ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 11 (2024) No. 3, pp. 21-36
DOI:10.33168/JLISS.2024.0302

Sentiment Analysis on Twitter Data for Depression Detection

Pavitra Sankar, Naveen Palanichamy*, Kok-Why Ng

Faculty of Computing and Informatics, Multimedia University,63100, Cyberjaya, Malaysia
p.naveen@mmu.edu.my (Corresponding author)

Abstract. Depression is currently the leading cause of disability worldwide, significantly increasing the disease burden. Depression impacts a person's thoughts, behavior, and quality of life. Since people nowadays tend to discuss their sentiments and thoughts regarding their mental health on social media, several researchers have recently investigated the analysis of social media material to identify and track sad users. Twitter is a popular platform for voicing people's opinions simply and directly. Therefore, numerous researchers have used Twitter to gain insights into depression. However, sentiment analysis (SA) becomes more complicated when the tweets combine two languages. This study aims to detect depression from tweets written in Malay and English languages. The data is retrieved from Twitter and pre-processed using the pre-processing approaches. Next, sentiments are extracted and labeled as positive, neutral, or negative. Bag of words (BoW) and Term Frequency- Inverse Document Frequency (TF-IDF) are the feature extraction techniques applied to the sentiments. Machine Learning (ML) classifiers such as Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and Long Short-Term Memory (LSTM) architecture, a Deep Learning (DL) technique, are used to analyze the dataset. The models' performance assessment includes the four standard measures, Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC). The result shows that Support Vector Machine is the ideally suited model for our ongoing study.

Keywords: Machine Learning, Deep Learning, Depression, Twitter

1. Introduction

Numerous mental diseases affect millions of people worldwide. These disorders affect the quality of life and their thoughts and behavior (Mathers & Loncar, 2006). One of the mental illnesses that result in a constant sensation of sorrow and disinterest is depression. There are many types of depression, but the most common depression is clinical depression condition which gives someone a constant sense of hopelessness and despair (Types of Depression, n.d.). It impacts how a person feels, thinks, and behaves and can lead to various mental and physical problems. As a result, people could struggle to carry out their daily activities and might occasionally believe that life is not worth living. It is essential to distinguish between depression and the changes in mood and temporary feelings people encounter daily. It may become a significant health issue if it continues for over two weeks. As a result, depressed individuals perform poorly and behave differently at work and home. A depressed person may commit suicide if they do not receive enough care. Social media give users a place to express their emotions. As social media usage advanced, people began to share their feelings, such as happiness, sadness, frustrations, loneliness, and more. Due to the veil of anonymity provided by social media, people are more open to discussing their mental health struggles without fearing the judgment they might face doing it in real life. Social media being easily accessible these days, along with the new-found awareness about mental health, has made these platforms a convenient and safe space for personal expression (Steinert & Dennis, 2022). These online spaces are perceived to be more accessible, less intimidating, and less stressful than real-life conversations with people. Moreover, social media data analysis helps subtlely detect depressed symptoms before developing into more severe phases of depression; this makes it possible to suggest approaches for early-stage depression therapy and prevention.

Twitter is the most frequently used social media platform that to identify depression. Researchers can gather a sizable sample of data from social media and use SA to identify potential patients and offer them more in-depth care. SA is a method for determining if a text is written in a positive, negative, or neutral tone. Every text word will receive a score from a SA algorithm based on the intended polarity; this makes it possible to determine whether a user is in a good or bad mood. Each tweet is subjected to the SA technique to determine its sentiment score, classifying it as positive, negative, or neutral. DL and ML techniques can be used to train SA models to understand text in ways beyond simple definitions, read for context, sarcasm, etc., and comprehend the writer's actual mood and feelings. The models can classify tweets as depressive or not depressive based on the sentiment score labels on tweets after the sentiment score via the SA approach. Then the detection performance is evaluated based on the accuracy of the ML and DL models.

The SA approach became an essential topic for studying Natural Language Processing (NLP). Researchers have done a significant amount of SA on social media data in the past few years. Both ML and DL techniques were used in SA, but they mainly concentrated on English since most natural language toolkits offer great datasets for the English language. However, Malay is a major language used in Malaysia, Brunei, Indonesia, Singapore, and southern Thailand, often known as Bahasa Melayu. The resources and tools for SA that are available in the Malay language are low, and the lack of data and research for Malay makes the data collection and labeling process often tedious. Besides, it takes a lot of time because some tasks have to be done manually while English language resources are easily accessible. Due to this, the study of SA in Malay is still relatively early. Thus, this paper presents a SA on Twitter data that contains a combination of Malay and English tweets to detect depression.

The paper begins with two related works in SA. Then, section 3 proposes a method of explaining ML and DL algorithms in greater depth. 4, findings, which describe how the study was conducted and how we used the data together with the results of the performance measures metrics, and 5, conclusion.

2. Related work

The prior section covered and supplied the context for this study. This section presents the researchers' numerous feature selection methods, ML techniques, DL techniques, and feature selection methods for SA.

2.1. Overview of Depression

Depression is one of the most prevailing mental illnesses in the world. Several types of depression exist, each with its unique characteristics. The most common kind of depression, known as major depressive disorder (MDD), is characterized by chronic sadness, a loss of interest in activities, and a generalized sensation of helplessness (Cleveland Clinic, 2017). A chronic but less severe form of depression is known as persistent depressive disorder (PDD). It is characterized by milder symptoms that persist for at least two years. The symptoms of seasonal affective disorder (SAD), which commonly manifests in the autumn and winter, include exhaustion, increased sleep, and a depressed mood. Some women experience postpartum depression (PPD), which causes extreme emotions of grief, worry, and tiredness. Major depressive symptoms are combined with psychotic characteristics like hallucinations or delusions in psychotic depression (Types of Depression - Beyond Blue, n.d.).

