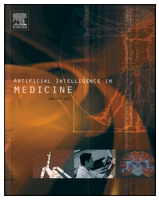
Artificial Intelligence In Medicine 147 (2024) 102716

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Depression detection for twitter users using sentiment analysis in English and Arabic tweets

AbdelMoniem Helmy *, Radwa Nassar, Nagy Ramdan

Department of Information Systems and Technology Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

ARTICLE INFO

Keywords:
Depression detection
Anxiety
Suicide
Machine learning
Social media
Text mining
Sentiment analysis

ABSTRACT

Since depression often results in suicidal thoughts and leaves a person severely disabled daily, there is an elevated risk of premature mortality due to mental problems caused by depression. Therefore, it's crucial to identify the patient's mental illness as soon as possible. People are increasingly using social media platforms to express their opinions and share daily activities, which makes online platforms rich sources of early depression detection. The contribution of this paper is multifold. *First*, it presents five machine-learning models for Arabic and English depression detection using Twitter text. The best model for Arabic text achieved an f1-score of 96.6 % for binary classification to depressed and Non_dep. For English text without negation, the model achieved 92 % for binary classification and 88 % for multi-classification (depressed, indifferent, happy). For English text with negation, an 87 %, and 85 % f1 score was achieved for binary and multi-classification respectively. *Second*, the work introduced a manually annotated Arabic_Dep_tweets_10,000 corpus of 10.000 Arabic tweets, which covered neutral tweets as well as a variety of depressed and happy terms. In addition, two automatically annotated English corpora, Eng_without_negation_60.000 corpus of 60,172 English tweets and Eng_with_negation_57.000 corpus of 57,392 English tweets. Both covered a wide range of depressed and cheerful terms; however, Negation was included in the Eng_with_negation_57.000 corpus. *Finally*, this paper exposes a depression-detection web application which implements our optimal models to detect tweets that contain depression symptoms and predict depression trends for a person either using English or Arabic language.

1. Introduction

1.1. Context and background

Depression, as defined by the World Health Organization (WHO), is a common mental health disorder characterized by persistent sadness, loss of interest or pleasure, feelings of guilt or low self-worth, disturbed sleep or appetite, tiredness, and poor concentration (WHO, 2021) [66]. Individuals who do not obtain therapy for depression in a timely manner will have worsening symptoms. More than 75 % of persons in the early stages of depression did not seek help from a psychologist, and their illnesses worsened [1,2].

Mental health is a critical aspect of overall well-being, with disorders such as depression and anxiety affecting millions of individuals worldwide. The timely detection and intervention of these mental health conditions are paramount in improving the quality of life for those affected [1,7]. Over the years, the field of mental health detection has witnessed significant advancements, driven in large part by the rapid proliferation of social media platforms and the vast amount of data they generate [3,4].

Recent research has explored the potential of utilizing social media data, particularly from platforms like Twitter, as a valuable resource for detecting signs and symptoms of mental health disorders [2,6]. This avenue of investigation has opened new possibilities for early intervention and support for individuals who may be experiencing mental health challenges [5].

However, while there have been notable strides in this field, there remain significant gaps and limitations in existing approaches to mental health detection through social media analysis. Many of the current methods rely on simplistic linguistic features and may not capture the nuanced expressions of mental distress [7]. Furthermore, most of the research in this area has focused on English-language content, limiting the applicability of these approaches to a global audience [3].

* Corresponding author.
E-mail address: abdelmoniem.hafez@cu.edu.eg (A. Helmy).

https://doi.org/10.1016/j.artmed.2023.102716
Received 13 January 2023; Received in revised form 6 November 2023; Accepted 8 November 2023
Available online 19 November 2023

1.2. Research objectives

In this context, our study seeks to address some of these critical limitations by introducing a methodology and datasets that bridge gaps in the current state of research. We not only extend our investigation to the Arabic language, which has received comparatively less attention in the field but is nonetheless crucial, but also propose new approaches for feature extraction and data resampling that have the potential to enhance the accuracy and reliability of mental health detection [2,5].

In this paper, we present our contributions, including a manually labeled Arabic depression corpus, two automatically labeled English depression corpora, and a comprehensive analysis of various machine learning algorithms applied to these datasets [6]. Additionally, we explore the impact of data resampling techniques, shedding light on their effectiveness in mitigating the challenges posed by imbalanced datasets [5].

By providing a comprehensive overview of our methodology and experimental results, we aim to contribute to the ongoing discourse on mental health detection via social media while addressing the gaps and limitations that have persisted in this field [2].

The rest of the paper is structured as follows: II. Related Work, III. Methodology, IV. Experiments & Results, V. Results Analysis, VI. Discussion, VII. Limitations, and VIII. Conclusion and Future Directions.

2. Related work

This section provides an overview of the existing research in the field of mental health detection, highlighting both the strengths and limitations of previous approaches.

1. **Lexicon-Based Approaches**: Cha, Kim, and Park (2022) proposed a lexicon-based approach to detect depression in Twitter data, particularly focusing on the university community [8]. This approach analyzed language patterns and utilized a lexicon to identify tweets indicative of depression symptoms. While this method offers valuable insights, it has limitations in terms of accuracy.
2. **Multimodal Analysis**: Safa, Bayat, and Moghtader (2021) [6] introduced an automatic detection method that combines multimodal analysis techniques to identify depression symptoms in tweets. By incorporating textual and visual cues, this approach aimed to enhance detection accuracy. However, there is room for improvement in multimodal feature integration.
3. **Literature Review Findings**: A systematic literature review [61] emphasized Twitter as the most extensively studied social media platform for detecting depression signs. Word embedding emerged as a popular linguistic feature extraction method, and support vector machine (SVM) was frequently used as a machine learning algorithm. This review provides valuable insights into the existing landscape but lacks specific publication details.
4. **Hybrid Deep Learning**: Kour and Gupta (2022) [5] proposed a hybrid deep learning approach for depression prediction from user tweets, combining feature-rich convolutional neural networks (CNN) and bi-directional long short-term memory (LSTM) models. This approach shows promise in improving prediction accuracy and addressing the limitations of previous models.
5. **Deep Learning Techniques**: Khafaga, Auvdaiappan, and Abouhawwash (2023) [62] explored the use of deep learning techniques for depression detection using Twitter data. Their deep learning model analyzed tweet content to identify signs of depression. This research contributes to the growing interest in leveraging deep learning for mental health detection.

Mustafa R.U et al. (2020) [51] mentioned the role of Preprocessing in cleaning data to convert the raw tweets into useful text involving data transformation, instance selection, normalization, and feature extraction. TF-IDF was used to assign weights for each token according to the relative impact. Later, LIWC was used which classified the words into fourteen psychological attributes. Each word categorized by LIWC was given a weight based on a happiness scale ranging from unsatisfied to cheerful (1–9). They identified the characteristics of symptoms for each of the three defined classes of depression, (H class) self-interest, feelings of worthlessness or guilt, problems with decision-making, and suicidal thoughts. (M class) (including PMDD sufferers): mood swings, anxiety, fatigue, irritation. (L class) (including SAD sufferers) and situational and atypical depression: some signs of fatigue or paranoia. For classification, they used four classifiers SVM, RF, 1DCNN, and NN. To build a classifier, they used the top 100 keywords used by depressed users. The optimal accuracy was with 1DCNN at 91 %. They state that 1D CNN works well when you want to extract relevant features from shorter chunks of a larger data set and when the feature's placement inside the chunk is not important and irrelevant.

Orabi et al. (2018) [13] pointed out that it the worth noting that social media platforms can mirror users' personal lives on a variety of levels. They advocated for the use of supervised machine learning techniques like deep neural networks. Given the limited amount (in comparison to most deep neural network architectures) of unstructured data, their primary goal was to detect depression using the most effective deep neural architecture from two of the most popular deep learning approaches in the field of natural language processing: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). A neural network-based approach for enhancing and optimizing word embeddings was presented. The improved embeddings were tested using three typical word embeddings: random trainable, skip-gram, and CBOW. Four different neural network models are utilized to assess depression detection accuracy. The first three models employ CNN, whereas the last model employs RNN. When compared to other models, the CNNWithMax models utilizing the optimized embedding had greater accuracy (87.957 %), F1 (86.967 %), AUC (0.951), precision (87.435 %), and recall (87.029 %).

In addition, Razak et al. (2020) [17] built a system for detecting depression in tweets. Vader Sentiment Analysis and two Machine Learning and deep learning approaches NB and CNN, are used in the system. The system's output is a percentage of positive and negative tweets from the users' accounts on the Twitter platform and the followers they have.

Also, F. M. Shah et al. (2020) [52] Experimented with a variety of feature sets including TrainableEmbed Features, GloveEmbed Feature, Word2Vec Embed Features, FastexEmbed Features and Metadata Features-LIWC. Embedded features were passed into (BiLSTM) layer of output dimension 600. At Risk Window 23, the W2VEmbed + Meta feature set has the greatest F1 Score of 0.81, with a precision of 0.78 and recall of 0.86. The maximum F Latency is 0.59 at Risk Window 15 and the lowest ERDE50 is 0.10 at Risk Window 10.

There are also two works concerned with other languages first, Xiaoxu Yao et al. (2021) [18] collected more than 100 thousand Chinese posts from the Sina Weibo platform. They utilized the deep neural network Attention-Based Bidirectional Long Short-Term Memory model (Att-BLSTM) to build their classifiers then they compared the performance of the ATT-BLSTM classifier to two baselines: Support Vector Machines (SVM) and Random Forests (RF).

Second work, Uddin et al. (2021) [19] collected more than 200 thousand public Norwegian posts from the ung.no website. They proposed an approach using Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN) to identify texts describing self-perceived symptoms of depression. LSTM and RNN achieved their best with the authors' proposed features + one hot encoder.

6. **General Discussions**: Several works (Chenhao Lin et al., 2020 [63]; Angskun, Tipprasert, and Angskun, 2022 [64]) discussed the development of methods to identify signs of depression in online social media users, emphasizing the importance of mental health and the

need for effective detection methods. However, these sources lack specific authorship details.

