Investigation of Depression Using Context Analysis

Salma Akter Asma[1], Sadik Hasan[2], Nazneen Akhter[3], Mehenaz Afrin[4], Afrina Khatun[5], Kazi Abu Taher[6]

Abstract

Depression is a major concern in today's time as it is becoming a pandemic worldwide. Nowadays people (especially the young generation) are using social media sites to share their feelings, emotions, and personal life activities. Their mental health condition can be analysed by reviewing their social media posts and activities. Recent research work in this field is trying to go beyond manual depression detection. Hence, an automated system is necessary for analysing depression symptoms from social media for the sake of society. For this purpose, in this work, a Machine Learning based depression detection technique has been proposed. To develop the model six Machine Learning (ML) classifiers namely Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbour (KNN), Passive Aggressive (PA), Random Forest (RF), and Bagging classifier have been used. To improve the performance of the classifiers a dimension reduction technique namely Latent Semantic Analysis (LSA) is used. A comparison among four-dimension reduction techniques such as Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Fast Independent Component Analysis (Fast ICA) is given to justify why LSA is considered a dimension reduction technique in this work. With LSA, the Bagging classifier reached the top performance with an accuracy of 94.62%, while the base classifier is RF.

Keywords: Machine Learning, Depression, Dimension Reduction, Contextual Meaning, LSA.

---

[1] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: aaliyaasma786@gmail.com

[2] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: sadik.h.emon@gmail.com

[3] Department of Computer Science and Engineering, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: nazneen.akhter@bup.edu.bd

[4] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: agroti202@gmail.com

[5] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: afrina.khatun@bup.edu.bd

[6] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh, Email: kataher@bup.edu.bd

1. Introduction

Nowadays, depression is the most alarming disorder that has spread mental illness globally. More than 264 million people of different ages have been suffering from this illness (Safiri et al., 2022).

Though depression often goes unnoticed by people for the maximum time, the consequences can be devastating. According to the World Health Organization (WHO) survey in 2012, nearly one million people commit suicide yearly due to depression (Hassan et al., 2017). They estimate that around 322 million people worldwide will have mental illness within 2030 (Hur et al., 2018). However, approximately 70% of patients would not consult a doctor in the primary state of depression which is the manifestation of extreme unconsciousness (Murray & Lopez, 1997). The governance of mental health issues is so different because of the dynamic variety of human psychology, complex to fetch the pattern of mental behavior, mental illness is noticed in the last stage, and people normally refuse it fearing culture and stigma (Gupta et al., 2021).

At present, more than half of the world's population is using social media sites. The trend of using social media has increased, especially in pandemic situations. By analyzing the posts on social media, a person's emotions can be detected. This detection system can allow them to be aware of their mental health. This information can also play a momentous role in the decision-making process of a psychologist. For clinical depression, normally a psychologist evaluates his patients by taking a depression test based on questionnaires and academic interviews and recording it. But sometimes, it is not enough for detecting depression properly (Wang et al., 2013). Pointedly, these records are restricted because of many factors, such as sex, age, privacy, etc. To go beyond the boundary of clinical data, text mining tools extract and analyze depression syndrome from social media platforms such as Twitter, Facebook, and Instagram (Ma et al., 2017). Different techniques like ML classifiers, hybrid classifier models, and some new classifier models are proposed.

This paper is organized as follows: Section 2.1, states the background work regarding depression detection. A brief description of the proposed model is presented in Section 3.1. The result and performance are addressed by answering some research questions in Section 4.1. Section 5 concludes the paper by highlighting future works.

2. Background of the Study

Researchers are proposing different tools and techniques to detect depression from social media posts. Machine Learning techniques are frequently used in this area. The following section briefly discusses the related work in this field.

Mustafa et al. (2020) proposed an ML-based depression detection technique from Tweets. They used four ML classifiers such as Neural Network (NN), SVM, RF, and 1D Convolutional Neural Network (1DCNN) to build the model. They identified some keywords from the dataset and assigned weights to them. Subsequently, they matched the weighted words with previously generated fourteen psychological attributes in Linguistic Inquiry and Word count (LIWC) to classify those words into their respective classes of emotions. They classified them into three levels of depression (High, medium, and low) with an accuracy of 91%.

An AD prediction model for detecting anxious depression prediction in real-time tweets was proposed by Kumar et al. (2019). According to user posting features, they set five-tuple vectors such as words, timing, frequency, sentiments, and contrast. They developed their model using four ML classifiers such as Multinomial Naive Bayes (MNB), Gradient Boosting (GB), and RF, and Voting (majority voting) classifier which gave 85.09% of accuracy. For detecting the targeted words and transforming them into vectors using one-hot encoding and Word Embedding including the Word2Vec method, Ma et al. (2017) used an ML-based depression detection technique. Koltai et al. (2021) targeted specific hashtags which indicated depression. They used different techniques like NLP, NN, Latent Dirichlet Allocation (LDA), and LSA to get negative, and positive parts for a post. For classifying emotions into six categories like happiness, sadness, fear, anger, surprise, and disgust by analyzing the social media posts Gaind et al. (2019) used two different approaches: NLP including textual features (emoticons, degree words, negations, part of speech, and grammatical analysis), and Machine Learning classification algorithms. Finally, they achieved 91.7% accuracy using J48, and 85.4% accuracy using the SMO classifier.