Globally, 4.4% of the world's population, or more than 300 million people, is affected, and the number is increasing daily (World Health Organization, 2021). Millions of people experience depression yearly, irrespective of their culture, gender, age, race, or financial situation. Depression symptoms can be separated into psychological, social, and physical categories. Although it is uncommon for depression sufferers to experience all these symptoms, they can indicate how severe the condition will be. Unfortunately, more than 70% of individuals with clinical depression do not seek treatment, highlighting the importance of accessible and advanced mental health care (Salas-Zárate et al., 2022).

Twitter, a popular social networking site where users share short messages called tweets, has become a platform for expressing opinions, including experiences with mental health issues. Researchers have started utilizing Twitter data to gather insights on mental health. Therefore, tweets that describe feelings with a negative connotation may express a negative emotion. The process of extracting feelings and views from tweets that users have posted on Twitter is known as SA. It has gained significant attention in recent years because it can be used in detecting earlier stages of depression by using Twitter data (Stephen & P., 2019).

2.2. Feature Extraction

Feature extraction is the process of transforming raw data into numerical features that can be used while maintaining the details of the original data set. It helps reduce the number of redundant data and speeds up the ML process. Kastrati et al. (2021) used BoW, TF, and TF-IDF in ML models to improve sentiment categorization performance. To determine the most effective method, three feature extraction methods, Word2Vec, TF-IDF, and BoW, and four classifiers are examined in the study by Jain et al. (2021). The task distribution uses a combined total of 35787 tweets from the two datasets that comprise the Twitter dataset. Palm (2019) used two of the most common feature extraction approaches, N-gram and TF-IDF, on previously gathered Twitter data.

Rabani et al. (2020) applied the TF-IDF and BoW to extract features. A TF-IDF implementation was used in WEKA to transform the text corpus into string attributes. Numerous features were derived from the data after applying the necessary filter. The author Ali (2020) used BoW for feature extraction and Information Gain (IG) as a filtering approach to improving classification performance. Moreover, Aljabri et al. (2021) gathered and pre-processed two different datasets on remote learning in the Arabic language, each with a large dataset size of over 70,000 and 92,00 tweets. The dataset was used to train and test the model's precision and was manually classified as having a positive or negative sentiment. Different SA models were developed for the experiment by combining various N-gram sizes (unigram and bigram) with the TF-IDF approach. The development of these models to convert text into numerical

variables enabled algorithms to process and analyze this information effectively.

TF-IDF and BoW are the two feature extractions that researchers use most frequently, as indicated in Table 1, out of the five feature extraction; this is because both approaches are straightforward to comprehend and apply and provide lots of flexibility for customization for specific text data.

Table 1: Summary of feature extraction used

Reference	BoW	TF-IDF	Word2Vec	N-gram	IG
Kastrati et al. (2021)	✓	✓			
Jain et al. (2021)	✓	✓	✓		
Rabani et al. (2020)	✓	✓			
Palm (2019)		✓		✓	
Ali (2020)	✓				✓
Aljabri et al. (2021)		✓		✓	

2.3. Machine Learning and Deep Learning Models

Recent research on detecting depression has found that social media users tweet about their feelings and mental state. Therefore, a few approaches have been to detect depression among social media users using ML and DL techniques. A pre-trained multilingual text representation model and DL transformers were fine-tuned by Kannan et al. (2021) to categorize the combined Tamil and English tweets into positive, negative, or neutral. A pre-trained mBERT model was employed and further tuned on the dataset. Models like the Perceptron (PC), Stochastic Gradient Descent (SGD), Ridge Classifier (RC), K-Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), Decision tree (DT), RF, Adaptive Boosting Classifier (AB), Gradient Boosting Classifier (GB), and SVM were tested and compared by Kannan et al. (2021). Following that, SVM, NB, DT, and RF were among the four classic ML models employed by Kastrati et al. (2021) to categorize a dataset of 10,742 manually labeled comments in Albanian. Rabani et al. (2020) employed five classification algorithms which are NB, Multinomial Naive Bayes (MNB), DT, Logistic Regression (LR), and SVM, using Weka.

Furthermore, Aljabri et al. (2021) employed six different classification methods: SVM, LR, KNN, NB, RF, and XGBoost (XGB). The result was assessed using four industry-accepted performance indicators. Similarly, Ali (2020) used NLP and ML algorithms for two datasets to extract sentiment using various classification techniques, including NB, MNB, KNN, LR, and SVM. The results were evaluated before and after using IG as a filtering method and BoW for feature extraction to improve classification performance. Moreover, Palm (2019) used classification models that are often used in research for SA. 12085 tweets were manually classified as positive, negative, or neutral, and SVM linear, SVM, Radial basis function (RBF), RF, and Multinomial logistic regression (MLR) classifications are used. Jain et al. (2021) collected 35787 tweets and used four classifiers in the paper. Linear SVM, RF, LR, and PC are studied to outline the best approach.

Besides, Katchapakirin et al. (2018) organized a study to build depression detection from Thai-language Facebook postings using NLP techniques. The authors employed a language translator from Google Cloud Translation API to first translate the postings from Thai to English because most models were only practiced in English. This technique could lead to inaccurate results. The authors evaluated various ML models on their projects to achieve the highest predicted performance, including SVM, RF, and DL models. MNB and Support Vector Regression (SVR) were two different models Arora & Arora (2019) used for their text-based depression detection experiment using Twitter tweets. The ratings were given as positive, neutral, and negative depending on the tweets because they used sentiment analyzers on the datasets. Thus, the accuracy of the sentiment analyzers was compared using both classifiers. Jalani et al. (2022) examined brand mentions on Twitter to analyze user sentiment towards three clothing brands using word embeddings in classification models including SVM, NB, RF, LR, and Multilayer Perceptron (MLP) in order to compare their accuracy performances., Asos, Uniqlo, and

Topshop.