Research done by Salma Almouzini et al. (2019) [12] more than 7000 Tweets from 97 users in the Gulf region have been collected. They mentioned four preprocessing steps as follows: Tokenization, Stemming, Stop words removal, and Elimination of speech effect. Bag-of-unigrams and negation handling were merged as features. The TweetToSparse-FeatureVector filter was used to add the NEGTOKEN- tag to words occurring in negated contexts. The outcome measures of the four machine learning algorithms used were compared through the percentage of recall and accuracy. Liblinear algorithm recorded an accuracy of 87.5 % and a recall of 87.5 % whereas the Ada algorithm recorded the lowest accuracy of 55.2 % and a recall of 55.3 %.

Raymond Chiong et al. (2021) [14] used two datasets for training and testing, three data sets for validation and four classifiers with BOW. First, the accuracy was 99 % on the two datasets with very low accuracy reaching 0 % on validation datasets due to overfitting. The author studied the impact of deleting the words "depression" and "diagnose" on the accuracy and the result reached 90 % on the validation datasets with the LR classifier. They conducted experiments to see if the sampling approach was impactful, and it was reported that both over- and under-sampling enhanced the detection of depression class with 92 % accuracy.

Hemanthkumar M et al. (2019) [15] the relevance of NLP techniques for getting useful keywords, which plays a big role in understanding emotions, was highlighted. The preprocessing included: Emoji Extraction, Hyperlink Removal, Slang substitution, Timestamp removal, Digits removal, Spelling correction, Shortening, Correction, Proper nouns removal, Lemmatization and Stop words removal. The author compared two algorithms SVM and NB with BOW for feature extraction. The best result was 0.7297 accuracy, 0.7504 recall, and 0.7458 precision using the Multinomial Nave Bayes algorithm.

Prof. S. J. Pachouly et al. (2021) [16] The relevance of data preprocessing as a necessary step in developing a Machine Learning model and relying on how well the information or text data has been preprocessed was highlighted. They preprocessed the corpus using Natural Language Processing (NLP) methods before using feature extraction methods and training the model. The Tokenization approach was used to divide up the tweets from Twitter into individual tokens as the first stage in pre-processing data. Second, remove any URLs, punctuation, and stop words. Third, emojis or emoticons were removed because they can provide essential information about the sentiment. Then they used stemmer to scale the words back to their origins and group terms that are related along the way. Foremost of feature extraction included were: Bag-of-Words, TF-IDF, and Parts of Speech (POS) Tagging. The model was built using NB and SVM Machine Learning Classification Techniques. With an accuracy of 87.5 %, the linear classifier was the most accurate.

In addition, Zunaira Jamil et al. (2017) [45] proposed a Twitter-based intelligent system that can detect at-risk users. They studied the impact of under-sampling and over-sampling on the original dataset and their effect on the accuracy rate. To classify tweets at the tweet level, an initial experiment was executed on 1. LSVM trained on the dataset in its first form before balancing 2. LSVM with trained on the dataset balanced using SMOTE 3. LSVM trained on the dataset balanced by under-sampling. The best performing is a Linear SVM classifier trained on a balanced dataset using SMOTE and the percentages for classifier evaluation were as follows: accuracy of 78.72 %, precision of 70.83 %, recall of 85 %, and F1 of 77.27 %.

As well as Suyash Dabhane et al. (2021) [20] who studied the impact of implementing algorithms individually and implementing ensemble learners. The models have been trained one by one and figured out how accurate they were. Some of these algorithms have been demonstrated to be effective on their own. However, the dilemma of overfitting was a common occurrence. As a result, to eliminate this issue of overfitting they trained their model using ensemble learning techniques and achieved an improved accuracy of roughly 87 %.

In another work, Adedeji (2019) [46] extracted 44,179 tweets through Twitter API. The author highlighted the importance of removing the features that would not contribute any importance to classifiers and the significance of applying essential text preprocessing steps. To obtain a balanced dataset the author combined the Twitter dataset with part of another public sentiment analysis to get more positive tweets. TFIDF, hashing, and N-gram methods were used for feature extraction. Seven classifiers were applied to detect depression (classification and regression trees, linear discriminant analysis, C5·O, regularized generalized linear model, adaptive boosting, extreme gradient boost, and random forest). The optimal result was a random forest algorithm with 0.83 f- measure, 0.88 recall, and 0.79 precision.

Furthermore, N. S. Alghamdi et al. (2020) [47] collected their data from Nafsany platform posts. They studied the impact of using four various feature extraction methods with six machine learning classification models. The TF-IDF technique, using SGD and either word-based or character-based models, had the highest accuracy rate, 73 %, according to the findings. After that, ADA and SVM come in second and third, respectively, with 72 % accuracy. When using BOW of characters as an NLP feature extraction strategy, ADA achieves this level of performance. When the TFIDF model was used for both words and characters, SVM performed the best. The researchers examined the effect of stemming on the difficulty of diagnosing depression. The results showed that stemming has a slight positive effect on the accuracy rate with TF-IDF and BOW.

On the other hand, there is another group of works that used an unsupervised learning approach where there is no labeled data [48–50]. For example, Yang et al. [50] surveyed 8063 Chinese middle and high school students. They proposed constructing depression classifications using an unsupervised machine-learning approach. The levels of depression were classified using K-means clustering. Furthermore [49] used supervised learning as mentioned in the previous section and unsupervised learning also. They generated and utilized the ArabDep lexicon to predict the depression symptoms from Nafsany Arabic-fetched posts.

7. **Big Data Analytics**: Angskun, Tipprasert, and Angskun (2022) [64] explored big data analytics for real-time depression detection on social networks, including twitter. While published in 2022, this work remains relevant, discussing methods to analyze social network data, such as tweets, for the detection of depression

After the reviewed articles on the depression detection domain, we found the following gaps:

- Many researchers neglect the importance of data balance and its impact on recall and precision rates since an imbalanced dataset leads to inconsistent accuracy, recall, and precision rates.
- Using corpora that contain negative words in the depression detection domain is a decisive issue. Negative words play a crucial part in altering the emotion from positive to negative or vice versa which in turn affects the model training which was not mentioned in any of the reviewed articles.
- There is a massive need for a public Arabic corpus in this domain. To our knowledge, until now there is no Arabic depression published corpus for the purpose of social media depression detection.

3. Methodology

The methodology outlined in this study for detecting depression in social media data, specifically Twitter, not only addresses existing gaps in depression detection but also integrates recent research trends to enhance the accuracy and effectiveness of the methodology [62].

The first step in the approach involves data collection from Twitter

using the Twitter API. This aligns with recent research that leverages large-scale social media data for mental health detection [65]. Twitter is chosen as a data source due to its widespread usage and accessibility for research purposes [65].

The next step involves text preprocessing techniques designed to handle the complexities of social media text data, including slang, abbreviations, and emoticons. These techniques ensure that the approach is capable of effectively capturing nuanced expressions related to depression, which has been identified as critical in the detection process [62].

The approach also incorporates feature extraction methods such as term frequency-inverse document frequency (TF-IDF) and Bag of Words (BOW), which serve as a strong foundation for establishing a performance baseline before delving into more complex approaches like deep learning and multimodal analysis [65].

In the realm of machine learning, the approach deploys supervised classifiers such as Support Vector Machine (SVM), Random Forest, Logistic Regression, and Light Gradient Boosting Machine (LightGBM), which are chosen based on recent research that indicates their continued relevance and effectiveness for text classification tasks [62,65]. Additionally, the approach adapts these models to specific modalities, further enhancing their performance for depression prediction [62].

The approach also emphasizes modality-specific adaptations, acknowledging the diversity in text data by accounting for linguistic nuances in English and Arabic. This ensures that the methodology remains up-to-date and adaptable to various languages and cultures [65].

Finally, a significant emphasis in the approach is on early detection, aligning with the growing recognition of the importance of early intervention in mental health [65]. By utilizing machine learning algorithms and modality-specific adaptations, the approach aims to provide timely and accurate predictions of depression severity, enabling proactive support and intervention for individuals in need.

Explaining our machine learning-based approach, the work proposed strategy for predicting depressed users from Arabic and English Twitter Tweets is listed in Fig. 1.

3.1. Data collection

Studies on the detection of various mental health disorders from publicly available social media content have focused on examining the linguistic variations between posts made by users who have a particular mental health disorder and posts made by a control group of users who are unaffected users in a variety of languages, including English, Norwegian, Chinese [18–20].