By analyzing the social media texts Hassan et al. (2017) determined the binary and multi-class sentiment classification. They did feature extraction using POS Removal of stop words unigram, stemming, negation checker, and sentiment analyzer. They made comparisons among SVM, Naive Bayes (NB), and Maximum Entropy (ME) classifiers. The comparisons among classifiers were based on depression measurements where the accuracy of SVM was 91%, NB was 83%, and ME was 80%. Wang et al. (2013) proposed an ML-based depression detection technique for Chinese text. They used Waikato Environment for Knowledge Analysis (Weka) to develop their model. They considered BayesNet, Trees (J48),

and Rules (Decision Table) classification techniques to detect depression. The average ROC of the three classifiers was 85%. Moreover, it was 80% acceptable for the psychologist to detect depressed users in SNS. Balabantaray et al. (2012) were concerned about opinion mining and sentiment analysis by Natural Language Processing (NLP) and text mining that deals with automated discovery and classified emotions into six categories such as positive, negative, fear, joy, surprise, hate, and disgust. They used an emotion classifier based on multi-class SVM kernels which converted the seam words into numeric data. They reported that the accuracy was 72.34%.

A novel supervised algorithm namely Sequential S3 (SS3) for early depression detection proposed by Burdisso et al. (2019). The SS3 algorithm takes less time to classify than the other individual classifiers for example SVM, MNB, and Neural Network. The F-Score and precision are 0.61 and 0.63 respectively. Detecting depression in Reddit social media Forum Tadesse et al. (2019) proposed an ML-based model. For feature extraction, they applied a combination feature such as LIWC dictionary, LDA topic, and N-gram. LDA was chosen to reduce the input of the text data and to extract topics (features) from the text. MLP classifier achieved the highest accuracy, which is 91%. Islam et al. (2018) proposed an ML-based model for detecting depression on Facebook. They used four classifiers DT, SVM, KNN, and Ensemble classifiers. All classifiers' accuracy was between 60 to 80 percent. Chiong et al. (2021) used seven different ML classifiers namely Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Multilayer Perceptron (MLP), Bagging Predictors (BP), Random Forest (RF), Adaptive Boosting, and Gradient Boosting (GB) for detecting depression from Twitter posts. Among them, Gradient Boosting (GB) classifier achieved the highest performance, with an accuracy of more than 98%. A hybrid algorithm which is dual classification with the fusion of SVM, and Naive Bayes (NB) algorithm used by Smys and Raj (2021). They have reported that hybrid classifiers (SVM, and NB) brought them higher accuracy than single classifiers (SVM, DT, RT, and NB).

Researchers are still facing many challenges regarding depression detection from text. However, recent trends in depression detection have adopted different techniques to enhance the performance of the ML classifiers. But they have not given enough focus on the contextual meaning of the text. In this work, we have mainly focused on the contextual meaning of the text whereas other dimension reduction techniques focus only to extract features and reduce dimensions. It can bring out the internal meaning of the text that helps to train the model more efficiently. Hence, in this work, a dimension reduction technique namely Latent Semantic Analysis (LSA) has been adopted to enhance the performance of the classifiers. This technique can extract the contextual meaning of the text which

helps the classifiers to achieve higher accuracy. In the meantime, a comparative study is given among four-dimension reduction techniques namely Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Fast Independent Component Analysis (Fast ICA). This study will help new practitioners to understand the importance of the dimension-reduction technique and help them to decide which dimension-reduction technique would be suitable for their model.

3. Depression Detection Model

In this paper, a depression detection technique using ML classifiers has been proposed. At first, data pre-processing techniques are applied to clean, transform and reduce the dimension of data which tends the model to work efficiently. Henceforth, models are evaluated using the six ML classifiers. Then, the outcome is reported based on the performance of the individual classifiers. Figure 1 shows the proposed model. The details are described in the following.

3.1 Machine Learning Classifiers

In this work, six ML classifiers namely SVM, DT, KNN, PA, RF, and Bagging have been used. The classifiers have different features to determine the optimal solution. A brief description of the classifiers is given below:

i. Support Vector Machine (SVM): It utilizes statistical learning theory to give optimized solutions. It fits the given dataset which returns a hyper-plane named 'best fit'. This hyper-plane segregates the dataset into classes. Utilizing the hyper-plane new classes are mapped into higher dimensional space and predicted what the class will be (Evgeniou & Pontil, 1999).

ii. Decision Tree (DT): This classifier consists of a root, internal node, branch, and leaf. To predict the class label, it starts working from root nodes, where the root node indicates the best attribute of the given dataset (Sharma & Kumar, 2016). The dataset is split into subsets. It compares the root attributes with the internal node attributes that represent a branch. It continues until it obtains the predicted class label at the leaf node.

iii. Random Forest (RF): It is an ensemble Decision Tree classifier. It adds randomness to the given dataset when building an individual decision tree and aggregates all of them. RF searches for the best feature while splitting the nodes among random subsets. All these combinations offer a more accurate and stable predicted class label (Biau & Scornet, 2016).

iv. K-Nearest Neighbor (KNN): It is a non-parametric algorithm. It is a method of finding the distance between the class and unknown class

v. of the dataset. It searches for the nearest neighbor between the classes and picks the class which gets the most votes. Afterward, this class is labeled as a predicted class (Novakovic et al., 2016).

vi. Passive Aggressive (PA): It is an online learning algorithm that works with margin base concept. It responds passive to correct classifications but is aggressive to the wrong classification. It penalized the model if it got an incorrect prediction. The model will make changes if the prediction is wrong (Crammer et al., 2006). It updates the classifier, adjusts it into the model, and labels it as the predicted class.

