Multimedia Tools and Applications (2024) 83:38819–38837
https://doi.org/10.1007/s11042-023-16221-z

Hybrid machine learning models to detect signs of depression

Shakir Khan[1] · Salihah Alqahtani[1]

Received: 6 October 2022 / Revised: 24 April 2023 / Accepted: 4 July 2023 /
Published online: 6 October 2023

Abstract
Depression is a prevalent mental illness that can only be diagnosed through self-reporting. Unfortunately, 70% of individuals do not seek medical help in the early stages of depression. With the increasing use of social media to share daily activities and emotions, it has become a valuable tool for identifying mental health issues. To this end, this paper proposes multiple hybrid machine-learning models for sentiment analysis to detect signs of depression by analyzing Twitter tweets. The study utilizes unsupervised approaches for feature extraction and supervised approaches as classifiers. The proposed models are evaluated on public sentiment tweets datasets using various performance metrics. The four modules proposed in this study can be used to detect signs of depression in text data. Module 1 uses BERT for feature extraction followed by an artificial neural network as a classification module. Module 2 involves data pre-processing followed by the TF-IDF feature extraction method and logistic regression as a classification algorithm. Module 3 also includes data pre-processing followed by the TF-IDF feature extraction method but uses a linear support vector machine as the classification algorithm. Finally, Module 4 uses the Spacy package with a small library as the feature extraction method and a linear support vector machine as the classification algorithm. Model 2 had the highest accuracy of 0.994, followed closely by model 3 with an accuracy of 0.992, while model 4 had a significantly lower accuracy of 0.868. Model 1 achieved an accuracy of 0.99.

Keywords Sentiment analysis · Depression detection · Social network · Twitter data analysis

1 Introduction

Depression is a serious mental health condition that affects millions of people worldwide. Early detection and intervention are critical for effective treatment, but unfortunately, depression often goes undiagnosed due to the stigma associated with mental health and

✉ Shakir Khan
sgkhan@imamu.edu.sa

[1] College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

the difficulty in recognizing symptoms [12]. Therefore, the development of reliable and accurate methods to detect signs of depression is of utmost importance. Machine learning (ML) techniques have been successfully applied to the diagnosis and detection of various diseases, including mental health conditions. However, due to the complexity and multifaceted nature of depression, single ML models may not be sufficient to accurately detect signs of depression. Hybrid machine learning models, which combine multiple ML algorithms or techniques, have the potential to improve the accuracy and reliability of depression detection [9].

The objective of this study is to develop and evaluate the performance of hybrid machine-learning models for detecting signs of depression in social media data. Social media platforms, such as Twitter and Facebook, are valuable sources of data for mental health research. In this study, we will collect a large dataset of social media posts from individuals with and without depression, and use it to train and test a variety of hybrid ML models. The research problem that this study aims to address is the need for more accurate and reliable methods for detecting signs of depression [13]. By developing and evaluating hybrid ML models for depression detection, this study will contribute to the growing body of research on using machine learning for mental health diagnosis and treatment. The findings of this study may have implications for the development of new tools and approaches for early detection and intervention for individuals with depression, which could ultimately improve outcomes and reduce the burden of this debilitating condition [18].

1.1 Problem statement

The problem statement for detecting signs of depression is to identify individuals who may be suffering from depression but have not yet been diagnosed or treated. Depression is a common mental health condition that can have serious consequences if left untreated, including suicide. However, many people with depression go undiagnosed and untreated due to a lack of awareness, stigma, or difficulty in recognizing the symptoms. Therefore, the goal of detecting signs of depression is to identify these individuals and connect them with appropriate mental health resources and support. This can be achieved through various means such as screening tools, diagnostic interviews, or the use of digital technologies such as mobile applications or chatbots.

1.2 This study's major contribution is exemplified below

- To detect signs of depression in Twitter tweets, one of the proposed hybrid machine learning models is Model 1, which utilizes BERT as the features extraction module (unsupervised method) followed by an artificial neural network (ANN) as the classification module.
- To detect signs of depression in Twitter tweets, another proposed hybrid machine learning model is Module 2, which involves a data pre-processing module followed by the use of Frequency Inverse Document Frequency (TF-IDF) as the feature's extraction module (unsupervised method) and Logistic Regression (LG) algorithm as the classification module.
- To detect signs of depression in Twitter tweets, Model 3 is proposed as another hybrid machine learning model. This model involves a data pre-processing module followed by the use of Frequency Inverse Document Frequency (TF-IDF) as the feature's extrac-

tion module (unsupervised method) and linear support vector machine (SVM) as the classification module.

- To detect signs of depression in Twitter tweets, Model 4 is proposed as another hybrid machine learning model. This model involves using the Spacy package with a small library as the feature's extraction module (unsupervised method), followed by a linear support vector machine (SVM) as the classification module.

2 Backgrounds

Predicting a person's psychological state based on his writing in a blog, for example, or his tweet on Twitter depends on Natural language processing (NLP) research includes sentiment analysis as one of its areas. NLP is a secondary branch of machine learning science, which is concerned with the interaction between computer and natural language text, NLP has many applications [17, 21] such as translation between different natural languages, intelligent chatbots, articles summarization, articles (newspaper, blogs, and so on) categories classification e.g., sport, economic, political, question Answering and sentiment analysis like classify the opinion of customers about a product, or recognize if a person has depression or not from analyzing his tweets. Sentiment Analysis is a process of analyzing natural language automatically, discovering essential opinions, feeling, and sentiments about any topic, and classifying their attitudes whether positive, negative, or natural [20]. In other words, it is the act of analyzing the perceptions views, and mental states of people based on what they post on social media sites. The authors in [18] pointed to those sentiments posted by users on social media platforms strongly express their deeper emotions and feeling. An analytical overview suggests that sentiments that are posted with a negative expression portray negative emotions. Therefore, it is possible to identify states such as happiness, sadness, or even depression of users based on the sentiments they post on Twitter. The advancements in technology especially artificial intelligence and machine learning (ML) provide a significant opportunity for researchers to extract crucial information through Twitter.

