Data Set Creation and Empirical Analysis for Detecting Signs of Depression from Social Media Postings

S. Kayalvizhi, Thenmozhi Durairaj

Data set creation and Empirical analysis for detecting signs of depression from social media postings

Kayalvizhi Sampath[1][0000−0002−5417−9910] and Thenmozhi Durairaj[1]

Sri Sivasubramaniya Nadar College Of Engineering, Chennai.
{kayalvizhis,theni_d}@ssn.edu.in

Abstract. Depression is a common mental illness that has to be detected and treated at an early stage to avoid serious consequences. There are many methods and modalities for detecting depression that involves physical examination of the individual. However, diagnosing mental health using their social media data is more effective as it avoids such physical examinations. Also, people express their emotions well in social media, it is desirable to diagnose their mental health using social media data. Though there are many existing systems that detects mental illness of a person by analysing their social media data, detecting the level of depression is also important for further treatment. Thus, in this research, we developed a gold standard data set that detects the levels of depression as 'not depressed', 'moderately depressed' and 'severely depressed' from the social media postings. Traditional learning algorithms were employed on this data set and an empirical analysis was presented in this paper. Data augmentation technique was applied to overcome the data imbalance. Among the several variations that are implemented, the model with Word2Vec vectorizer and Random Forest classifier on augmented data outperforms the other variations with a score of 0.877 for both accuracy and F1 measure.

Keywords: Depression · Data set · Data augmentation · Levels of depression · Random Forest

1 Introduction

Depression (major depressive disorder) is a common and serious medical illness that negatively affects the way one feels, thinks and acts [1]. The rate of depression is rapidly increasing day by day. According to Global Health Data Exchange (GHDx), depression has affected 280 million people worldwide [3]. Detecting depression is important since it has to be observed and treated at an early stage to avoid severe consequences[1]. The depression was generally diagnosed by different methods modalities clinical interviews [5][12], analysing the behaviour[6], monitoring facial and speech modulations[19], physical exams with Depression scales

[1] https://www.healthline.com/health/depression/effects-on-body

[14][28], videos and audios [18], etc. All these methods of diagnosing involves more involvement of an individual or discussion about their feeling in person.

On the other hand, social media is highly emerging into our lives with a considerable rate of increase in social media users according to the statistics of statista [4]. Slowly, the social media became a comfortable virtual platform to express our feelings. And so, social media platform can be considered as a source to analyse people's thoughts and so can also be used for analysing mental health of an individual. Thus, we aim to use social media texts for analysing the mental health of a person.

The existing works collect social media texts from open source platforms like Reddit [32], Facebook[13], Twitter [27][30][11][16], Live journals [20], blog posts[31], Instagram [26] etc. and used them to detect depression.

Research gaps:
All these research works concentrate on diagnosing depression from the social media texts. Although detecting depression has its own significance, detecting the level of depression also has its equal importance for further treatment. Generally, depression is classified into three stages namely mild, moderate and severe [2]. Each stage has its own symptoms and effects and so detecting the level of depression is also a crucial one. Thus, we propose a data set to detect the level of depression in addition to detection of depression from the social media texts. This paper explains the process of data set creation that detects the levels of depression along with some baseline models.

Our contributions in this research include:

1. Creating a new bench mark data set to detect the sign of depression from social media data at postings level.
2. Developing base line models with traditional learning classifiers.
3. Analysing the impact of data augmentation

2 Related Work

The aim of our research work is to create a data set that identifies the sign of depression and detect the level of depression and thus, the existing works are analysed in terms of data collection, modalities and methodologies of detecting depression.

2.1 Modalities and methodologies of depression detection:

For detecting depression, the data was collected by various methods like clinical interviews [5][12], analysing the behaviour[6], monitoring facial and speech modulations[19], physical exams with Depression scales [14][28], videos and audios [18], etc. Since, the social media users are rapidly increasing day by day, social media data can also be considered as a main source for detecting the mental

health. This key idea gave rise to the most utilized data set E-Risk@CLEF-2017 pilot task data set [17] that was collected from Reddit. In addition to this data set, many other data sets such as DAIC corpus [5], AVEC [18], etc. also evolved that detects depression from the social media data. Though few benchmark data set exists to detect depression, more researchers tend to collect data from social media and create their own data sets.