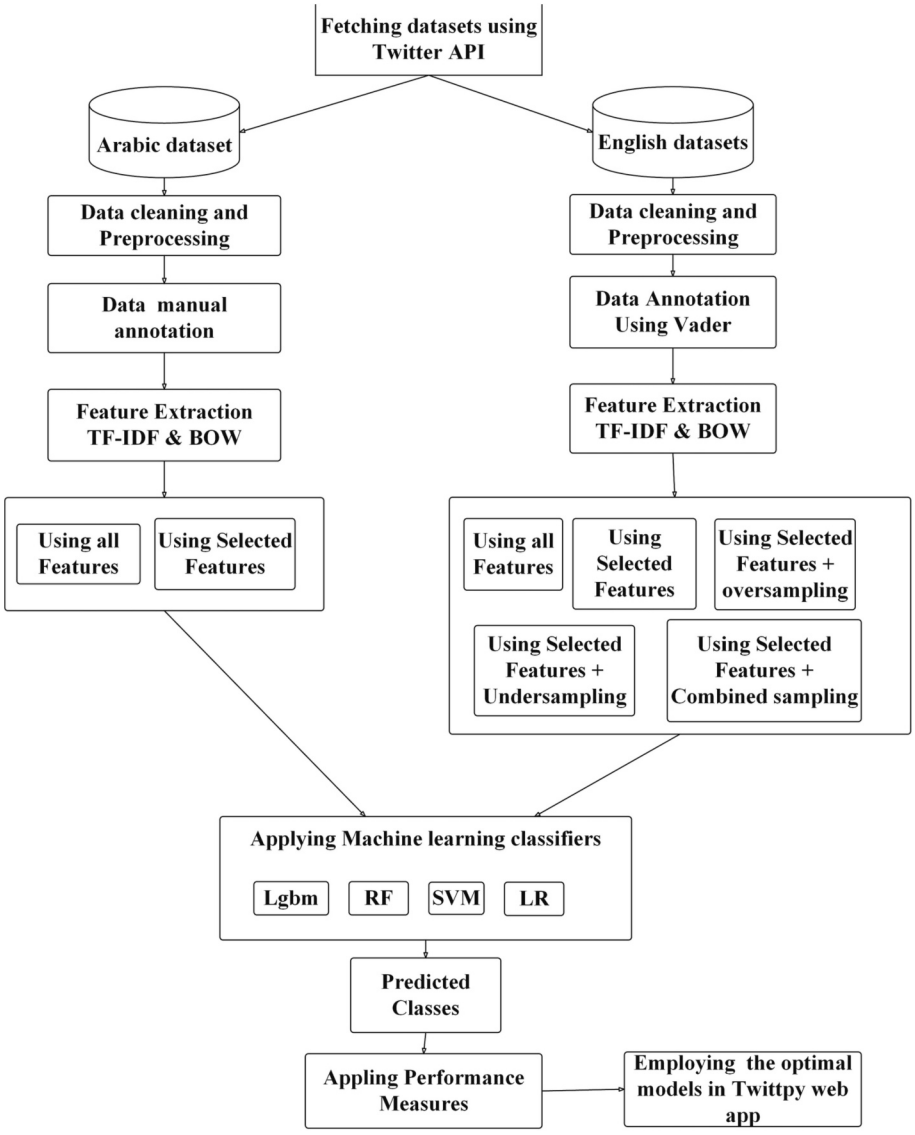


Fig. 1. Depression detection proposed methodology

To the highest of our knowledge, no public Twitter corpora for the analysis of depression were found at the time when we were looking for a depression dataset in either English or Arabic. Only the sentiment analysis corpora that the researchers used to classify depression were available. In this study, English and Arabic data were fetched to support the prediction of depression symptoms. To build a corpus for studying how the English and Arabic languages are used in cases of depression, we thoroughly explored a variety of publically accessible web platforms as our sources.

This paper considered Twitter as one of the many freely accessible online sources for our corpora since it is one of the most widely used social networking sites, with 353 million users with daily access reach to 187 million users [21]. It is worth mentioning that Twitter is the most preferred platform for sharing suicidal thoughts and some people write blogs on depression and anxiety [22].

3.1.1. Preparation of Arabic data set

The first gap we are concerned about bridging is introducing a public Arabic depression corpus. Our Arabic dataset Arabic_Dep_tweets_10,000 [54] consists of tweets fetched using the Twitter Application Programming Interface (API). Tweets posted between 1st January 2019 and 15th April 2022 have been collected. Tweets in our study were a mix of Modern Standard Arabic (MSA), Egyptian dialect and Gulf dialects. A total of 57,391 Tweets have been collected and divided into first, the data was obtained using a combination of words that express mental illness such as depression proclivity like "مكتئب","الكتئاب","انتحار","عايز أموت","موت","طبيب نفسي","قلق","حزين","حزن","اكابة" in a total of 29,326 tweets. And on the other hand combination of words that reflect happiness like "سعيدة","سعيد","فرح","فرحان","متفاءل","متحمس","حياة","حب","امل" in a total of 31,093 tweets. (Fig. 2) illustrate a word cloud that displays the most common words associated with different classes for our Arabic data set.

3.1.2. Preparation of English data set

As mentioned above in the Arabic section the two English corpora consist of tweets also collected using the Twitter Application Programming Interface (API). From 1 January to 30 July 2022, tweets were gathered. The utilization of two different corpora, one with negations and one without negations, serves a specific purpose in our research, shedding light on the importance of determining the scope and sequence of words influenced by negations. This choice significantly impacts our research in several ways and provides valuable insights into the field of sentiment analysis and depression detection.

1. **Scope of Negations**: The choice of using one corpus with negations and another without negations allows us to explore the impact of negations on sentiment analysis and depression detection. Negations, such as "not sad" or "no depression," can reverse the polarity of sentiments in text. By including these negations in the "Eng_negation_60.000" corpus, we aim to highlight how the presence or absence of negations can affect the accuracy of sentiment analysis and, consequently, depression detection.
2. **Linguistic Features**: Negations are essential linguistic features that influence the meaning of text. Understanding how different classifiers and feature extraction methods handle negations is crucial, as it can significantly affect the accuracy of results. Our research delves into this aspect, emphasizing the importance of robust algorithms that can appropriately deal with negations in text.
3. **Impact on Results**: By comparing the performance of classifiers on both corpora, we can assess the impact of negations on depression detection. This comparison allows us to draw conclusions about the necessity of handling negations effectively in sentiment analysis tasks.

3.1.2.1. DataSet I.

First corpus Eng_without_negation_60.000 [55,56] comprised of 60,172 tweets. It was gathered by combining keywords such as "depression", "depressed", "depress", "low self-esteem", "sad", "sadness", "cry", "suicidal", "suicide", "alone", "anxiety", "desperate", "despair", "feeling bad", "feeling down", "sorrow", "struggle", "pessimistic", "wanna die", "dying", "dead inside" and the happiness keywords like "happy", "happiness", "good", "glad", "optimistic", "hope", "hopeful", "enjoy", "cheerful", "well", "love", "life". The following figure (Fig. 3) demonstrates word cloud for Eng-Dep-60,000 dataset.

3.1.2.2. DataSet II.

In the second Eng_with_negation_57.000 [57,58] corpus negation has been taken into account to be added to our corpus. It consists of 30,000 from the first corpus and 27,392 was fetched by negation words like "not good", "not optimistic", "not glad", "not cheerful", "not happy", "unhappy", "no hope", "no energy", "unmotivated", "no motivation", "not sad", "not sorrow", "not suicidal", "not depressed", "not despair", "no depression", "no appetite", "not enjoying", "hopeless". (Fig. 4) is a word cloud for Eng_Dep_57,000 dataset.

3.2. Data pre-processing

3.2.1. Arabic data sets

Text cleaning is our first mission, this stage includes removing @ mentions, all Arabic and English punctuations [.?!,:;-[]{}() /' "], URLs [https://www….], # hash tags, non-characters, short words, meaningless words, repeating characters, symbols. Then emoji handling: all emojis have been converted into their meaning text. And Arabic stop word removal which does not add any meaning or value to the text and will affect the analysis was eliminated like [etc،أصبح، أن، أنت، أنا، هو، هي…هذا، هؤلاء] since they used in both depression and non-depression cases. In addition to Tashkeel removal [] and character Normalization also an important pre-processing process for example [أ إ آ to be ا] ,[ة to be ه] ,[ى to be ي]

Fig. 2. Arabic_Dep_tweets_10,000 dataset word cloud.

Fig. 3. Eng_without_negation_60.000 dataset word cloud.

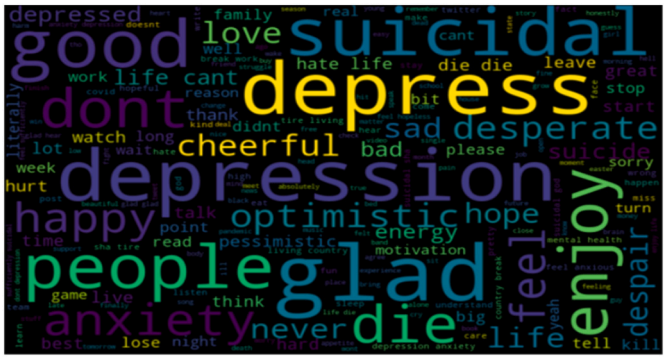


Fig. 4. Eng_with_negation_57.000 dataset word cloud.

be ى] So that the algorithm does not treat the words that contain diacritics or not normalized as another word. Next Tokenization converts the normal text strings into a list of tokens. Finally, the stemming to return the words to their original roots for example (,(عدس <– عيدس رحن <–ارحتان،رحتان)).

3.2.2. English data sets

As mentioned in the Arabic dataset our English text cleaning also includes: removing @ mentions, all punctuations [.?!,:;-[]{}()/' "], URLs [https://www….], # hash tags, non-characters, short words, meaningless words, repeating characters, symbols.

Additionally, the entire text has been changed to lowercase so that the algorithm does not perceive the same word in different cases as different.

Then, the Removal of stop words: Stop words are the most often used terms that are meant to be ignored because their use does not improve efficiency but instead has the potential to degrade it. Stop words include words like "he," "are," "she," etc. Finding a list of efficient stop words is a crucial task because stop words are not universal and depend on the case.

In addition one of the most crucial tasks in the field of natural language processing (NLP) is part-of-speech (POS) tagging. A word's POS tagging is influenced by its position, the words around it, and its POS tags in addition to the word itself [28]. Part-of-speech (POS) to comprehend a word's function within a phrase, its grammatical category must be assigned through the process of tagging Adverbs, conjunctions, nouns, verbs, and other conventional components of speech [25].

Finally, lemmatization is the process of assembling various inflected forms of a word, or lemma. It converts several words into one common root. A valid word is the result of lemmatization; ordinary suffix removal wouldn't have the same result as it is in stemming [25]. The function of lemmatization depends on the part of speech since some words have distinct meanings depending on that part of speech. For example, [am, is, are –> be], [playing, plays, plays–> play].

3.3. Data annotation and contextualization

The practice of marking data in various formats, such as text, photos, or video, so that computers can understand it is known as data annotation. Labeled datasets are essential for supervised machine learning since ML models need to comprehend input patterns to interpret them and generate reliable outputs [53].

3.3.1. Arabic_Dep_10,000 annotation

Each tweet is manually annotated into 1 for the "Depressed" class or 0 for the "Non depressed" class. After removing duplicated tweets, excluding tweets that are considered confusing and difficult to be labeled the total of tweets were 5000 depressed tweets and 4000 Non_dep tweets. From the 40,000-Egyptian-tweets corpus [23] 1000 tweets were utilized to convey neutrality to be considered as Non_dep tweets after reviewing their annotation in the context of depression detection to be the final corpus of 10,000 Arabic tweets.

3.3.2. English datasets annotation

After the corpus fetching, cleaning and preprocessing stages, the turn of the annotation phase comes. Valence Aware Dictionary and sEntiment Reasoner or VADER lexicon has been used for our two English corpora annotation to be binary labeled into "Depressed" and "Non_dep" and multi-labeled into "Depressed", "indifferent" and "happy". Vader is a lexicon and rule-based sentiment analysis tool attuned to sentiments expressed in social media and has been used previously by Razak et al. [17] in depression detection. The number of tweets for each class for binary and multi-classifications for both Eng-Dep-60,000 and Eng_with_negation_57.000 dataset is as mentioned in (Table 1), (Table 2), (Table 3)and (Table 4).

