Depression Detection from Social Media Data

Prajakta Kumbhar[1], Vaishnavi Kothari[2], Divya Patil[3], Bhakti Pawar[4], Prof. Shilpa P. Khedkar[5]
Students, Department of Computer Engineering[1,2,3,4]
Guide, Department of Computer Engineering[5]
Modern Education Society's College of Engineering, Pune, Maharashtra, India

Abstract: *Through social networking sites, users can gather information about themselves and their posts. This data can then be analysed and used to develop effective monitoring systems that can detect users with psychological disorders. Psychological illnesses' symptoms are frequently viewed passively.In this study, the author argues that the use of social behaviour extraction can help identify these conditions at an early stage. Our proposed methodology aims to provide a new approach to the treatment of psychological disorders by identifying and addressing the causes of confusion. It can be utilized in informal organizations. We analyse the characteristics and apply machine learning to large-scale data sets to analyse the characteristics of different forms of psychological diseases using different algorithms like ANN, RNN and Naive Bayes and output is generated and users are classified according to their features (happy, low, gloomy, etc.).*

Keywords: Sentimental Analysis, social media, feature extraction, Machine Learning

I. INTRODUCTION

"Mental Pain is less dramatic than physical pain, but it is more common and also harder to bear", said by C.S Lewis. Depression is a prevalent mental condition that can lead to suicide. It is also a primary cause of disability worldwide. In a given year, one out of every 15 adults suffers from depression, and women are twice as likely as men to suffer from it.[1]. More than 280 million individuals worldwide, from all walks of life, suffer from depression. Depression affects women more than it does males[2]. However, in the early stages of depression, 70% of patients refuse to see a doctor, putting their condition at risk of progressing. Despite the fact that depression is common, the condition can be different for different people. Some of the contributing factors include genetics, exercise, and diet. Changing your diet can help lower your risk of depression[3]. Social media applications provides a platform for users like us to share our views, thoughts, emotions or opinions on random topics which we can relate easily. The text which we share through our social media handles contains some valuable intentions that can be used for some real time entities like healthcare, politics, environment, view communication, entertainment, and tourism. These contents or views created by us users is the data that is beneficial for the analysts to analyse the state of our minds. Since the abundance of advancements in these social media platforms, instead of keeping their mental health struggles private, people begin to share them with the world via social media. This data, derived from the texts in the user's account, can be evaluated to determine the user's mental health status and provide a means of assisting them in their recovery.

Since the spreading negative emotions on the social handles people are affected negatively, resulting in sadness and other mental diseases. So the output is that it provides analysts a source , to detect whether that user is depressed or not based on the comments that they make on their social media accounts. This paper focuses on the different ways used for classifying a given piece of language text in line with the opinions expressed in it.Sentiment analysis (also known as opinion mining) is a natural language processing (NLP) technique for determining the positivity, negativity, or neutrality of data. Then one type of semantic analysis is Emotion detection that allows you to identify emotions other than polarity, such as happiness, frustration, anger, and sadness.

DOI: 10.48175/568

A specific system will assign a score for every text word which the user writes based on the designed polarity which will allow identifying the state of the user whether they are in a positive mood or negative mood. Then the analysts implement machine-learning algorithms like ANN, RNN and Naive Bayes to detect the depression based on the information that is acquired and was labelled with the sentiment scores. The competence of detection is evaluated based on the accuracy of the machine learning algorithms. This process includes:

1. Collecting the data from social media;
2. Because the data is typically made in posts format, text pre-processing techniques are used to improve quality.
3. The extraction of features from the sentences;
4. the incorporation of these characteristics into machine learning algorithms to create a judgment mechanism capable of predicting whether the user is depressed or not.

II. LITERATURE SURVEY

There has been a lot of research in the subject of depression detection during the last few years, and various research publications have been published in this topic.

The ‘DAIC-WOZ’ dataset was used in [1] which has class imbalances. To solve the class imbalance , the Synthetic Minority Oversampling technique(SMOTE) was used. Logistic Regression, Random Forest and SVM are used for classifying the users into depressed and non-depressed users. SVM with SMOTE analysis gave the best result with 93% of accuracy. An application named “CureD” was created where users can check their depression levels by answering a standard PHQ-8 questionnaire, and their voice was also recorded.