2.2 Data collection from social media:

The social media texts were collected from open source platforms like Reddit [32][29], Facebook[13], Twitter [27][30][11][16], Live journals [20], blog posts[31], Instagram [26] etc. The data from twitter was collected using API's and annotated into depressed and not depressed classes based on key words like "depressed, hopeless and suicide" [11], using a questionnaire [30], survey [27], etc. The data was also scrapped from groups of live journals [20], blog posts[31] and manually annotated into depressed and not depressed.

Among these social media platforms, Reddit possess large amount text discussion than the other platforms and so Reddit has become widely used platform to collect social media text data recently.

The data were collected from these platforms using Application Programming Interface (API) using hashtags, groups, communities, etc. The data from reddit was collected from Subreddits like "r/depression help, r/aww, r/AskReddit, r/news, r/Showerthoughts, r/pics, r/gaming, r/depression, r/videos r/todayilearned r/funny" and annotated manually by two annotators into depressed and not depressed class [32]. The data was also from subreddits like "r/anxiety, r/depression and r/depression_help" and annotated into a data set [24]. A data set was created with classes depression, suicide_watch, opiates and controlled which was collected using subreddits such as "r/suicidewatch, r/depression", opioid related forums and other general forums [33]. A survey was also done based on the studies of depression and anxiety from the Reddit data [8].

From the Table 1, it is clear that all these research works have collected the social media data only to detect the presence of depression. Although, diagnosing depression is important, detecting the level of depression is more crucial for further treatment. And thus, we propose a data set that detects the level of depression.

3 Proposed Work

We propose to develop a gold standard data set that detects the levels of depression as not depressed, moderately depressed and severely depressed. Initially, the data set was created by collecting the data from the social media platform, Reddit. For collecting the data from archives of Reddit, two way communication is needed, which requires app authentication. After getting proper authentication, the subreddits from which the data must be collected are chosen and the data was extracted. After extracting the data, the data is pre-processed and exported

Table 1: Comparison of existing data sets

Existing system	**Social Media Platform**	**Class Labels**
Eichstaedt et.al [13]	Facebook	Depressed and not depressed
Nguyen et.al [20]	Live journal	Depressed and control
Tyshchenko et. al [31]	Blog post	Clinical and Control
Deshpande et.al [11]	Twitter	Neutral and negative
Lin et.al [16]	Twitter	Depressed and not depressed
Reece et.al [27]	Twitter	PTSD and Depression
Tsugawa et.al [30]	Twitter	Depressed and not depressed
Losada et.al [17]	Reddit	Depression and Not depression
Wolohan et.al [32]	Reddit	Depressed and not depressed
Tadesse et.al [29]	Reddit	Depression indicative and standard
Pirina et.al [24]	Reddit	positive and negative
Yao et.al [33]	Reddit	Depression, Suicide watch, Control and Opiates
Proposed Data set	**Reddit**	**Not depressed, moderately depressed & severely depressed**

in the required format which forms the data set. The data were then annotated into levels of depression by domain experts following the annotation guidelines. After annotation, the inter-rater agreement is calculated to analyze the quality of data and annotation. Then, the corpus is formed using the mutually annotated instances.Baseline models were also employed on the corpus to analyze the performance. To overcome the data imbalance problem, data augmentation technique was applied and their impact on performance was also analyzed.

3.1 Data set creation:

For creating the data set, a suitable social media platform is chosen initially and data is scraped using suitable methods. After scraping the data, the data is processed and dumped in a suitable format.

Data collection: For creating the data set, the data was collected from Reddit[2], an open source social media platform since it has more textual data when compared to other social media platforms. This data will be of postings format which includes only one or more statements of an individual. The postings data are scraped from the Reddit archives using the API "pushshift".