The use of artificial intelligence algorithms in the process of analyzing feelings from a text written in one of the natural languages is not an easy matter, and it requires a lot of data processing, extracting data features, and choosing the appropriate classifier algorithm. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings. The maximum limit of characters that are allowed on Twitter is 140. Predicting that someone is depressed by analyzing the feelings of their tweets is a complex process and is prone to errors due to context such as declarative or interrogative, using terms that can lead to misunderstanding and other complexities of natural languages. Many issues come with sentiment analysis and natural language processing, such as informal style of writing, sarcasm, and some language challenges. Sarcasm and irony in the tweets are one of the most faced challenges that have the attention of researchers [20].

In this paper, many issues need to be solved, selecting and implementing appropriate data pre-processing methods, choosing unsupervised text features extraction mechanisms, and selecting appropriate classification algorithms. Supervised learning models in the machine learning approach are more commonly used due to their good results [6, 11, 20] Therefore, we study using a hybrid machine learning model. However, this paper aims to determine if Twitter users have depression based on sentiment analysis of user tweets using a hybrid machine learning model.

The paper aims to detect signs of depression in tweets using natural language processing, Hybrid models will be used, an unsupervised approach for features extraction and a supervised approach as a classifier. Where multiple hybrid models will be proposed and evaluated on the public sentiment tweets dataset, and many performance metrics will be used to evaluate the proposed models. In addition, discussing if data pre-processing methods and data features extraction models are efficient to determine if Twitter users have depression based on sentiment analysis of user tweets.

2.1 Related works

Sentiment analysis of social media has been greatly explored by several researchers with the use of diverse models. The computation of different models for sentiment analysis yields different results that create space for further investigation. Sentiment analysis takes diverse forms, which allow researchers to investigate the establishment of different computational models and approaches to depict emotions and moods based on user posts.

In 2020, Chiu and Lane et al. [3] proposed the detection of depression has gained significant attention in recent years, with research focusing on sentiment analysis of social media platforms such as Instagram. A proposed multimodal system uses a combination of image, text, and behavior features to classify posts as depressed or non-depression. Supervised models are utilized to carry out this classification task.

In 2017, Hassan and Hussain et al. [14] developed a computing model that uses a supervised learning algorithm to identify signs of depression in users with post-traumatic stress disorders. However, their study has limitations as it heavily relies on the selection of good training examples, which may lead to bias in the type of data used for analysis.

In 2019, Arora et al. [7] used a multi-classifier approach using a voting technique was applied to classify data into different categories. Three classifiers, SVM, NB, and ME, were used as inner learners. SVM classifier outperformed NB and ME classifiers with an accuracy of up to 91%. Each feature was assigned several votes to determine the label with the highest votes.

In 2019, Uddin and Bapery et al. [1] used a novel sentiment analysis approach proposed by combining Multinomial Naive Bayes and Support Vector Regression classifiers, and their performance is compared. This is the first study to evaluate these two classifiers in combination for sentiment analysis tasks.

In 2020, Kamite and Kamble et al. [19] proposed the Long Short-Term Memory (LSTM) deep learning technique to analyze depression in the Bangla language. They achieved high accuracy by tuning the model with an LSTM size of 128 and a batch size of 25, making it effective for small datasets. This research demonstrates the potential of LSTM for analyzing mental health in non-English languages.

In 2016, Rosa and Rodriguez et al. [10] used supervised machine learning models, such as Naive Bayes and Random Forest, to classify syntactical markers related to depression symptoms. They emphasized the significance of feature extraction to develop an effective model. By utilizing these models, they were able to analyze and classify data accurately.

In 2018, Islam and Kamal et al. [15] introduced a depression monitoring system that was developed to notify medical staff and relatives about the emotional behavior of users with psychological disorders. The system considers factors such as age, gender, and medical information to obtain the sentiment intensity of a phrase containing the user's mood.

The sentiment intensity value is then corrected, and a classifying method is applied to determine the main characteristics of depression. Various methods were also defined to send messages to authorized subjects.

In 2018, Islam and Kamal et al. [2] used it on classifying tweets into depressive and non-depressive users, the authors employed SVM and random forest classifiers but achieved accuracy below 80%. They noted the importance of incorporating additional features and fine-tuning the model to improve its performance. It was observed that using only one ML model at a time yielded suboptimal results.

In 2021, Saha and Marouf et al. [8] introduced that in detecting depression using emotions in Facebook posts, six KNN classifiers were used: Fine KNN, Medium KNN, Coarse KNN, Cosine KNN, Cubic KNN, and Weighted KNN. The dataset was divided into two groups - positive and negative emotions. The KNN models utilized emotional, linguistic, and temporal features. Among all the KNN classifiers, Coarse KNN showed the best result in detecting depression.

In 2018, Devlin and Chang et al. [16] proposed that they collected depression data from various social media platforms and verified it with psychologists. They extracted linguistic features and applied ten supervised machine learning models including Naïve Bayes, Random Forest, and Multilayer Perceptron. However, Random Forest showed the highest accuracy of 60.54%, which is lower compared to other studies. Show the given below Fig. 1 Smart art Chart format of the Literature Review.

