ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Detection of depression-related posts in reddit social media forum

[1] A.N. RAMAMANI, [2] CHITTURI SINDHURI

[1](Assistant Professor), MCA, S.V.K.P & Dr K.S. Raju Arts & Science College

[2]MCA, scholar, S.V.K.P & Dr K.S. Raju Arts & Science College

ABSTRACT

Depression is viewed as the largest contributor to global disability and a major reason for suicide. It has an impact on the language usage reflected in the written text. The key objective of our study is to examine Reddit users' posts to detect any factors that may reveal the depression attitudes of relevant online users. For such purpose, we employ the Natural Language Processing (NLP) techniques and machine

learning approaches to train the data and evaluate the efficiency of our proposed method. We identify a lexicon of terms that are more common among depressed accounts. The results show that our proposed method can significantly improve performance accuracy. The best single feature is bigram with the Support Vector Machine (SVM) classifier to detect depression with 80% accuracy and 0.80 F1 scores. The strength and effectiveness of the combined features (LIWC+LDA+bigram) are most successfully demonstrated with the Multilayer Perceptron (MLP) classifier resulting in the top performance for depression detection reaching 91% accuracy and 0.93 F1 scores. According to our study, better performance improvement can be achieved by proper feature selections and their multiple feature combinations.

1.INTRODUCTION

Depression as a common mental health disorder has long been defined as a single disease with a set of diagnostic criteria. It often co-occurs with anxiety or other psychological and physical disorders; and has an impact on feelings and behavior of the affected individuals [1]. According to the WHO study, there are 322 million

people estimated to suffer from depression, equivalent to 4.4% of the global population. Nearly half of the in-risk individuals live in the South-East Asia (27%) and Western Pacific region (27%) including China and India. In many countries depression is still under-diagnosed and left without any adequate treatment which can lead into a serious self-perception and at its worst, to suicide [2]. In addition, the social stigma surrounding depression prevents many affected individuals from seeking an appropriate professional assistance.

As a result, they turn to less formal resources such as social media. With the development of Internet usage, people have started to share their experiences and challenges with mental health disorders through online forums, micro-blogs or tweets. Their online activities inspired many researchers to introduce new forms of potential health care solutions and methods for early depression detection systems. Using different Natural Language Processing (NLP) techniques and text classification approaches, they tried to succeed in a higher performance improvement. Some studies use single set features, such as bag of words (BOW) [3], [4], N-grams [5], LIWC [6] or LDA [7], [8] to identify depression in their posts. Some other papers compare the performance of individual features with various machine learning classifiers [9] [12]. Recent studies examine the power of single features and their combinations such as N-grams+LIWC [13] or BOW+LDA and TF-IDF+LDA [14] to improve theaccuracy results. They experiment with a smarter text pre-processing, and introduce different substitute words depending on the nature of the original string. For instance, Tyshchenko et al. [14] suggested categorizing the stop words and adding LIWC-like word categories as an extra feature to an already designed method (BOW+TFIDF+LIWC). In addition, he applied multiple feature combinations to increase the perfor- 1mance using Convolutional Neural Networks (CNN) which consist of neurons with learnable weights and differ in terms of their layers. CNNs are very similar to simple feed-forward neural networks and state of the art method in the text and sentence classification tasks.

A meta-analysis by Guntuku et al. [15] summarizes several iterations of depression detection tasks in computational linguistics. Another interesting review for mental health support and intervention in social media is written by Calvo et al. [16] who reviewed the taxonomy of data sources, NLP techniques and computational methods to detect various mental health applications. Even with this significant progress, challenges still remain. This paper aims to search for a solution to a performance increase through a proper features selection and their multiple feature combinations. First, we choose the most beneficial linguistic features applied for depression identification to characterize the content of the posts. Second, we analyze the correlation significance, hidden topics and word frequency extracted from the text. Regarding the correlation, we focus on the LIWC dictionary and its three feature types (linguistic dimensions, psychological processes and personal concerns). For the topic examination, we choose the LDA method as one of the successful features. For the word frequency, we use unigrams and bigrams by leveraging the vectors based on TF-IDF scheme. Finally, we set five text classifying techniques and conduct their execution using the extracted data to detect depression. We compare the performance results based on three single feature sets and their multiple feature combinations. In our experiment, we use data collected from the Reddit social media platform. It was chosen as the

data source as it allows longer posts. Targeting technical approaches towards detection tasks, our paper follows the lines of Calvo et al. research [17].