3.4. Features extraction

Feature extraction techniques are a necessary step for the text classification process. To characterize and describe the data, feature extraction is utilized to extract the most distinctive features from a dataset. Features are created to extract information that a machine learning system can understand and are essential for precise prediction [26,27]. The paper experimented with *Term Frequency–Inverse Document Frequency* Tf-Idf and Bag of Word BOW techniques.

3.4.1. Tf-Idf

Term frequency (TF) and inverse document frequency (IDF) are the two components that make up the TF-IDF scheme. A word's frequency in a document and its inverse document frequency over a group of documents are multiplied to achieve the words' weights [28].

3.4.2. BOW

A textual illustration of word recurrence in a document is called a "bag of words". it doesn't pay attention to grammatical conventions or word order; it only keeps track of word counts. It is referred to as a "bag" of words because any details regarding the arrangement or structure of the words within the document are ignored. The model doesn't care where in the document recognized terms appear; it is only interested in whether known words occur in the document [29].

The total number of features for our datasets was 13,950 for the Eng_without_negation_60.000 dataset, 12,647 features for the Eng_with_negation_57.000 dataset, and 22,012 features for the Arabic_Dep_10,000 dataset. After using all features that were extracted; the selectpercentile method which Selects features according to a percentile of the highest scores was used for feature selection to keep only 10 % of features.

3.5. Addressing data imbalance

As shown in (Table 1), (Table 2), (Table 3) and (Table 4) our two English datasets are slightly imbalanced. To tackle the class imbalance, the paper experimented with a variety of resampling techniques, including first, oversampling which creates additional instances for the minority class using the Synthetic Minority Oversampling Technique (SMOTE) method [40]. Second, undersampling which in turn randomly removes instances from the majority class using the OneSidedSelection

Table 1
Number of tweets per class for Multi-classification in Eng_without_negation_60.000 dataset.

Class	Number of Tweets
Depressed	17,149 Tweet
Indifferent	23,152 Tweet
Happy	19,871 Tweet

Table 2
Number of tweets per class for Binary-classification in Eng_without_negation_60.000 dataset.

Class	Number of Tweets
Depressed	21,802 Tweet
Non_dep	38,369 Tweet

Table 3
Number of tweets per class for Multi-classification in Eng_with_negation_57.000dataset.

Class	Number of Tweets
Depressed	24,662 Tweet
Indifferent	11,923 Tweet
Happy	20,806 Tweet

Table 4
Number of tweets per class for Binary-classification in Eng_with_negation_57.000 dataset.

Class	Number of Tweets
Depressed	25,520 Tweet
Non_dep	31,871 Tweet

method [41]. Finally, combined sampling which merges the two previous resampling methods using SMOTETomek [42] bridges the resampling gaps that we found in related works.

3.6. Utilizing supervised ML techniques

The Python 3.10.0 sci-kit-learn package has been used to build all the machine-learning models. Our proposed models are developed using text classification and sentiment analysis algorithms which are Light Gradient Boosting Machine, Random Forest, Support Vector Machine with both kernels linear and Rbf, and Logistic Regression on all datasets to determine the binary and multi-classifications. Our three datasets were used to train and test the ML models using hold out method with a ratio of 70:30 and using a 5-fold CV also with a ratio of 80:20 to avoid over-fitting. The selected ML algorithms are described in the next subsections.

3.6.1. Light gradient boosting machine (Lgbm) [30–32]

Lgbm is belongs to "Boost" ensemble learning techniques. Boosting is a sequential procedure in which each new model using a portion of the data tries to fix the flaws in the preceding model. The preceding model serves as a foundation for the following models. Overall outcomes are improved by the boosting method, which combines several weak learners to create a strong learner. It is based on the decision tree method. Runtime speed and accuracy are primarily optimized by Light gbm in two methods.

- It uses a histogram-based technique to divide continuously varying data into various buckets (rather than sorting them individually). This significantly enhances runtime.
- The level-wise tree development approach is substituted with the leaf-wise tree growth method (used by most other decision tree-based methods).

3.6.2. Random forest

The random forest algorithm is an extension of the bagging method which belongs to ensemble learning techniques. It makes an uncorrelated forest of decision trees using bagging and feature randomization and combines the results of various decision trees to get a single outcome [33].

3.6.3. Support vector machine (SVM)

SVM is a non-probabilistic classifier that is capable of determining the ideal border for every instance. Each post or document is represented by an SVM model as a vector in space. The spacing between the points and a hyperplane is then calculated [34]. SVM tries to increase the distances between the classes and the separating hyperplane. When a hyperplane that divides the two classes with the greatest distance to the closest data points has been found, the ideal separation has been achieved [35] so the larger the distance, the lower the error generated by the classifier [39]. This is a linear SVM. Unlike the linear kernel, which cannot handle the situation where the relationship between class labels and attributes is nonlinear, Radial Basis Function (Rbf) kernel nonlinearly maps samples into a higher dimensional space [36]. SVM is basically used with binary classification but the same method is applied for multiclass classification when the multi-classification problem is divided into various binary classification problems [38].

3.6.4. Logistic regression

The probability of a dependent categorical variable is predicted by logistic regression. The dependent's binary variables have yes/no codes. A large sample size is more effective for logistic regression. The logistic function is a sigmoid function that produces a number between zero and one for any real input x [39].

4. Experiments

After preprocessing and featuring steps had been applied, the work implemented the five ML classifiers. For the Arabic dataset, each classifier's results were noted based on (I) using all dataset features and (II) using feature selection. For our two English datasets, each classifier's results were noted based on (I) using all dataset features, (II) using feature selection, (III) oversampling, (IV) under-sampling, and (V) combined-sampling. All experiments were conducted using Tf-idf and BOW approaches, as well as using the holdout and 5-fold CV methods. The following subsections detail each dataset result. All classifiers were evaluated by three performance measures as shown below [15,43]:

1. Confusion Matrix: In the context of classification metrics, "P" typically represents Precision, and "R" represents Recall. Let me provide a clear explanation of these terms:
   - Precision (P): Precision is a metric that measures the accuracy of positive predictions made by a classification model. It calculates the ratio of true positive predictions (correctly predicted positive cases) to all positive predictions made by the model. In other words, precision tells us how many of the predicted positive cases were actually correct. The formula for precision is:

$$P = TP + FPTP$$

Where:

- TP (True Positives) is the number of correctly predicted positive cases.
- FP (False Positives) is the number of incorrectly predicted positive cases (negative cases that were predicted as positive).

Precision provides insight into the model's ability to avoid false positives. A high precision indicates that when the model predicts a positive case, it is highly likely to be correct.

- Recall (R): Recall is a metric that measures the ability of a classification model to identify all relevant instances of the positive class. It calculates the ratio of true positive predictions to all actual positive cases. In essence, recall tells us how many of the actual positive cases were correctly identified by the model. The formula for recall is:

$$R = TP + FNTP$$

Where:

- TP (True Positives) is the number of correctly predicted positive cases.
- FN (False Negatives) is the number of positive cases that were incorrectly predicted as negative.

Recall is particularly important in situations where missing a positive case (a false negative) is costly or has serious consequences. A high recall indicates that the model is effective at capturing most of the positive cases.

2. Accuracy: A model's accuracy is measured by how many correct predictions it has made overall compared to all other predictions. Only when the dataset is balanced can accuracy be used as an acceptable evaluation metric.

$$\text{Accuracy} = TP + TN/(TP + FP + FN + TN)$$

3. F1-Score: It is the harmony of recall and precision. It is, in other words, the harmonic mean of recall and precision.

$$F1 - \text{Score} = (2^{*}P^{*}R)/(P + R)$$

Another assessment metric that was added to the proposed method to better examine its effectiveness is the Receiver Operating Characteristic (ROC) chart. The true positive rate (TPR) versus the false positive rate (FPR) for various thresholds is plotted on the ROC curve. The optimum ROC chart has a larger area under the curve (AUC) [44].

This section is divided into first, *A.* English corpora experiments which contain the binary and multi results for each corpus. Second, *B.* Arabic corpus experiments results and, finally, *C.* The web application.

4.1. English corpora

Experiments on English corpora illustrated the performance of all classifiers and the effect of the absence or presence of negation. In addition, the effect of the count of classes to be classified on the classifiers' accuracy, precision, recall, F1 score, and AUC. The following subsections show the result of binary and multi-classifications.

Preliminary experiments were conducted first which include: ML classifiers being applied after selecting 10 % of the original features of the imbalanced two corpora by using the chi-square method. 10 % of the features have been used since they provide the highest results.

Second, data resampling experiments were conducted; the paper then applied selected features with data resampling techniques including three steps over-sampling using SMOTE, under-sampling using onesideselection, and SMOTE+ Tomek for combined-sampling to balance the original corpora.

4.1.1. Twets without negation

Experiments on Eng_without_negation_60.000 illustrated the performance of all classifiers with the absence of negation in this corpus. The following sub sections show the result of binary and multi classifications.

4.1.1.1. Binary classification. This work trained our proposed ML classifiers to distinguish between "depressed" and "Non_dep" classes on the original imbalanced status and after resampled classes. Table 5 illustrates the count of tweets for each class after all resampling techniques.

4.1.1.1.1. TF-IDF. Table 6: the average of the f1 score for all models using first, feature selection (imbalanced), second over-sampling, third under-sampling and finally combined-sampling with Tf-Idf. In the table below, the bold f1 scores refer to the highest ones among the classifiers.

Table 5
Eng_without_negation_60.000 tweets counts for each class after data resampling in the binary classification experiment.

Class		Over sampling	Under sampling	Combined sampling
Depressed	Tf-Idf	38,369	21,802	38,014
Non_dep		38,369	33,123	38,369
Depressed	BOW	38,369	21,802	38,369
Non_dep		38,369	36,167	38,277

Table 6
Eng_without_negation_60.000 F1-scores average of binary classification for all classifiers using Tf-Idf.

