A Machine Learning Approach to Analyze Mental Health from Reddit Posts

Chapter · June 2022

DOI: 10.1007/978-981-16-8739-6_33

Smriti Nayak
Silicon Institute of Technology
3 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Debolina Mahapatra
National Institute of Technology Patna
7 PUBLICATIONS **33** CITATIONS
SEE PROFILE

Riddhi Chatterjee
12 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Shantipriya Parida
Silo AI
97 PUBLICATIONS **444** CITATIONS
SEE PROFILE

Chapter 33
A Machine Learning Approach to Analyze Mental Health from Reddit Posts

Smriti Nayak, Debolina Mahapatra, Riddhi Chatterjee, Shantipriya Parida, and Satya Ranjan Dash

Abstract Reddit is a platform with a heavy focus on its community forums and hence is comparatively unique from other social media platforms. It is divided into sub-Reddits, resulting in distinct topic-specific communities. The convenience of expressing thoughts, a flexibility of describing emotions, inter-operability of using jargon, the security of user identity makes Reddit forums replete with mental health-relevant data. Timely diagnosis and detection of early symptoms are one of the main challenges of several mental health conditions for which they have been affecting millions of people across the globe. In this paper, we use a dataset collected from Reddit, containing posts from different sub-Reddits, to extract and interpret meaningful insights using natural language processing techniques followed by supervised machine learning algorithms to build a predictive model to analyze different states of mental health. The paper aims to discover how a user's psychology is evident from the language used, which can be instrumental in identifying early symptoms in vulnerable groups. This work presents a comparative analysis of two popular feature engineering techniques along with commonly used classification algorithms.

S. Nayak
Silicon Institute of Technology, Bhubaneshwar, India

D. Mahapatra
National Institute of Technology Patna, Patna, India

R. Chatterjee
Heritage Institute of Technology, Kolkata, India

S. Parida
Idiap Research Institute, Martigny, Switzerland
e-mail: shantipriya.parida@idiap.ch

S. R. Dash (✉)
KIIT University, Bhubaneswar, India
e-mail: sdashfca@kiit.ac.in

S. Dehuri et al. (eds.), *Biologically Inspired Techniques in Many Criteria Decision Making*,
Smart Innovation, Systems and Technologies 271,
https://doi.org/10.1007/978-981-16-8739-6_33

33.1 Introduction

The evolution in social networking has augmented opportunities for people to communicate on the Internet. The number of social media users increased from 2.31 billion users in 2016 to 4.33 billion users in 2021. The global social penetration rate has surged by 87% in the last five years. Social media allows us to connect with people, share and discover ideas, find and review businesses, trade goods, services, shop online and even converse anonymously. Social media is widely perceived to have provided humans a safe space to express themselves. Social media platforms furnish a peer support network, facilitate social interaction and promote engagement and retention in treatment and therapy. Reddit [1] is a network of communities based on people's interests. It has been graded as the 18th-most-visited Web site in the world. It is one of the most popular social networks with over 100,000 active communities and 52 million daily users. Unlike Facebook and Twitter, Reddit does not restrict its users on the length of their posts. Reddit segregates its contents via sub-Reddits that are dedicated to a specific topic, including various communities, thus making it a very valuable information pool.

Mental health ailments are broadly defined as health conditions that alter an individual's thoughts, perceptions, emotional states and behaviors causing profound distress, thus disrupting the effective functioning of their personal, professional and social lives. According to a World Health Organization report, 450 million people around the world experience a mental illness in their lifetime. The report also estimates that roughly 1 in 4 people is likely to have a psychiatric disorder or a neurological disorder [2]. Unfortunately, mental health conditions continue to be underreported and under-diagnosed [3]. This increases the risk of early symptoms being unacknowledged and is further exacerbated due to lack of sufficient awareness, negative attitudes and prevalent social stigma which act as barriers to care. Approximately, two-thirds of the individuals do not take any professional assistance. Certain studies have found that a significant number of people use social media as a medium to share personal thoughts and experiences, vent out their negative emotions, search for information about mental health disorders and treatment/therapy options. These people give and receive support from others, often anonymously, facing similar challenges.