Authors of [4] used the user data of Sina Weibo and a deep integrated support vector machine(DISVM) to classify students. A depression detection can be more accurately made with DISVM since it categorizes input data before recognizing depression. The result is that the recognition model becomes more stable and accurate. Additionally , the proposed depression recognition scheme can detect potential depression patients among college students by using Sina Weibo.

[5]Authors created a methodology where they used a combination of machine learning algorithms to maximize results. For the first model, they employed KNN, Support vector machine and logistic regression algorithms, while for the second model, a decision tree algorithm, Naïve Bayes classifier, Support Vector Machine algorithms, were used. The average accuracies were calculated since here was no consistency in results since all of the predictions were voted out. Out of the two models, model 1 yielded the best result with an average accuracy of 89.6%.

[6]Paper discussed a two-stage method in which , in the first stage, sentimental analysis was used to predict binary classes(i.e. depressed / not depressed) based on a person's tweets and then a deep learning module long short term memory (LSTM) and CNN was employed. In the second stage, the dataset was divided into train and test set, and then three vectorizers for vectorizing tweets were used : count vectorizer, TD-IDF, and n-grams.

[7]Developed a system that converts videos into frames and then passes through a network for face detection and a CNN model for feature extraction. The resulting emotion vector is then analysed using the Beck Depression InventoryII. The solution uses visual graphics to generate a correlation between the emotion vector and inventory vector. It then displays the detected depression level in a document.

[8]In this paper, the authors proposed a content-based ensemble method (CBEM) that aims to improve the accuracy of the system. Due to the ease of recording and non-invasiveness of the data, electroencephalogram and eye movements data are commonly used for depression detection. Both free-view eye tracking and resting-state EEG datasets were used for the validation of the method, and both datasets had 36,34 subjects each. In both datasets, CBEM achieved 82.5% and 92.65% accuracy. According to the results, CBEM outperforms traditional classification methods.

[9]Authors employed an approach which adopts a multi-stage machine learning pipeline. the first step, they project mobility features of the majority class (undepressed users) and then, using a One-Class SVM algorithm, they classify a test set of users as either depressed (anomalous) or not depressed (inliers).

[10]Authors developed a multilayer deep learning algorithm that can classify users with depression. For training, CNNs, Gated Recurrent Units (GRUs), and Multilayer Perceptrons (MLPs) were used. The training was done in 2 steps. In the first, each post was classified as either general or mental health-related. In the second step, we categorized users into depression and non-depression groups based on their behaviour.

A hybrid model combining a factor graph model (FGM) with a convolutional neural network (CNN) was proposed in paper[11]. The posting behaviour and social interaction of the user were the attributes of the dataset. Also algorithms like Support Vector Machine (SVM), Gradient Boosted Decision Tree (GBDT), Deep Neural Network (DNN) were applied and the results were compared with the hybrid model.

[12]Authors used the AVEC dataset which contains webcam recordings of people during a certain task. Principal Components Analysis (PCA) was used for feature selection for both audio and video recordings. Fusion and classification techniques were used. Fusion involves a) concatenating two feature sets (video and audio) into a single vector, and b) combining the classification results from different classifiers, trained separately on video and audio data, through the AND and OR operands. A gender-based and a gender-independent classification were performed.

[13]Using a twitter API, 10,000 tweets were gathered. Classification was performed using Multinomial Naive Bayes (MNB) and Support Vector Machines (SVM).Through the use of POS Tagger, the tokenized text is assigned the respective parts of speech in order to improve the quality of the training data. SVM scored 79.73 while MNB achieved 83.29 for its F1 score.

A method was developed by authors of [14] that accounted for the differences in speech types and emotions between males and females and their respective contributions to depression detection.

Authors of paper[15] used three models: (a) using machine learning classifiers and WEKA, (b) using imaging and machine learning methods, and (c) using risk factors. Machine learning classifiers like Bayes Net Classifier (BN), Multilayer perceptron (MLP), Logistic regression, Decision Trees, Sequential Minimal Optimisation were used. Bayes Net classifier gave the best results with accuracy of 95%.

Authors of paper[16] implemented a technique in which phrases were extracted from social networks which were passed through a sentimental analysis based on a lexical dictionary. The audio, messages and interactions of users with other users were considered. The sentiment intensities were calculated for the phrases. The sequential minimal optimization(SMO) algorithm was applied to classify the mood phrases which gave 86.1% of correct results .