Classifier	Imbalanced corpus	Over sampling	Under sampling	Combined sampling
L gbm	86 %	89 %	86 %	89 %
RF	86 %	90 %	87 %	90 %
L-svm	87 %	90 %	88 %	90 %
Rbf -svm	**88 %**	**92 %**	**89 %**	**92 %**
LR	85 %	87 %	85 %	87 %

It can be seen from Table 6 that both over-sampling and combined sampling improved the F1 score of all models compared to preliminary results and under sampling. The rates of f1 score improvement in the over and combined sampling were 4 % with the RBF-SVM as the best model, whereas, An under-sampling archived slight improvement for only RF, L-SVM, and RBF-SVM with 1 %.

We found that the under-sampling technique OneSidedSelection (OSS) does not fully balance the dataset in this experiment as shown in Table 1. Since it eliminates many instances from the majority class, including redundant examples and ambiguous examples. It means that there are no redundant or ambiguous instances the corpus contains to remove to be balanced. This led us to investigate why over-sampling and combined-sampling are better than under-sampling.

In this experiment, Tf-Idf with RBF-SVM achieved the highest F1 score with 92 %.

We consider that recall is more crucial to depression detection problem. As a result, we strive for high recall, especially for depressed class. A false positive (FP) which represents precision is a user who is predicted to have depression but does not actually have it, according to the definition of the term in the context of depression detection. A user who is truly depressed but is predicted to not have depression is known as a false negative (FN) which represents recall.

Recall is important in this domain because early detection and identifying signs of depression in individuals is very helpful in curing the depression and save many costs either for individuals or society and even can saves a person's life. On the other hand, if a person is predicted as willing to have depression but he/she is not, will not have much effect.

Table 7 illustrates the confusion matrixes for RBF-SVM where it achieved the highest recall with Tf-Idf in this binary experiment with 94.7 % for the depressed class.

Fig. 5 displays a receiver operating characteristic (ROC) chart for a

Table 7
Confusion matrix of RBF-SVM using Tf-Idf and combined resampling of Eng_without_negation_60.000 in binary classification experiment.

		Actual		
		Depressed	Non_dep	
Predicted	Depressed	7202	739	*P=90.6%*
	Non_dep	399	6937	*P=94.5%*
		R= 94.7%	*R= 90.3%*	

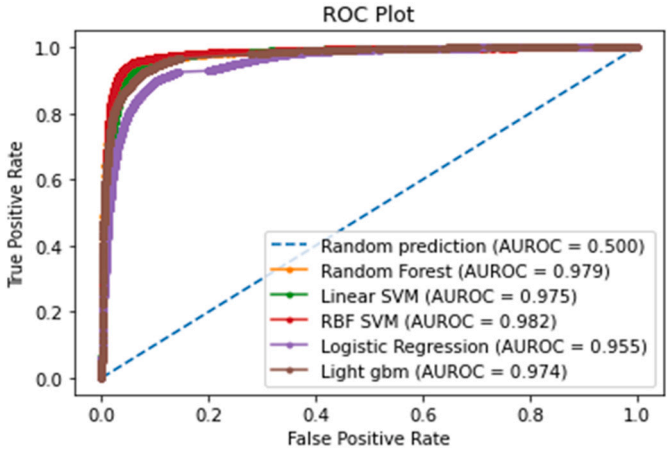


Fig. 5. All classifiers ROC curves and AUCs for Eng_without_negation_60.000 binary classification with combined resampling using Tf-Idf.

binary experiment with a combined-resampling technique as an additional evaluation metric to further examine the performance of the suggested approaches.

The ability of a classifier to differentiate between classes is measured by the Area Under the Curve (AUC), which is used as a summary of the ROC curve. The model performs better at differentiating between the depressed and Non_dep classes the higher the AUC. Since the AUCs for all classifiers in Fig. 1 are close to 1, the diagnostic test is perfect for differentiating between depressed and Non_dep users. In the experiments with Tf-Idf, RBF-SVM achieved the highest AUC with 0.982, While LR achieved the lower AUC with 0.955.

4.1.1.1.2. BOW. Table 8 illustrates the BOW feature extraction technique experiments with resampling techniques especially over and combined sampling which achieved a simple enhancement when compared with Tf-Idf. The classifier with the highest scores are indicated in bold in the table below.

The best classifiers with applying BOW were L-SVM and LR with 89 % F1-score.

Table 9 illustrates the confusion matrixes for L-SVM, where it achieved the highest recall with 89.1 % for the depressed class with BOW in this binary experiment.

A receiver operating characteristic (ROC) chart for binary experiments with a combined-resampling technique is displayed in Fig. 6. In the experiments with BOW, LR achieved 0.969 AUC which is the highest; while Lgbm is the lowest one with 0.963.

4.1.1.2. Multi-class classification. This work trained our proposed ML classifiers to distinguish between "depressed", "indifferent", and "happy" classes on the original imbalanced status and after resampled classes. Table 10 illustrates the count of tweets for each class after all resampling techniques.

4.1.1.2.1. TF-IDF. As Table 11 illustrates, the results of preliminary experiments slightly improved after both over and combined sampling. The F1 score of all models increased except LR and L-SVM while using Tf-Idf feature extraction technique. The bold f1 scores in the table below denote the classifier with the highest scores.

In the experiment with Tf-Idf, RBF-SVM achieved the highest F1 score of 88 %.

Table 12 illustrates the confusion matrixes of multi classification experiment with recall and precision of RBF-SVM with Tf-Idf which achieved the highest percentages of 91.2 % recall for the depressed class.

Table 8
Eng_without_negation_60.000 F1 score average of binary classification for all classifiers using BOW.

Classifier	Imbalanced corpus	Over sampling	Under sampling	Combined sampling
L gbm	86 %	87 %	86 %	87 %
RF	86 %	88 %	86 %	88 %
L-svm	**88 %**	**89 %**	**89 %**	**89 %**
Rbf-svm	87 %	88 %	88 %	88 %
LR	87 %	89 %	88 %	89 %

Table 9
Confusion matrix of Linear-SVM using BOW and combined resampling of Eng_without_negation_60.000 in binary classification experiment.

		Actual		
		Depressed	Non_dep	
Predicted	Depressed	6845	726	*P=90.4%*
	Non_dep	829	6948	*P=89.3%*
		R= 89.1%	*R= 90.5%*	

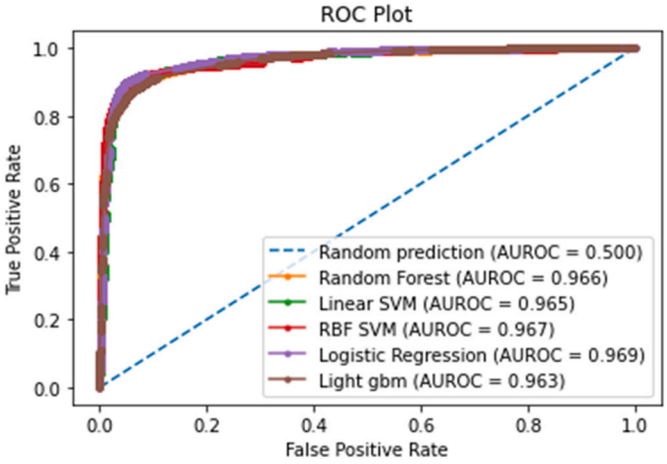


Fig. 6. ROC curves and AUCs for Eng_without_negation_60.000 binary classification with combined resampling using BOW.

Table 10
Eng_without_negation_60.000 tweets counts for each class after data resampling in the multi classification experiment.

Class		Over sampling	Under sampling	Combined sampling
Depressed	Tf-Idf	23,152	17,149	23,152
Indifferent		23,152	18,394	23,152
Happy		23,152	15,655	22,635
Depressed	BOW	23,152	17,149	23,152
Indifferent		23,152	21,087	23,152
Happy		23,152	14,877	23,060

Table 11
Eng_without_negation_60.000 F1-scores average of multi-classification for all classifiers using Tf-Idf.

Classifier	Imbalanced corpus	Over sampling	Under sampling	Combined sampling
Lgbm	82 %	82 %	80 %	82 %
RF	83 %	84 %	82 %	84 %
L- svm	**88 %**	**88 %**	**87 %**	**88 %**
Rbf-svm	84 %	85 %	84 %	85 %
LR	82 %	83 %	81 %	83 %

Table 12
Confusion matrix of RBF-SVM using Tf-Idf and combined resampling of Eng_without_negation_60.000 in multi classification experiment.

		Actual			
		Depressed	Indifferent	Happy	
Predicted	Depressed	4226	368	42	*P= 91.1%*
	Indifferent	355	3791	401	*P= 83%*
	Happy	49	371	4084	*P= 91%*
		R=91.2%	*R=84%*	*R=90.2%*	

Fig. 7: ROC curves for RBF-SVM that were applied with combined sampling and Tf-Idf in the multi-classification experiment.

In The following graph, the model has three curves representing each class 2 for the "depressed" class, 1 for the "indifferent" class and 0 for the "happy" class using the one vs. all approach. RBF-SVM achieved the highest AUCs compared to the rest classifiers with 0.98 for class 0, 0.95 for class 1 and 0.98 for class 2.

4.1.1.2.2. BOW. With the BOW feature extraction technique in the multi-classification experiment, resampling techniques especially over and combined sampling achieved a simple enhancement. Table 13: that over and combined sampling achieved a 1 % improvement with RF, RBF-SVM, and LR. While Lgbm and L-SVM didn't achieve any enhancement percentage. Under-sampling achieved a decrease for 4 models; 2 % with Lgbm and 1 % with RF, L-SVM, and LR while RBF-SVM didn't achieve any improvement with under-sampling. The best classifier with applying BOW was L-SVM with an 88 % F1 score. The bold f1 scores below refer to that the RBF SVM classifier achieved the best scores.

Table 14 illustrates the confusion matrixes of the multi-classification experiment with recall and precision of L-SVM with BOW which achieved the highest percentages of 87.4 % recall for depressed class.

Fig. 8: ROC curves for L-SVM that were applied with combined sampling and BOW. L-SVM was the best in differentiation between classes since it differentiated class 0 and class 2 with 0.97 but class 1 with 0.93 AUCs which is the highest.