Now considered as a new medical research paradigm, artificial intelligence and natural language processing have been useful in detecting and predicting mental health conditions from social media data. This work emphasizes the role of social media as a potentially feasible intervention platform for providing support to people with any mental ailment, boosting engagement and retention in care and augmenting existing mental health services, all while keeping safety in mind [4]. In this paper, we have summarized the current research on the use of social media data for the mental health analysis of individuals. Various machine learning models have been implemented to analyze and classify Reddit posts into related categories of psychiatric conditions.

The paper aims to serve as a preliminary study in this domain which can be further expanded by using more sophisticated algorithms. It is organized as follows: Sect. 2

provides the literature review, and Sect. 3 states the methodology along with the experimental results and analysis. Finally, the paper ends with the conclusion, and future work is given in Sect. 4.

33.2 Literature Review

Natural language processing (NLP), an interdisciplinary branch born out of the trinity of computational linguistics, computer science and artificial intelligence, is committed to the comprehension, interpretation and analysis of human language. The overarching objectives of mental health applications are to aid in understanding mental health, act as a channel of support and promote the well-being of individuals. These are realized through textual mining, sentiment detection, sentiment analysis, emotion classification and building technological interventions [5]. As NLP techniques attempt to bridge and facilitate human–computer interactions, it, therefore, becomes a veritable tool for achieving the said objectives.

Seal et al. [6] constructed an emotion detection method that is comprised of text preprocessing, keyword extraction from sentences using POS-tagging and keyword analysis to categorize the revealed emotional affinity. Herzig et al. [7] introduced an ensemble methodology of bag of words (BOW) and word embedding-based classifier, to perform emotion detection. In the latter classifier, document representations that were exercised are continuous bag of words (CBOW) and term frequency-inverse document frequency (TF-IDF). The field of machine learning (ML) has considerably augmented the understanding and scope of research in the vast domain of mental health. The usage of several computational and statistical methods to build better systems for diagnosis, prognosis and treatment of mental ailment symptoms has been accelerated by ML techniques [8]. These techniques are employed to extract psychological insights from predominantly four arenas of data pools, namely data, sensors, structured data and multi-modal technology interactions [9].

A significant portion of the literature is focused on the early detection and monitoring of indicators of depression. Zhou et al. [10] built a model using computer vision and data mining techniques to project a continuous, multifaceted view of one's mental health. Support vector machine (SVM) classifier and logistic regression have been utilized for emotion inference in their work. Fatima et al. [11] identified depressive posts and the degree of depression from user-generated content by investigating the linguistic style and associated sentiment of the posts. Random forest (RF) classifier was utilized for the aforementioned purpose. In the context of social media analysis, a novel framework to detect users prone to depression via conducting a temporal analysis of eight basic emotions exhibited by Twitter posts was performed by Chen et al. [12]. Their methodology focused on examining the textual contents of tweets by employing SVM and RF classifiers with optimized parameters. The classification accuracy of detecting emotions was augmented by 8% for the SVM classifier and 3% for the RF classifier when temporal measures of emotions were included in the prediction task.

Similarly, Suhasini et al. [13] employed Naive Bayes (NB) and k-nearest neighbor algorithm (KNN) to execute textual mining of emotions of Twitter posts and label them into four distinct emotional categories, by exploring the characteristics of the tweets. The NB algorithm performed better than KNN, showing an accuracy of 72.6%. Gaind et al. [14] devised a composition scheme of NLP techniques like emotion-words set (EWS) and ML classification algorithms like sequential minimal optimization (SMO) algorithm and J48 algorithm to detect, classify and quantify Twitter posts. They presented six classification categories of emotions, namely happiness, sadness, surprise, anger, disgust and fear.