App authentication: For scraping the data from Reddit achieves, Python Reddit API Wrapper(PRAW) is used. The data can be only scraped after getting authentication from the Reddit platform. This authentication process involves creation of an application in their domain, for which a unique client secret

[2] `https://www.reddit.com`

key and client id will be assigned. Thus, PRAW allows a two way communication only with these credentials of user_agent (application name), client_id and client_secret to get data from Reddit.

Subreddit selection Reddit is a collection of million groups or forums called subreddits. For collecting the confessions or discussion of people about their mental health, data was scraped from the archives of subreddits groups like "r/Mental Health, r/depression, r/loneliness, r/stress, r/anxiety".

Data extraction: For each posting, the details such as post ID, title, URL, publish date, name of the subreddit, score of the post and total number of comments can be collected using PRAW. Among these data, PostID, title, text, URL, date and subreddit name are all collected in dictionary format.

Data pre-processing and exporting: After collecting these data, the text and title part are pre-processed by removing the non-ASCII characters and emoticons to get a clean data set. The processed data is exported into a Comma Separated Values (.csv) format file with the five columns. The sample of the collected postings is shown in Table 2.

Table 2: Sample Postings data

Post ID	Title	Text	Url	Publish date	Subreddit
g69pqt	Don't want to get of bed	I'm done with me crying all day and thinking to myself that I can't do a thing and I don't what to get out of bed at all	https://www.reddit.com/r/depression/comments/g69pqt/dont_want_to_get_of_bed/	2020-04-23 02:51:32	depression
gb9zei	Today is a day where I feel emptier than on other days.	It's like I am alone with all my problems. I am sad about the fact I can't trust anyone and nobody could help me because I feel like nobody understand how I feel. Depression is holding me tight today..	https://www.reddit.com/r/depression/comments/gb9zei/today_is_a_day_where_i_feel_emptier_than_on_other/	2020-05-01 08:10:06	depression

3.2 Data Annotation

After collecting the data, the data were annotated according to the signs of depression. Although all the postings were collected from subreddits that exhibit

the characteristics of mental illness, there is a possibility of postings that do not confess or discuss depression. Thus, the collected postings data were annotated by two domain experts into three labels that denote the level of signs of depression namely "Not depressed, Moderate and Severe". Framing the annotation guidelines for postings data is difficult since the mental health of an individual has to be analyzed using his/her single postings. For annotating the data into three classes, the guidelines were formatted as follows:

Label 1 - Not depressed : The postings data will be annotated as "Not Depressed", if the postings data reflect one of the following mannerism:

- If the statements have only one or two lines about irrelevant topics.
- If the statements reflect momentary feelings of present situation.
- If the statements are about asking questions about any or medication
- If the statement is about ask/seek help for friend's difficulties.

Example:

I struggled to count to 20 today : For some context I work in mcdonalds and I was on the line finishing the stuff and almost every 20 box I think I sent was wrong. I just couldn't focus and do it quick and it was awful, and I ended up doing almost 2 hours of overtime because I can't say no to people. I fucking hate it, the worst part is I was an A grade maths student less than 5 years ago. Now I can't count to 20, I can't do basic maths without a calculator, I can barely focus at times. I feel like I'm just regressing as a person in every way but still aging, I just wanna end it all before it gets worse and I become a fucking amoeba of a person.

Label 2 - Moderately depressed : The postings data will be annotated as "moderately depressed", if the postings falls under these conditions:

- If the statements reflect change in feelings (feeling low for some time and feeling better for some time).
- If the statement shows that they aren't feeling completely immersed in any situations
- If the statements show that they have hope for life.

Example :

If I disappeared today, would it really matter?
I'm just too tired to go on, but at the same time I'm too tired to end it. I always thought about this but with the quarantine I just realised it is true. My friends never felt close to me, just like the only two relationships I have ever been in. They never cared about me, to the point where I even asked for help and they just turned a blind eye. And my family isn't any better. I don't know what to do, and I believe it won't matter if I do something or not. I'm sorry if my English isn't good, it isn't my first language.