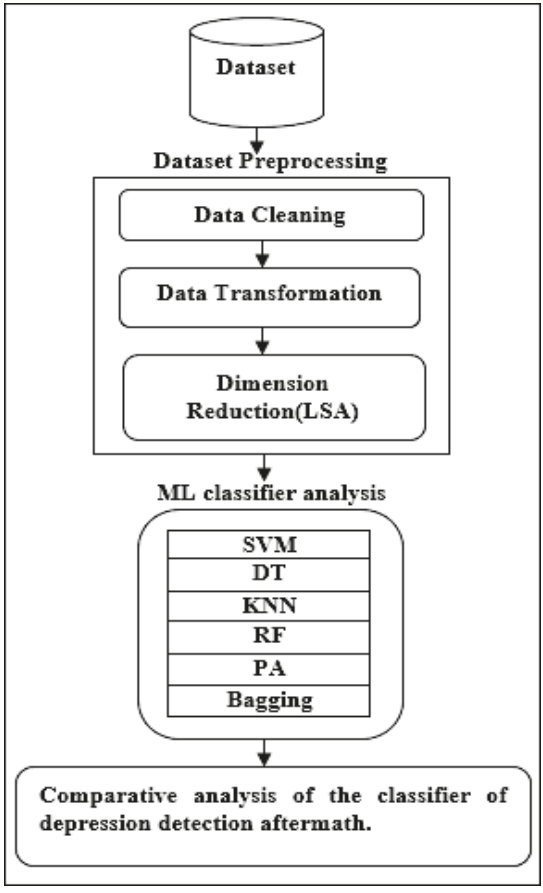

Figure 1: Flow diagram of depression detection technique using ML classifiers

vii. Bagging: It is an ensemble classifier that utilizes multiple models of the base classifier. It trains every model on a different set of data that

follows a technique: raw sampling with replacement (B uhlmann, 2012). Subsequently, it aggregates all the trained models using a voting method to build a stronger classifier whose predictive power is greater than the individual classifier. It reduces the variance of the dataset and gets rid of the trouble of over-fitting.

3.2 Data Pre-processing

In text analysis, data pre-processing is an important part as this can remove noise and unwanted elements from the text data and make the dataset more convenient to use. In this experiment, many data pre-processing techniques have been used for different purposes. They are discussed below.

3.2.1 Data Cleaning and Transformation

At first, URLs, duplicate words, extra spaces, and user mentions are removed to save time while analyzing the dataset. All the punctuations are erased to precise the data as it removes unnecessary signs which do not carry any information regarding depression. Then, all the stop-words are removed as they carry negligible information. For removing the suffix or prefix of a word or to find a root word, the Porter stemming technique has been applied. Further, tokenization is done with the RegexpTokenizer toolkit to split the tweets into words and then convert them into lowercase.

Frequency-Inverse Document Frequency (TF-IDF) vectorizer has been used to deal with the most frequent word in the dataset. It transforms the data with encoded numbers that carry the weights of the word and counts their frequent appearance. For contextual analysis, the uni-bigram model has been applied with the TF-IDF vectorizer.

3.2.2 Dimension Reduction Technique

For data pre-processing, dimension reduction is one of the most robust methods to reduce the data size as well as keep the variation of a dataset as much as possible. Among all the variations of data pre-processing, dimension reduction is the best choice for increasing the accuracy level because it avoids the over-fitting problem and takes care of multi-co-linearity. Besides, it helps with data visualization and is also useful for factor analysis.

i. Latent Semantic Analysis (LSA): In this work, A dimension reduction technique namely Latent Semantic Analysis (LSA) is used. It is also known as Latent Semantic Indexing (LSI). It reduces the dimension of the matrix. As LSA is language-independent, the dataset does not need to maintain the grammatical or auxiliary structure. While users are on social media expressing their thoughts, most of them are not maintaining sentence structure rules. LSA utilizes the sample vector directly, which

organizes the text structure semantically that helps the user's request for matching more accurately (Dumais et al., 1988) (Halko et al., 2011).