An Emotion recognition system was designed by authors[17] for classrooms using a deep Convolutional Neural Networks (CNN) architecture. Student's faces were recognised during the teaching sessions and the images were converted into LBP codes to make the system more robust to illumination variations. And then multiple CNNs were used to predict the class.

[18]Authors designed a behaviour recommendation system which considers the person's current behavior and health and recommends some changes in his/her behaviour to improve his/her health. Contrast pattern mining was used to recommend behaviour changes. For comparison purposes, the RANDKN and TOPCP approach was used. A Canadian Community Health Care Survey Data was taken as a dataset.

James Pennebaker and Laura King's stream-of-consciousness essay dataset was used by the authors[19]. With CNN, a multiple layer perceptron (MLP) with one hidden layer was trained. SVM applied with CNN did not improve the results, but applying MLP alone did.

[20]Authors implemented a linear regression technique to examine social network platforms like facebook and twitter to examine how official statistical institutes interact with citizens. Authors have found out that twitter is more powerful than facebook .

A Sentimeter-Br2 based on Sentimeter-Br is used by authors of paper[21] which improved the performance of music recommendation systems. Sentimeter-Br2 considers adverbs, n-grams ,removes stopwords and the differing value of sentiments depending on the verbal tenses. The music was recommended on the basis of the user's current social media data. If there was no data updated by the user , then music of his/her style was recommended.

A song recommendation system was proposed by the authors of [22]. It recommends songs on the basis of user's Proposed a system that recommends songs based on the user's content based audio information with the contextual emotion information mined from user-generated articles and listening history. The authors used a factorization machine technique. Effective contextual text information is more effective instead of just simple word counts of the articles the users write as the context feature.

III. PROBLEM STATEMENT

Social networks have been developed as a great point for its users to communicate with their interested friends and share their opinions, photos, and videos reflecting their moods, feelings and sentiments. This creates an opportunity to analyze social network data for user's feelings and sentiments to investigate their moods and attitudes when they are communicating via these online tools.

IV. PROPOSED SYSTEM

Machine learning algorithms are trained on datasets, and then a model is created for analysis. The machine learning technique is appropriate based on the model's accuracy. Supervised learning, unsupervised learning, and reinforcement learning are the three strategies used in machine learning algorithms. The model is trained via labelled information that contains each input and results in supervised learning.The sections of the process square measure the coaching section and testing phase. Unsupervised learning strategies don't use coaching information or labeled information. It finds the hidden structures or patterns from unlabeled information.

Depression based on their social media data which is in text format. The suggested framework contains five major phases, as shown in Figure 1:

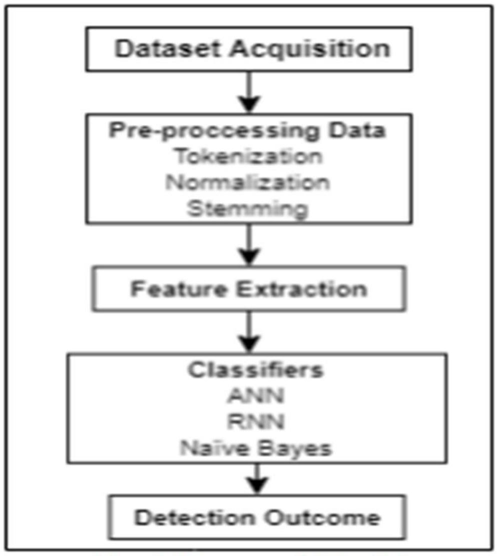

Fig.1 Proposed Model

1. Data is gathered from a well-known dataset.
2. Data pre-processing procedures such as tokenization, normalisation, and stemming on the dataset.

3. From the user's post, extract the various aspects in light of psychogeriatric metrics.
4. The data is processed using classifier approaches such as ANN, RNN, and Nave Bayer.
5. The result of the model is presented on the screen.

V. METHODOLOGY

The initial stage in the process is to collect data from a dataset obtained from Reddit dataset/Kaggle. This dataset contains social media comments and responses in text format. To begin, all posts for both depressed and non-depressed accounts are retrieved, as well as information about user accounts and activities such as number of followers, number of followers, time of posts, number of mentions, and number of reposts.