Saha et al. [15] attempted to infer expressions of stress from Reddit posts of college communities using a stress classifier. They adopted a transfer learning approach to build a binary SVM classifier that scrutinized and determined posts as either “high stress” or “low stress”. The classifier achieved an accuracy of 82%. Further, suicidal ideation detection (SID) methods, from a computational approach, can greatly benefit from engaging ML classifiers. This is illustrated by a study conducted by Pestian et al. [16]. Their robust ML classifier, based on unweighted SVM, could compute the risk of suicide for individuals before them showing acute suicidal tendencies. By extracting structural, lexical and semantic information from manually labeled suicide notes, the authors in [17] built an automatic emotion detection system to determine 15 classes of emotions.

A noteworthy amount of feature engineering is requisite for conventional ML models to show optimal performance. This step of preprocessing impedes their efficiency, owing to being a tedious and resource-consuming process. State-of-the-art deep learning (DL) algorithms aim to directly map the input data features through a multi-layer network structure [18]. This inevitably leads to models showing superior performance.

Ragheb et al. [19] applied deep transfer learning for sentiment analysis and detection in textual conversations. Self-attention mechanisms and turn-based conversation modeling were used. Gkotsis et al. [17] used a convolutional neural network (CNN) that demonstrated a good accuracy of 91.08% to determine Reddit posts in the domain of mental health discussions and further classified those posts, according to the type of symptoms exhibited, with an accuracy of 71.37%. Sekulic et al. [20] used a hierarchical attention network (HAN) model [21] to predict if a social media user has a specific mental condition, out of nine others that are prevalent. Dheeraj et al. [22] presented a mechanism that inspects emotions from psychiatric texts to discern negative emotions. Multi-head attention with bidirectional long short-term memory and convolutional neural network (MHA-BCNN) was utilized for the said purpose. Their choice of model was influenced due to the selection of long text sequences. Kim et al. [23] collated sub-Reddit posts from the mental health community and developed a CNN model to accurately ascertain a user’s potential mental state and the kind of mental ailment one could be facing, out of the category of depression, anxiety, bipolar disorder, borderline personality disorder (BPD), schizophrenia and autism.

Table 33.1 Sample of the dataset

Title	Text	Sub-Reddit
Exposure does not work!	“I have struggled with social anxiety from childhood and the main. …”	Anxiety
Paranoia	“Does anyone here deal with paranoia on a daily basis?….”	BPD
Depression coming back?	“So the thing is that for some years I’ve been on and off with self-harm …”	Depression
I’m a sociopath…	“My therapist said I have sociopathic tendencies so does that mean I ….”	Mental health
Unwanted obsessive and upsetting thoughts?	“I’ve been stressing out so much lately over things that haven’t happened …”	Bipolar
Auditory hallucinations (non-verbal)	“I’ve realized lately that I have auditory hallucinations …..”	Schizophrenia
Am I autistic?	“I get super anxiety and am uncomfortable even going into the store…”	Autism

33.3 Methodology

33.3.1 Data Collection

We have used the dataset collected by Kim et al. [11] in their work. The dataset comprises posts under the following six sub-Reddits associated with the subject of mental health: r/depression (258,495 posts), r/Anxiety (86,243 posts), r/bipolar (41,493 posts), r/BPD (38,216 posts), r/schizophrenia (17,506 posts), r/autism (7143 posts) as well as r/mental health (39,373 posts).

