ONLINE DEPRESSION DETECTION APPLICATION

[1]BURRA SANJU, [2]D.SAI KRISHNA
[1]MCA Student, [2] Assistant Professor
Department Of MCA
Sree Chaitanya College of Engineering, Karimnagar

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 the accuracy 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.

The rest of the paper is organized as follows. In section II, we discuss related work in depression detection. In section III, we define the properties of the Reddit dataset. In section IV, we introduce the methodology and conduct data preprocessing followed by feature extraction. In section V, we compare and analyze the feature sets and examine the results as well as the most powerful machine learning technique for depression detection. We conclude our study and provide a direction for future work in section VI.

2. LITERATURE SURVEY

The mental health concern for the new millennium. Cyberpsychol

Surveyed 35 therapists who have treated clients suffering from cyber-related problems to gather outcome information. Respondents reported an average caseload of 9 clients who they classified as Internet-addicted, with a range between 2 and 50 clients treated within the past year. Five general subtypes of Internet addiction were categorized based on the most problematic types of online applications, and they include addictions to Cybersex, Cyber-relationships, online stock trading or gambling, information surfing, and computer games. Treatment strategies included cognitive–behavioral approaches, sexual offender therapy, marital and family therapy, social skills training, and pharmacological interventions. Based on their client encounters, efforts to initiate support groups and recovery programs specializing in the treatment of Internet addiction were being considered. Finally, based upon the findings, this article examines the impact of cyber-disorders on future research, treatment, and public policy issues for the new millennium. (psycinfo Database Record (c) 2023 APA, all rights reserved)

The emergence of a new clinical disorder, Cyberpsychol

Anecdotal reports indicate that some on-line users are becoming addicted to the Internet resulting in academic, social, and occupational impairment. This study investigated the existence of Internet addiction and the extent of problems caused by such potential misuse. Of all the diagnoses referenced in the %DSM-IV%, Pathological Gambling (PG) was viewed as most akin to the pathological nature of Internet use. By using PG as a model, addictive Internet use can be defined as an impulse-control disorder that does not involve an intoxicant. Therefore, this study developed a brief 8-item questionnaire referred to as a Diagnostic Questionnaire (DQ) that modified criteria for PG to provide a screening instrument for classification of participants. On the basis of these criteria, case studies of 396 dependent Internet users and 100 nondependent Internet users were classified. Qualitative analyses suggest significant behavioral and functional usage differences between the 2

groups such as the types of applications utilized, the degree of difficulty controlling weekly usage, and the severity of problems noted (e.g., academic, relationship, financial, etc). Clinical and social implications of pathological Internet use and future directions for research are discussed. (psycinfo Database Record (c) 2023 APA, all rights reserved)

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 Pyszvzynsky 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 perceptrons 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.

4. PROPOSED SYSTEM

- 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].

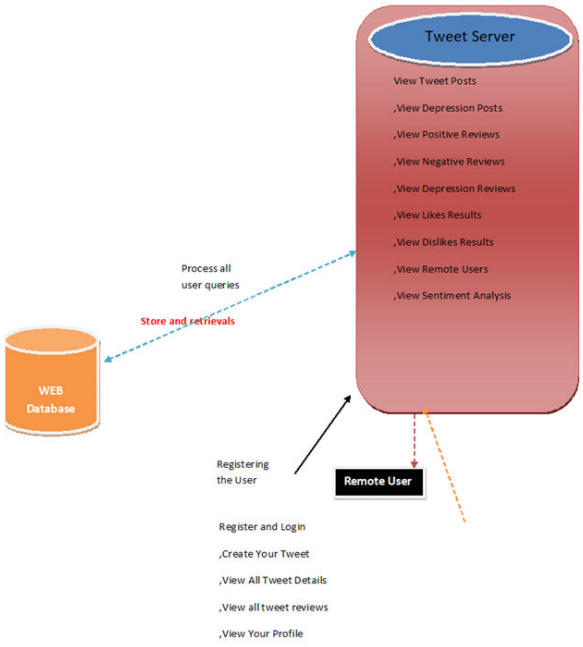
- The proposed system has four specific contributions: first, to examine the relationship between depression and user's language usage; second, to design three LIWC features for our specific research problem; third, to evaluate the power of N-grams probabilities, LIWC and LDA as single features for performance 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.

5. SYSTEM DESIGN

Architecture Diagram



6. IMPLEMENTATION

MODULES

Admin

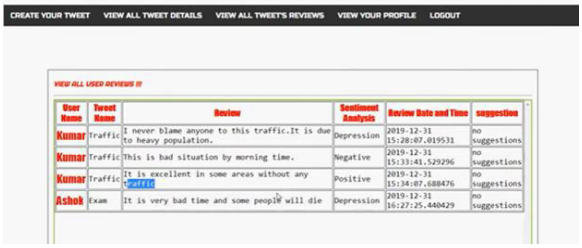
In this module, the Admin has to login by using valid user name and password. After login successful he can do some operations such View Tweet Posts,View Depression Posts,View Positive Reviews,View Negative

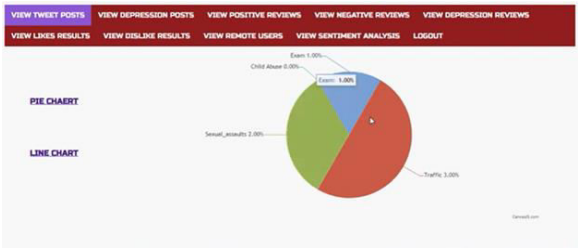
Reviews,View Depression Reviews,View Likes Results,View Dislikes Results,View Remote Users,View Sentiment Analysis.