We have collected a total of 3000 Posts for the generation of the training module and testing module for our model. Data collection is followed by pre-processing which is applied on text posts. We used a tokenization method for splitting the sentences to each word,the tokenized sentences are then processed for normalization and lemmanization which helped to remove stop words and eliminate the affixes from the word.

We extract several features based on psycholinguistic assessments from the user's post in feature extraction to characterise and demonstrate differences between depressive and non-depressive postings. Feature extraction process helped in the selection of keyword like from these is statement “I am happy” here the keyword is happy, so these technique is used for , remove keyword and send to trained module. These keywords are classified according to 3 levels of depression: first low level, second medium level and third high level depressed level by using algorithm ANN, RNN and naïve bayer and it will train the keywords.

Artificial Neural Network (ANN): In the science of artificial intelligence, an Artificial Neural Network seeks to duplicate the network of neurons that make up a human brain so that computers can understand things and make decisions in a human-like fashion. Computers are programmed to act like interconnected brain cells in order to create an artificial neural network.

Naïve Bayes Classifier Algorithm: The Naïve Bayes method is a supervised learning technique for addressing classification issues that is based on the Bayes theorem.It is mostly utilised in text classification tasks that require a large training dataset. The fast machine learning models such as the Naive Bayes Classifier can produce speedy predictions.It's a probabilistic classifier, which means it makes predictions based on an object's probability. Spam filtration, sentiment analysis, and article classification are all common uses of the Naïve Bayes Algorithm.

Recurrent Neural Network (RNN): Recurrent neural networks (RNNs) are artificial neural networks that are mostly employed in speech recognition and natural language processing (NLP). Deep learning and the construction of models that mimic the activity of neurons in the human brain use RNN.

Text, genomes, handwriting, the spoken word, and numerical time series data from sensors, stock markets, and government agencies are all examples of data that recurrent networks are meant to recognise patterns in.

A recurrent neural network resembles a traditional neural network except that the neurons have a memory state. A basic memory will be used in the computation.

The last stage would predict the result whether the person is depressed or non-depressed. If a person is highly depressed, we will send a motivational message . Final output mainly contains the accuracy of the model which is compared with pre-trained data and the result is displayed.

VI. CONCLUSION

A classification problem is characterised as determining whether or not a person is depressed based on their social media profile behaviour. We develop a predictive model to predict whether user posts are depressed or not based on detecting depressed users using a machine learning approach and sentiment analysis.Various machine learning algorithms are used, and various feature datasets are investigated. Data preparation and alignment, data labelling, and feature extraction are just a few of the preprocessing procedures. The machine learning methods ANN, RNN, and Naive

DOI: 10.48175/568

Bayes were chosen and applied to the post dataset to determine the algorithm's accuracy in categorising depressed and non-depressed users. This research might be viewed as a first step in developing a comprehensive social media-based platform for analysing and predicting mental and psychological difficulties in users and recommending treatments.

VII. FUTURE SCOPE

In future, the work can be enhanced by including some additional features of online users on social media behavior .e.g time of post and interaction with other users. However the analysis is limited to text only ,further research can improve the system by adding features like text in different languages ,audio messages ,image and video based depression detection.

ACKNOWLEDGMENT

We did like to take this occasion to thank our mentor Prof.S.P.khedkar for all of her assistance and direction, as well as for providing numerous resources for our project, such as a laboratory with all necessary software platforms and a constant Internet connection.

REFERENCES

[1] Yalamanchili, Bhanusree, Nikhil Sai Kota, Maruthi Saketh Abbaraju, Venkata Sai Sathwik Nadella, and Sandeep Varma Alluri. "Real-time Acoustic based Depression Detection using Machine Learning Techniques." In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-6. IEEE, 2020.

[2] W. H. Organisation, "Depression," World Health Organisation, 13 September 2021. [Online]. Available: https://www.who.int/newsroom/fact-sheets/detail/depression.

[3] N. Schimelpfening , "Why Some People Are More Prone to Depression Than Others," Verywellmind, 26 March 2021. [Online]. Available: Why Some People Are More Prone to Depression Than Others.

[4] Ding, Yan, et al. "A depression recognition method for college students using deep integrated support vector algorithm." IEEE Access 8 (2020): 75616-75629.