3 Proposed methodologies

Several machine learning techniques have been explored for sentiment analysis; most of these algorithms are classified as supervised or unsupervised, with supervised learning algorithms being employed in the majority of approaches. The suggested technique involves employing a hybrid model to create numerous combinations of supervised and

Year	No of the Papers Reviewed
2016	1
2017	1
2018	2
2019	1
2020	2
2021	2

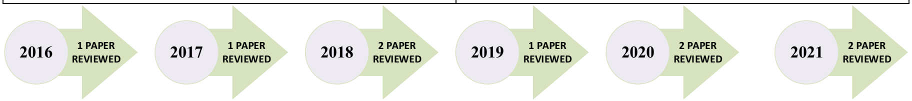

Fig. 1 Smart art Chart format of the Literature Review

unsupervised machine-learning algorithms for sentiment analysis to identify indicators of depression.

3.1 Proposed hybrid models

By applying each of them to a selected sentiment tweets dataset, many hybrids were constructed to evaluate the findings and recognize the best-proposed models. We developed our models as a combination of three modules Compact graphical format shows in below Fig. 2, which are:

1. Data pre-processing module (optional)
2. Features extraction module
3. Classification module

The proposed modules are:

1. Model 1 -BERT & ANN: BERT as features extraction module (unsupervised method) follows by an artificial neural network as a classification module.
2. Module 2 - Pre-processing & TF-IDF & LG: there is a data pre-processing module followed by Frequency Inverse Document Frequency as a feature's extraction module (unsupervised method) follows by Logistic Regression algorithm as a classification module.
3. Model 3 -Pre-processing & TF-IDF & SVM: there is a data pre-processing module followed by Frequency Inverse Document Frequency as a feature's extraction module (unsupervised method) follows by a linear support vector machine as a classification module.

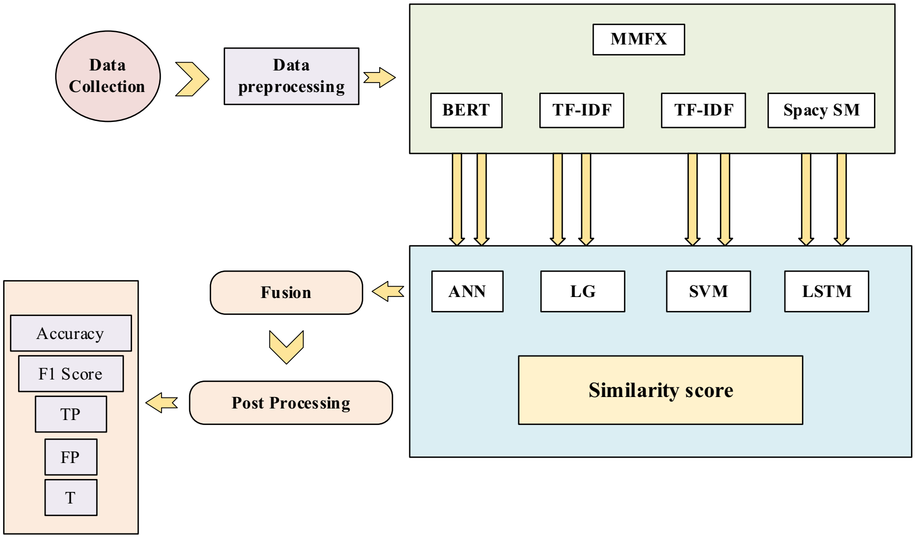

Fig. 2 Compact graphical format

4. Model 4 – SpacySM & SVM: Spacy package with the small library as features extraction module (unsupervised method) follows by linear support vector machine as classification module.

3.1.1 BERT & ANN

The first hybrid model uses Pre-training of Deep Bidirectional Transformers or Bidirectional Encoder Representations from Transformers (BERT) from Google [4] unsupervised machine learning algorithm as a features extraction module, and it uses Artificial Neural Network (ANN) as a classification module.

The proposed ANN consists of an input layer, one hidden layer, output layer. The hidden layer is spending Rectified Liner Unit (RELU/relu) stimulation utility, which is very mutual and well-known in neural networks.

3.1.2 Pre-processing & TF-IDF & LG

The second hybrid model has a pre-processing module, and the Frequency Inverse Document Frequency vectorize (TF-IDF) unsupervised algorithm is used for features extraction, TF-IDF is an unsupervised machine learning algorithm, and it is followed by Logistic Recognition (LG) classification algorithm.

3.1.3 Pre-processing & TF-IDF & SVM

The third hybrid model has a pre-processing module, and the Frequency Inverse Document Frequency vectorize (TF-IDF) unsupervised algorithm is used for features extraction, TF-IDF is an unsupervised machine learning algorithm, and it is followed by Linear Support Vector Machine (LSVM).

3.1.4 Spacy SM & SVM

The fourth proposed hybrid model uses Spacy which is advanced an open-source software python library used in advanced natural language processing and machine learning for features extraction, this proposed model uses a small web package for the English natural language and follows by linear support vector machine algorithm.

3.2 Dataset pre-processing module

In this paper, a dataset called sentiment tweets [5] is used for evaluating the proposed models. It includes the following fields: User ID, message (Tweets), and label (0 means there is no depression, and 1 means there is depression). The CSV (Comma Separated Value) file contains 10,414 data of which 8251 data records are used for training and 2063 records are used for testing.