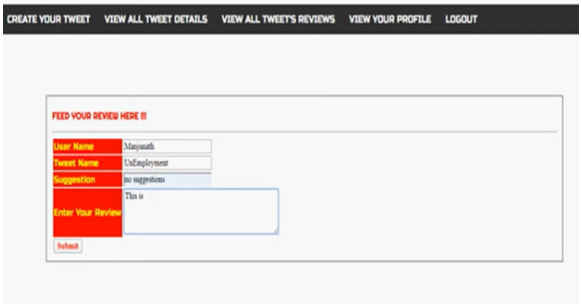
User

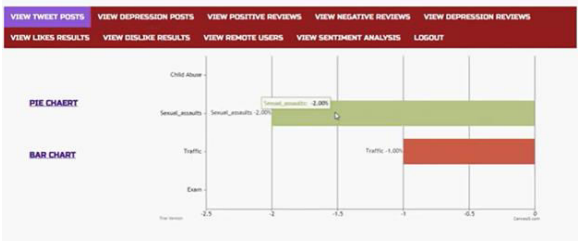
In this module, there are n numbers of users are present. User should register before doing some operations. After registration successful he has to wait for admin to authorize him and after admin authorized him. He can login by using authorized user name and password. Login successful he will do some operations like Create Your Tweet,View All Tweet Details,View all tweet reviews,View Your Profile.

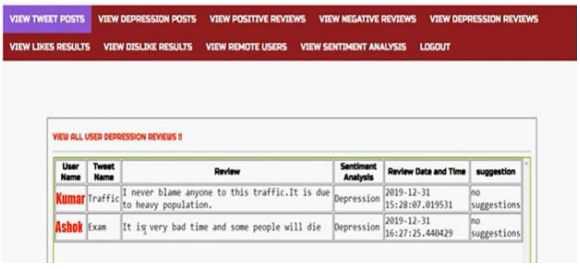
7. **SCREEN SHOTS**

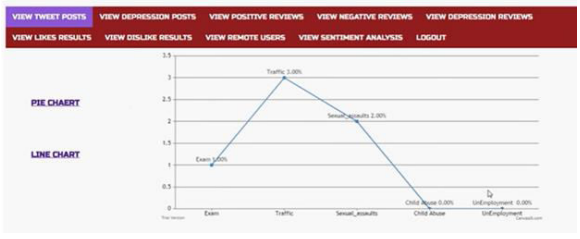












8. CONCLUSION

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 classification techniques. We identified a lexicon of words more common among the depressed accounts. According to our findings, the language predictors of depression contained the words related to preoccupation with themselves, feelings of sadness, anxiety, anger, hostility or suicidal thoughts, with a greater emphasis on the present and future. To measure the signs of depression, we examined the performance of both single feature and combined feature sets using various text classifying methods. Our results show that a higher predictive performance is hidden in proper features selection and their multiple feature combinations. The strength and effectiveness of combined features are demonstrated with the MLP classifier reaching 91% accuracy and 0.93 F1 score achieving the highest performance degree for detecting the presence of depression in Reddit social media conducted in our study. Additionally, the best feature among the single feature sets is bigram; with SVM classifier it can detect depression with 80% accuracy and 0.79 F1 score. Considering LIWC and LDA features, LIWC outperformed topic models generated by LDA. Although our experiment shows that the performances of applied methodologies are reasonably good, the absolute values of the metrics indicate that this is a challenging task and worthy of further exploration. We believe this experiment could further underline the infrastructure for new mechanisms applied in different areas of healthcare to estimate depression and related variables. It can be beneficial for the individuals suffering from mental health disorders to be more proactive towards their fast recovery. In our future work, we will try to examine the relationship between the users' personality [65] and their depression-related behavior reflected in social media.

REFERENCES

[1] W. H. Organization, "Depression and other common mental disorders: 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. abs/ 1607.07384, 2016.

[4] S. Paul, S. K. Jandhyala, and T. Basu, "Early detection of signs of anorexia and depression over social media using effective machine learning frameworks," 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.
[6] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From adhd to sad: Analyzing the language of mental health on twitter through selfreported diagnoses," in Proceedings of the 2ndWorkshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015, pp. 1–10.
[7] D. Maupomé and M.-J. Meurs, "Using topic extraction on social media content for the early detection of depression," in Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018., 2018.
[8] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, "Beyond lda: Exploring supervised topic modeling for depression-related language in twitter," in Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015, pp. 99–107.
[9] D. Preotiuc-Pietro, J. C. Eichstaedt, G. J. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. H. Ungar, "The role of personality, age, and gender in tweeting about mental illness," in CLPsych@HLT-NAACL, 2015.
[10] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 217–226, 2014.
[11] H. A. Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar, "Towards assessing changes in degree of depression through facebook," in Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 118–125.
[12] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from twitter activity," in Proceedings of the 33rd annual ACM conference on human factors in computing systems. ACM, 2015, pp. 3187–3196.
[13] J. Wolohan, M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard, "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with nlp," in Proceedings of the First International Workshop on Language Cognition and Computational Models, 2018, pp. 11–21.
[14] Y. Tyshchenko, "Depression and anxiety detection from blog posts data." University of Tartu Institute of Computer Science Computer Science Curriculum, 2018.
[15] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," Current Opinion in Behavioral Sciences, vol. 18, pp. 43–49, 2017.