model across all cases. This superior performance is due to MentalBERT being pre trained not only on general English sentences from unpublished books but also on an extensive dataset of mental health-related posts collected from Reddit. Consequently, MentalBERT is better suited to our specific task, demonstrating the importance of contextual relevance in model training.

References

[1] Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. In npj Digital Medicine (Vol. 5, Issue 1). Nature Research. https://doi.org/10.1038/s41746-022-00589-7.

[2] Yan, W. J., Ruan, Q. N., & Jiang, K. (2023). Challenges for Artificial Intelligence in Recognizing Mental Disorders. Diagnostics, 13(1). https://doi.org/10.3390/DIAGNOSTICS13010002.

[3] Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do Models of Mental Health Based on Social Media Data Generalize?. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3774–3788, Online. Association for Computational Linguistics.

[4] Kayalvizhi S and Thenmozhi D. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. In arXiv preprint arXiv:2202.03047. 5. American Psychiatric Association. (2022). Depressive disorders. In Diagnostic and statistical manual of mental disorders (5th ed., text rev.)

[5] American Psychiatric Association. (2022). Depressive disorders. In Diagnostic and statistical manual of mental disorders (5th ed., text rev.)

[6] Yates, A., Cohan, A., & Goharian, N. (2017). Depression and Self-Harm Risk Assessment in Online Forums. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 310-315).

[7] Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2016). Towards Universal Paraphrastic Sentence Embeddings. In Proceedings of the International Conference on Learning Representations (ICLR).

Work plan

1. **Paper and academic articles research**

   - We decided to take a rigorous approach while developing our project; therefore, every step or decision we took had to be justified by some already existing research. This implies that before coming up with this project, we had to discard many possible ideas due to a lack of precise information in the current literature or because many experimental designs required a skill set too advanced for our current knowledge.

     [*Responsible: Guarini Tomas, with the partecipations of Gorni Matteo, Morandi Arianna, Nemanja Ilic & Spagnolo Beatrice*]

2. **Data Processing Pipeline development**

   - We created a python script to correctly perform different data processing tasks and implement a word2vec continuous bag of words model.

     [*Responsible: Gorni Matteo, with the partecipations of Guarini Tomas, Morandi Arianna & Spagnolo Beatrice*]

   - SMOTE Augmentation: Correctly implement the Synthetic Minority Over-sampling Technique (SMOTE) to address imbalanced datasets.

     [*Responsible: Nemanja Ilic*]

3. **Machine Learning Models Implementation**

   - Build and evaluate three different machine learning models to test our hypothesis.

     [*Responsible: Davide Nieto*]

4. **Deep Learning Models Implementation**

   - Test our hypothesis using two state-of-the-art transformer models, BERT and MentalBERT.

     [*Responsible: Beatrice Spagnolo, with the partecipations of Gorni Matteo, Guarini Tomas, Nemanja Ilic & Morandi Arianna*]

   - Tune and evaluate these models on our two datasets.

     [*Responsible: Guarini Tomas, with the partecipations of Gorni Matteo, Morandi Arianna & Spagnolo Beatrice*]

   - Assess the performance by testing on a dataset different from the one used for training.

     [*Responsible: Morandi Arianna, with the partecipations of Gorni Matteo, Guarini Tomas & Spagnolo Beatrice*]

5. **Project report**

   - Writing project report.

     [*Responsible: Gorni Matteo, with the partecipations of Guarini Tomas, Morandi Arianna & Spagnolo Beatrice*]

6. **Presentation**

   - Creating slides to present the project.

     [*Responsible: Morandi Arianna, with the partecipations of Gorni Matteo, Guarini Tomas & Spagnolo Beatrice*]