By collecting tweets about the LGBTQ community, a SA was performed on a Tamil dataset (K. & Navaneethakrishnan, 2022). mBERT was then fine-tuned using the dataset, and its performance was evaluated against that of baselines, NB, and LSTM. It was shown that the multi-head attention mechanism and contextual comprehension capability of mBERT helped it perform the best. In addition to that, Razak et al. (2020) combined tweets in English and Malay and used NB, Convolutional neural network (CNN), and Feedforward neural network (FNN) for a total of 10,882 tweets. Malay tweets and tweets containing several languages are incompatible with both the NLP and ML models, which predicts that the tweets are neutral. However, when applying a DL model, the system can foresee tweets in Malay and tweets in multiple languages.

Many papers have described their analyses as limited to individual languages, making it impossible to apply such approaches globally. This is one of the research gaps noted when analyzing related work. For instance, a tweet written in English contains various words that are not customary and do not fall under the exact spelling, making it difficult to do a broad SA. Thus, this project's investigation will be narrowed down to Malay and English combined. Based on Table 2, it is possible to conclude that most past studies used RF, NB, and SVM. It can also be noted that all of them perform well. As for the DL models in Table 3, since LSTM has not been utilized frequently for SA on multi-language data, for this project, the SA model will be developed using the three well-known ML models RF, NB, and SVM and one DL model LSTM to evaluate which one has the highest accuracy.

Table 2: Summary of ML models used

Reference	SVM	GB	AB	RF	LR	DT	GNB	KNN	MNB	NB	RC	SGD	XGB	RBF	SVR	MLR	PC
Kannan et al. (2021)	✓	✓	✓	✓		✓	✓	✓			✓	✓					✓
Rabani et al. (2020)	✓				✓	✓			✓	✓							
Kastrati et al. (2021)	✓			✓		✓				✓							
Aljabri et al. (2021)	✓			✓	✓			✓		✓			✓				
Ali (2020)	✓				✓			✓	✓	✓							
Palm (2019)	✓			✓										✓		✓	
Jain et al. (2021)	✓			✓	✓												✓
Katchapakirin et al. (2018)	✓			✓													
Arora & Arora (2019)									✓						✓		
K. & Navaneethakrishnan (2022)										✓							
Razak et al. (2020)										✓							
Jalani et al. (2022)	✓			✓	✓					✓							

Table 3: Summary of DL models used

Reference	FNN	LSTM	CNN
K. & Navaneethakrishnan (2022)		✓	
Razak et al. (2020)	✓		✓

Furthermore, several researchers did not employ feature extraction or selection techniques for the SA model. Implementing feature extraction or selection will allow the SA model to obtain high accuracy scores. As a result, two standard feature extraction algorithms will be used in this study: TF-IDF and BoW. Based on previous work, the models will be evaluated using the four widely used assessment metrics. Additional metric evaluations, such as ROC and AUC, will be used, as both will provide an

overall picture of the model's adequacy.

3. Research Methodology

The experiment's general layout is shown in Figure 1. Sections 3.1 and 3.2 detail the experiment's dataset collection and the pre-processing methods used for data cleaning. Section 3.3 describes the data labelling procedure. The feature extraction techniques used in this paper were covered in Section 3.4, while the classification models are covered in Section 3.5. The details of model building can be seen in Section 3.6. Section 3.7 discusses the procedure of performance evaluation metrics. The workflow starts with the data extraction from Twitter into the dataset of tweets about depression. The data is then cleaned up using various pre-processing methods and labeled using a lexicon-based method with the proper sentiment labels. The labeled dataset is then split into training and testing sets to train and test ML models. The BoW and TF- IDF techniques are used in this project. Figure 1 shows the research's basic flow.

3.1. Data Collection

In consideration of potential limitations associated with labeling the crawled dataset solely based on sentiment analysis, which might not fully capture the complexity of depression cases, a comprehensive data collection strategy was employed. The purpose was to ensure a well-rounded dataset reflective of depression-related content. The Twitter API is required to retrieve tweets from Twitter. This strategy involved leveraging the Twitter API to extract tweets by employing carefully curated keywords indicative of depression and mental health concerns. After getting authorization, tweets are extracted and saved in CSV files. To facilitate targeted data gathering, a carefully curated set of keywords was meticulously chosen to align with the investigation's focus. The dataset was gathered by querying Twitter for tweets containing keywords indicative of depression and mental health concerns, such as “depressed,” “depression,” “sad,” “suicide,” “suicidal,” “want to die,” “wanna die,” “kill myself,” “depress,” “misery,” “unhappy,” “miserable” and their equivalents in Malay, “sedih,” “murung,” “koyak,” “penat”, “bunuh diri”, “marah”, “nangis”, “nak mati” were used to crawl the data. These carefully selected keywords ensured that the collected tweets were directly related to the topic under investigation, enabling a focused dataset acquisition process.

Then, since requests for more than 2000 tweets per minute from Twitter will probably be rate limited, a query for 2000 tweets is made. The request is executed and saved the tweets every hour for a day to produce a 10000-tweet dataset. The dataset used in this study has a shape of (10,000, 3), indicating that it contains 10,000 rows with three columns. The columns include Date and Time, Locations, and Tweets. The Date and Time column shows the time the tweets were posted, and the Locations column shows where the tweets were posted. The text of the tweets is displayed in the Tweets column. The initial column, Date and Time, employs the datetime () data type to capture temporal information accurately. The Locations column is reserved for categorical data, specifically indicating the geographical origins of the tweets. Lastly, the third column, Tweets, employs the string data type to encompass the textual content of each individual tweet.

