A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter

Swati Jain, Suraj Prakash Narayan, Rupesh Kumar Dewang, Utkarsh Bhartiya, Nalini Meena and Varun Kumar

***Abstract*—Depression as a disorder has been a great concern in our society and has been perpetually a hot topic for researchers in the world. Despite the massive quantity of analysis on understanding individual moods together with depression, anxiety, and stress supported activity logs collected by pervasive computing devices like smartphones, foretelling depressed moods continues to be an open question. In this paper, we have proposed a depression analysis and suicidal ideation detection system, for predicting the suicidal acts based on the level of depression. We collected real time data from students and parents by making them fill questionnaires similar to PHQ-9 (Parent health questionnaire) consisting of questions like What's your age? or Are you regular in school/college? and processed it into meaningful data with related features like age, sex, regularity in the school, etc. Then, classification machine algorithms are used to train and classify it in five stages of depression depending on severity - Minimal or none, mild, moderate, moderately severe and severe. Maximum accuracy i.e. 83.87 % was achieved by using XGBoost classifier in this dataset. Also, data was collected in the form of tweets and were classified into whether the person who tweeted is in depression or not using classification algorithms. Logistic Regression classifier gave the maximum accuracy i.e. 86.45 % for the same.**

***Index Terms*—Twitter, Tweets, Reddit, Suicide, Depression, Social Media, Machine Learning, Classification.**

I. Introduction

Depression is a disorder of major public health importance, in terms of its prevalence and therefore the suffering, dysfunction, morbidity, and economic burden [1]. It's a serious enfeebling disorder which might have an effect on folks from all ages which might lead to low mood, feelings of guilt, insomnia, and cause problems like hurting chronic back pain, and bilateral medicine symptoms and might be fatal typically if left untreated. According to the World Health Organization (WHO), roughly 350 million human-being square units are suffering from depression nowadays [2]. United Nations agency ranks depression mutually of the foremost devastating diseases within the world [3]. Additionally, the two-third fraction of depressed folks do not look for applicable treatments, that cause major consequences[8]. The medical science relies on asking the patients questions about their situations, that doesn't diagnose depression in a very precise way [4]. According to the Global Burden of Disease Study, it's calculable that if the current increasing rate of the amendment within the pattern of mortality and disease continue, by 2020 depression can account to 5.7 % of all the diseases and it might be the second leading explanation for incapacity worldwide, after the heart diseases [1]. With an endless increase in the menace of depression, there is a demand to develop automatic techniques for the detection of the presence and extent of depression thereby stopping new events to occur. Therefore, the motivation of this paper, is to explore the whole different sources of information, like social media posts, blogs, language, and action cues, to predict the severity of depression. Data was collected from tweets and questionnaires prepared. While doing so, we also investigated different feature representation including pie charts and bar charts and modeling techniques such as supervised classifiers including support vector machine and Random forest classifiers corresponding to each modality for improving the performance of automatic prediction thereby helping the needful in identifying the depression in the early stage which will help in preventing the catastrophic outcomes of the same.

The rest of the paper is structured as follows. Section II introduces most relevant related works. Section III provides a detailed description of the proposed methodology. Observations and results are described in Section IV. Finally, Section V concludes the paper.

II. Related Work

Depression studies came much earlier and was a major focus than that of Internet. Detecting depression from documents in particular has become an increasingly important research area, with interesting methods and results reported for Facebook, Twitter and various other forum posts [5]. Based upon the questionnaire survey throughout the world, many widely-accepted ways and criteria have been developed. For example, in one of the related work by Lenore Sawyer Radloff, CES-D Scale contains 20 questions about the mental conditions like users bad feelings and sleep conditions [6]. The questions either have several options aligned with different scores or require users to feedback the degree of their situations. The depression level is diagnosed according to the scale of the total score. In Another example, there are 21 categories about users mental and physiological state in

Swati Jain, Suraj Prakash Narayan, Rupesh Kumar Dewang, Utkarsh Bhartiya, Nalini Meena and Varun Kumar are with Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology, Allahabad, Prayagraj-211004, India (E-mail: swatijain345@gmail.com, sprksh.narayan@gmail.com, rupeshdewang@mnnit.ac.in, bhartiyautkarsh9695@gmail.com, nalini61097@gmail.com, varunmaurya37@gmail.com)

978-1-7281-0577-2/19/$31.00 2019 IEEE

Becks Depression Inventory such as mood ,sense of failure, lack of satisfaction, irritability, feeling of guilt ,sense of punishment, self-hate, self accusations and inhibition of work [7].