For example: instead of writing (I am a good kind of person), they write (I kinda good person). Although they are not properly making their sentences, we can easily extract the raw information from it through LSA. It does not depend on a specific word, string, or meaningful phrase. It can adapt to any kind of new, changing, or emerging thing. It is not sensitive to noise (unspelled data, arbitrary string) at all, any type of data can be read. It can perform context-based categorization (Landauer et al., 1998).

ii. Principal Component Analysis (PCA): Principal component analysis (PCA) is a method used to minimize the dimension of large datasets. It converts a huge quantity of variables into a small one and protects most of the information. It is like a characteristic expulsion method where we create new independent characteristics from the old. Initially, it normalizes the data. Thereafter, it calculates the covariance matrix. The next step calculates eigenvalues and eigenvectors. After completing the calculation, it chooses the components and forms the features vector. Finally, it forms principal components (Abdi & Williams, 2010).

iii. Latent Dirichlet Allocation (LDA): Initially, LDA concentrates on projecting the features in higher dimension space to the lower dimension. First, it is needed to calculate the separability between classes which means the distance between the mean of the different classes called between-class variance. Then, it calculates the distance between the mean and sample of each class called within-class variance. In the end, it fabricates the lower-dimensional space that can maximize the between-class variance and minimizes its within-class (Tran et al., 2019).

iv. Fast Independent Component Analysis (Fast ICA): It finds the latent topic in the text document. It presents all the hidden latent variables as linear combinations; those are statistically independent. In the beginning, it was used for signal processing but later it was found that it is also good for text analysis. The hidden latent variables are the text document topics, which provide the probability distributions on the universe of terms (Qi et al., 2001).

4. Result Analysis

To evaluate the effectiveness of detecting depression from social media (Twitter post), individual classifiers such as Decision Tree (DT), Support Vector Machine (SVM), K- Nearest Neighbour (KNN), Passive Aggressive (PA), Random Forest (RF), and Bagging have been used along with four-dimension reduction technique

namely Latent Semantic Analysis (LSA), Principal component analysis (PCA), Latent Dirichlet Allocation (LDA), and Fast Independent Component Analysis (Fast ICA). The result is reported based on four evaluation metrics namely precision, recall, accuracy, and F1-score. This work is conducted with an Intel Core i5 processor, 4GB RAM, 64-bit operating system, and Windows 10 education. Spyder (python 3.8) is used to implement models. Experimental details are discussed in the following.

4.1 Data Set for Depression Detection Model

The experiment is conducted with an existing Twitter dataset provided by Shen et al. (2017). They have collected their dataset from the Twitter API's. This dataset has been categorized into three parts namely D1, D2, and D3. D1 contains 292,564 depressed-related tweets of 1,402 users. D1 dataset has been collected from the tweets between 2019 to 2016. Here the tweets are marked as depressed when some words such as I'm, I was, I am found in the tweets. D2 contains 300 million users and 10 billion non-depressed related tweets, and it is fetched from the December 2016 posted tweets. The author labeled D2 dataset as non-depressed when the 'depress' word was never found in the user tweets. D3 consists of 36,993 depressed candidate users and more than 35 million tweets and these tweets are also from December 2016. Each dataset contains three subsets such as timeline, tweets, and users. The timeline subset consists of each user one monthly tweet post. The tweet subset contains unique tweets from the users. Finally, the user's subset contains information about each user. In this experiment, we have used the tweet subset. There are 6493, 5384, and 58900 tweets in D1, D2, and D3 tweet subsets respectively. The tweets are in different languages. D1, D2, and D3 tweets subset consist of 10, 78, and 91 types of language.

To train our model, 5000 tweets were randomly selected from both D1 and D2 datasets tweet subsets. Also, a similar number of depressed and non-depressed tweets has been collected from the D3 dataset tweet subset for testing our model. The D3 tweet dataset has been labeled with the **Table 1** lexicon words. **Table 1** presents 20 depression-related words. These words are the most frequent in the D1 dataset which is significantly related to depression. If the tweet contains any of these words, it is labeled as '1' means depression otherwise '0' means non-depression. Before comparing with **Table 1** there some pre-processing techniques are applied such as lower casing, hash-tag removal, and URL removal. After labeling it was found that there were 50,710 depression-related tweets and 8,190 normal tweet

Table-1: Word list for labeling of D3 dataset.

Abandon, abuse, problem, suffer, loser, fail, painful, depressed, depression, diagnosed, suicidal, broke, helpless, tired, torture, sick, ugly, insomnia, PTSD, destroy.

4.2 Performance Evaluation Metrics

Performance evaluation metrics mean the measurement of the quality of ML models using various measuring quantities. Evaluation metrics assure the legitimation and appeasement of a model (Botchkarev, 2019). For providing parallelism between the techniques, these metrics are used thoroughly. Here, four evaluation metrics are used namely precision, recall, F1-score, and accuracy. In this section, a descriptive study of these metrics is given.

i. True Positives (TP): It provides the number of states where the depression detection model can find out the real response to depression called true positive (T P) (Flach, 2003).

ii. True Negatives (TN): It presents the number of statues in which the model cannot find any depression, and no depression happens called the true negatives (TN) (Awoyemi et al., 2017).

iii. False Positives (FP): It indicates the number of times where the model detects depression, but it did not happen in the actual case (Zalpour et al., 2020).

iv. False Negatives (FN): It placed the falsely detected depression in the model, but it did not happen in the real case (Hemdan et al., 2020).

v. Precision (P): It provides the proportional positive measurement accurately (Pecorelli et al., 2019).

$$\mathrm{P} = \frac{TP}{TP + FP} \qquad (1)$$

vi. Recall (R): It grants the calculative result of actual positive value which are detected perfectly (Kurtanovi c & Maalej, 2017).

$$\mathrm{R} = \frac{TP}{TP + FN} \qquad (2)$$

vii. F1-Score (F): It approves the combination result of both precision and recall into a single meter (Ban et al., 2019).

$$\mathrm{F} = \frac{2 * P * R}{P + R} \qquad (3)$$

viii. Accuracy: The accuracy of an ML classifier means how it classifies data points accurately (Garc ıa et al., 2009). Accuracy refers to how many data point is perfectly predicted from all the data points. In other words, it is detected as the ratio of summation of the number of true positives and true negatives and the summation of the number of true positives, true negatives, false positives, and false negatives.