3.2. Data Pre-processing

The gathered tweets are raw data with links, emojis, and punctuation. During pre-processing, particular parts are removed and replaced, such as punctuation, numbers, special characters, and emoji, removing URLs, and transforming capital letters to lowercase. Human intervention is necessary to ensure that the dataset is clean and impure-free. The intervention tasks are segmenting hashtags and removing the id and user id columns.

3.2.1. Translation

Before translation, the dataset underwent meticulous cleaning to manage noisy data effectively. Notably,

a dictionary-based approach was employed to expand Twitter-specific abbreviations, enhancing the readability and comprehension of the translated tweets. This involved mapping abbreviations to their complete word or phrase counterparts; for instance, the abbreviation 'tlg' was expanded to 'tolong,' meaning 'please.' Given the limited resources for sentiment analysis (SA) in Malay, the Google Translate API facilitated the translation process. The Google Translate console script seamlessly translated the Malay words into English

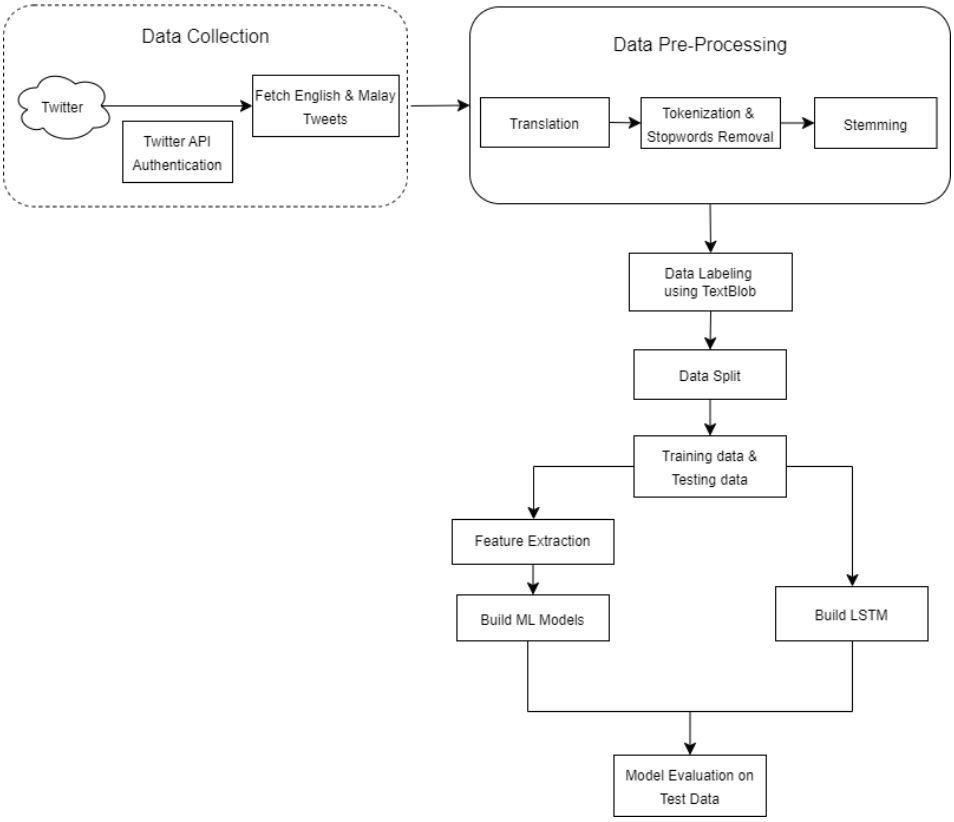

Fig. 1: Overview of the proposed methodology

For the assessment of translation quality, two distinct methodologies were employed. Human evaluation offered subjective judgments encompassing quality, fluency, and accuracy aspects. Concurrently, the BLEU score provided an objective quantitative metric grounded in n-gram overlap. Nonetheless, both evaluation methods have inherent limitations in capturing cultural nuances, fluency, and contextual meanings. To mitigate these limitations, a comprehensive evaluation approach was adopted by combining both human and automated assessments, thereby enhancing the overall translation quality appraisal. For the evaluation, humans rated the translated tweets with an average quality score of 8.0 on a scale of 1 to 10. Simultaneously, the BLEU score yielded a result of 0.85, reflecting the lexical overlap for the translated texts. It is worth noting that despite the quantifiable BLEU score, human evaluation observed that specific cultural nuances in Malay were not entirely preserved in the English translation.

3.2.2. Tokenization and Stopwords Removal

A pre-processing method called tokenization separates a text stream into tokens, such as words, sentences, symbols, or other essential elements. Decoupling words from URLs, hashtags, and at-mentions is a process known as tokenization. It is crucial for SA because it tells us about the components of a tweet. Counting the occurrences of terms is the primary goal of this method. Prepositional phrases and other stop-words are common but do not affect the text's overall sentiment and should be removed during pre-processing.

3.2.3. Stemming

Stemming is reducing a word's complexity from its original state. The model can fully comprehend the text's meaning by reducing the words' complexity. The word "like" in its infinitive form, for instance, is simpler for a machine to comprehend than the words "liked," "liking," and "likes."

3.3. Data Labeling

After the tweets have been pre-processed, the dataset will be labeled using the TextBlob library function in Python. TextBlob is a vocabulary technique that can be applied to many NLP tasks, including SA, paraphrase mining, sorting, and other activities. It will also classify them as either positive, negative, or neutral. Each tweet's polarity is produced using the TextBlob library, where the polarity score ranges from 1 to -1. Tweets with polarity values below 0 are viewed negatively, tweets above or equal to 0 are viewed neutrally, and tweets above 0 are viewed positively. Figure 2 shows the tweets with positive and negative sentiments after labeling the data.