Another work by Richardson studied the performance characteristics and validity of the Patient Health Questionnaire - 9 item (PHQ-9) as a tool for depression among adolescents [8]. Our system uses questionnaire similar to PHQ-9 that is, an enhanced version of it which covers all aspects or factors and symptoms leading to depression. Tzirakis et al. [9] inspired by the use of deep learning in detecting emotion, presented an approach to analyze emotion as well depression state of the person based on deep learning. They used Deep residual Network of 50 layers on visual data and Convolution Neural Network (CNN) on audio .

Recently, Rafiqul Islam performed depression analysis on data from Facebook collected from an online public source through machine learning technique as an efficient and scalable method [10].

III. Proposed Methodology

In this section, we have given the detail of used datasets, features extraction, proposed algorithms and model.

A. DataSet I from Questionnaire

Real time data is collected from students and parents by making them fill questionnaires similar to PHQ-9. Questionnaires were prepared keeping in mind the symptoms observed in a student while he/she is suffering through any level of depression and to what level, parents are involved in this scenario. Also, we consulted various counselors to make our dataset as effective as possible in determining severity of depression. The missing data is handled by filling out the spaces with the utmost possible answer. Following are the few features which were focused during the preparation of questionnaire. –

1. Age
2. Sex
3. Regularity in school/college
4. Feeling tired/ having little energy
5. Feeling down or hopeless.
6. Degree of insomnia
7. Poor appetite
8. Trouble concentrating on things.
9. Thoughts of getting dead
10. Intentionally overdosed on drugs.
11. Suffered from any physical/mental abuse

In total, 18 features were used for this dataset and 5 for the documentation purpose which includes timestamp that is the time at which the response has been recorded and comments section in which students can give additional information regarding the same. After circulating the form in various schools and colleges, 619 responses were recorded which were then preprocessed like handling missing values and converted to dataset.

For example, If the question is- Are you regular in school/college and has 4 choices, then these choices are encoded by the label encoder as in the Table I -

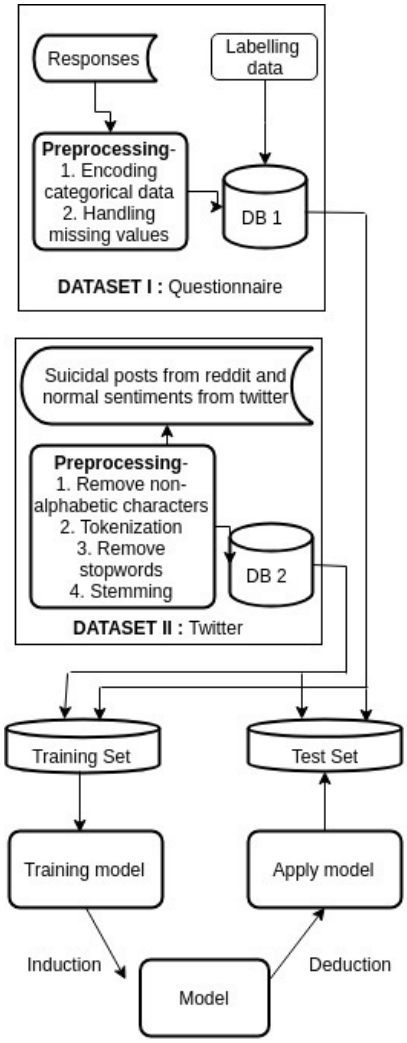


Fig. 1: Proposed Model of Depression Analysis and Suicidal Ideation Detection System

TABLE I: Encoded Values of the options

Options	Value
Not at all	0
Several Days	1
More than half the days	2
Nearly everyday	3

B. DataSet II from Twitter

We utilized PRAW for getting dataset(containing user posts related to depression and suicidal ideation) from Reddit, which stands for Python Reddit API wrapper - an opensource Python library for accessing the Reddit content. The Twitter posts(containing positive and negative sentiments but not related to suicide or depression) were collected from available dataset. As Reddit has dedicated sections called sub-reddits for suicide and depression related posts, we scrapped those posts from the sub-reddits. We also took positive sentiment

Algorithm 1 Pre-processing for Dataset I

procedure PREPROCESS($data$)
  for x in $data$.columns **do**
    if $data$[x].type == *object* **then**
      $data$[x] = transform($data$[x].values) ▷ Label Encoder is used for transformation
    end if
  end for
  return $data$

and negative sentiment (doesnt contain suicide related vocab) labeled dataset from Kaggle. After collecting posts from Reddit and Twitter, we removed excess white space and then changed the text to lower case. Then data cleaning is carried out, for that we carry out the following procedure. The first step involves removal of all non-alphabetic characters. The we removed the stopwords, for that we utilized NLTK stopwords corpus. Then data set was created by stemming the words (carried out primarily for feature reduction), we used porter stemmer for the task.