4.3 Research Questions and Evaluation

In this section, the model is empirically satisfied by addressing three questions RQ1, and RQ2, RQ3. RQ1 demonstrates how the classifiers perform to detect depression. RQ2 states how LSA influences the performance of the classifiers. RQ3 determines LSA is the best among these four-dimension reduction techniques. The following discussion covers the whole description of the above-mentioned research questions along with their assessment.

RQ1: How do the classifiers perform to detect depression from the text?

Among all six classifiers Bagging classifier performs the best in terms of accuracy whereas RF is the base classifier. This mixture helps to keep progressing the accuracy level. Bagging and RF both hold the bootstrap sampling feature and joining that similar feature makes the classifier more potential. The accuracy of the Bagging classifier with RF is 94.62%. **Table 2** shows that with Bagging, the RF classifier also achieves the highest precision level which is 0.95. Figure 2 shows the performance of the Bagging classifier with four different dimension reduction techniques.

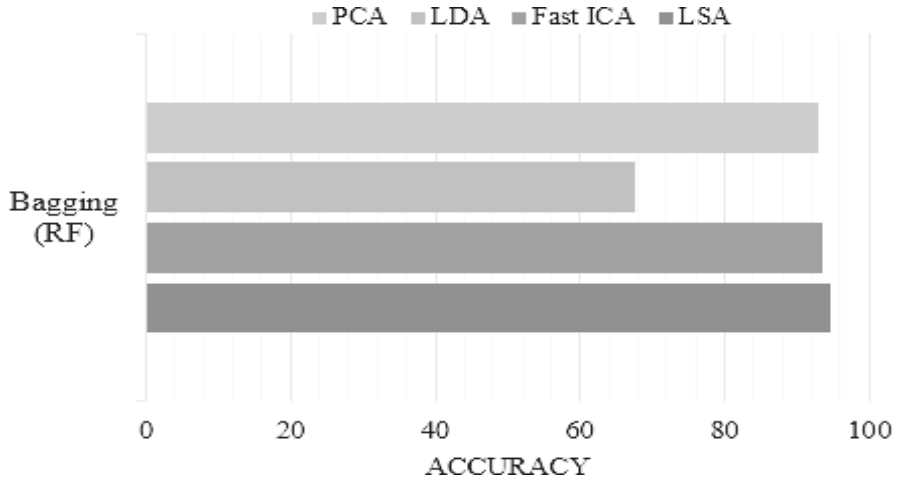

Figure 2: Performance of the dimension reduction techniques for Bagging classifier in terms of accuracy

Afterward, **Table 2** shows that ensemble classifier RF reveals a competitive result of 94.39%, as RF prevents overfitting problems with the help of multiple trees. **Table 2** shows that PA achieves the next highest result obtaining 93.88% accuracy. PA reacts passively for the right classification. KNN plays a substantial role as it is stable for the higher K number (k=8) for majority voting. Since KNN decreased the feature space taken as input, it obtains good accuracy.

Table-2: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score with LSA.

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	92.11	0.92	0.92	0.92
DT	92.73	0.92	0.92	0.92
KNN	70.04	0.78	0.7	0.67
RF	94.39	0.94	0.94	0.94
PA	93.88	0.94	0.93	0.93
Bagging (RF)	94.62	0.95	0.94	0.94

Table 3 shows that KNN obtained 93.53% accuracy. The fifth best result is 92.73% which is achieved by the usual DT classifier since it minimizes the characteristics of a tree and gives a better prediction. SVM performs with an accuracy of 92.75%, while the kernel is Radial Basis Function (RBF). It executes the linear manipulations for mapping points into higher dimensional space, which makes it easier to separate the classification to make predictions.

Table-3: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score with Fast ICA.

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	92.75	0.929	0.928	0.928
DT	92.09	0.921	0.921	0.921
KNN	93.53	0.937	0.935	0.935
RF	93.47	0.936	0.935	0.935
PA	89.51	0.909	0.895	0.894
Bagging (RF)	93.58	0.937	0.936	0.936

RQ2: How dimension reduction technique influences the performance of the classifiers?

Table 2, Table 3, Table 4, Table 5, and Table 6, illustrate the performance of LSA, Fast ICA, LDA, PCA, and without dimension reduction technique with ML

classifiers in terms of precision, recall, F1-score, and accuracy respectively. Dimension reduction is one of the most robust methods to reduce the data size as well as keep the variation of a dataset as much as possible. It takes less time to analyze the data.

Table-4: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score with LDA.

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	62.48	0.643	0.625	0.613
DT	68.08	0.733	0.681	0.662
KNN	61.05	0.684	0.611	0.567
RF	67.38	0.759	0.674	0.645
PA	63.80	0.640	0.638	0.637
Bagging (RF)	67.50	0.763	0.675	0.640

By analysing all the dimension reduction techniques this paper came up with the decision that dimension reduction techniques have a great impact on the performance of the classifiers. The dimension reduction technique can boost the performance of the classifiers such as LSA and can decrease the performance, such as LDA.