The tweets records of depression occupy about 22.43% of total records, while tweets without depression are about 77.56% of total records.

Pre-process text functions include converting capital letters to small letters except for the name of countries or cities like London, removing stop words, removing gerund ‘ing’ from a verb, converting any verb conjugation to its original present form, changing from plural to singular and numbers and digits.

3.3 Experiment methodology

Every proposed approach is written in a discrete Python file, there are four approached hybrid models and there is one dataset. The performance metrics for proposed hybrid models are:

- Accuracy for testing
- F1-score for testing
- True positive (TP)
- False positive (FP)
- True negative (TN)
- False negative (FN)

The flowchart of the experiment approach is shown in Fig. 3.

1. Go over the data.
2. Pre-processing of data sets
3. Separate the datasets (20 for test, 80% for training)
4. Create a model for the proposed strategy.
5. Educate the model
6. Put the model to the test
7. Show performance metrics for the model

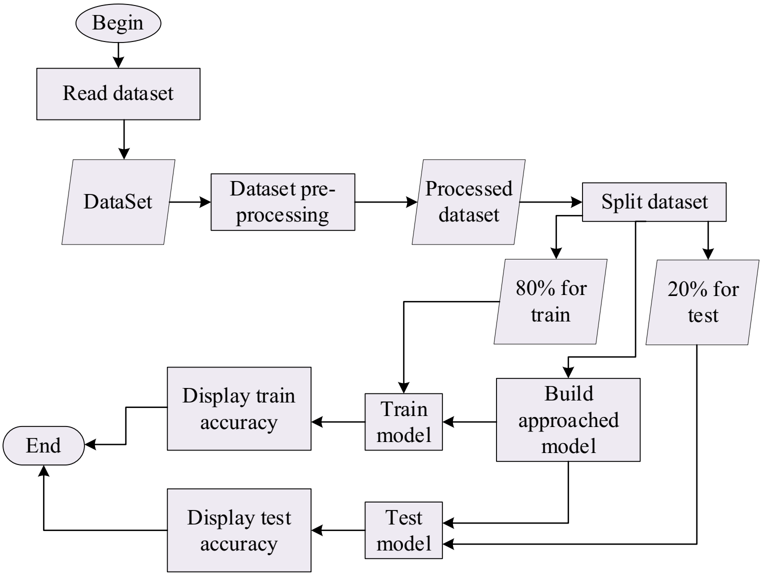

Fig. 3 Experiment methodology flowchart

3.4 Experiment parameters

The optimal parameters utilized with various architectures for various experiments are shown in this section.

3.4.1 BERT & ANN

This proposed hybrid model has no data pre-processing module, the BERT algorithm is applied to the dataset to extract features, first, we need to download a pre-trained model called 'distillery-base-nil-mean-tokens'. It is a sentence-transformers model: It maps sentences & paragraphs to a 768-dimensional dense vector space and can be used for tasks like clustering or semantic search.

After the feature's extraction process, the ANN is applied to generated features to train the model, the proposed ANN has the following structure:

1. Input layer with data dimension d = 768
2. 1 hidden layer with 256 neurons and the activation function is Relu
3. Output layer: one neuron and the activation function is sigmoid

The classification algorithm is trained for 20 iterations, the optimization algorithm is adam, the loss function is binary_crossentropy and the metrics are accuracy, also batch_size = 16, the ANN model was trained on 80% of data and validated over the remainder 20%.

3.4.2 Pre-processing & TF-IDF & LG

The proposed model contains a data pre-processing module, follows by unsupervised features extraction TF-IDF algorithm, which extracts 20,056 features, after that the model applies logistic regression with the solver method being liblinear and penalty l1, the proposed classification algorithm was trained over 80% of data and was validated by remainder 20% of data.

3.4.3 Pre-processing & TF-IDF & SVM

The proposed hybrid model contains a data pre-processing module, follows by unsupervised features extraction TF-IDF algorithm, which extracts 20,056 features, after that the model applies a Multinomial Linear support vector machine algorithm, the proposed classification algorithm was trained over 80% of data and was validated by remainder 20% of data.

3.4.4 SpacySM & SVM

The proposed hybrid model has no data pre-processing module, it uses a spacy small core English language for web to extract features En_core_web_sm is an unsupervised algorithm and extracts 96 features, after that the model applies linear support vector machine, the proposed classification algorithm was trained over 80% of data and was validated by remainder 20% of data.

3.5 Benefits of the research for the society at large concerning evaluating its key determinant

The proposed hybrid machine learning models for sentiment analysis to detect signs of depression by analyzing Twitter tweets have significant benefits for society at large. One of the key determinants of the success of these models is the early detection of depression, which is crucial for effective treatment and management. Depression is a prevalent mental health issue that affects millions of people worldwide, and early detection is critical for preventing long-term negative outcomes such as suicidal ideation and chronic depression. With the increasing use of social media platforms like Twitter, these models can provide a non-invasive and accessible way to detect signs of depression in real-time, potentially leading to earlier interventions and better mental health outcomes. By using unsupervised approaches for feature extraction and supervised approaches as classifiers, the proposed models can effectively analyze large amounts of Twitter data to identify individuals who may be experiencing symptoms of depression. This can be particularly beneficial for individuals who may not be aware that they are experiencing depression or may be hesitant to seek medical help due to the stigma associated with mental health issues. Additionally, the proposed models can be used to identify trends and patterns in depression-related tweets, providing valuable insights into the prevalence and distribution of depression symptoms across different demographics and geographical locations. This can be used to inform public health policy and interventions to improve mental health outcomes. Overall, the proposed hybrid machine learning models for sentiment analysis have the potential to significantly improve the early detection and treatment of depression, ultimately leading to better mental health outcomes for individuals and society as a whole.

