Social Media Posts as a Window into Mental Health: A Machine Learning Approach

Aadil Ganie (aadilganiganie@gmail.com)
University of Miskolc

Samad Dadvandipour
University of Miskolc

Research Article

Keywords:

Posted Date: January 31st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2518185/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Mental health is a crucial factor influencing the overall well-being of humans, which has gained significant attention in recent times due to the high prevalence of mental health disorders and their detrimental effects on individuals and society. In an attempt to tackle this pressing issue, researchers have explored the possibility of using the copious amounts of data available on social media platforms to predict and classify mental health status. In our study, we analyzed three datasets: the first one comprising 7 classes (depression, anxiety, autism, mental health, schizophrenia, BPD, and bipolar), the second dataset comprising 2 classes (positive and negative), and the third dataset comprising 2 classes (suicide and non-suicide). The final dataset included 14 classes, with 7 belonging to the non-suicidal subset and 7 belonging to the suicidal subset. We employed logistic regression, support vector machines, and multinomial naive Bayes for classification and prediction, and evaluated the performance of our models using receiver operating characteristic (ROC) curves and confusion matrices. The logistic regression model outperformed the other models, achieving an accuracy of 80%. Our models have been deployed using streamlit, providing a user-friendly interface for predicting mental health status and risk for suicidal ideation. If the prediction of the social media post falls within the suicide subset class, a chatbot (GPT2) will be activated in an effort to engage the individual with suicidal ideation and reduce the likelihood of suicide. Our research serves as a helpful tool for mental health professionals and has the potential to be extended to other platforms, addressing the urgent need to detect and address mental health issues and suicidal ideation.

Introduction

Mental health, a subset of psychological well-being, is a complex and multifaceted aspect of human existence that encompasses an individual's cognitive, emotional, and behavioural states, as well as their subjective experiences and perceptions of their surroundings. It is influenced by a wide range of social, biological, and environmental factors, and is essential for overall physical and psychological functioning. As such, mental health is a critical determinant of human well-being and requires ongoing attention and care.[1]. The prevalence of mental health disorders and the negative impact they can have on individuals and society make it imperative to develop effective strategies for detecting and addressing these issues. In this vein, researchers have sought to leverage the vast amount of data available on social media platforms as a means of predicting and classifying mental health status.[2][3]. One promising approach involves the use of deep learning algorithms to analyze the content of social media posts and extract relevant features that can be used to predict mental health status[4].

In this research, we employed machine learning techniques to build models that can predict the mental health status of individuals based on their social media posts. Our dataset consists of posts from individuals with various mental health conditions, including anxiety, borderline personality disorder, autism, bipolar disorder, depression, and schizophrenia, as well as posts from individuals with normal mental health. To assess the risk of suicidal thoughts in our dataset, we used another dataset containing posts from individuals with suicidal and non-suicidal thoughts. We used the suicidal class only to check
the severity of the mental health of a person. We calculated the cosine similarity between the mental health conditions in our dataset and the suicidal class in this dataset, we prefix the label with 'suicide' if the cosine similarity is greater than .6, e.g, the rows of anxiety class for which the cosine value is greater than .6 become 'suicideanxiety' and so on.[5]. This allowed us to classify the posts in our dataset as either suicidal or non-suicidal, initially, we have 8 classes but after calculating the cosine similarity and prefixing the data in each class, the number of classes in the final dataset doubled, and the dataset we prepared from different sources contains a lot of noise we cleaned the data using techniques like stopword removal, stemming, expanding contractions, spell check etc. We then used logistic regression, support vector machines [6], and multinomial naive Bayes [7], for classification, and evaluated the performance of our models using the receiver operating characteristic (ROC) curve and confusion matrix.