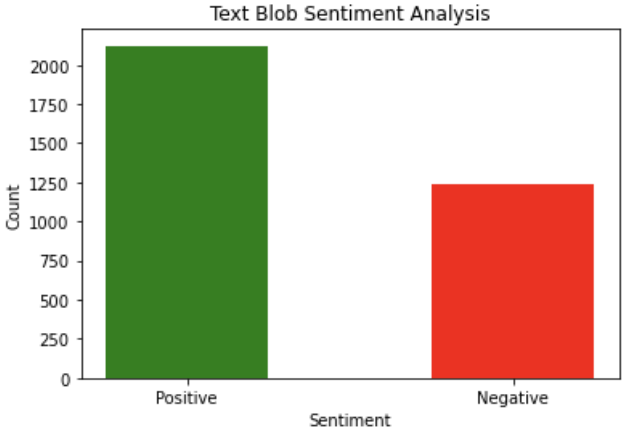

Fig. 2: TextBlob sentiment analysis

3.4. Feature Extraction

Since text cannot be computed, ML algorithms cannot directly process natural language text. Text input is therefore transformed into numerical values that the techniques can recognize and handle utilizing a features extraction approach. The two most popular feature extraction techniques for extracting features from tweets, BoW and TF-IDF, will be used in this project to achieve excellent accuracy.

To extract features from shortened words or information, a method called BoW is applied. The BoW counts the number of times each term appears, determines the document's keywords based on each word's frequency, and generates a frequency histogram. In short, the BoW is employed to increase the lexicon of all unmatched phrases and train models based on their frequency. A well-known method for information extraction and NLP is called TF-IDF. Finding the frequency of terms in a sizable document database is the goal of TF-IDF. In a nutshell, TF-IDF is a feature extraction method that takes weighted features from textual data and extracts them. It provides the weight of each phrase in the corpus to improve the performance of learning models. Common terms in the corpus are shown by smaller TF-IDF values, which suggest that they are unimportant. On the other hand, the larger TF-IDF values indicate fewer frequent words in the corpus and are, therefore, significant.

3.5. Classification Model

Three ML models, SVM, RF, and NB, are implemented in this paper. For the DL model, LSTM was used. The following sections explain the ML and DL models in depth.

3.5.1. SVM

SVM is an effective ML that is utilized for classification and regression problems. Maximizing the

difference margin between classes or target values, they arrive at the ideal decision limits. The linear model is represented by the SVM formula y(x) = wT * x + b, where x is the input data point, w is the weight vector, b is the bias component, and y(x) is the projected result (Saini, 2021). The kernel approach, which transforms data into higher-dimensional spaces, allows SVMs to handle both linearly and non-linearly separable data. SVMs can now accurately forecast the future and capture complicated patterns. SVMs are frequently employed because of their versatility and capacity to identify the best decision limits.

3.5.2. RF

RF is a well-known supervised learning technique that enhances model performance through ensemble learning. To produce predictions, it integrates several decision trees that have been trained on various subsets of the dataset (E R, 2021). RF achieves improved predictive accuracy and prevents overfitting by averaging the forecasts from each tree and depending on a majority vote. It is a popular choice for handling challenging problems and achieving robust results. In RF, the Gini index is used to measure attribute defilement concerning classes. It calculates the probability (f(ci, x) / x) of a specific class category (ci) existing in a randomly chosen category (pixel) from the training set (x). The Gini index helps determine the optimal attribute for splitting the data in decision tree construction within Random Forest.

$$\sum\sum_{j\neq i} = \left(\frac{f(ci,x)}{x}\right) f\ cj, \frac{x}{x}) \tag{1}$$

3.5.3. NB

The NB method is also a supervised learning method for classifying, and it is based on the Bayes theorem (Ray, 2019). It is mainly employed in text categorization with an extensive training set. The NB classifier is one of the most simple and effective classification algorithms. NB helps create quick ML models capable of giving good prediction accuracy. It provides predictions based on the likelihood of an object occurring because it is a probabilistic classifier. The Bayes' theorem, often known as Bayes' law or Bayes' rule, is used to determine how likely a hypothesis is given previous knowledge. This is decided by conditional probability. Given that another event has already occurred, the possibility of an event happening is known as conditional probability. The Bayes theorem's formula is as follows:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{2}$$

The posterior probability, or P(A|B), calculates the likelihood that a given hypothesis (A) will occur. P(B|A) refers to Likelihood Probability, which calculates the likelihood that a given hypothesis is true based on available evidence. Prior probability, or P(A), is the likelihood of a hypothesis before observing the evidence. A marginal probability, or P(B), represents the likelihood of the evidence.

3.5.4. LSTM

Recurrent neural network (RNN) architectures such as LSTM are made to handle sequential data and capture long-term dependencies (What Is LSTM - Introduction to Long Short Term Memory, n.d.). It solves the vanishing gradient issue by using memory cells, input gates, output gates, and forget gates. These gates control information flow inside the network, enabling LSTM to remember or forget information over extensive time intervals selectively. With its capacity to gather and retain critical information over time, LSTM has gained popularity for various tasks, including time series analysis, speech recognition, and natural language processing.