4 Results

Depression is a significant public health issue that affects millions of people worldwide. Despite its prevalence, diagnosis, and treatment can be challenging due to the stigma surrounding mental illness and the lack of awareness and education about depression. Additionally, many individuals who suffer from depression may not seek medical help or report their symptoms until they have reached an advanced stage.

To address this problem, researchers have turned to social media platforms such as Twitter to identify individuals who may be experiencing symptoms of depression. Twitter has become a valuable tool for analyzing emotions and sentiment in large-scale datasets, making it an ideal source of data for detecting signs of depression. The main problem with using social media data for depression detection is the lack of an accurate and efficient way to analyze and interpret the large volume of unstructured data generated on these platforms. Traditional manual analysis methods are time-consuming and costly, making them impractical for large-scale datasets. Machine learning algorithms have been widely used to analyze social media data due to their ability to automatically learn patterns and relationships from large datasets.

In this research, the authors propose a set of hybrid machine-learning models for sentiment analysis to detect signs of depression by analyzing Twitter tweets. The proposed models aim to improve the accuracy of depression detection while also reducing the time and cost associated with manual analysis. The authors propose four modules for depression detection, each with a different feature extraction and classification algorithm. Module 1 uses BERT for feature extraction and an artificial neural network as the classification algorithm. Module 2 involves data pre-processing followed by the TF-IDF feature extraction method and logistic regression as the classification algorithm. Module 3 also includes data pre-processing followed by the TF-IDF feature extraction method but uses a linear support vector machine as the classification algorithm. Finally, Module 4 uses the Spacy package with a small library as the feature extraction method and a linear support vector machine as the classification algorithm.

The proposed models were evaluated on public sentiment tweets datasets using various performance metrics such as accuracy, precision, recall, and F1-score. Model 2 had the highest accuracy of 0.994, followed closely by model 3 with an accuracy of 0.992, while model 4 had a significantly lower accuracy of 0.868. Model 1 achieved an accuracy of 0.99.

The contribution of this research is three-fold. First, the authors proposed multiple hybrid machine-learning models for depression detection that utilize both unsupervised and supervised approaches for feature extraction and classification. This approach allows for a more accurate and efficient analysis of social media data for depression detection. Second, the proposed models were evaluated on publicly available sentiment tweets datasets, making the results of this research easily replicable and applicable to real-world scenarios. The performance metrics used in the evaluation provide a comprehensive measure of the accuracy and effectiveness of the proposed models.

Finally, the authors' proposed models can be used to detect signs of depression in text data, which can be integrated into existing monitoring systems to provide early detection of depression symptoms in individuals. Early detection of depression can lead to earlier intervention and treatment, which can improve the overall prognosis for individuals suffering from depression.

Overall, this research highlights the potential of machine learning algorithms for the detection of depression using social media data. The proposed models provide a more accurate and efficient method of depression detection, which can be integrated into existing monitoring systems to provide early detection of depression symptoms in individuals. Future research can explore the effectiveness of these models in different languages and cultures, as well as in other mental health conditions such as anxiety and post-traumatic stress disorder.

4.1 Dataset used

Public sentiment tweets dataset has been used in this model https://www.kaggle.com/gargmanas/sentimental-analysis-for-tweets.

4.2 Evaluation metrics

The suggested technique is compared with the other technique using several performance metrics such as sensitivity, specificity, and accuracy.

Accuracy: it is equal to the number of correct predictions divided by the total predictions.

F1-Score: is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

True Positive (TP): the total number of when the classifier classifies a tweet as a depression and it is correctly a depression.

True Negative (TN): the total number of when the classifier classifies a tweet as not a depression and it is correctly not a depression.

False Positive (FP): the total number of when the classifier classifies a tweet as a depression, but it is not a depression.

False Negative (FN): the total number of when the classifier classifies a tweet as not a depression, but it is a depression.

Precision: It is implied as the measure of the correctly identified positive cases from all the predicted positive cases. Thus, it is useful when the costs of False Positives are high (this will be used to calculate F1-score).

Recall: It is the measure of the correctly identified positive cases from all the actual positive cases. It is important when the cost of False Negatives is high (this will be used to calculate F1-score).

- Accuracy

The probability of the right classification is known as accuracy (ACC), and it is described as:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \times 100\%$$

- **True-negative (TN)** - The percentage of clean signals that are accurately categorized as clean signals
- **True-positive (TP)** - Stego signals that are correctly identified as such out of all those that are sent.
- **False-negative (FN)** - The percentage of stego signals that are mislabelled as clean signals.
- False Positive (FP): the total number of when the classifier classifies a tweet as a depression, but it is not a depression.