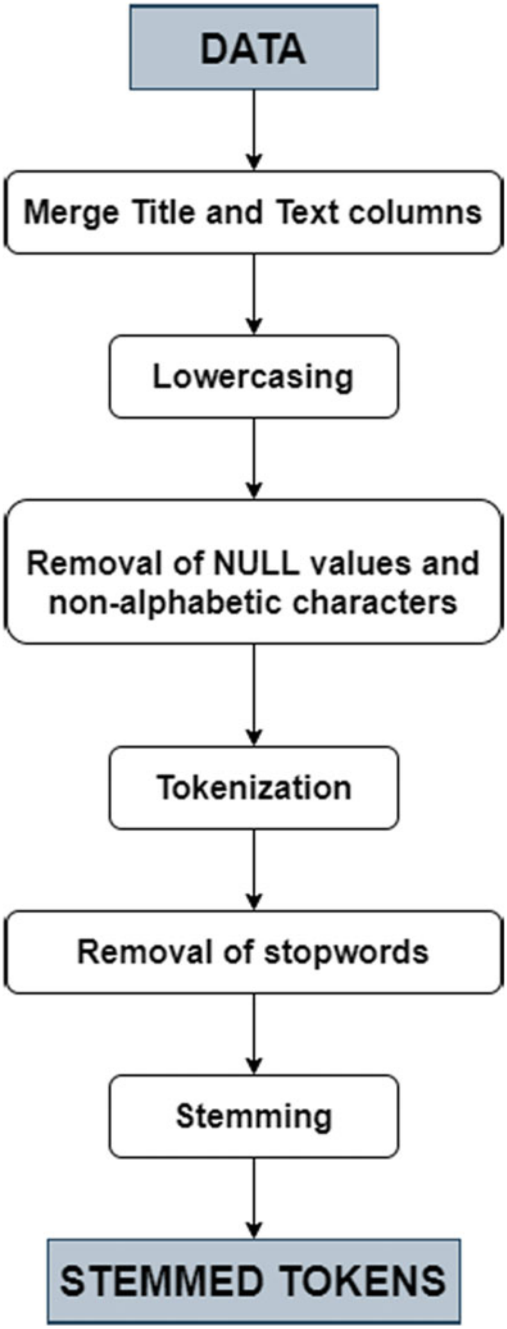
A brief snapshot of the dataset has been provided in Table 33.1.

33.3.2 Preprocessing

The dataset originally contains three columns: title of the post, text content of the post and the sub-Reddit category associated with it. The raw data collected must be preprocessed before training, evaluating and using machine learning models. We have used the following preprocessing pipeline as given in Fig. 33.1.

Initially, the title and the text columns were concatenated, followed by lowercasing to normalize the text in each of the Reddit posts. The unnecessary punctuation marks, white spaces, null values and other non-alphabetic characters were removed from each post. Tokenization separates words in a sentence via space characters. Certain words like *a*, *an*, *the*, *is*, *are*, etc., are commonly used words that do not carry any significant importance in the text. Hence, all the stop-words present are removed from

Fig. 33.1 Text preprocessing pipeline


DATA
Merge Title and Text columns
Lowercasing
Removal of NULL values and non-alphabetic characters
Tokenization
Removal of stopwords
Stemming
STEMMED TOKENS

the post. This step is followed by stemming which is a process to remove morphological affixes from words. It converts the words to their root form and decreases the number of word corpus. For example, the stemmed form of eating, eats and eaten is “eat”. The natural language toolkit (NLTK) [24] is used to perform various preprocessing steps.

33.3.3 Feature Extraction

For classifying our posts into their respective mental health category, we first need to extract features from them. Text data cannot be directly given to a machine learning model as they accept numeric feature vectors as input. It is important to convert textual data into vector representation before feeding them into a machine learning model. The features are numerical attributes that exhibit abstraction of the raw data at the level of whether (or to what extent) a particular characteristic is present for a given post. This entails using some NLP techniques to process the raw data, convert text to numbers to generate useful features in the form of vectors. In this work, we have used bag of words (BoW) and term frequency-inverse document frequency (TF-IDF) for feature engineering.

Bag of Words

The simplest encoding of text into vectors is achieved by using bag of words. In this approach, a single feature vector is built using all of the terms in the vocabulary obtained by tokenizing the sentences in the documents. Each text document (or post) present in the dataset is represented uniquely by converting it into a feature vector representation. Each word is treated as a distinct property. As a result, the number of features in the vocabulary is equal to the number of unique words.

Every Reddit post in the dataset is considered a sample or record. If the word appears in the sample, it stores the “frequency” of the word, which is regarded as the feature, and if the word does not appear, it is zero. A word is represented by each column of a vector. In this process of word encoding, the representation of the word takes precedence over the order of the words. This approach returns an encoded vector with a total vocabulary length and an integer count of how many times each word appears in the document.