3.6. Model Building

Two different feature extraction techniques were combined in the research's models. The following parameters were chosen for the RF model: 'n_estimators = 500' to use 500 decision trees for predictions; 'max_depth = 300' to limit each tree's depth to 10 levels; 'criterion = entropy' to choose the splitting

criteria based on Gini impurity; and 'random_state = 42'. Five parameters make up the Support Vector Machine (SVM) model. The loss function used to reduce loss and boost accuracy is determined by the "loss = log" expression. With comparable results, "l1_ratio = 0.15" generates models that are a little less centralized. Maximum iterations are restricted to 300 predictors and a maximum depth of 300 when "max_iter = 300" is used. The quantity of CPUs used for training is specified by "n_jobs = 4". To avoid overfitting, the command "random_state = 12" trains the models four times using 12 random seeds.
Then, Var smoothing was used in the NB model, with an alpha value of 0.5. By adjusting the class-conditional probabilities by a small amount, this method tackles the problem of zero probability. The model compromises prior knowledge and flexibility with an alpha set to 0.5. Additionally, for the LSTM model, a batch size of 16 epochs was used to train. Figure 3 shows the architecture of the LSTM model. The testing dataset was used to assess the model's performance.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 55, 64)            128000

 lstm (LSTM)                 (None, 16)                5184

 dense (Dense)               (None, 1)                 17

=================================================================
Total params: 133,201
Trainable params: 133,201
Non-trainable params: 0
_________________________________________________________________
None
```

Fig. 3: The architecture of the LSTM model

3.7. Performance Evaluation Metrics

Model evaluation is an essential step in the model creation process. After the models are completed, the four standard performance evaluation metrics and a classification report will be generated. The AUC is then plotted after the ROC Curve. Firstly, accuracy is the percentage of actual outcomes concerning the total instances studied. Precision shows how often a model's precision in predicting positive labels is correct. Then, recall measures the percentage of real positives a model successfully identified (True Positive). Recall should be used when the cost of a false negative is high. F1-Score is the mean of the precision and recall values. The true-positive rate is represented on the Y-axis of the ROC curve, which has the false-positive rate plotted on the X-axis. Finally, the AUC metric measures how successfully positive classes are differentiated from negative class probabilities.

In addition to model evaluation, the study incorporates an essential error analysis process, which involves meticulously examining the errors or inconsistencies made by the model during predictions or classifications. This analysis enhances the understanding of error patterns and underlying causes, facilitating iterative improvements in the model's performance. Moreover, error analysis not only aids in refining predictive capabilities but also sheds light on challenges specific to sentiment analysis for detecting depression indicators, guiding potential enhancements to the approach. The comprehensive insights gained from model evaluation and error analysis contribute to a deeper comprehension of the model's accuracy and categorization capabilities.

Table 4: Evaluation metrics formula

Evaluation Metrics	Formula
Accuracy	$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$
Precision	$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$
F1-Score	$F1 - score = \frac{2\,x\,precision\,x\,recall}{(precision + recall)} = \frac{2TP}{2TP + FP + FN}$
Recall	$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$
ROC curve	$True\ Positive\ Rate = \frac{True\ Positive}{True\ Positive + False\ Negative}$ $False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + True\ Negative}$
AUC	$AUC = \frac{True\ Positive\ Rate + True\ Positive\ Rate}{2}$

4. Results and Discussion

This section evaluates the effectiveness of machine learning models using the BoW and TF-IDF features. Building upon the foundation laid by previous studies (Kastrati et al., 2021), the performance of the models in detecting depression indicators from social media texts is accessed. In Sections 4.1 and 4.2, the results of the ML models for BoW and TF-IDF were presented. In Section 4.3, the results of the LSTM model are presented. Finally, the overall conclusions are covered in Section 4.4.

4.1. Results of BoW

The BoW results for all three models are displayed in Table 5. The score for SVM is 91%, 93%, 90%, and 92%, while for NB, it is 90%, 90%, 92%, and 91% in all four measures. Then, RF scores 89%, 89%, 91%, and 90%, the lowest score compared to the other ML models. Even though all models had good BoW performance, SVM with BoW performance was superior and more dependable, as shown in the evaluation metrics.

Table 5. Models' Performance Using BoW

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.91	0.93	0.90	0.92
RF	0.89	0.89	0.91	0.90
NB	0.90	0.90	0.92	0.91

4.2. Results of TD-IDF

Table 6 shows the results of using the feature extraction TF-IDF for the classification algorithms in predicting the sentiments of tweets. Like Bow, SVM using TF-IDF produced the highest values of 90%, 90%, 92%, and 91%, and NB received 89%, 90%, 89%, and 90% in all four measures, respectively. Meanwhile, RF obtains slightly lower results in all four measures with values of 88%, 87%, 90%, and 89%, respectively.