F1-Score: is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Precision: It is implied as the measure of the correctly identified positive cases from all the predicted positive cases. Thus, it is useful when the costs of False Positives are high (this will be used to calculate F1-score).

$$Precision = \frac{TP}{TP + FP}$$

Recall: It is the measure of the correctly identified positive cases from all the actual positive cases. It is important when the cost of False Negatives is high (this will be used to calculate F1-score).

$$Recall\ (Sensitivity)\ =\ \frac{TP}{TP+FN}$$

4.3 Twitter sentiment dataset results

Table 1 illustrates the performance metrics results for applying the proposed hybrid models on the selected Twitter sentiment dataset, where these models have the following notations:

- BERT & ANN denoted as model 1
- Pre-processing & TF-IDF & LG denoted as model 2
- Pre-processing & TF-IDF & SVM denoted as model 3
- Spacy SM & SVM denoted as model 4

The results of the proposed multiple hybrid machine learning models for sentiment analysis in Table 1 are to detect signs of depression by analyzing Twitter tweets and were evaluated using various performance metrics. The four models were evaluated based on accuracy, F1-score, true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Model 2 had the highest accuracy of 0.994, followed closely by model 3 with an accuracy of 0.992, while model 4 had a significantly lower accuracy of 0.868. Model 1 achieved an accuracy of 0.99. The F1-score also showed a similar trend, with model 2 having the highest F1-score of 0.991, followed by model 3 with an F1-score of 0.988. Model 4 had the lowest F1-score of 0.8, indicating that it has difficulty in correctly identifying true positives and true negatives. The number of TP, FP, TN, and FN was also calculated to evaluate the performance of the models. Model 2 had the highest number of TP, with 1576 correctly identified as depressive, followed by model 3 with 1596. Model 4 had the lowest number of TP with 1496. In terms of FP, model 4 had the highest number with 77 incorrectly identified as depressive, while model 2 had 0 FP. For TN, model 1 had the highest number of correctly identified non-depressive tweets with 1606, followed by model 2 with 475. Finally, model 4 had the highest number of FN with 195, indicating that it had difficulty in correctly identifying depressive tweets. In conclusion, the proposed multiple hybrid machine learning models for sentiment analysis to detect signs of depression by analyzing Twitter tweets achieved high accuracy and F1-score. Model 2, which used data pre-processing followed by the TF-IDF feature extraction method and logistic regression as a classification algorithm, had the highest accuracy and F1 score. However, it is important to note that each model has its strengths and weaknesses, and the selection of the most appropriate model will depend on the specific requirements and goals of the application. This study provides a valuable contribution to the field of mental health by demonstrating the potential of machine learning techniques to identify signs of depression from text data.

Table 1 Twitter sentiment dataset performance metrics results

Metrics	Model 1	Model 2	Model 3	Model 4
Accuracy	0.99	0.994	0.992	0.868
F1-score	0.986	0.991	0.988	0.800
TP	438	1576	1596	1496
FP	8	0	0	77
TN	1606	475	451	295
FN	11	12	16	195

The Table 1 metrics represent the performance of four different models on a Twitter sentiment dataset performance metrics result, which could be classification or prediction of some kind. Each model is evaluated based on its accuracy, F1-score, true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Accuracy is the percentage of correctly classified instances out of the total number of instances. Model 2 has the highest accuracy with 0.994, indicating that it has the highest number of correct predictions overall. F1-score is the harmonic mean of precision and recall, where precision is the number of true positives divided by the total number of positive predictions, and recall is the number of true positives divided by the total number of actual positives. Model 2 also has the highest F1-score of 0.991, indicating that it has the best balance between precision and recall. True positives (TP) are the number of instances that are correctly classified as positive by the model, while false positives (FP) are the number of instances that are incorrectly classified as positive by the model. Model 2 has the highest number of true positives (1576), indicating that it has correctly identified a large number of positive instances.

True negatives (TN) are the number of instances that are correctly classified as negative by the model, while false negatives (FN) are the number of instances that are incorrectly classified as negative by the model. Model 4 has the highest number of false negatives

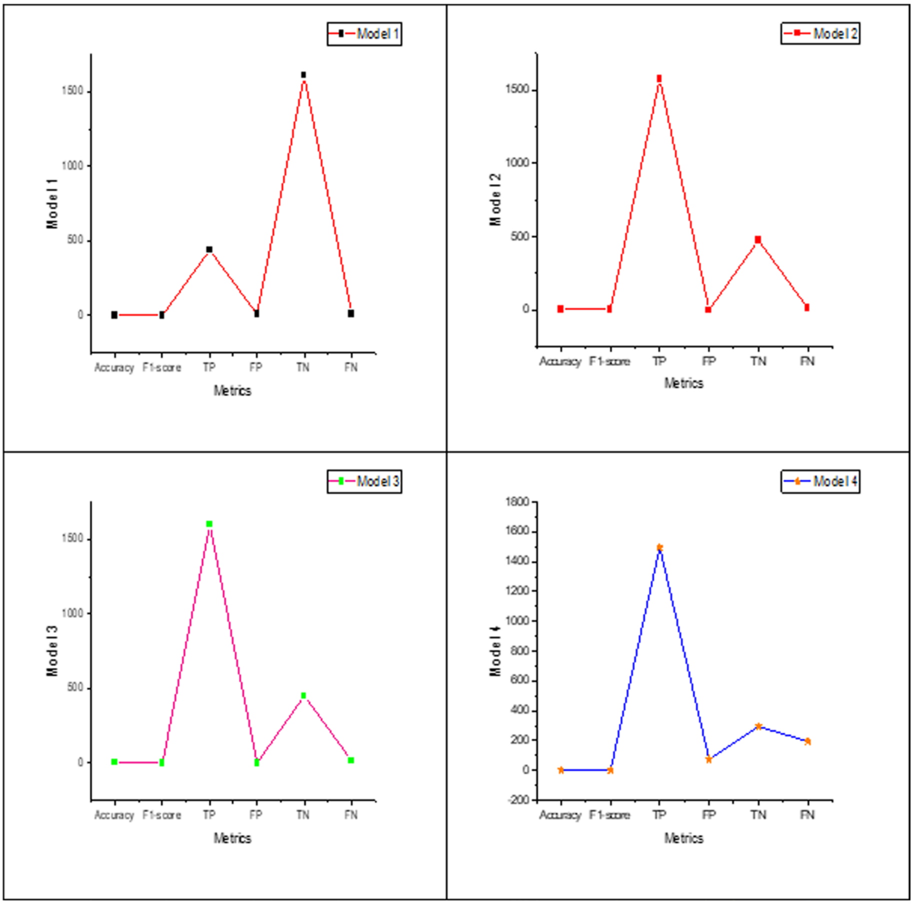

Fig. 4 Graphical representation of Twitter sentiment dataset performance metrics results