Our study has four specific contributions: first, to examine the relationship between depression and user's language usage; second, to design three LIWC features for our specific research problem; third, to evaluate the power of N-grams probabilities, LIWC and LDA as single features for performance accuracy; fourth, to show the predictive power of both single and combined features with proposed classification approaches to achieve a higher performancein depression identification tasks.

2.LITERATURE SURVEY

Huijie Lin, Jepia Jia*, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-SengChua

published a paper titled "Detecting Dression Based on Social Interactions in Social Networks." Due to social media's widespread use, people are accustomed to posting about their everyday activities and chatting with friends there, making it possible to use this data to identify depression. This study uses a sizable dataset from actual social platforms to comprehensively examine the relationship between users' Depression moods and social interactions. We discover that a user's Depression condition is closely related to that of his or her social media friends. We establish a set of textual, visual, and social features that are associated to depression initially, and then we suggest a novel hybrid model. Using a factor graph model coupled with a convolutional neural network to use the information from tweets and social interactions for depression detection. Applied methodology: The system's phases are as follows: • Support Vector Machine (SVM): This popular binary classifier has been demonstrated to perform well on a variety of classification problems. In our issue, SVM with RBF kernel is applied. • Random Forest (RF): This ensemble learning technique for decision trees entails building a number of decision trees using attributes that are randomly chosen and bagging the outcomes for categorization. • Gradient Boosted Decision Tree (GBDT): It trains a gradient boosted decision tree model using features connected to each user. • Deep Neural Network (DNN) for Depression Detection at the User Level a neural network using convolutions To deal with the issue of user-level Depression detection, (CNN) with cross autoencoders is given. This is the actual baseline method that we should compare our suggested improvements to. Paper 2: M. S. Neethu and R. Rajshri's sentiment analysis of tweets using machine learning techniques Abstract: Finding and classifying the opinions or sentiments included in a source text is the technique of sentiment analysis. The amount of sentiment-rich data produced by tweets, status updates, blog posts, and other social media platforms is substantial. Finding the consensus can be greatly aided by sentiment analysis of this user-generated data. Twitter sentiment analysis is more complicated than traditional sentiment analysis due to the use of slang terms and misspelt words. 140 characters is the limit permitted on Twitter. they are Machine learning and knowledge

base approaches are strategies for analysing emotions in text. In this study, we examine tweets on electrical devices like laptops and smartphones using a machine learning technique. By performing sentiment analysis in a particular domain, it is feasible to ascertain the impact of domain information on sentiment classification. It is provided a new feature vector for. Methodology Used: The following actions were taken: Symbolic Approaches: Neethu M. S. and Rajasree R. first presented symbolic techniques, commonly referred to as knowledge-based approaches, in July 2013. In this method, available lexical resources are used. This sentiment analysis technique uses the bag-of-words method. Due to its focus on the word list or string of words, the BOW paradigm is unable to Consider the sentence's setting. This model consists of a list of terms that each have their own value when discovered in the provided text. This writing style has little regard for grammatical conventions and is just concerned with the words. Techniques for machine learning Contrary to knowledge-based systems, machine learning techniques classify data using a training set and a test set rather than a lexical resources list. The training set includes input vectors and the corresponding class labels for the network's training. The model is then tested against the test set's class labels against unidentified feature vectors. Among other methods, machine learning includes SVM, maximum entropy, and Nave Bayes. As a result, the algorithm can maintain its flexibility in the face of ongoing modifying the lingo on social networks. In this approach, a classification model is constructed using a training set and attempts to assign the input feature vectors to appropriate class labels. Combine the outcomes of knowledge-based techniques and machine learning techniques to ensure a thorough examination of the dataset.