Table 6. Models' Performance Using TD-IDF

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.90	0.90	0.92	0.91
RF	0.88	0.87	0.90	0.89
NB	0.89	0.90	0.89	0.90

4.3. Results of LSTM

The results of using the LSTM model for predicting the sentiments of tweets are displayed in Table 7. The outcomes for LSTM are the lowest compared to the other ML models despite the small differences. The score received by the LSTM model is 86%, 82%, 93%, and 87%.

Table 7. Models' Performance Using LSTM

Model	Accuracy	Precision	Recall	F1-Score
LSTM	0.86	0.82	0.93	0.87

4.4. Overall Findings

As for the overall findings, the results show that for both feature extractions, the SVM model beats the RF and NB model. The SVM with the BoW model consistently earns the highest score in all four performance evaluation measures and is the best model with the best performance. Thus, the SVM model with Bow is the best model to detect depression from tweets. The table shows that all models, using both feature extraction, perform well across all four performance evaluation metrics. This demonstrates that it has been shown that the ML models used can perform SA more effectively. There could be several reasons for their relatively lower performance. RF, being an ensemble model, relies on the combination of decision trees, and it may not have captured the complex relationships and patterns present in the data as effectively as SVM. NB, a probabilistic model, assumes that features are independent of one another, which may not be true for the given dataset and results in less-than-ideal outcomes. For LSTM, a deep learning model, additional training data, or a new architecture may be needed to capture the sentiment patterns in the text data successfully. Furthermore, the feature extraction methods may have hampered these models' performance, or the hyper-parameter values they used may not have been the best ones for the task at hand. A problem that has been discovered with sentiment analysis models has been discovered is their capacity to parse ambiguous language, as revealed by the error analysis. These models often struggle with interpreting the frequent use of sarcasm and irony in social media posts.

In the pursuit of evaluating the effectiveness of the ML models for depression detection through sentiment analysis, it is essential to delve beyond conventional metric calculations. While the primary focus has been on metrics such as accuracy, precision, recall, and F1-score to gauge model performance, an additional layer of analysis has been applied to explore the connection between sentiment-based indicators and the actual level of depression. This involved categorizing sentiment predictions into distinct levels of depression severity based on the obtained results. By closely examining the alignment between sentiment scores and established patterns of depressive content, the goal was to compare sentiment analysis outcomes with the depression level directly. This comprehensive analysis provides a nuanced perspective on the models' performance, offering insights into their potential to identify indicators related to depression. This approach constructs a bridge between sentiment analysis and a more profound understanding of depression detection.

Table 8 shows the AUC values for all the models, and the ROC curves in Figures 4, 5, 6, and 7 can also be used to evaluate each model's performance accurately. Each model gives a result that is close to one, indicating a good threshold, a higher TPR, and a lower FPR. The SVM model performs better than the RF and NB models, demonstrating that it performs better.

Table 8. Models' AUC values

Feature Extraction	BoW			TF-IDF			-
Model	SVM	RF	NB	SVM	RF	NB	LSTM
AUC	0.96	0.93	0.95	0.97	0.92	0.95	0.93

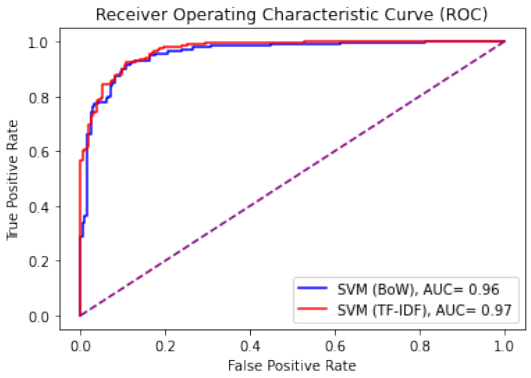

Fig. 4: ROC curve for SVM

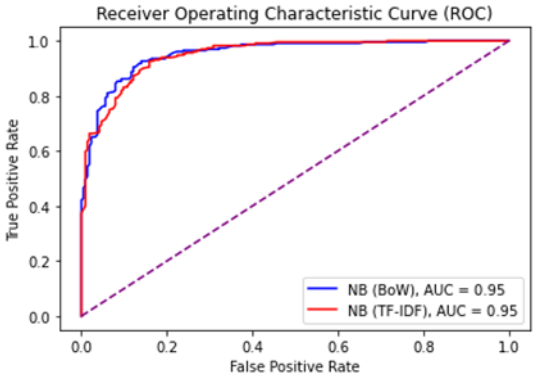

Fig. 5: ROC curve for NB

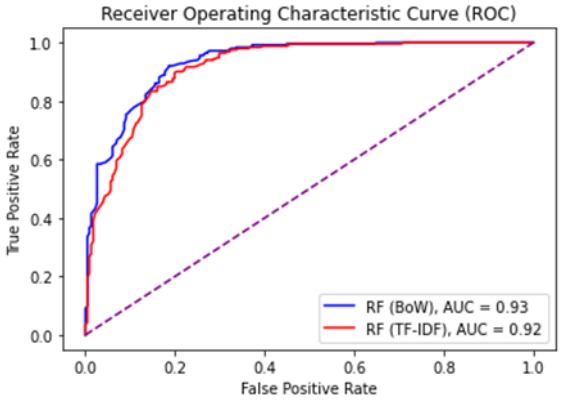

Fig. 6: ROC curve for RF

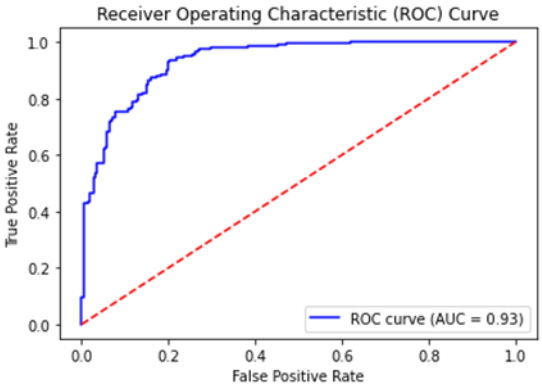

Fig. 7: ROC curve for LSTM

5. Conclusion

This study analyzed pre-processed Twitter data to identify depression in Malay and English tweets. BoW and TF-IDF were used to extract sentiments and categorize them as positive, neutral, or negative. ML classifiers like SVM, RF, NB, and the DL architecture of LSTM were used. Evaluation metrics such as accuracy, precision, recall, F1-score, ROC curve, and AUC were used to determine the best-performing model. The results show that the SVM with the BoW features extraction method performs the best. Together, these models can identify depression in its earliest stages and help by providing help to those affected. This data was gathered from positive, neutral, and negative tweets tweeted by individuals who may or may not have depression. It is essential to understand the relevant aspects of the current dataset labeling strategy based on sentiment analysis and to suggest alternative approaches for future research. Though sentiment analysis offered helpful insights, it was unable to capture the variety of depression symptoms fully. Additional metrics, such as self-reported assessments of depression symptoms, should be added to the dataset to increase its accuracy and reliability. To fully understand depression, using several data sources and incorporating clinical records or expert annotations into the labeling process is essential. These techniques enable researchers to identify cases of depression more precisely while also increasing the validity of their findings.