Table-5: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score with PCA**.**

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	90.81	0.921	0.908	0.907
DT	89.92	0.903	0.899	0.899
KNN	63.44	0.764	0.634	0.583
RF	93.13	0.934	0.931	0.931
PA	90.42	0.912	0.904	0.904
Bagging (RF)	92.90	0.932	0.929	0.929

Figure 3 shows that among all the classifiers Bagging performs best along with LSA which is 94.62%. **Table 6** shows that before applying reduction techniques, Bagging achieved an accuracy of 94.45%. **Table 3** illustrates that adopting Fast ICA also enhances the result in terms of accuracy as compared to without dimension reduction techniques though it performs less than LSA.

Fast ICA transforms the text into independent components, which makes it easier to separate the classification. Feeding PCA into the model does not give a good result. PCA does not work with finding the important patterns of the dataset all the

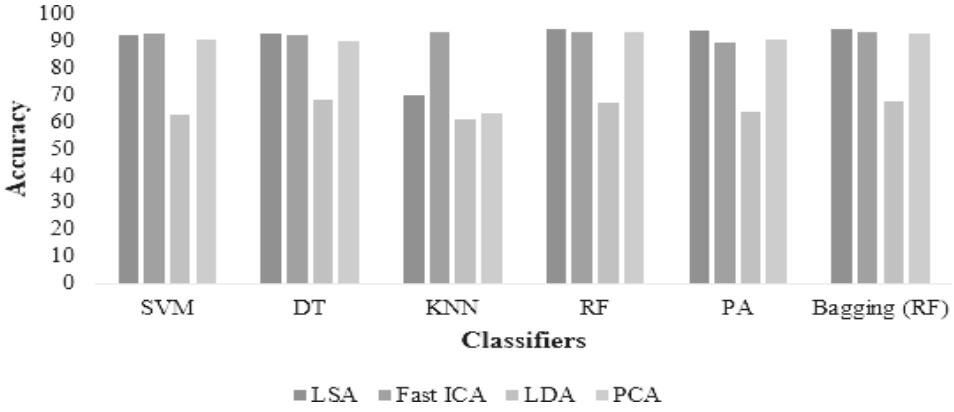

Figure 3: Performance of the dimension reduction techniques for ML classifiers in terms of accuracy

time. LSA always finds the important patterns of the dataset so that it gets the best prediction. LDA focuses on topic modeling whereas LSA emphasizes reducing the matrix size of a dataset while solving problems. **Table 4** and Figure 3 indicate that LDA as a reduction technique performs worse.

Table 2 shows that the accuracy of Bagging with RF is 94.62% whereas **Table 6** shows that, without the dimension reduction technique this drops down to 94.45%. **Table 6** shows that the accuracy of the PA classifier is 91.95%. **Table 2** shows that after adopting LSA, it becomes 93.88% and Table V shows that after adopting PCA it becomes 90.42%.

Table-6: Performance of the classifiers in terms of accuracy, precision, recall, and F1-score without any dimension reduction technique.

Classifier	Accuracy (%)	Precision	Recall	F1-Score
SVM	90.96	0.922	0.910	0.909
DT	94.01	0.942	0.942	0.940
KNN	58.51	0.756	0.585	0.502
RF	94.28	0.946	0.943	0.943
PA	91.95	0.926	0.919	0.919
Bagging (RF)	94.45	0.948	0.945	0.945

In SVM the best case of dimension reduction technique is Fast ICA. With Fast ICA, SVM also leads to greater accuracy which is 92.75%. Previously without any dimension reduction technique, it was 90.96%. According to the accuracy level, the performance of SVM, KNN, RF, and PA has increased after adopting LSA, and Fast ICA. The accuracy of KNN without any dimension reduction technique is 58.51%. The accuracy of KNN with LSA, Fast ICA, PCA, and LDA is 70.04%, 93.53%, 63.44%, and 61.05% respectively. LDA decreased all the classifier's performances except KNN. As KNN works to find the point in the nearest neighbor to make a prediction, Fast ICA also helps to find and separates the independent components of the dataset, so it greatly boosts the performance of KNN.

From the result analysis, the DT performs worst with the dimension reduction technique. Without the dimension reduction technique, it achieves an accuracy of 94.01% whereas with LSA, PCA and LDA, and Fast ICA. it achieves 92.73%, 89.92%, 68.08%, and 92.09%. **Table 6** shows that DT achieves the highest precision, recall, and F1-score level without any dimension reduction technique. As the dataset contains multi-languages it doesn't work great with LDA. LDA works worst when a single topic doesn't discuss coherently.

RQ3: Is LSA the best among these four-dimension reduction techniques?

Figure 3 demonstrates the comparison of classifiers with dimension reduction techniques in terms of accuracy. Among all the dimension-reduction techniques LSA does a tremendous job to attain accuracy with the maximum classifiers. Figure 4 shows the performance of classifiers in terms of accuracy where LSA is used as a dimension-reduction technique.