3. EXISTING SYSTEM

Sigmund Freud [18] wrote about Freudian slips or linguistic mistakes to reveal the secret thoughts and feelings of the writers. With the development of sociology and psycholinguistic theories, various approaches towards the relationship between depression and its language have been defined. For instance, according to Aaron Beck et al. [19]'s cognitive theory of

depression, affected individuals tend to perceive themselves and their environment in mostly negative terms. They often express themselves through negatively valenced words and first-person pronouns. Their typical feature is self-preoccupation defined by sky and Greenberg [20] which can develop into an extreme self-criticism stage. According to Durkheim's [21] social integration model, people suffering from depression often feel detached from their social life and have a difficulty to integrate into society.

Rude et al. [26] who examined the linguistic patterns of the essays written by currently-depressed, formerly-depressed and never depressed college students. According to his results, depressed students used more negatively valenced words and less positive emotion words. Zinken et al. [27] studied the psychological relevance of syntactic structures to predict the improvement of depressive symptoms. He supposed that a written text might barely differ in its word usage; however, may differ in its syntactic structure, especially in the construction of relationships between the events. Analyzing a causation and insight words tasks, he found out that in the text written by depressed individuals there was a decreased use of complex syntax in comparison to non-depressed ones.

De Choudhury et al. [29] used linguistic features to train a classifier to examine Twitter posts that indicated depression. Coppersmith et al. [6] looked for tweets that explicitly stated ``I was just diagnosed with depression" sentences.

Preotiuc-Pietroi et al. [9] applied broader textual features such as LIWC, LDA and frequent 1-3 grams on the Twitter data to examine the personality of the users with self declared post-traumatic stress (PTSD) disorders. His results show that the users suffering from PTSD were both older and more conscientious in comparison to depressed individuals. Since the language predictive of depression and PTSD had a large overlap with the language predictive of personality, the authors conclude that the users with a particular personality or demographic profiles tend to share their mental health diagnosis on social media, and thus the results may not generalize to other sources of autobiographical text. Resnik et al. [8] proved that the LDA model can

uncover a meaningful and potentially useful latent structure for the automatic identification of important topics for depression detection.

Tsugawa et al. [12] predicted depression from Twitter data in a Japanese sample where he showed that the features based on a topic modeling are useful in the tasks for recognizing depressive and suicidal users. Bentoni et al. [5] demonstrated the effectiveness of multi-task learning (MTL) models on mental health disorders with a limited amount of target data. He used feed-forward multi-layer perceptions and feed-forward multi-task models trained to predict each task separately as well as to predict a set of conditions simultaneously . They experimented with a feed-forward network against independent logistic regression models to test if MTL would have performed well in the domain.

- Reece et al. [22] found out that the first stage of depression may be detectable from Twitter data several months prior to its diagnosis with 0.87 AUC of performance probability.

Disadvantages

- In the existing work, the system doesn't provide effective and strong data classification techniques.
- In the existing system, problem of non preprocessing data absence.

The proposed system aims to search for a solution to a performance increase through a proper features selection and their multiple feature combinations. First, we choose the most beneficial linguistic features applied for depression identification to characterize the content of the posts. Second, we analyze the correlation significance, hidden topics and word frequency extracted from the text. Regarding the correlation, we focus on the LIWC dictionary and its three feature types (linguistic dimensions, psychological processes and personal concerns). For the topic examination, we choose the LDA method as one of the successful features.