Algorithm 2 Pre-processing for Dataset II

procedure PREPROCESSII($Tweets$)
  for x in $Tweets$ **do**
    Remove non-alphabetic characters in x
    Do word tokeniszation
    Remove stopwords
    Do stemming
  end for
  return $Tweets$

C. Feature Extraction

*1) **DataSet I**:* Before training the collected data, data is preprocessed. Prepared questionnaire refers to PHQ-9 (Patient health questionnaire) and is the extended version of it. Questions asked for the documentation purpose including Email-address or school/college name which do not contribute in predicting depression stages are removed. Finally, scores are allotted to all the options depending upon the level to which they contribute in depression using LabelEncoder which is used to transform categorical labels to numerical labels. Then, on the basis of total scores of each data entry, depression stages are labeled to create a data set as shown in the table II.

We partitioned the data set in a 80-20 split where 80 % of the data is reserved for training and 20 % is marked for testing. Training set is shown in figure 2 by a PCA [1] plot where dataset characterized by more than 15 dimensions or features is plotted as points in a plane. PCA discover a new coordinate system with each point having a new (x,y) value. The axes in the system don't actually mean anything physical. In fact, they're combinations of features called "principal components" that are chosen to give one axes lots of variation. Since there are 5 stages of depression, classification machine algorithms are applied to train and test the data.

[1]Principal Component Analysis (PCA) is used to reduce a large set of features to a small set that still contains most of the information in the large set that is it is a dimension-reduction tool

Algorithm 3 Feature Labeling

procedure PROCESS($data$)
  Divide $data$ into $training-data$ and $testing-data$
  for x in $training-data$ **do**
    $sum \leftarrow 0$
    $n \leftarrow Number-of-features(columns)$
    $i \leftarrow 0$
    while $i \leq n$ **do**
      $sum \leftarrow$sum+x_i
    end while
    $x_{label} \leftarrow Label-according-to-the-sum$
  end for
  return $training-data$

TABLE II: Labels on the basis of scores for students dataset

Score	Depression severity	Label/Stage
0-5	Minimal or none	0
5-14	Mild	1
14-27	Moderate	2
27-39	Moderately severe	3
>30	Severe	4

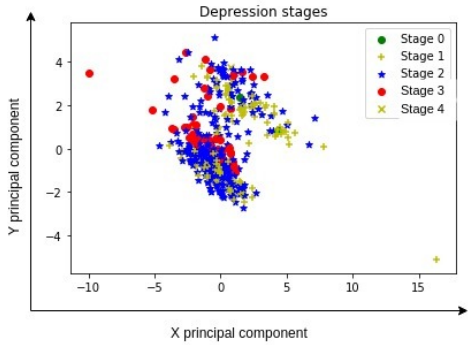


Fig. 2: PCA plot for Training DataSet of Questionnaire

For example, suppose the collected preprocessed data of two students after encoding the selected options from the questionnaire into numbers is shown in Table III- Now, according to responses, each column in the processed data will be labeled from 0 to 4 depending upon the severity of depression. Total score for each student will be calculated by the formula -

$$s_j = \sum_{i=1}^{n} a_i \tag{1}$$

where,
s_j = Score of the j^{th} student
n = Total no. of features
a_i = Value of i^{th} feature for j^{th} student in the data

For labeling , Table II is referred. In the above example, total score for student 1 is 12 and for student 2 , it is 26.