References

Advanced, in. (2015). *Sentiment Analysis on Twitter Data*. Academia.edu. https://www.academia.edu/10671663/Sentiment_Analysis_on_Twitter_Data

Ali, M. M. (2021). Arabic sentiment analysis about online learning to mitigate covid-19. *Journal of Intelligent Systems*, *30*(1), 524–540. https://doi.org/10.1515/jisys-2020-0115

Aljabri, M., Chrouf, S. Mhd. B., Alzahrani, N. A., Alghamdi, L., Alfehaid, R., Alqarawi, R., Alhuthayfi, J., & Alduhailan, N. (2021). Sentiment Analysis of Arabic Tweets Regarding Distance Learning in Saudi Arabia during the COVID-19 Pandemic. *Sensors*, *21*(16), 5431. https://doi.org/10.3390/s21165431

Arora, P., & Arora, P. (2019). *Mining Twitter Data for Depression Detection*. IEEE Xplore. https://doi.org/10.1109/ICSC45622.2019.8938353

Cleveland Clinic. (2017). *Depression Symptoms, Causes, & Treatment* . Cleveland Clinic; Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/9290-depression

D, E. (2019, December 8). *Accuracy, Recall & Precision*. Medium. https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d

E R, S. (2021, June 17). Random Forest | Introduction to Random Forest Algorithm. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Jain, V., Kumar, V., Pal, V., & Vishwakarma, D. K. (2021). Detection of Cyberbullying on Social Media Using Machine learning. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. https://doi.org/10.1109/iccmc51019.2021.9418254

Jalani, M. S., Ng, H., Yap, T. T. V., & Goh, V. T. (2022). Performance of Sentiment Classification on Tweets of Clothing Brands. *Journal of Informatics and Web Engineering*, 1(1), 16-22.

Kannan, R., Swaminathan, S., Anutariya, C., & Saravanan, V. (2021). Exploiting Multilingual Neural Linguistic Representation for Sentiment Classification of Political Tweets in Code-mix Language. *The 12th International Conference on Advances in Information Technology*. https://doi.org/10.1145/3468784.3470787

Kastrati, Z., Ahmedi, L., Kurti, A., Kadriu, F., Murtezaj, D., & Gashi, F. (2021). A Deep Learning Sentiment Analyser for Social Media Comments in Low-Resource Languages. *Electronics*, *10*(10), 1133. https://doi.org/10.3390/electronics10101133

Katchapakirin, K., Wongpatikaseree, K., Yomaboot, P., & Kaewpitakkun, Y. (2018). Facebook Social Media for Depression Detection in the Thai Community. *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. https://www.semanticscholar.org/paper/Facebook-Social-Media-for-Depression-Detection-in-Katchapakirin-Wongpatikaseree/7f96019570c5c29a516620aa03ff8fa8dfd8e736

K., L. S., & Navaneethakrishnan, S. C. (2022, July 1). *Building Tamil Text Dataset on LGBTQIA and Offensive Language Detection using Multilingual BERT*. IEEE Xplore. https://doi.org/10.1109/ICICT54344.2022.9850904

LSTM networks - Keras Deep Learning Cookbook [Book]. (n.d.). Www.oreilly.com. https://www.oreilly.com/library/view/keras-deep-learning/9781788621755/d6480eab-dfca-45ae-9758-eb801bc0891c.xhtml

Mathers, C. D., & Loncar, D. (2006). Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Medicine*, *3*(11), e442. https://doi.org/10.1371/journal.pmed.0030442

Rabani, S. T., Khan, Q. R., & Khanday, A. M. U. D. (2020). Detection of Suicidal Ideation on Twitter using Machine Learning & Ensemble Approaches. *Baghdad Science Journal*, *17*(4), 1328. https://doi.org/10.21123/bsj.2020.17.4.1328

Ray, S. (2019, September 3). 6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Razak, C. S. A., Zulkarnain, M. A., Hamid, S. H. A., Anuar, N. B., Jali, M. Z., & Meon, H. (2020). Tweep: A System Development to Detect Depression in Twitter Posts. *Lecture Notes in Electrical Engineering*, 543–552. https://doi.org/10.1007/978-981-15-0058-9_52

Saini, A. (2021, October 12). *Support Vector Machine(SVM): A Complete guide for beginners*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Salas-Zárate, R., Alor-Hernández, G., Salas-Zárate, M. del P., Paredes-Valverde, M. A., Bustos-López, M., & Sánchez-Cervantes, J. L. (2022). Detecting Depression Signs on Social Media: A Systematic Literature Review. *Healthcare*, *10*(2), 291. https://doi.org/10.3390/healthcare10020291

Sentiment classification of Swedish Twitter data Niklas Palm. (n.d.). Retrieved February 13, 2023, from https://www.utn.uu.se/sts/student/wp-content/uploads/2019/07/1907_Niklas_Palm.pdf

Steinert, S., & Dennis, M. J. (2022). Emotions and Digital Well-Being: on Social Media's Emotional Affordances. *Philosophy & Technology*, *35*(2). https://doi.org/10.1007/s13347-022-00530-6

Stephen, J. J., & P., P. (2019). Detecting the magnitude of depression in Twitter users using sentiment analysis. *International Journal of Electrical and Computer Engineering (IJECE)*, *9*(4), 3247. https://doi.org/10.11591/ijece.v9i4.pp3247-3255

Types of depression - Beyond Blue. (n.d.). Www.beyondblue.org.au. https://www.beyondblue.org.au/mental-health/depression/types-of-depression

Types of Depression. (n.d.). Depression and Bipolar Support Alliance. https://www.dbsalliance.org/education/depression/types-of-depression/#:~:text=Major%20depressive%20disorder%20and%20persistent

What is LSTM - Introduction to Long Short Term Memory. (n.d.). Intellipaat.com. Retrieved June 13,

2023, from https://intellipaat.com/blog/what-is-lstm/?US#:~:text=

World Health Organization. (2021, September 13). *Depression*. World Health Organization; World Health Organization. https://www.who.int/news-room/fact-sheets/detail/depression