4.1.2. *Twets with negation*

Eng_with_negation_57.000 experiments illustrated the performance of all classifiers with the presence of negation in this corpus. The following sub-sections show the result of binary and multi

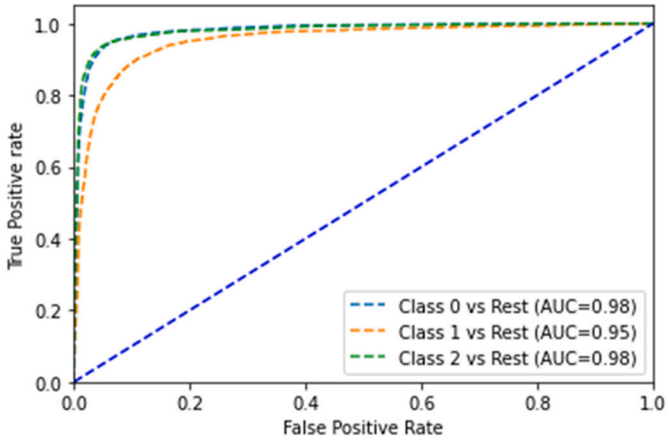


Fig. 7. ROC curves and AUCs for RBF-SVM classifier in multi classification experiment with combined resampling using Tf-Idf for Eng_without_negation_60.000 corpus.

Table 13
Eng_without_negation_60.000 F1-scores average of multi-classification for all classifiers using BOW.

Classifier	Imbalanced corpus	Over sampling	Under sampling	Combined sampling
Lgbm	82 %	83 %	81 %	83 %
R F	82 %	85 %	81 %	85 %
L-svm	85 %	85 %	84 %	85 %
Rbf -svm	**87 %**	**88 %**	**86 %**	**88 %**
LR	78 %	79 %	77 %	80 %

Table 14
Confusion matrix of L-SVM using BOW and combined resampling of Eng_without_negation_60.000 in multi classification experiment.

		Actual			
		Depressed	Indifferent	Happy	
Predicted	Depressed	4051	475	57	*p= 88.3%*
	Indifferent	493	3758	492	*P= 79.2%*
	Happy	87	397	4063	*P= 89.3%*
		R=87.4%	*R=81.1%*	*R=88.3%*	

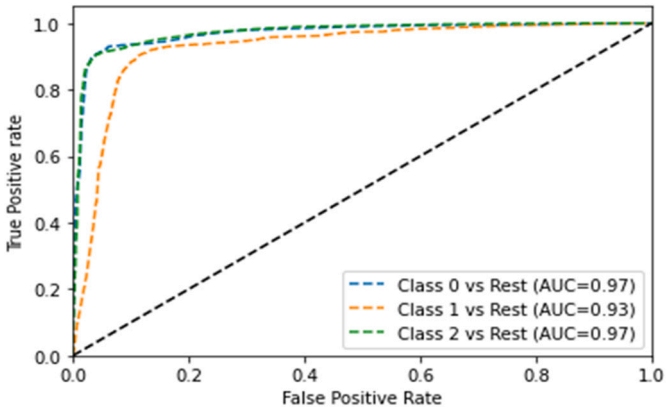


Fig. 8. ROC curves and AUCs for L-SVM classifier in multi classification experiment with combined resampling using BOW for Eng_without_negation_60.000 corpus.

classifications.

4.1.2.1. Binary classification. This paper used the initial imbalanced status and the resampled classes to train our proposed ML classifiers to distinguish between the two classes. Table 15 illustrates the count of tweets for each class after all resampling techniques.

4.1.2.1.1. TF-IDF. It can be seen in Table 16, that both over-sampling and combined sampling improved the preliminary results with Tf-Idf. Whereas, under-sampling didn't achieve any improvement

Table 15
Eng_with_negation_57.000 tweets counts for each class after data resampling in the binary classification experiment.

Class		Over sampling	Under sampling	Combined sampling
Depressed	Tf-Idf	31,871	25,520	31,871
Non_dep		31,871	25,924	30,968
Depressed	BOW	31,871	25,520	31,871
Non_dep		31,871	28,591	31,803

Table 16
Eng_with_negation_57.000 Accuracy average and F1 average score of binary classification for all classifiers using Tf-Idf.

Classifier	Imbalanced	Over sampling	Under sampling	Combined sampling
Lgbm	80 %	83 %	80 %	83 %
R F	**82 %**	**86 %**	**82 %**	**87 %**
L-svm	74 %	78 %	73 %	79 %
Rbf -svm	82 %	85 %	82 %	86 %
LR	74 %	79 %	74 %	79 %

for all classifiers. In this experiment, RF achieved the highest F1 score with an 87.9 % but L-SVM and LR were the lowest with a 79 % F1 score according to combined sampling results. The RF classifier obtained the highest scores in this experiment, as indicated by the bolded f1 scores below.

Table 17 illustrates the confusion matrixes for RF which achieved the highest recall of 91.1 % for depressed class.

Fig. 9 illustrates that the AUCs for all classifiers are close to 1 but the AUCs are less than all classifiers AUCs of Eng_without_negation_60.000. This indicates the influence of the classifiers' performance by the presence of negation in the corpus. In the experiments with Tf-Idf, RF achieved the highest AUC with 0.953 which means that it is the best at distinguishing between classes with the presence of negation in the corpus; While L-SVM achieved the lower AUC with 0.88.

4.1.2.1.2. BOW. It is clear from Table 18 that Over and combined sampling achieved 4 % improvement with Lgbm, 6 % with L-SVM and LR, and 2 % with RBF-SVM but FR achieved 3 % with over-sampling and 2 % with combined. Under-sampling achieves improvement for 3 models; 1 % with L-SVM, RBF-SVM and 2 % with Lgbm. The best classifiers with applying BOW were RBF-SVM and RF with 86 % F1 scores according to over sampling results. The bolded f1 scores below indicate the highest scores that have been achieved by the RBF-SVM classifier.

And, Table 19 illustrates the confusion matrixes for RBF-SVM which achieved the highest recall with BOW by 87.8 % for the depressed class.

In the experiments with BOW, RF achieved 0.94 AUC which is the highest; while L-SVM is the lowest one with 0.862 as shown in Fig. 10. Overall, it can be noticed that all classifiers' performance especially SVM with its two kernels was influenced by negation presence in the corpus since they achieved high AUCs with Eng_without_negation_60.000 corpus which does not contain any negation.

4.1.2.2. Multi-class classification. Preliminary experiments were applied to the original imbalanced status. In addition to resampling three classes. Table 20 illustrates the count of tweets for the three classes after all resampling techniques.

4.1.2.2.1. TF-IDF. As illustrated in Table 21, both over and combined sampling achieves the F1 score improvement of all models while using Tf-Idf feature extraction technique. The over-sampling enhancements were limited, while the combined sampling achieved better enhancement in this experiment. The rates were 2 % with L-gbm and RF, 4 % with L-SVM, and 3 % with LR. Whereas under-sampling has not archived any improvement for all classifiers; On the contrary, it decreased the F1 score by 5 % with L-gbm and RF and 2 % with L-SVM and LR except RBF-SVM did not achieve any decreases or increases. In the experiment with Tf-Idf, RF achieved the highest F1 score of 85 %

Table 17
Confusion matrix of RF using Tf-Idf and combined resampling of Eng_with_negation_57.000 binary classification experiment

		Actual		
		Depressed	Non_dep	
Predicted	Depressed	5812	1018	*P=85%*
	Non_dep	561	5076	*P=90%*
		R= 91.1%	*R= 83.2%*	

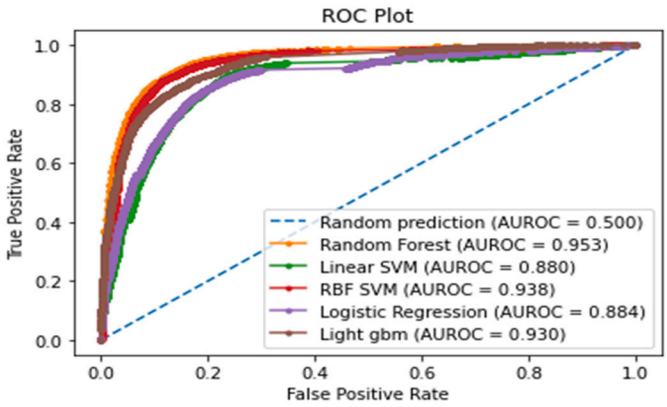


Fig. 9. All classifiers ROC curves and AUCs for Eng_with_negation_57.000 binary classification with combined resampling using Tf-Idf.

Table 18
Eng_with_negation_57.000 F1-scores average of binary classification for all classifiers using BOW.

Classifier	Imbalanced	Over sampling	Under sampling	Combined sampling
Lgbm	79 %	83 %	81 %	83 %
RF	83 %	86 %	83 %	85 %
L-svm	73 %	79 %	72 %	79 %
Rbf- svm	**84 %**	**86 %**	**83 %**	**86 %**
LR	73 %	79 %	73 %	79 %

Table 19
Confusion matrix of RBF-SVM using BOW and combined sampling of Eng_with_negation_57.000 binary classification experiment.

		Actual		
		Depressed	Non-dep	
Predicted	Depressed	5624	1002	*P=84.8%*
	Non_dep	781	5359	*P=87.2%*
		R= 87.8%	*R= 84.2%*	

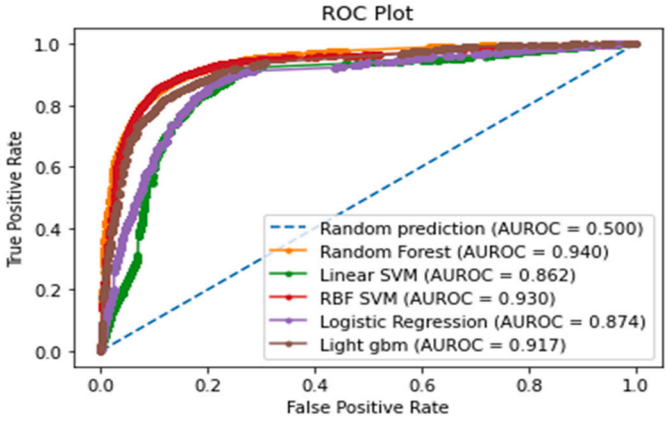


Fig. 10. All classifiers ROC curves and AUCs for Eng_with_negation_57.000 binary classification with combined resampling using BOW.