Term Frequency-Inverse Document Frequency

TF-IDF is an acronym for *term frequency-inverse document frequency* and is used to determine the relevance of a term in a particular corpus by calculating the weight of each term. Term frequency is used to calculate the occurrence of a particular term in the entire corpus. The frequency value of a term is normalized by dividing it by the total number of words in the corpus. Document frequency is a metric to assess the significance of a document in the context of the entire corpus.

33.3.4 Experimental Results

Machine learning approaches can open new avenues for learning human behavior patterns, recognizing early symptoms of mental health disorders and risk factors, making illness progression predictions and customizing and improving therapies. Each Reddit post in the dataset has a label associated with it which determines the related mental health condition of the individual. By using machine learning algorithms on this dataset, the predictive model learns different associations between the post and its corresponding label.

The dataset underwent two types of feature extraction—bag of words and term frequency-inverse document frequency. Following this stage, feature vectors of size 2000 were obtained, respectively. This data was split into 80% for training and 20% for testing. Machine learning algorithms like MultinomialNB, decision tree, random forest classifier, logistic regression and ensemble techniques like AdaBoost and XGBoost were applied to both the feature sets.

The accuracy of the machine learning models is used to evaluate the performance of each of the feature engineering techniques used, as shown in Fig. 33.2 and stated below in Table 33.2.

It can be inferred that multinomial logistic regression, when implemented using TF-IDF feature vector, provides the highest accuracy of 77% and performs better than the other machine learning models. Among the ensemble techniques used, XGBoost exhibits a good performance with nearly 76% on both feature sets.

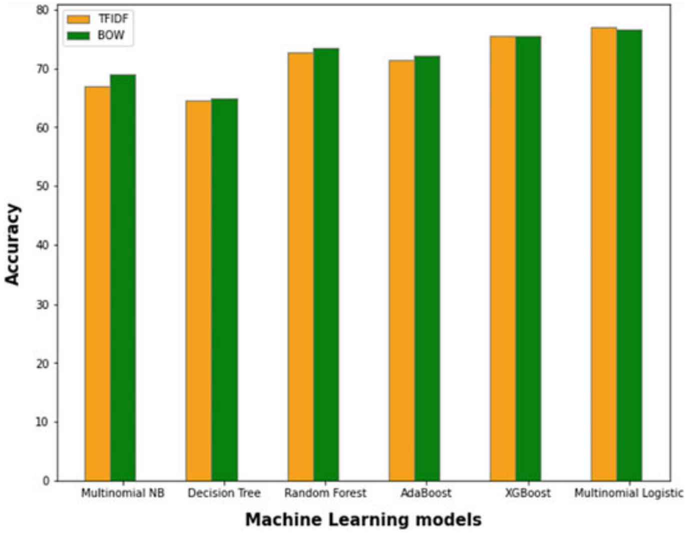


Fig. 33.2 Graph for performance analysis of machine learning models

Table 33.2 Performance analysis of machine learning models

Machine learning model	TF-IDF features	BoW features
Multinomial Naïve Bayes	0.67	0.69
Decision tree	0.646	0.65
Random forest classifier	0.728	0.735
AdaBoost classifier	0.714	0.7225
XGBoost classifier	0.7547	0.7553
Multinomial logistic regression	0.7710	0.7666

33.4 Conclusions and Future Work

Reddit has several concentrated and structured mental health forums which provides its users an opportunity to anonymously share their experiences and engage in peer-to-peer support groups. The main objective of this paper is to use these posts and apply NLP techniques to build a predictive machine learning model targeted to identify a possible mental health condition. Previous works based on this dataset have focused on binary classification task for detection of each mental health category individually. In this work, we have treated it as a multi-class classification problem and have reported the performance of machine learning models with feature engineering which will serve as a preliminary study on using Reddit data for analyzing mental health.

Future work shall be done on building more complex and efficient predictive models that can help us to resolve the imbalanced data problem. Deep neural network-based techniques can be applied to obtain refined results using automated feature extraction supported by these models. The insights gained from our work will help researchers build better predictive models on this dataset using some more sophisticated approach.