In this work, applying LSA makes a higher accuracy level possible with classifiers by reducing the dimension of the dataset. It understands the logic behind the text to classify tweets. It works fast as it uses less time and space to analyze users. **Figure 3** and **Table 2** shows that among all the classifiers Bagging performs best along with LSA which is 94.62%. With PCA, LDA and Fast ICA Bagging classifiers achieve accuracy of 92.9%, 67.5%, and 93.58% while the base classifier is RF. RF classifier achieved its highest accuracy of 94.39% with LSA. As RF and Bagging classifiers both have the bootstrapping technique, it helps LSA to interpret the text more correctly. Due to this, it achieves higher accuracy. PA also achieves its best accuracy, 93.88% with LSA. Though SVM, KNN, and DT don't achieve the highest accuracy with LSA, they gave a good performance with LSA. KNN, SVM, and DT achieve their second-highest accuracy of 70.04%, 92.11%, and 92.73% with LSA.

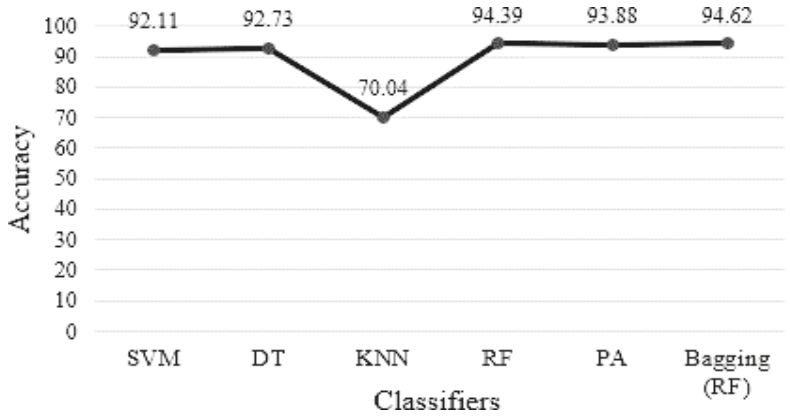

Figure 4: Performance of the classifiers in terms of accuracy with LSA

Comparing **Table 2, Table 3, Table 4, Table 5, and Table 6** shows that SVM and KNN do better with the Fast ICA algorithm. The Decision Tree does better without any dimension reduction technique. LSA also achieves the highest precision level of 0.95 with the Bagging classifier while the base classifier is RF.

Table-7: Performance of our and other authors' comparative proposed model on social media dataset.

Author Name	Classifiers	Accuracy (%)
Mustafa et al. (2020)	SVM, RF	91, 83
Kumar et al. (2019)	RF, Voting	81.04, 85.09
Gaind et al. (2019)	SVM, J48	91.7, 85.4
Hassan et al. (2017)	SVM	91
Proposed Model	SVM, RF	92.11, 94.39

The dataset contains multi-languages which makes it difficult for other dimension-reduction techniques to interpret the tweets. As LSA is language-independent, the dataset does not need to maintain the grammatical structure and it can interpret multi-language datasets. LSA removes the multi-collinearity of the classifiers. It also helps to represent the contextual meaning of the text. Due to all of these, it achieves higher accuracy with most of the classifiers. LSA helps to extract the raw information from the text which helps the algorithm to make correct predictions. After the study, it is proved that the dimension reduction technique (LSA) enhances the performance of the classifiers. Existing research works have used different techniques and ML classifiers. Most of the authors have used SVM and RF classifiers to develop their models. In **Table 7**, a comparison among some

existing research work has been given. It is noticeable that most of the researchers ignored dimension reduction techniques. In this work, SVM and RF with LSA perform 92.11% and 94.39% respectively in terms of accuracy. It is clear that LSA enhances the performance of the classifiers.

5. Conclusion

Most people over the world would not consult doctors at an early stage of depression because of negligence, and embarrassment. Besides, people share their feelings & emotions on social media platforms which is very helpful for detecting their mental health. In this paper, an ML-based depression detection technique from text is proposed. Four-dimension reduction techniques namely LSA, PCA, LDA, and Fast ICA are used to evaluate the performance of the algorithm. It is clear from the result that LSA and ICA help to increase the performance of the classifiers. With LSA, the Bagging classifiers perform the best in terms of accuracy while RF is used as the base classifier. The accuracy of the Bagging classifier with RF is 94.62%. LDA performs the worst among them. From the result analysis, it can be said that ML classifiers perform well to detect depression from the text. In the future, users' full profiles will be considered to detect depression. As we worked with only text-based analysis in this paper, in the future, user profiles, posted pictures, images, age, profession, and tweets will be analysed further to detect depression analysis more precisely. The deep learning model will be applied in the following version.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), (pp. 433–459).

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 international conference on computing networking and informatics (iccni) (pp. 1–9).

Balabantaray, R. C., Mohammad, M., & Sharma, N. (2012). Multi-class twitter emotion classification: A new approach. International Journal of Applied Information Systems, 4(1), (pp. 48–53).

Ban, X., Liu, S., Chen, C., & Chua, C. (2019). A performance evaluation of deep-learnt features for software vulnerability detection. Concurrency and Computation: Practice and Experience, 31(19).

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197–227.

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 45.