Table 20
Eng_with_negation_57.000 tweets counts for each class after data resampling in the multi classification experiment.

Class		Over sampling	Under sampling	Combined sampling
Depressed	Tf-Idf	24,662	17,669	24,662
Indifferent		24,662	11,923	24,662
Happy		24,662	14,128	23,884
Depressed	BOW	24,662	22,559	24,662
Indifferent		24,662	16,571	24,662
Happy		24,662	11,923	24,585

Table 21
Eng_with_negation_57.000 F1-scores averages of multi-classification for all classifiers using Tf-Idf.

Classifier	Imbalanced corpus	Over sampling	Under sampling	Combined sampling
Lgbm	77 %	77 %	72 %	79 %
R F	**83 %**	**83 %**	**78 %**	**85 %**
L-svm	73 %	73 %	71 %	77 %
Rbf- svm	77 %	77 %	77 %	84 %
LR	70 %	70 %	68 %	73 %

with combined resampling. The top scores attained by the RF classifier are indicated by the bolded f1 scores below.

Table 22 illustrates the confusion matrixes with recall and precision for RF which achieved the highest percentages with Tf-Idf by 80.3 %.

Fig. 11: ROC curves and AUCs for RF that applied with combined sampling and Tf-Idf Since it achieved the highest AUC for all classes with 0.96 for class 0, 0.98 for class 1 and 0.96 for class 2.

4.1.2.2.2. BOW. When the BOW feature extraction technique has been used with resampling techniques especially over and combined sampling achieved remarkable enhancement. The best classifiers with applying BOW are bolded in the Table 23. RF and RBF-SVM achieved 84 % F1 score with over and combined sampling as shown below.

Although RBF-SVM and RF outperformed the rest of the classifiers with the same f-scores, RBF-SVM achieved 81.2 % recall while RF achieved 77.3 % for the depressed class according to Table 24. So RBF-SVM exceeded the performance of the RF classifier in distinguishing the depressed class.

Fig. 12: ROC curves for RF that was applied with combined sampling and BOW. It achieved the best performance in differentiating between classes among the other classifiers.

4.2. Arabic corpus

The experiments of the Arabic_Dep_tweets_10,000 corpus are a binary experiment that includes running all ML classifiers with the entire extracted features and with 10 % of the original features. Our Arabic corpus is balanced thus no need for applying data resampling techniques.

Table 22
Confusion matrix of RF using Tf-Idf and combined resampling of Eng_with_negation_57.000 in multi classification experiment.

		Actual			
		Depressed	Indifferent	Happy	
Predicted	Depressed	3962	127	661	*P=83.4 %*
	Indifferent	278	4654	215	*P=90.4 %*
	Happy	693	171	3884	*P=81.8 %*
		R=80.3%	*R=93.9%*	*R=81.5%*	

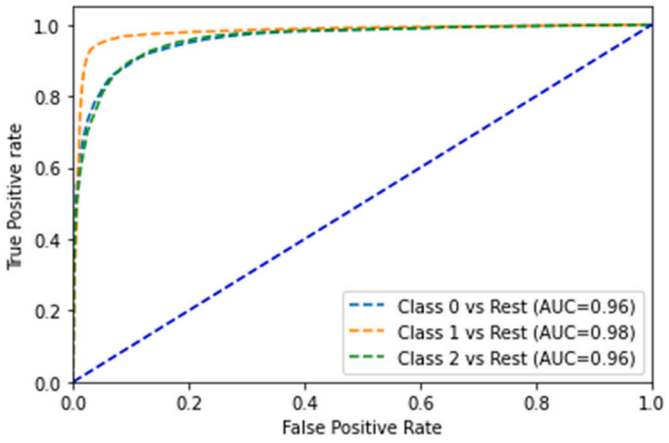


Fig. 11. ROC curves and AUCs for RF classifier with combined sampling and Tf-Idf for multi classification of Eng_with_negation_57.000 corpus.

Table 23
Eng_with_negation_57.000 F1- scores averages of multi classification for all classifiers using BOW.

Classifier	Imbalanced corpus	Over sampling	Under sampling	Combined sampling
Lgbm	76 %	79 %	75 %	79 %
RF	**81 %**	**84 %**	**79 %**	**84 %**
L-svm	74 %	76 %	72 %	77 %
Rbf -svm	**81 %**	**84 %**	**79 %**	**84 %**
LR	71 %	74 %	69 %	74 %

Table 24
Confusion matrix of RBF-SVM using BOW and combined resampling of Eng_with_negation_57.000 in multi classification experiment.

classification experiment

		Actual			
		Depressed	Indifferent	Happy	
Predicted	Depressed	4005	193	792	*P=80.2 %*
	Indifferent	138	4543	211	*P=92.8 %*
	Happy	789	197	3914	*P=79.8 %*
		R=81.2%	*R=92%*	*R=79.6%*	

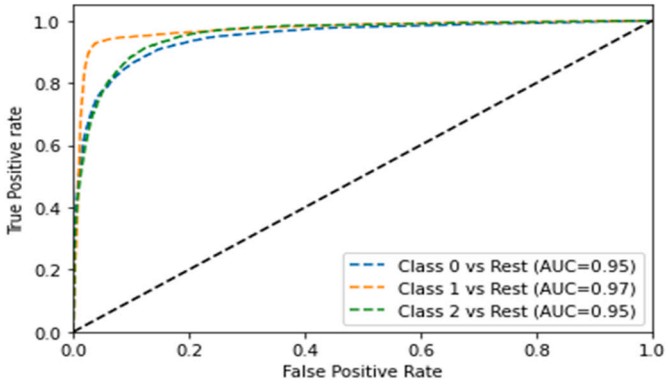


Fig. 12. ROC curves and AUCs for RF with combined sampling and BOW for multi classification of Eng_with_negation_57.000.

4.2.1. TF-IDF

As shown in Table 25, with Tf-idf, the experiment with feature selection enhanced Lgbm's f1 score by 0.1 % and 0.4 % with RF while slightly reducing L-SVM, RBF-SVM, and LR's f1-score by 0.2 %, 0.1 %, 0.1 % respectively. RBF-SVM fulfilled the highest f1 score by using all features and Tf-Idf by 96.6 %.

Table 26 illustrates the confusion matrixes for RBF-SVM which achieved the highest recall for depressed class by 95 %.

Fig. 13 illustrates that the AUCs for all classifiers are almost equal to 1. This indicates that all models accurately distinguished between classes. RBF-SVM with Tf-Idf fulfilled 0.996 AUC which was the highest.

4.2.2. BOW

As cleared from Table 27, with BOW experiments feature selection enhanced RF and L-SVM f1-score by 0.4 % and 0.2 % while reducing the others by 0.1 %. RBF-SVM fulfilled the highest f1 score by using all features and Tf-Idf by 96.6 % but LR was the best with BOW. The LR classifier has produced the highest results, which are indicated by the bolded f1 scores below.

Table 28 illustrates the confusion matrixes for LR which achieved the highest recall for depressed class by 94.9 %. Fig. 14 illustrates that the AUCs for all classifiers with BOW were almost equal to 1. This indicates that all models accurately distinguish between classes. RF with BOW fulfilled 0.993 AUC which is the optimal.

5. Analysis

In this section, we discuss the experimental results obtained from our research, which aimed to evaluate the performance of various machine learning classifiers on different datasets and feature representations. The experiments were conducted on English and Arabic datasets, with a focus on binary and multi-class classification tasks. Additionally, we explored the impact of various resampling techniques and feature selection on the classifiers' performance. We also analyzed the Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) as an additional evaluation metric.

5.1. English corpora experiments

5.1.1. Binary classification with Eng_without_negation_60.000

In the binary classification experiments on the Eng_without_negation_60.000 corpus, we observed that over-sampling and combined sampling techniques significantly improved the F1 scores of all models when using Tf-Idf feature representation. This indicates that balancing the dataset through these techniques effectively enhances the classifiers' ability to distinguish between “depressed” and “Non_dep” classes.

Furthermore, we found that the RBF-SVM classifier achieved the highest F1 score of 92 % with Tf-Idf. Recall, a crucial metric in depression detection, was particularly important, as early detection of depression is crucial. The RBF-SVM also exhibited the highest recall of 94.7 % for the “depressed” class. The ROC analysis further demonstrated the effectiveness of the classifiers, with RBF-SVM achieving the highest AUC of 0.982.

However, when using the Bag of Words (BOW) representation, the improvement in F1 scores was less pronounced, with L-SVM and LR achieving the highest F1 scores of 89 %. The ROC analysis revealed that LR achieved the highest AUC of 0.969 in this case.

Table 25
All ML classifiers' f1-scores averages of Arabic_Dep_tweets_10,000 experiments with Tf-Idf.

Classifier	All features	Feature selection
Lgbm	96.1 %	96.2 %
RF	95.6 %	96 %
L- svm	96.4 %	96.2 %
Rbf -svm	**96.6 %**	**96.5 %**
LR	96.3 %	96.2 %

Table 26
Confusion matrix of RBF-SVM using Tf-Idf of Arabic_Dep_tweets_10,000 using all features.

		Actual		
		Depressed	Non_dep	
Predicted	Depressed	1485	25	*P=98.3%*
	Non_dep	78	1411	*P=94.7%*
		R= 95%	*R= 98.2%*	

5.1.2. Multi-class classification with Eng_with_negation_60.000

In the multi-class classification experiments on the Eng_without_negation_60.000 corpus, we noted that over-sampling and combined sampling techniques improved F1 scores for most models when using Tf-Idf feature extraction technique. Again, the RBF-SVM classifier performed well, achieving the highest F1 score of 88 %. The ROC analysis showed that RBF-SVM had the best ability to differentiate between classes, with AUC values of 0.98, 0.95, and 0.98 for the three classes.

With BOW representation, similar improvements were observed, and L-SVM and LR achieved the highest F1 scores of 88 %. The ROC analysis confirmed that LR achieved the highest AUC of 0.969.