Label - 3 : Severely depressed : The data will be annotated as "Severely depressed", if the postings have one of the following scenarios:

- If the statements express more than one disorder conditions.

– If the statements explain about history of suicide attempts.

Example:

Getting depressed again?
So I'm 22F and I have taken antidepressants the last time 4 years ago. I've had ups and downs when I got off and with 19 I was having a rough time for two months - started drinking and smoking weed a lot. Kinda managed to get back on track then and haven't been feeling too bad until now. Lately I've been feeling kinda blue and started making mistakes or have to go through stuff multiple times to do it correctly or to be able to remember it. Currently I'm having a week off and have to go back to work on monday. I just don't know I feel like I'm getting worse and want to sleep most of the time and at first I thought it's because I'm used to working a lot, but when I think about having to go back soon I feel like throwing up and at the same time doing nothing also doesn't sit well with me.I guess I'm kinda scared at the moment because I don't want to feel like I was feeling years ago and I still don't feel comfortable with my own mind and don't trust myself that I'm strong enough to pull through if depression hits me again.

3.3 Inter-rater agreement

After annotating the data, inter-rater agreement was calculated between the decisions of two judges using kappa coefficient estimated using a per-annotator empirical prior over the class labels [7]. Inter-rater agreement[3] is the degree of agreement among independent observers who rate, code, or assess the same phenomenon. The inter rater agreement is measured using Cohen's kappa statistics [10].

Table 3: Landis & Koch measurement table of inter rater agreement

Kappa value (κ)	**Strength of agreement**
< 0	Poor
0.01 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 0.99	Almost perfect agreement

The inter-rater agreement between the annotations was calculated using sklearn [22]. For our annotation, the kappa value (κ) is 0.686. According to Landis & Koch [15] in the Table 3, the κ value denotes substantial agreement between the annotators, which proves the consistency of labeling according to the annotation guidelines. Thus, the mutually annotated instances form the corpus.

3.4 Corpus Analysis

Initially 20,088 instances of postings data were annotated, out of which 16,613 instances were found to be mutually annotated instances by the two judges,

[3] https://en.wikipedia.org/wiki/Interrater_reliability

and thus they were considered as instances of data set with their corresponding labels. Table 4 shows the complete statistics of the corpus.

Table 4: Postings data analysis

Category	**Count**
Total number of instances annotated	20,088
Data set instances (*number of instances mutually annotated)*	16,632
Total number of sentences	1,56,676
Total number of words	26,59,938
Total number of stop-words	12,47,016
Total number of words other than stop-words	14,12,922
Total number of unique words	28,415
Total number of unique stop-words	150
Total number of unique words other than stop-words	28,265
Range of sentences per instance	1 - 260
Range of words per instance	1 - 5065
Average number of sentences per posting instance	9.42
Average number of words per posting instance	159.92

The whole corpus has 1,56,676 sentences with 26,59,938 words which shows the size of the corpus created. In the corpus, each posting with its labels is considered as each instance in the corpus. An instance in the corpus will have an average of 9.42 sentences each that varies in the range of 1 to 260 sentences with an average of 159.92 words that lies between 1 to 5065 words.

The distribution of the three class labels in the data set is shown in Figure 1. As

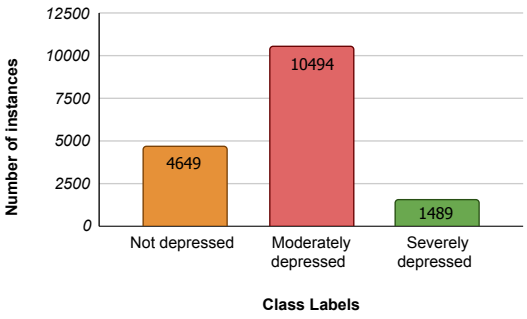


Fig. 1: Class wise distribution of the data set

shown in figure, the data set is unbalanced with 10,494 instances of “moderately

depressed" class, 1489 instances of "severely depressed" class and 4649 instances of "Not depressed" class which includes some duplicate entries.