(195), indicating that it has incorrectly identified a large number of negative instances. In summary, Model 2 appears to be the best-performing model overall based on these metrics, with high accuracy, F1-score, and a high number of true positives.

All proposed hybrid models have highly accurate and high F1-score, the best model is model 2 which is Pre-processing & TF-IDF & LG, and the worse model is model 4 which is Spacy SM & SVM.

The large English core of spacy technology with 300 feature dimensions is better than the small core with 96 feature dimensions, model 1 which BERT & ANN has a very close performance to winner model 2 without data pre-processing module and with a simple artificial neural network. Show the given below Figs. 4 and 5.

5 Discussion

Four hybrid models were applied to the target dataset, model 2 which is pre-processing & TF-IDF & LG is the best model, while model 3 which is pre-processing & TF-IDF & SVM have the same pre-processing procedure and the same features extraction algorithms but differ only in the classifier.

In light of the previous result, solving natural language processing problems using logistic regression classifier is more powerful than using Multinomial Naïve Bayes classifier or using a linear support vector machine classifier, and the multinomial naïve Bayes algorithm is the worst classifier.

Using a small spacy English language library is the worst case, and using the large one gives better results, the thesis recommends using the large library rather than the small one.

BERT has a very good result and it is very close to models 2 and 3, and using BERT has enhancement over spacy libraries. Show the given below Table 2 Comparison Analysis of recent paper.

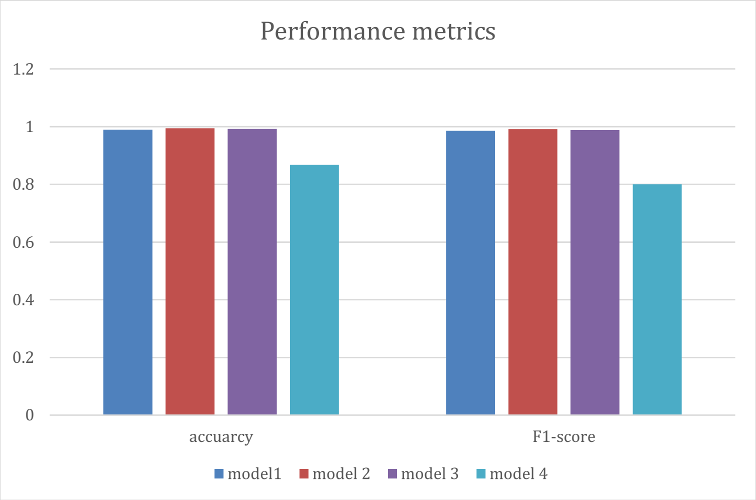

Fig. 5 Performance metrics comparison

Table 2 Comparison analysis of recent paper

REF	Author	Technique	Findings	Result
1	Mirza et al., [12]	Patient Health Questionnaire (PHQ-9), Generalized Anxiety Disorder Scale (GAD-7)	demanding academic workload, lack of sleep, long working hours, financial stress, and the pressure to perform well.	88%
2	Kamite et al., [9]	Machine learning model	To analyze social media posts for language patterns and behavioral cues associated with depression, such as negative sentiment, self-disclosure, and social isolation.	70%
3	Park et al., [13]	Machine learning model	To highlight key differences between the two groups in terms of perception towards online social media and behaviors within such systems	85%
4	proposed	multiple hybrid machine-learning models	multiple hybrid machine-learning models for sentiment analysis to detect signs of depression by analyzing Twitter tweets	99%

6 Policy suggestions

The policy suggestions for detecting signs of depression in a separate section because it is important to not only identify the problem but also provide potential solutions to address the issue. The policy suggestion that could be made is to explore the use of social media platforms, specifically Twitter, as a potential tool for detecting signs of depression. The study proposes the use of hybrid machine learning models for sentiment analysis to analyze Twitter tweets for signs of depression. By leveraging the vast amount of data available on social media platforms, this approach could help to identify individuals who are at risk for depression and connect them with appropriate mental health resources and support. Policymakers and healthcare providers could work together to develop guidelines for the ethical use of social media data for mental health purposes and promote the use of these tools to improve the detection and treatment of depression. Additionally, more research could be conducted to determine the effectiveness of this approach and to explore other potential applications of social media data for mental health purposes.

