Detection of depression-related posts in reddit social media forum

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.

**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.

**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.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).