3.5 Base line models

The data set has been evaluated using traditional models which are considered as baseline models. The data set has four columns namely id, title, text and class label. For implementation, the title data and text data are initially combined. The combined text data is pre-processed, extracted features, balanced, classified using traditional classifiers and evaluated by cross validation.

Data Pre-processing: The title and text column are combined together as a single text data column by filling the "NA" instances of both title and text data. The combined text data is cleaned by converting the words to lower case letters and removing unwanted punctuation, "[removed]" tags, web links, HTML links, stop words and small words(words with length less than two). After cleaning, the instances are tokenized using regexptokenizer [21], stemmed using porter stemmer [25] and lemmatized using wordnet lemmatizer.

Feature extraction: The features were extracted using three vectorizers namely Word2Vec, Term Frequency - Inverse Document Frequency (TF-IDF) vectorizer and Glove [23] vectorizer.

- **Word2Vec:** It produces a vector that represents the context of the word considering the occurrence of the word. The vectors are generated using Continuous Bag Of Words.
- **TF-IDF:** It produces a score considering the occurrence of the word in the document. It is based on the relevance of a topic in a particular document. The vectors are calculated using four grams considering a maximum of 2000 features.
- **Glove:** It produces the word embeddings considering the occurrence and co-occurrence of the words with reduced dimensionality. The words are mapped to a word embedding using 6 Billion pre-trained tokens with 100 features each.

Classifiers: Twelve different classifiers that include Ada Boost Classifier, Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbour, linear-Support Vector Machine, Linear Deterministic Analysis, Logistic Regression, Multi-layer Perceptron, Qua-
dratic Deterministic Analysis, Radial Basis Function - Support Vector Machine and Random Forest of Scikit-learn [21] were used for classification.

- **Ada Boost Classifier(ABC):** The Adaptive Boosting algorithm is a collection of N estimator models that assigns higher weights to the mis-classified samples in the next model. In our implementation, 100 estimator models with t0 random state at a learning rate of 0.1 were used to fine tune the model.

- **Decision Tree (DT):** The decision tree classifier predicts the target value based on the decision rules that was formed using features to identify the target variable. The decision rules are formed using gini index and entropy for information gain. For implementing the decision trees, the decision tree classifier was fine tuned with two splits of minimum samples of one leaf node each by calculating gini to choose the best split and random state as 0.
- **Gaussian Naive Bayes (GNB):** The Gaussian normal distribution variant of Naive Bayes classifier that depends on the Bayes theorem is Gaussian Naive Bayes.
- **K-Nearest Neighbour(KNN):** KNN classifies the data point by plotting them and finding the similarity between the data points. In implementation, number of neighbours were set as three with equal weights and euclidean distance as metric to calculate distance.
- **Logistic Regression (LR):** The probabilistic model that predicts the class label based on the sigmoid function for binary classification. As our data set are multi-class data sets, multi-nominal logistic regression was used to evaluate the data sets. For implementation, the classifier was trained with a tolerance of 1e-4, 1.0 as inverse of regularization strength and intercept scaling as 1.
- **Multi-layer Perceptron (MLP):** The artificial neural network that is trained to predict the class label along with back propagation of error. The multi-layer perceptron of two layers of 100 hidden nodes each was trained with relu activation function, adam optimizer, learning rate of 0.001 for a maximum 300 iterations.
- **Discriminant Analysis:** The generative model that utilizes Gaussian distribution for classification by assuming each class has a different co-variance. For implementation, the co-variance is calculated with threshold of 1.0e-04. Linear DA (LDA) and Quadratic DA (QDA) both were implemented.
- **Support Vector Machine:** The supervised model that projects the data into higher dimensions and then classifies using hyper-planes. The model was trained with RBF kernel (RBF-SVM) and linear kernel (L-SVM) function of three degree, 0.1 regularization parameter without any specifying any maximum iterations.
- **Random Forest (RF):** Random Forest combines many decision trees as in ensemble method to generate predictions. It overcomes the limitation of decision trees by bagging and bootstrap aggregation. It was implemented with 100 number of estimators.