For the word frequency, we use unigrams and bigrams by leveraging the vectors based on TF-IDF scheme. Finally, we set five text classifying techniques and conduct their execution using the extracted data to detect depression. We compare the performance results based on three single feature sets and their multiple feature combinations. In our

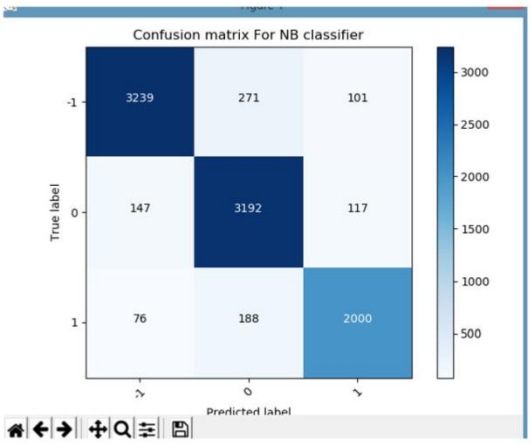
experiment, we use data collected from the Reddit social media platform. It was chosen as the data source as it allows longer posts. Targeting technical approaches towards detection tasks, our paper follows the lines of Calvo et al. research [17].

The proposed system has four specific contributions: first, to examine the relationship between depression and user's language usage; second, to design three LIWC features for our specific research problem; third, to evaluate the power of N-grams probabilities, LIWC and LDA as single features for performance accuracy; fourth, to show the predictive power of both single and combined features with proposed classification approaches to achieve a higher performance in depression identification tasks.

Advantages

- The system is more effective since it is implemented strong features extraction techniques.
- In the proposed system, the systems propose a Latent Dirichlet Allocation model for data classification to find depression.

4. OUTPUTSCREENS



5. CONCLUSION

In this paper, we tried to identify the presence of depression in Reddit social media; and searched for affective performance increase solutions of depression detection. We characterized a closer connection between depression and a language usage by applying NLP and text classificationtechniques. We identified a lexicon of words more commonamong the depressed accounts. According to our findings,the language predictors of

depression contained the wordsrelated to preoccupation with themselves, feelings of sadness,anxiety, anger, hostility or suicidal thoughts, with a greateremphasis on the present and future.To measure the signs of depression, we examined theperformance of both single feature and combined featuresets using various text classifying methods. Our results showthat a higher predictive performance is hidden in properfeatures selection and their multiple feature combinations.The strength and effectiveness of combined features aredemonstrated with the MLP classifier reaching 91% accuracyand 0.93 F1 score achieving the highest performance degreefor detecting the presence of depression in Reddit socialmedia conducted in our study.Additionally, the best feature among the single feature setsis bigram; with SVM classifier it can detect depression with80% accuracy and 0.79 F1 score. Considering LIWC andLDA features, LIWC outperformed topic models generatedby LDA.Although our experiment shows that the performancesof applied methodologies are reasonably good, the absolutevalues of the metrics indicate that this is a challenging taskand worthy of further exploration. We believe this experimentcould further underline the infrastructure for new mechanismsapplied in different areas of healthcare to estimatedepression and related variables. It can be beneficial forthe individuals suffering from mental health disorders to bemore proactive towards their fast recovery. In our futurework, we will try to examine the relationship between theusers' personality [65] and their depression-related behavior reflected in social media.

6. REFERENCE

1] W. H. Organization, "Depression and other common mental dis-orders: Global health estimates. geneva: World health organization;2017. licence: Cc by-nc-sa 3.0 igo." http://www.who.int/en/news-room/fact-sheets/detail/depression, 2017.

2] M. Friedrich, "Depression is the Leading Cause of Disability Around the
World Depression Leading Cause of Disability Globally Global Health,"
JAMA, vol. 317, no. 15, pp. 1517–1517, 2017.

[3] M. Nadeem, “Identifying depression on twitter,” CoRR, vol. ab-s/1607.07384, 2016.

4] S. Paul, S. K. Jandhyala, and T. Basu, “Early detection of signs of anorexiaand depression over social media using effective machine learning frame-works,” in CLEF, 2018.

[5] A. Benton, M. Mitchell, and D. Hovy, “Multi-task learning for mental
health using social media text,” CoRR, vol. abs/1712.03538, 2017