Buhlmann, P. (2012). Bagging, boosting and ensemble ¨ methods. In Handbook of computational statistics, Springer (pp. 985–1022).

Burdisso, S. G., Errecalde, M., & Montes-y Gomez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications, 133, (pp. 182– 197).

Chiong, R., Budhi, G. S., & Dhakal, S. (2021). Combining sentiment lexicons and content-based features for depression detection. IEEE Intelligent Systems, 36(6), (pp. 99–105).

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006) Online passive aggressive algorithms.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In Proceedings of the sigchi conference on human factors in computing systems (pp. 281–285).

Evgeniou, T., & Pontil, M. (1999). Support vector machines: Theory and applications. In Advanced course on artificial intelligence (pp. 249–257).

Flach, P. A. (2003). The geometry of roc space: understanding machine learning metrics through roc isometrics. In Proceedings of the 20th international conference on machine learning (icml-03) (pp. 194–201).

Gaind, B., Syal, V., & Padgalwar, S. (2019). Emotion detection and analysis on social media. arXiv preprint arXiv:1901.08458.

Garc´ıa, S., Fernandez, A., Luengo, J., & Herrera, F. (2009). ´ A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Computing, 13(10).

Gupta, D., Bhatia, M., & Kumar, A. (2021). Real-time mental health analytics using iomt and social media datasets: Research and challenges. In Proceedings of the international conference on innovative computing & communication (icicc).

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review, 53(2), (pp. 217–288).

Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017). Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression. In 2017 international conference on information and communication technology convergence (ictc) (pp. 138–140).

Hemdan, E. E.-D., Shouman, M. A., & Karar, M. E. (2020). Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:2003.11055.

Hur, N. W., Kim, H. C., Waite, L., & Youm, Y. (2018). Is the relationship between depression and c reactive protein level moderated by social support in elderly? -korean social life, health, and aging project (kshap). Psychiatry investigation, 15(1), 24.

Islam, M., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., Ulhaq, A., et al. (2018). Depression detection from social network data using machine learning techniques. Health information science and systems, 6(1), (pp. 1–12).

Koltai, J., Kmetty, Z., & Bozsonyi, K. (2021). From durkheim to machine learning: Finding the relevant sociological content in depression and suicide-related social media discourses. In Pathways between social science and computational social science, Springer (pp. 237–258).

Kumar, A., Sharma, A., & Arora, A. (2019). Anxious depression prediction in real-time social data. arXiv preprint arXiv:1903.10222.

Kurtanovic, Z., & Maalej, W. (2017). Automatically classifying functional and non-functional requirements using supervised machine learning. In 2017 ieee 25th international requirements engineering conference (re) (pp. 490–495).

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259–284.

Ma, L., Wang, Z., & Zhang, Y. (2017). Extracting depression symptoms from social networks and web blogs via text mining. In International symposium on bioinformatics research and applications (pp. 325–330).

Murray, C. J., & Lopez, A. D. (1997). Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. The lancet, 349(9064), 1498– 1504.

Mustafa, R. U., Ashraf, N., Ahmed, F. S., Ferzund, J., Shahzad, B., & Gelbukh, A. (2020). A multiclass depression detection in social media based on sentiment analysis. In 17th international conference on information technology–new generations (itng 2020) (pp. 659– 662).

Novakovic, J., Veljovic, A., Ilic, S., & Papic, M. (2016). Experimental study of using the k-nearest neighbour classifier with filter methods. In Proceedings of the conference on computer science and technology (pp. 91– 99).

Pecorelli, F., Palomba, F., Di Nucci, D., & De Lucia, A. (2019). Comparing heuristic and machine learning approaches for metric-based code smell detection. In 2019 ieee/acm 27th international conference on program comprehension (icpc) (pp. 93–104).

Qi, Y., Doermann, D., & DeMenthon, D. (2001). Hybrid independent component analysis and support vector machine learning scheme for face detection.

In 2001 ieee international conference on acoustics, speech, and signal processing. proceedings (cat. no. 01ch37221) (Vol. 3, pp. 1481–1484).

Safiri, S., Pourfathi, H., Eagan, A., Mansournia, M. A., Khodayari, M. T., Sullman, M. J., . . . others (2022). Global, regional, and national burden of migraine in 204 countries and territories, 1990 to 2019. Pain, 163(2).

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4), 2094– 2097.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., . . . Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. In Ijcai (pp. 3838–3844).

Smys, S., & Raj, J. S. (2021). Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. Journal of trends in Computer Science and Smart technology (TCSST), 3(01), 24–39.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. IEEE Access, 7, 44883–44893.

Tran, B. X., Latkin, C. A., Sharafeldin, N., Nguyen, K., Vu, G. T., Tam, W. W., . . . Ho, R. C. (2019). Characterizing artificial intelligence applications in cancer research: a latent dirichlet allocation analysis. JMIR Medical Informatics, 7(4).

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013). A depression detection model based on sentiment analysis in micro-blog social network. In Pacific-asia conference on knowledge discovery and data mining (pp. 201–213).

Zalpour, M., Akbarizadeh, G., & Alaei-Sheini, N. (2020). A new approach for oil tank detection using deep learning features with control false alarm rate in highresolution satellite imagery. International Journal of Remote Sensing, 41(6), 2239–22˙