4 Implementation and Results

The features extracted in subsection 3.5 are classified using the above classifiers in subsection 3.5 and evaluated using stratified k-fold sampling of Scikit-learn [21]. In this validation, data are split into 10 folds and the evaluation results with respect to weighted average F1-score is tabulated in Table 5. From the Table 5, it is clear that the model with Random Forest Classifier and Multi-Layer Perceptron (MLP) applied on the features extracted using Glove performs

Table 5: Performance of baseline models

F1 - score	TF- IDF	Glove	Word2Vec
ABC	0.263	0.496	0.451
DT	0.273	0.614	0.469
GNB	0.271	0.415	0.302
KNN	0.258	0.604	0.594
L-SVM	0.273	0.309	0.273
LDA	0.270	0.395	0.391
LR	0.270	0.329	0.395
MLP	0.269	**0.647**	0.625
QDA	0.276	0.459	0.368
RBF -SVM	0.273	0.560	0.452
RF	0.272	**0.647**	0.456

Table 6: F1 score

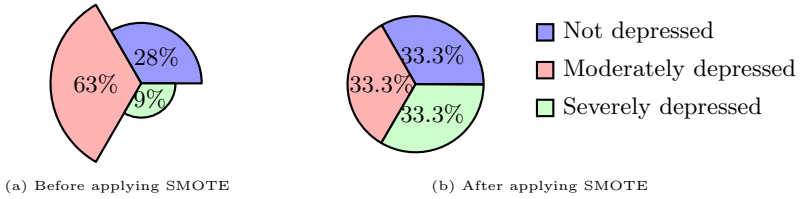
Accuracy	TF- IDF	Glove	Word2Vec
ABC	0.384	0.654	0.616
DT	0.388	0.697	0.579
GNB	0.351	0.464	0.351
KNN	0.379	0.717	0.694
L-SVM	0.388	0.646	0.623
LDA	0.388	0.659	0.619
LR	0.387	0.650	0.619
MLP	0.386	0.754	0.700
QDA	0.393	0.499	0.485
RBF -SVM	0.388	0.733	0.667
RF	0.388	**0.760**	0.695

Table 7: Accuracy

equally well with an F1-score of 0.647. The performance of the models with accuracy as metric is shown in Table 5. From the table, it is clear than the model with Random Forest classifier and Glove vectorizer performs better with an accuracy of 0.760.

4.1 With Data augmentation

The postings data is populated with more "moderately depressed" instances and thus, the data has to be balanced before classification for better performance. For balancing the data, Synthetic Minority Oversampling Technique (SMOTE) [9] was applied after vectorization. The effect of data augmentation can be observed in Figure 2.



(a) Before applying SMOTE (b) After applying SMOTE

Fig. 2: Effect of data augmentation in the data

The features extracted in subsection 3.5 are augmented using SMOTE and then classified using the classifiers in subsection 3.5. The performance of these models in terms of F1-score and accuracy after data augmentation are shown in Table 9 and 10 respectively. From the tables, it is clear that the performance was improved and model with Random Forest classifier applied on the features extracted using Word2Vec performs well with a score of 0.877.

Table 8: Performance of baseline models after data augmentation

F1 - score	TF- IDF	Glove	Word2Vec
ABC	0.451	0.622	0.559
DT	0.469	0.772	0.721
GNB	0.290	0.449	0.389
KNN	0.549	0.814	0.834
L-SVM	0.273	0.570	0.642
LDA	0.391	0.550	0.540
LR	0.395	0.544	0.551
MLP	0.625	0.775	0.852
QDA	0.368	0.592	0.477
RBF -SVM	0.452	0.762	0.788
RF	0.449	0.854	**0.877**

Table 9: F1-score

Accuracy	TF- IDF	Glove	Word2Vec
ABC	0.616	0.628	0.562
DT	0.579 0.781	0.728	
GNB	0.351	0.479	0.427
KNN	0.695	0.839	0.854
L-SVM	0.623	0.575	0.642
LDA	0.619	0.550	0.550
LR	0.619	0.547	0.559
MLP	0.700	0.780	0.857
QDA	0.485	0.615	0.497
RBF -SVM	0.667	0.769	0.792
RF	0.689	0.864	**0.877**