TABLE III: Labeling data

SNo	Features	Student 1	Student 2
1	What is your age?	2	2
2	Sex? (Male/Female)	1	0
3	Are you regular in school/college?	0	2
4	Feeling down or hopeless?	1	2
5	Insomnia, or sleeping too much?	1	3
6	Having little energy or feeling tired?	1	2
7	Overeating or poor appetite?	0	1
8	Feeling bad about yourself?	1	3
9	Trouble focusing on things?	1	2
10	Moving or speaking so slowly that other people could have noticed?	0	1
11	Thoughts that you would be better off dead?	0	2
12	Intentionally overdosed on drugs?	0	0
13	Have suffered any form of physical or mental abuse?	0	1
14	Little interest or pleasure in doing things?	1	2
15	How much time do you spend on social media?	2	2
16	Have you seen a student psychologist before?	0	0
17	Where do you stay?	0	1
18	Do you share everything with your parents?	1	1

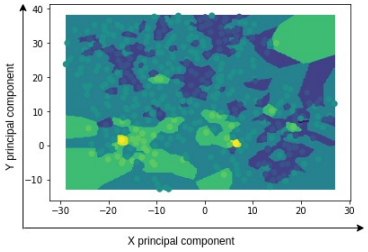
Therefore, we can label these as 1 i.e Mild depression and 2 i.e Moderate depression respectively.

*2) **DataSet II**:* We used Tf-idf that is Term Frequency Inverse Document Frequency, weighted word count feature extraction to form our feature vectors. We feed the test data to the tfidf vectorizer for creating feature vectors for the test set which will be used by our classifiers to predict. We now partition the data set II similar to data set I in a 80-20 split ratio where 80 % of the data is reserved for training and the rest 20 % is reserved for testing. We have ensured that there is equal representation of sentiments from both classes in training and testing set by performing random shuffling before partitioning.

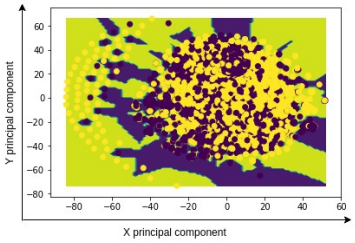
D. Machine Learning Supervised Algorithms

In order to determine suicidal ideation with the data set, text classification and sentiment analysis algorithms are utilized for both the datasets. For categorization into five levels of depression severity, supervised classification algorithms are used which includes Logistic Regression, Decision tree classifier and XGBoost algorithm on dataset I and similar algorithms on dataset II for categorizing them into yes or no are used [11]. PCA plots have been plotted for both datasets against all the applied algorithms. For dataset I, yellow color represent people having minimal or none depression that is 1^{st} stage, purple for mild depression, light green spots on dark green refers to moderate depression, dark green refers to moderately severe and light green refers to severe depression. For dataset II, yellow color refers to suicidal posts whereas purple refers to normal post.

1) Logistic Regression: It is one of the most widely used classifier in machine learning. Here the variable y, that we want to predict is discrete value. Ex spam or not spam, online transactions fraudulent or not [12]. It can be used for binary classification problem as well as multiclass classification problem. Decision boundary for both data sets are shown in figure 3.

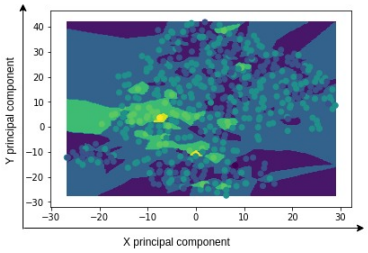


(a) Decision boundary for data set I

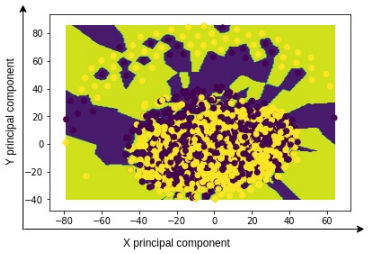


(b) Decision boundary for data set II. The two colors yellow and blue correspond to the two classes risky and not-risky respectively.

Fig. 3: Logistic Regression Decision Boundary



(a) Decision boundary for data set I



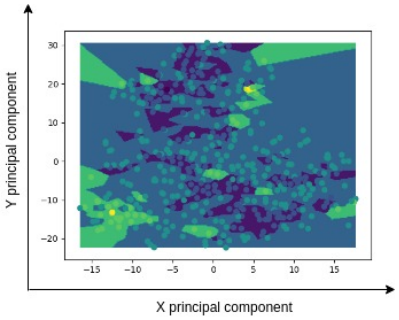
(b) Decision boundary for data set II.

Fig. 4: Random Forest Decision Boundary

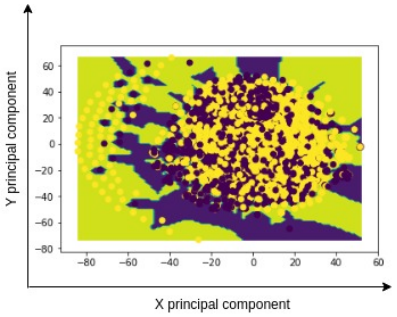
3.