According to mental health statistics 2022 around 13% [8] of the world's population is facing mental health issues. In Hungary, Hungarian men can expect good health until the age of 60 and Hungarian women until age 62, In contrast to Ireland, Sweden and Malta, both men and women are in good health until age 70 or more according to research conducted by center for economic and regional studies [9]. According to data from Kepios, there were 4.74 billion social media users worldwide in October 2022, representing 59.3% of the global population. In the past year, 190 million new users have joined social media, representing an annualized growth rate of 4.2%, or an average of 6 new users per second. These figures suggest that more than 90% of internet users now use social media on a monthly basis [10]. The proliferation of social media usage among a significant proportion of the global population has made it a prime source of data for researchers, particularly those working with language models. Leveraging this trend, we sought to investigate the potential utility of social media posts for predicting users' mental health. The first dataset, which was collected from Reddit by [11], comprised 7 classes (depression, anxiety, autism, mental health, schizophrenia, BPD, and bipolar). However, this dataset was insufficient for determining whether individuals suffering from mental health issues also harbored suicidal ideation or whether they were normal. Therefore, we supplemented the data with a "positive" class from the second dataset, resulting in a total of 8 classes: depression, anxiety, autism, mental health, schizophrenia, BPD, bipolar, and normal.

To further our research, we utilized another dataset containing two classes (suicide and non-suicide) and calculated the cosine similarity between the 8 classes described above and the suicide class in order to predict suicidal ideation. This resulted in the number of classes being doubled (further explanation can be found in the results and discussion section). After training a machine learning model on the 16 resulting classes, we implemented an additional feature: a chatbot utilizing GPT2. If the prediction of a social media post falls within the "suicide" subset, the chatbot will be activated and, based on the prompt, generate responses resembling human conversation. The aim of this chatbot is to mitigate the risk of suicidal ideation by engaging the user in conversation, while simultaneously using a sentiment analyzer and key words to monitor the chatbot. If the sentiment score of the user's questions is "neg" < .5 or if there are words such as "suicide" or "crisis," we recommend that the user seek medical attention and provide them with national helplines for mental health or call for emergency services.
The deployment of our models using streamlit provides a user-friendly interface through which individuals can input their own social media posts and receive a prediction of their mental health status and level of risk for suicidal thoughts. This research serves as an assistant to psychiatrists and other mental health professionals, helping to detect mental health issues at an early stage and identify individuals at risk for suicidal thoughts [12]. It also has the potential to be expanded to other platforms, addressing the pressing need to address mental health issues and suicidal thoughts in our society. There are several areas for future research that could build upon the work presented here. One is to incorporate additional data sources, such as demographic information or physiological data [13], to enhance the predictive power of our models. Another interesting avenue for future research is to study the ethical and social implications of using social media data to predict mental health status [14]. This could include examining the privacy concerns surrounding the collection and use of such data, as well as exploring the potential biases that may be introduced by using social media posts as a sole source of information.

Finally, it would be valuable to explore the potential applications of our models beyond the domain of mental health [15]. For example, our models could be used to predict other health-related outcomes, such as the risk of developing chronic diseases or the likelihood of engaging in unhealthy behaviours [14]. Overall, there is a wealth of potential research directions that could build upon the work presented in this study, and we believe that our approach has the potential to make a significant impact in the field of mental health and beyond.

Literature Review

Mental health is a paramount aspect of human well-being that has garnered increasing attention in recent years due to the high prevalence of mental health disorders and their detrimental effects on individuals and society [1]. Social media platforms, such as Twitter and Facebook, have become increasingly popular for individuals to express their thoughts and emotions, making them a valuable source of data for predicting and classifying mental health status [2] [16] [17] [18] [19] [20] [21] [22]. Since 2013, research has been able to identify and evaluate the existence of clinically significant mental health conditions such as major depression [2][23][24], suicidality [25][26][27], eating disorders [28][29], and schizophrenia [30], through the utilization of various methodologies and techniques such as machine learning, artificial intelligence, natural language processing, and human-computer interaction. The utilization of such techniques have been implemented by various social media platforms, such as Facebook, as a means of preventing suicide [31][32] and improving mental health outcomes. This has given rise to a burgeoning field known as "digital psychiatry" [33] which aims to harness these predictive signals for the betterment of mental health service provision. It has been demonstrated through various studies that individuals suffering from a plethora of mental disorders such as depression, psychosis, and other severe mental illnesses, tend to utilize social media platforms at rates comparable to the general population. This usage ranges from approximately 70% among middle-aged and older individuals to as high as 97% among younger individuals [34] [35][36] . Furthermore, exploratory research has revealed that a significant number of individuals with mental illness tend to utilize social media as a medium to share
their personal experiences, gain information pertaining to their mental health and treatment options, and provide and receive support from others facing similar mental health challenges [37] [38].