Table 10: Accuracy

The significance of improvement in the performance by incorporating data augmentation in terms of F1 score was measured using Benefit Cost Ratio (BCR) value. In general, BCR value is calculated by dividing the proposed total benefit cost by the proposed total cost. If the calculated value is greater than one, then the proposed cost is proven as significant one. In terms of performance, the metric is calculated by dividing the performance score of proposed model by the performance of existing model. Since, F1 score is considered to be the suitable performance metric, BCR value is calculated by dividing F1 score of model with data augmentation by that of model without data augmentation as shown below.

$$\text{BCR metric}_{(f1)} = \frac{\text{F1 score of proposed model (with data augmentation)}}{\text{F1 score of model without data augmentation}}$$

The BCR values calculated with F1 score are plotted in a graph and is shown in

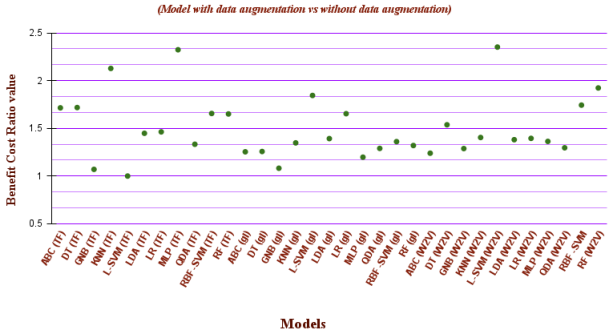


Fig. 3: BCR values of F1 scores

Fig.3. In the figure, vertical axis represents the scores, horizontal axis represents the different baseline models built and the data points represent the distribution of BCR values. From fig. 3, it is clear that the improvement in performance data augmentation is significant with respect to F1 scores since all the BCR values are greater than one.

5 Research insights

The researchers can further extend this work by implementing the following methods:

- Extend the data set by considering the images along with text data.
- Implement deep learning models in the data set.
- Implement other methods of data augmentation to improve performance.

6 Conclusions

Depression is a common mental illness that has to be detected and treated early to avoid serious consequences. Among the other ways of detecting, diagnosing mental health using their social media data seems much more effective since it involves less involvement of the individual. All the existing systems are designed to detect depression from social media texts. Although detecting depression is more important, detecting the level of depression also has its equal significance. Thus, we propose a data set that not only detects depression from social media but also analyzes the level of depression. For creating the data set, the data was collected from subreddits and annotated by domain experts into three levels of depression, namely not depressed, moderately depressed and severely depressed.

An empirical analysis of traditional learning algorithms was also done for evaluating the data sets. Among the models, the model with Glove vectorizer and Random Forest classifier performs well with a F1-score of 0.647 and accuracy of 0.760.

While analyzing the data set, "the moderately depressed" class seems to be highly populated than the classes and so, a data augmentation method named SMOTE was applied, and the performance is analyzed. Data augmentation improved the performance by 23% and 12% in terms of F1-score and accuracy respectively, with both F1-score and accuracy of 0.877. The significance of improvement in performance by incorporating data augmentation was also proved using BCR values.

The data set can also be extended by considering the images along with texts for more accurate detection. The work can be extended further by implementing other traditional learning and deep learning models. Other augmentation techniques can also be experimented with for improving the performance of the model.

Data set availability

The data set is available to the public in a repository of a Github in the link: https://github.com/Kayal-Sampath/detecting-signs-of-depression-from-social-media-postings.