[5] Kumar, Piyush, Rishi Chauhan, Thompson Stephan, Achyut Shankar, and Sanjeev Thakur. "A Machine Learning Implementation for Mental Health Care. Application: Smart Watch for Depression Detection." In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 568-574. IEEE, 2021.

[6] Shetty, Nisha P., et al. "Predicting depression using deep learning and ensemble algorithms on raw twitter data." International Journal of Electrical and Computer Engineering 10.4 (2020): 3751.

[7] Mulay, Akshada, Anagha Dhekne, Rasi Wani, Shivani Kadam, Pranjali Deshpande, and Pritish Deshpande. "Automatic Depression Level Detection Through Visual Input." In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 19-22. IEEE, 2020.

[8] Zhu, Jing, Zihan Wang, Tao Gong, Shuai Zeng, Xiaowei Li, Bin Hu, Jianxiu Li, Shuting Sun, and Lan Zhang. "An improved classification model for depression detection using EEG and eye tracking data." IEEE transactions on nanobioscience 19, no. 3 (2020): 527-537.

[9] Gerych, Walter, Emmanuel Agu, and Elke Rundensteiner. "Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach." In 2019 IEEE 13th International Conference on Semantic Computing (ICSC), pp. 124-127. IEEE, 2019.

[10] Wongkoblap, Akkapon, Miguel A. Vadillo, and Vasa Curcin. "Classifying depressed users with multiple instance learning from social network data." In 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 436-436. IEEE, 2018.

[11] Lin, Huijie, et al. "Detecting stress based on social interactions in social networks." IEEE Transactions on Knowledge and Data Engineering 29.9 (2017): 1820-1833.

[12] Pampouchidou, A., O. Simantiraki, C-M. Vazakopoulou, C. Chatzaki, M. Pediaditis, A. Maridaki, K. Marias et al. "Facial geometry and speech analysis for depression detection." In Engineering in Medicine and Biology Society (EMBC), 39th Annual International Conference of the IEEE, pp. 1433-1436. IEEE, 2017.

[13] Deshpande, Mandar, and Vignesh Rao. "Depression detection using emotion artificial intelligence." In 2017 international conference on intelligent sustainable systems (iciss), pp. 858-862. IEEE, 2017.

[14] Jiang, Haihua, Bin Hu, Zhenyu Liu, Lihua Yan, Tianyang Wang, Fei Liu, Huanyu Kang, and Xiaoyu Li. "Investigation of different speech types and emotions for detecting depression using different classifiers." Speech Communication 90 (2017): 39-46.

[15] Hooda, Madhurima, Aashie Roy Saxena, and Babita Yadav. "A Study and Comparison of Prediction Algorithms for Depression Detection among Millennials: A Machine Learning Approach." In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp. 779-783. IEEE, 2017.

[16] Rosa, Renata L., et al. "Monitoring system for potential users with depression using sentiment analysis." 2016 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2016.

[17] Sahla, K. S., and T. Senthil Kumar. "Classroom Teaching Assessment Based on Student Emotions." In The International Symposium on Intelligent Systems Technologies and Applications, pp. 475-486. Springer International Publishing, 2016.

[18] Chen, Yan, et al. "Contrast pattern based collaborative behavior recommendation for life improvement." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2017.

[19] Majumder, Navonil, et al. "Deep learning-based document modeling for personality detection from text." IEEE Intelligent Systems 32.2 (2017): 74-79..

[20] Glavan, Ionela-Roxana, Andreea Mirica, and Bogdan Narcis Firtescu. "The Use of Social Media for Communication In Official Statistics at European Level." Romanian Statistical Review 64, no. 4 (2016): 37-48.

[21] Rosa, Renata L., Demsteneso Z. Rodriguez, and Graça Bressan. "Music recommendation system based on user's sentiments extracted from social networks." IEEE Transactions on Consumer Electronics 61.3 (2015): 359-367.

[22] Chen, Chih-Ming, Ming-Feng Tsai, Jen-Yu Liu, and Yi-Hsuan Yang. "Using emotional context from article for contextual music recommendation." In Proceedings of the 21st ACM international conference on Multimedia, pp. 649-652. 2013.

DOI: 10.48175/568