Numerous studies have utilized machine learning techniques to classify mental health status based on social media data. One study employed support vector machines and random forests to classify depression and anxiety in Twitter users, achieving an accuracy of 73% [2]. Another study employed a combination of natural language processing and machine learning techniques to classify depression, anxiety, and stress in social media users, achieving an accuracy of 72.5% [16]. Other studies have employed decision tree algorithms [39] convolutional neural networks [40][41], and deep learning approaches [42][43] to classify mental health conditions in social media data. Other studies have focused on predicting suicidal ideation on social media platforms. One study utilized a combination of machine learning and network analysis to identify individuals at risk of suicidal ideation on Twitter, achieving an accuracy of 83% [2]. Another study employed a deep learning approach to predict suicidal ideation in Reddit posts, achieving an accuracy of 82.8% [44]. Other studies have used machine learning techniques to predict suicidal behavior in social media data [45][46] and to identify individuals at risk of suicide [47] [48][49].

In this study, we utilized advanced machine learning techniques to predict the mental health status of individuals based on their social media posts. We classified the posts in our dataset as either suicidal or non-suicidal using another dataset containing posts from individuals with suicidal thoughts and the cosine similarity measure. We applied logistic regression, support vector machines.

**Results and Discussions**

For this study, we analyzed three datasets which are mentioned below with a description. This section has been divided into several sub-sections which are as follows:

1. Data Preparation and preprocessing
2. Insights from data using exploratory data analysis (EDA)
3. Model building, testing and evaluation
4. Model deployment using streamlit API
5. Data preparation and preprocessing: The first dataset has been studied in [1], but this study only considered 'Anxiety', 'BPD', 'autism', 'bipolar', 'depression', 'mental health', and 'schizophrenia', this study seems to be incomplete as normal/positive sentiment is not included. In this study we added the positive sentiment from the famous dataset [2], One more thing that we add to this data is suicide and non-suicide dataset, but from this dataset we took the suicide part only, to calculate the severity of mental illness by mapping the above mental cases into suicide data and calculate cosine similarity. Initial data distribution of the above three datasets.

The data in the second dataset is highly imbalanced, so we performed both under-sampling and over-sampling, since the depression class is the one with the highest number of counts we downsample it to 100k.
After balancing the data, it needs to be cleaned for analysis. The data went through the cleaning processes like stopword removal, stemming, expanding contractions, removing noise etc. We select the suicide class from our suicide non-suicide dataset and convert it into a long string for calculating the cosine similarity.

2. Insights into the data using EDA: The reason we calculate the cosine similarity between the different classes in the first dataset with suicide is, we want to check the severity level of different mental health classes like BPD, autism, depression etc. So, we map the word from the suicide string with this class and if the cosine similarity is greater than .6 added suicide to the label.

The number of classes doubled from 8 to 16, we believe that there are some words in normal classes which are present in other classes as well and that is not exceptional, however, we can increase the threshold so that only suicide labels can be added to other classes except normal.

From the above heatmap, we concluded that the classes ‘suicideAnxiety’, ‘suicideBPD’, ‘suicidedepression’, ‘suicidementalhealth’, ‘suicideschizophrenia’, ‘suicidebipolar’, and ‘suicideautism’ are very severe i.e., if the social media post falls under one of the classes, then that person is prone to suicide as compared to other classes, from the heatmap above it seems the threshold of .6 works very well but you can fine-tune it. Most frequent words for some classes have been plotted using wordcloud as below.