References

1. American psychiatric association. https://www.psychiatry.org/patients-families/depression/what-is-depression, (Accessed: 2021-11-17)
2. Healthline. https://www.healthline.com/health/depression/mild-depression, (Accessed: 2021-11-17)
3. Institute of health metrics and evaluation. global health data exchange (ghdx). http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b, (Accessed: 2021-11-17)
4. Statista statistics. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, (Accessed: 2021-11-17)
5. Al Hanai, T., Ghassemi, M.M., Glass, J.R.: Detecting depression with audio/text sequence modeling of interviews. In: Interspeech. pp. 1716–1720 (2018)
6. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Hyett, M., Parker, G., Breakspear, M.: Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. IEEE Transactions on Affective Computing **9**(4), 478–490 (2016)
7. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008)
8. Boettcher, N., et al.: Studies of depression and anxiety using reddit as a data source: Scoping review. JMIR Mental Health **8**(11), e29487 (2021)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (Jun 2002). https://doi.org/10.1613/jair.953, http://dx.doi.org/10.1613/jair.953
10. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
11. Deshpande, M., Rao, V.: Depression detection using emotion artificial intelligence. In: 2017 international conference on intelligent sustainable systems (iciss). pp. 858–862. IEEE (2017)
12. Dibeklioğlu, H., Hammal, Z., Yang, Y., Cohn, J.F.: Multimodal detection of depression in clinical interviews. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 307–310 (2015)
13. Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A.: Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences **115**(44), 11203–11208 (2018)
14. Havigerová, J.M., Haviger, J., Kučera, D., Hoffmannová, P.: Text-based detection of the risk of depression. Frontiers in psychology **10**, 513 (2019)
15. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
16. Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., Leung, H.: Sensemood: Depression detection on social media. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 407–411 (2020)

17. Losada, D.E., Crestani, F., Parapar, J.: erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 346–360. Springer (2017)
18. Morales, M.R., Levitan, R.: Speech vs. text: A comparative analysis of features for depression detection systems. In: 2016 IEEE spoken language technology workshop (SLT). pp. 136–143. IEEE (2016)
19. Nasir, M., Jati, A., Shivakumar, P.G., Nallan Chakravarthula, S., Georgiou, P.: Multimodal and multiresolution depression detection from speech and facial landmark features. In: Proceedings of the 6th international workshop on audio/visual emotion challenge. pp. 43–50 (2016)
20. Nguyen, T., Phung, D., Dao, B., Venkatesh, S., Berk, M.: Affective and content analysis of online depression communities. IEEE Transactions on Affective Computing **5**(3), 217–226 (2014)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
22. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
24. Pirina, I., Çöltekin, Ç.: Identifying depression on Reddit: The effect of training data. In: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. pp. 9–12. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-5903, https://aclanthology.org/W18-5903
25. Porter, M.F.: An algorithm for suffix stripping. Program (1980)
26. Reece, A.G., Danforth, C.M.: Instagram photos reveal predictive markers of depression. EPJ Data Science **6**, 1–12 (2017)
27. Reece, A.G., Reagan, A.J., Lix, K.L., Dodds, P.S., Danforth, C.M., Langer, E.J.: Forecasting the onset and course of mental illness with twitter data. Scientific reports **7**(1), 1–11 (2017)
28. Stankevich, M., Latyshev, A., Kuminskaya, E., Smirnov, I., Grigoriev, O.: Depression detection from social media texts. In: Data Analytics and Management in Data Intensive Domains: XXI International Conference DAMDID/RDCL'2019. p. 352 (2019)
29. Tadesse, M.M., Lin, H., Xu, B., Yang, L.: Detection of depression-related posts in reddit social media forum. IEEE Access **7**, 44883–44893 (2019). https://doi.org/10.1109/ACCESS.2019.2909180
30. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems. pp. 3187–3196 (2015)
31. Tyshchenko, Y.: Depression and anxiety detection from blog posts data. Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia (2018)
32. Wolohan, J., Hiraga, M., Mukherjee, A., Sayyed, Z.A., Millard, M.: 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. pp. 11–21 (2018)

33. Yao, H., Rashidian, S., Dong, X., Duanmu, H., Rosenthal, R.N., Wang, F.: Detection of suicidality among opioid users on reddit: Machine learning–based approach. Journal of medical internet research **22**(11), e15293 (2020)