2) Random Forest Classifier: A random forest classifier is a meta estimator that fits a number of decision tree classifiers on various sub- samples of the dataset .It improves the predictive accuracy using averaging and control over- fitting by ignoring outliers [13]. The PCA plot obtained for the classifier is shown in figure 4 in which 5 stages are visible in form of 5 different colors for dataset I.

3) XGBoost Classifer: XGBoost classifier is used for supervised machine learning problems, where the training data



(a) Decision boundary for data set I



(b) Decision boundary for data set II.

Fig. 5: XGBoost Decision Boundary

including multiple features x_i is used to predict a target variable y_i [14]. It is designed for speed and performance and is an implementation of gradient boosted decision trees . Decision boundary for both datasets are shown in figure 5.

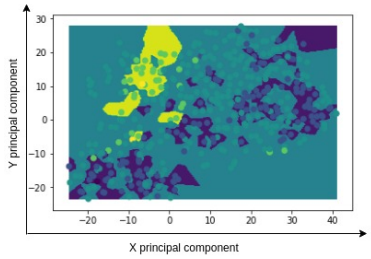
4) Support Vector Machine: Support vector machines are supervised machine learning algorithms that can perform non-linear classifications by mapping data to higher dimensions through the use of the kernel trick.The PCA plot obtained for the classifier is shown in figure 6 for both datasets.

IV. Results

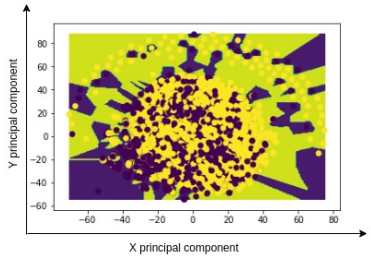
We tested classifiers using our pre-processed Test data sets. To assess the performance of different classifiers, we computed the accuracy of each for both datasets. A confusion matrix is a table that is used to describe the performance of a machine learning classifier on a set of test data in terms of accuracy, precision, Recall and F-measure for which the true values are known. Hence, we have calculated the accuracy using this matrix by the formula-

$$Accuracy = \frac{(TN + TP)}{(FN + FP + TN + TP)} \quad (2)$$

where,
TP = True positive;
TN = True negative;
FP = False positive
FN = False negative
For example, on applying Logistic regression on data set II, figure 7 shows the confusion matrix formed.



(a) Decision boundary for data set I



(b) Decision boundary for data set II.

Fig. 6: Support Vector Machine Decision Boundary

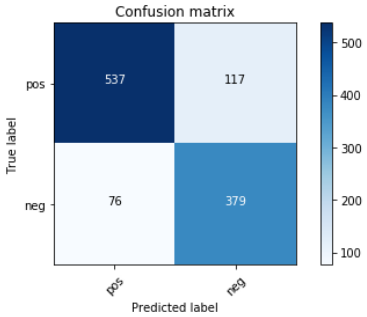


Fig. 7: Confusion matrix describing the performance of Logistic Regression on dataset II

TABLE IV: Accuracy against each algorithm applied on the dataset

Classifier	Accuracy for dataset I (%)	Accuracy for dataset II (%)
Random Forest Tree Classifier	76.34	82.05
XGBoost Classifier	83.87	84.02
Logistic Regression Classifier	59.22	86.45
Support Vector Machine	76.50	85.44

For dataset I, XGBoost classifier gave the highest accuracy i.e 83.87 % and Logistic Regression gave the lowest unlike dataset II where Logistic regression gave the highest accuracy of 86.45 % because dataset I consist of 18 features which were more than the words on which tweets were extracted and model was trained. Logistic regression starts to falter when there are large number of features and good chunk of missing data. Also, too many categorical variables are also a problem for it.

Graphs are designed to help one communicate the survey results. A series of graph are shown from figure 8 to 10 -

Maximum people facing depression are of age group 19-21

according to bar graph obtained as shown in Fig. 8.