Model Development: Logistic regression has been proven very efficient for classification tasks; this study also used the logistic regression algorithm for prediction. The feature extraction technique used for this study is both count vectorizer and tfidf, we also experimented with Randomforest and support vector machine. The model has been tested and evaluated, we choose ROC, and confusion matrix as a measure for evaluating the performance of the models. The reason for choosing ROC as an evaluation measure is, The ROC curve is a useful evaluation measure because it is insensitive to the class distribution and can be used to compare classifiers with different distributions of positive and negative instances. It is also useful because it provides a visual representation of the trade-off between TPR and FPR, allowing users to easily compare the performance of different classifiers.

True positive rate (TPR) = True positives / (True positives + False negatives)
False positive rate (FPR) = False positives / (False positives + True negatives)

A. Logistic Regression Results:

The true positive rate is more than 90% for every class except for ‘mentalhealth’ class. This suggests our model has performed well but there is room for improvement.

The logistic regression algorithm performed well for some classes like ‘Normal’, ‘depression’, and ‘suicidedepression’, however, there is a high false positive rate in ‘suicideschizophrenia’, ‘suicidemental’, ‘suicideautism’, and ‘schizophrenia’ as model wrongly classified almost 24%, 24%, 30%, and 26% into ‘mentalhealth’, ‘suicidebipolar’, ‘autism’, and ‘depression’ classes respectively. One thing to note here is
although logistic regression misclassified some data almost all misclassified data has been classified into the particular subset of classes which belongs to suicide except for the 'suicideautism' which has been classified into the 'autism' class. We can say that recall for each subset class i.e., suicide subset and non-suicide subset is very high, from this we can conclude that although we may not be able to correctly classify the mental health of a person into particular classes we can correctly classify whether the mental health of a person is suicidal or non-suicidal.

MultinomialNB:

From the ROC and confusion matrix of multinomialNB, we conclude that although the model did not perform very especially for 'suicideautism' class. The same evidence can be found in ROC as well, most of the value's range between 70 to 80.

Conclusion

In conclusion, this study has demonstrated the potential of using advanced machine learning techniques to predict the mental health status of individuals based on their social media posts. Our research has demonstrated the feasibility of using social media data to predict mental health status and risk for suicidal ideation. By leveraging cosine similarity, we were able to expand the number of classes in our dataset and improve the performance of our machine learning model. Additionally, the implementation of a GPT2 chatbot has shown promise as a means of engaging individuals with suicidal ideation and mitigating the risk of self-harm. Our results show that logistic regression outperformed support vector machines and multinomial naive Bayes in predicting mental health status, achieving an accuracy of 80%. Deploying our models using streamlit provided a user-friendly interface for predicting mental health status and risk for suicidal ideation. This research serves as an assistant to mental health professionals and has the potential to be extended to other platforms, addressing the pressing need to detect and address mental health issues and suicidal ideation. Future research directions could include the exploration of other machine learning algorithms, the incorporation of additional data sources, the use of transfer learning techniques, and the examination of the ethical and social implications of using social media data to predict mental health status. Overall, our approach has the potential to make a significant impact in the field of mental health and beyond.

References


Figures

Figure 1

Emoticon dataset [50], 0 indicates negative sentiment and 4 Positive sentiment
Figure 2

Mental Health Dataset [11]

![Figure 2](image1)

Figure 3

Suicide_nonsuicide dataset [3]

![Figure 3](image2)

Figure 4

Data after adding Normal sentiment

![Figure 4](image3)
Figure 5

Downsampling of depression class to 100k

Figure 6

Oversampling of minority class to majority class, almost balanced dataset
**Figure 7**

Figure legend not available with this version.

**Figure 8**

Heatmap Subreddit vs Cosine Similarity
Figure 9

Figure legend not available with this version.

Figure 10

Roc logistic regression
Figure 11

CM Logistic regression

Figure 12

MutinomialNB Roc
Figure 13

MultinomialNB Confusion Matrix