References

1. Reddit. https://www.reddit.com/
2. The World health report: 2001: Mental health: new understanding, new hope (2001). World Health Organization: Institutional Repository for Information Security. https://apps.who.int/iris/handle/10665/42390
3. Ritchie, H.: Global mental health: five key insights which emerge from the data. Our World in Data (2018). https://ourworldindata.org/global-mental-health
4. Naslund, J.A., Bondre, A., Torous, J., Aschbrener, K.A.: Social media and mental health: benefits, risks, and opportunities for research and practice. J. Technol. Behav. Sci. **5**, 245–257 (2020)
5. Calvo, R.A., Milne, D.N., Hussain, M.S., Christensen, H.: Natural language processing in mental health applications using non-clinical texts. Nat. Lang. Eng. **23**(5), 649–685 (2017)

6. Seal, D., Roy, U.K., Basak, R.: Sentence-level emotion detection from text based on semantic rules. In: Information and Communication Technology for Sustainable Development, pp. 423–430. Springer, Singapore (2020)
7. Herzig, J., Shmueli-Scheuer, M., Konopnicki, D.: Emotion detection from text via ensemble classification using word embeddings. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, pp. 269–272 (2017)
8. Ryan, S., Doherty, G.: Fairness definitions for digital mental health applications
9. Thieme, A., Belgrave, D., Doherty, G.: Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems. ACM Trans. Comput.-Hum. Interact. (TOCHI) **27**(5), 1–53 (2020)
10. Zhou, D., Luo, J., Silenzio, V.M., Zhou, Y., Hu, J., Currier, G., Kautz, H.: Tackling mental health by integrating unobtrusive multimodal sensing. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
11. Fatima, I., Mukhtar, H., Ahmad, H.F., Rajpoot, K.: Analysis of user-generated content from online social communities to characterise and predict depression degree. J. Inf. Sci. **44**(5), 683–695 (2018)
12. Chen, X., Sykora, M.D., Jackson, T.W., Elayan, S.: What about mood swings: identifying depression on twitter with temporal measures of emotions. In: Companion Proceedings of the Web Conference 2018, pp. 1653–1660 (2018)
13. Suhasini, M., Srinivasu, B.: Emotion detection framework for twitter data using supervised classifiers. In: Data Engineering and Communication Technology, pp. 565–576. Springer, Singapore (2020)
14. Gaind, B., Syal, V., & Padgalwar, S.: Emotion detection and analysis on social media (2019). arXiv preprint arXiv:1901.08458
15. Saha, K., De Choudhury, M.: Modeling stress with social media around incidents of gun violence on college campuses. Proc. ACM Hum.-Comput. Interact. **1**(CSCW), 1–27 (2017)
16. Pestian, J., Santel, D., Sorter, M., Bayram, U., Connolly, B., Glauser, T., DelBello, M., Tamang, S., Cohen, K.: A machine learning approach to identifying changes in suicidal language. Suicide Life-Threat. Behav. **50**(5), 939–947 (2020)
17. Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T.J., Dobson, R.J., Dutta, R.: Characterisation of mental health conditions in social media using Informed Deep Learning. Sci. Rep. **7**(1), 1–11 (2017)
18. Su, C., Xu, Z., Pathak, J., Wang, F.: Deep learning in mental health outcome research: a scoping review. Transl. Psychiatry **10**(1), 1–26 (2020)
19. Ragheb, W., Azé, J., Bringay, S., Servajean, M.: Attention-based modeling for emotion detection and classification in textual conversations (2019). arXiv preprint arXiv:1906.07020
20. Sekulić, I., Strube, M.: Adapting deep learning methods for mental health prediction on social media (2020). arXiv preprint arXiv:2003.07634
21. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)
22. Dheeraj, K., Ramakrishnudu, T.: Negative emotions detection on online mental-health related patients texts using the deep learning with mha-bcnn model. Expert Syst. Appl. **182**, 115265 (2021)
23. Kim, J., Lee, J., Park, E., Han, J.: A deep learning model for detecting mental illness from user content on social media. Sci. Rep. **10**(1), 1–6 (2020)
24. Natural Language Toolkit. NLTK 3.6.2 documentation. https://www.nltk.org/.