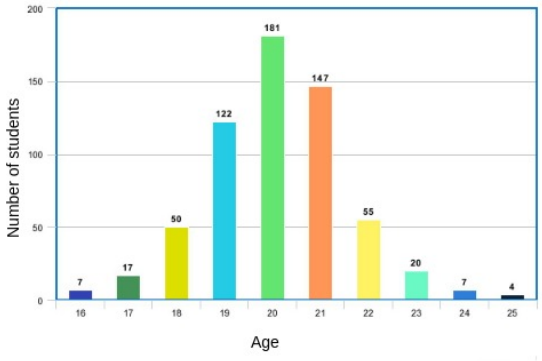


Fig. 8: Number of students in different age groups

Around 33.7percent people feels bad about themselves-that they are a failure or let their family down as shown in Fig. 9.

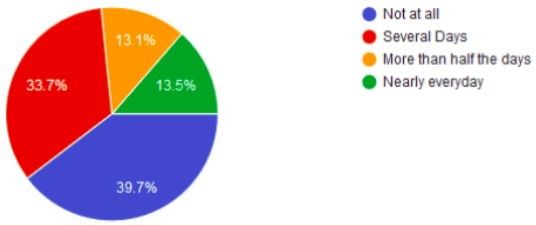


Fig. 9: Number of students feeling bad about themselves.

According to the graph formed by dataset , 48.2 % people used to feel hopeless and depressed on several days as shown in Fig. 10.

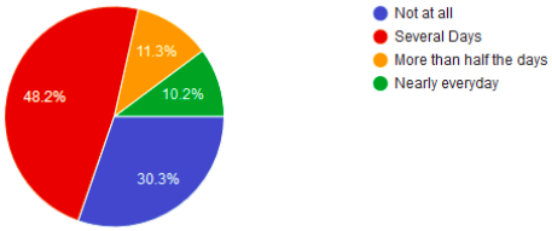


Fig. 10: Number of students feeling hopeless

V. Conclusion and Future Work

Now a days Depression is leading to catastrophic outcomes such as suicide which could be life taking as well. Studies have shown that life events preceding the onset of depression play a significant role in depression. We have analyzed social media posts (especially twitter),conducted questionnaire and asked students and parents to give their opinion and also scrapped blogs on internet .Major factors of depression among the age group of 15-29 which we found during the course of the project are parental pressure, love, failures, bullying, body shaming, inferiority complex, exam pressure, peer pressure, physical and sexual abuse etc. Depression being a recurrent type of illness, repeated episode of the same are common. Finally, little is known about the prevention and identification of the disorder at an early stage. Among future directions, we hope to understand how social media behavior analysis can help in leading to development of methods for analyzing depression at scale.

References

[1] S. Grover, A. Dutt, and A. Avasthi, "An overview of indian research in depression," *Indian journal of psychiatry*, vol. 52, no. Suppl1, p. S178, 2010.

[2] M. Reddy, "Depression: the disorder and the burden," *Indian journal of psychological medicine*, vol. 32, no. 1, p. 1, 2010.

[3] "Depression." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[4] M. M. Aldarwish and H. F. Ahmad, "Predicting depression levels using social media posts," in *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*. IEEE, 2017, pp. 277–280.

[5] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[6] "Depression and suicide: Overview, etiology of depression and suicidality, epidemiology of depression and suicide," Feb 2019. [Online]. Available: https://emedicine.medscape.com/article/805459-overview

[7] L. RADLOFF, "Scale: A self-report depression scale for research in the general population." *J Clin Exp Neuropsychol*, vol. 19, pp. 340–356, 1997.

[8] L. P. Richardson, E. McCauley, D. C. Grossman, C. A. McCarty, J. Richards, J. E. Russo, C. Rockhill, and W. Katon, "Evaluation of the patient health questionnaire-9 item for detecting major depression among adolescents," *Pediatrics*, vol. 126, no. 6, pp. 1117–1123, 2010.

[9] O. Whooley, "Diagnostic ambivalence: psychiatric workarounds and the diagnostic and statistical manual of mental disorders," *Sociology of Health & Illness*, vol. 32, no. 3, pp. 452–469, 2010.

[10] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health information science and systems*, vol. 6, no. 1, p. 8, 2018.

[11] A. Sabeeh and R. K. Dewang, "Comparison, classification and survey of aspect based sentiment analysis," in *International Conference on Advanced Informatics for Computing Research*. Springer, 2018, pp. 612–629.

[12] R. K. Dewang and A. K. Singh, "State-of-art approaches for review spammer detection: a survey," *Journal of Intelligent Information Systems*, vol. 50, no. 2, pp. 231–264, 2018.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[14] "A gentle introduction to xgboost for applied machine learning," Sep 2016. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/