7 Conclusions

In conclusion, this study has proposed and evaluated multiple hybrid machine-learning models for sentiment analysis to detect signs of depression by analyzing Twitter tweets. The results have shown that the proposed models achieved high accuracy in detecting signs of depression, with Model 2 (pre-processing & TF-IDF & LG) being the best-performing model. However, the study has some limitations, including the use of a single dataset and the lack of a control group to compare the performance of the proposed models. To remedy these limitations, future studies could use multiple datasets and include a control group for comparison purposes. Overall, the study has achieved its objectives of proposing multiple hybrid machine-learning models for detecting signs of depression in Twitter tweets and evaluating their performance. The proposed models have shown high accuracy and could potentially be used as a tool for the early detection of depression. In terms of future scope, this study opens up possibilities for further research in the field of mental health and social media analysis. Future studies could explore the use of other feature extraction methods and machine learning algorithms, as well as the integration of other sources of data, such as images and videos. Furthermore, the proposed models could be applied to other social media platforms to detect signs of depression and other mental health issues.

8 Future work

Sentiment analysis is a hot topic in NLP and requires many enhancements to get more accurate results, in general, sentiment analysis depends on three modules or subsystems which are:

Data pre-processing module.
Features extraction module.
Classification algorithms

For future work, using the Arabic dataset is one of the available choices to complete more features of this thesis, developing more hybrid models for sentiment analysis using deep neural networks approaches like Recursive neural network (RNN) and Recurrent neural network (RNN) with combination with CNN network with different parameters is a very interesting area.

Using AraBERT which is the Arabic version of the BERT approach, seems promising to solve many of the problems of the Arabic language resulting from the complexities of the language, and in terms of my reading of many types of research, they have many uses, including sentiment analysis.

Note that the test sample differs in each experiment according to the random state in the splitting dataset process, for example in the experiment of model 1 the test samples with un-depression are more than the depression samples which is the opposite in the remainder experiments where the depression samples are more than un-depression samples.

Acknowledgements The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University for funding and supporting this work through grant number IMSIU-RG23056.

Funding This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-RG23056).

Declarations

Conflict of interest The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Arora P, Arora P (2019) Mining Twitter data for depression detection. In: 2019 international conference on signal processing and communication (ICSC)
2. Azam F, Agro M, Sami M, Abro MH, Dewani A (2021) Identifying depression among Twitter users using sentiment analysis. In: 2021 international conference on artificial intelligence (ICAI)
3. Chiu Y, Lane HY, Koh JL, Chen AL (2020) Multimodal depression detection on Instagram considering time interval of posts. J Intell Inf Syst 56(1):25–47
4. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 1:4171–4186
5. Garg M Sentimental Analysis for Tweets, Kaggle, 03-May-2021. [Online]. Available: https://www.kaggle.com/gargmanas/sentimental-analysis-for-tweets. Accessed 1 Feb 2022
6. Haq AU, Li JP, Ahmad S, Khan S, Alshara MA, Alotaibi RM (2021) Diagnostic approach for accurate diagnosis of COVID-19 employing deep learning and transfer learning techniques through chest X-ray images clinical data in E-Healthcare. Sensors 21(24):8219
7. Hassan 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)
8. Islam MR, Kamal AR, Sultana N, Islam R, Moni MA, ulhaq A (2018) Detecting depression using K-nearest neighbors (KNN) classification technique. In: 2018 international conference on computer, communication, chemical, material and electronic engineering (IC4ME2)
9. Kamite SR, Kamble VB (2020) Detection of depression in social media via twitter using machine learning approach. 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), pp 122–125
10. Kamite SR, Kamble VB (2020) Detection of depression in social media via Twitter using Machine Learning Approach. In: 2020 international conference on smart innovations in design, environment, management, planning, and computing (ICSIDEMPC)

11. Khan S, Kamal A, Fazil M, Alshara MA et al (2022) Hcovbi-caps: hate speech detection using a convolutional and bi-directional gated recurrent unit with capsule network. IEEE Access 10(1):7881–7894
12. Mirza A, Baig M, Beyari GM, Halawani MA, Mirza AA (2021) Depression and anxiety among medical students: a brief overview. Adv Med Educ Pract 12:393–398
13. Park M, McDonald D, Cha M (2021) Perception Differences between the Depressed and Non-Depressed Users in Twitter. Proceedings of the International AAAI Conference on Web and Social Media 7(1):476–485. https://doi.org/10.1609/icwsm.v7i1.14425
14. Reece AG, Reagan AJ, Lix KLM et al (2017) Forecasting the onset and course of mental illness with Twitter data. Sci Rep 7:13006. https://doi.org/10.1038/s41598-017-12961-9
15. Rosa RL, Rodriguez DZ, Schwartz GM, de Campos Ribeiro I, Bressan G (2016) Monitoring system for potential users with depression using sentiment analysis. In: 2016 IEEE international conference on consumer electronics (ICCE)
16. Saha A, Marouf AA, Hossain R (2021) Sentiment analysis from depression-related user-generated contents from social media. In: 2021 8th international conference on computer and communication engineering (ICCCE)
17. Spacy industrial-strength natural language processing in python. Industrial-strength Natural Language Processing in Python. [Online]. Available: https://spacy.io/. Accessed 22 Feb 2022
18. Stephen JJ, Prabu P (2019) Detecting the magnitude of depression in twitter users using sentiment analysis. Int J Electr Comput Eng (IJECE) 9(4):3247
19. Uddin AH, Bapery D, Arif AS (2019) Depression analysis from social media data in Bangla language using Long short-term memory (lstm) recurrent neural network technique. In: 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2)
20. Wankhade M, Rao ACS, Kulkarni C (2022) A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55:5731–5780. https://doi.org/10.1007/s10462-022-10144-1
21. Zhao W, Peng H, Eger S, Cambria E, Yang M (2019) Towards scalable and reliable capsule networks for challenging NLP applications. In: Proceedings of the 57th annual meeting of the association for computational linguistics

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.