5.1.3. Binary classification with Eng_with_negation_57.000

In the binary classification experiments on the Eng_with_negation_57.000 corpus, we observed that over-sampling and combined sampling techniques significantly improved F1 scores for most models when using both Tf-Idf and BOW representations. RF achieved the highest F1 score of 87.9 % with Tf-Idf, while RBF-SVM and RF achieved the highest F1 score of 84 % with BOW.

The ROC analysis showed that all classifiers performed well in distinguishing between classes, with AUC values close to 1. However, RF achieved the highest AUC of 0.953 with Tf-Idf, indicating its effectiveness in the presence of negation in the corpus.

5.1.4. Multi-class classification with Eng_with_negation_57.000

In the multi-class classification experiments on the Eng_with_negation_57.000 corpus, we observed that over- sampling and combined sampling techniques significantly improved F1 scores for most models when using both Tf-Idf and BOW representations. RBF-SVM and RF achieved the highest F1 scores of 84 % with BOW.

The ROC analysis confirmed that all classifiers effectively differentiated between classes, with AUC values close to 1. RF achieved the highest AUC of 0.96 for class 0, 0.98 for class 1, and 0.96 for class 2 with Tf-Idf.

5.2. Arabic corpus experiments (Arabic_Dep_tweets_10,000)

In the experiments on the Arabic corpus, which was already balanced, we found that feature selection had a minor impact on F1 scores, with RBF-SVM achieving the highest F1 score of 96.6 % using Tf-Idf feature extraction technique. The ROC analysis demonstrated excellent performance, with RBF-SVM achieving an AUC of 0.996.

Similarly, with BOW representation, feature selection had a minor impact on F1 scores, and RBF-SVM again achieved the highest F1 score of 96.6 %. The ROC analysis showed that all classifiers effectively distinguished between classes, with RF achieving the highest AUC of 0.993.

In conclusion, our experiments demonstrated that the choice of feature representation and resampling techniques can significantly impact the performance of machine learning classifiers in depression detection tasks. RBF-SVM and RF consistently performed well across

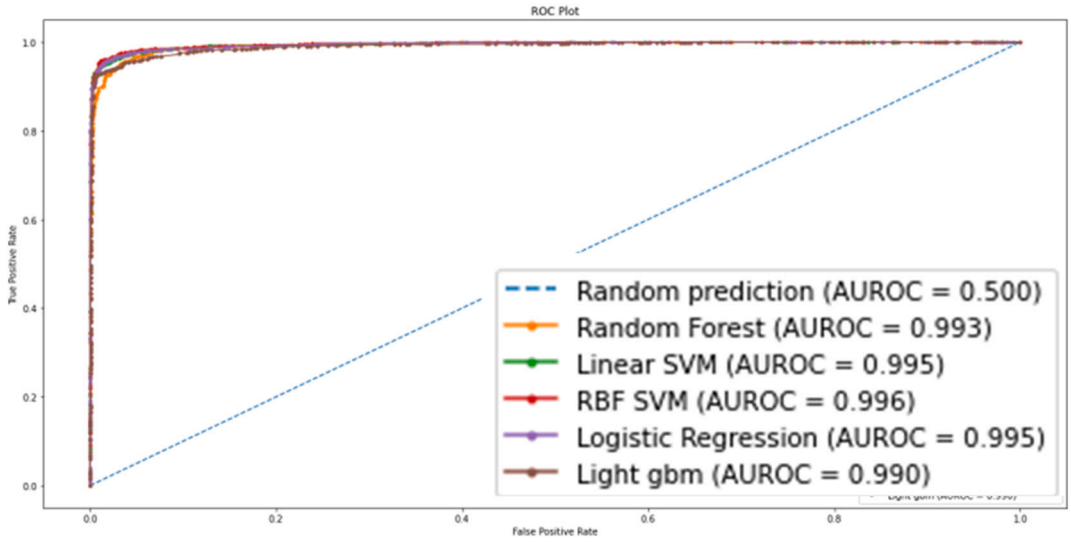


Fig. 13. ROC curves and AUCs for *Arabic_Dep_tweets_10,000* using Tf-Idf with all features.

Table 27
All ML classifiers' f1-scores average of Arabic_Dep_tweets_10,000 experiments with BOW.

Classifier	All features	Feature selection
Lgbm	96.3 %	96.2 %
RF	95.7 %	96.1 %
L-svm	95.9 %	96.1 %
Rbf -svm	96.2 %	96.1 %
LR	**96.4 %**	**96.3 %**

Table 28
Confusion matrix of LR using BOW of Arabic_Dep_tweets_10,000 using all features.

		Actual		
		Depressed	Non_dep	
Predicted	Depressed	1499	28	*P=98.1%*
	Non_dep	79	1396	*P=94.6%*
		R= 94.9%	*R= 98%*	

different datasets and representations, making them strong candidates for depression detection applications. The results also highlighted the importance of considering recall, especially for the "depressed" class, in depression detection, as early detection is crucial for effective intervention and support. Table 29 provides a comparative analysis with more recent works in the field to provide a broader perspective. The table includes a summary of various studies related to depression detection on social media platforms, including whether they provided a new dataset, the languages used, whether they handled negation, and their respective F1 scores. Our work, which includes a new dataset, utilizes both English and Arabic languages, handles negation and achieves an F1 score of 87 % and 96.6 %, highlighting its contribution to the field of depression detection on social media.

6. Application

Twittpy is the system that is developed in our present work. The main goal of developing this app is to test our best models on real live tweets that are new for them. It enables us to first, test the model by checking the status of the specific Twitter user that its username has been entered. Second, an individual tweet can be checked. These two features can be done in Arabic or English language.

For English prediction, the system uses random forest with the Tf-Idf model which is trained on Eng_with_negation_57.000 since this model achieved the highest scores and is well-trained on negation words. Although RBF-SVM outperforms other classifiers for Arabic prediction, Random Forest with Tf-Idf was employed because it performed well in manual testing with new tweets whereas RBF-SVM did not.

Figs. 15 and 16 showed English and Arabic depression detection pages that allow us to predict live depressed tweets of specific Twitter users by entering his/her username and the count of tweets that are required to be examined. In addition, enables us to enter individual tweets to check.

In live tweet prediction, after entering the username and the count of tweets click "predict" on the English page or "اكشف" on the Arabic page. The models will run, and the tweets will be displayed in a table with their classification result and tweet time. Additionally, a pie chart displaying the percentage of each class as well as the percentage of depression will be included.

7. Discussion

Recent research findings have indicated a growing interest in deep learning and multimodal analysis for improving the accuracy of depression detection [9]. While these approaches show promise, they also pose challenges such as data acquisition and model complexity. Our study, on the other hand, continues to leverage machine learning techniques due to their proven effectiveness and practicality in large-scale data analysis [10].

One of the identified gaps in the literature is the lack of sufficient Arabic corpora for depression detection. To address this gap, we introduced a manually labeled Arabic depression corpus of 10,000 tweets. This not only enhances the diversity of data sources but also makes strides in overcoming language-specific limitations in the field [5].

Additionally, we examined various text preprocessing techniques, including term frequency-inverse document frequency (TF-IDF) and Bag of Words (BOW), as well as modeling techniques such as Support Vector Machine (SVM), Random Forest, Logistic Regression, and Light Gradient Boosting Machine classifiers.

Our approach aligns with recent trends by offering a comprehensive analysis of these techniques, allowing us to identify the most effective combination of preprocessing and modeling for improved prediction accuracy [11].

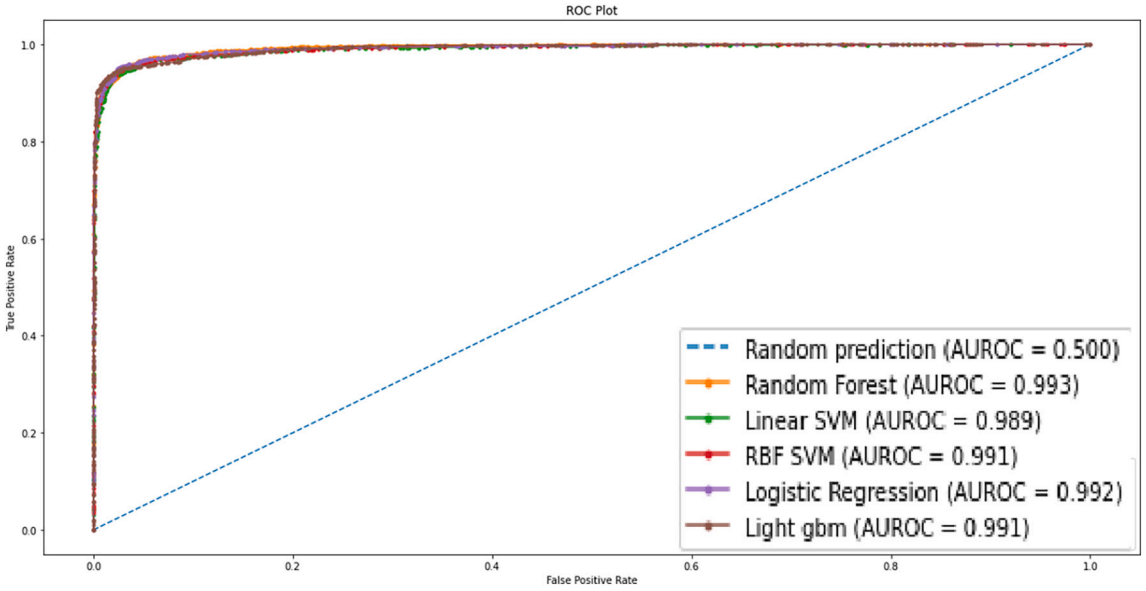


Fig. 14. ROC curves and AUCs for Arabic_Dep_tweets_10,000 using BOW with all features

Table 29
comparison of our work with the previous works.

Ref	Providing new dataset	Used language	Negation handling	F1 score
Zunaira Jamil et al. (2017) [45]	No	English	–	78 %
Orabi et al. (2018) [13]	No	English	No	87.957 %
Salma Almouzini et al. (2019) [12]	No	Arabic	Yes	87 %
Hemanthkumar M et al. (2019) [15]	No	English	–	75 %
Adedeji (2019) [46]	No	English	–	83 % F1
Razak et al. (2020) [17]	No	English	No	–
N. S. Alghamdi et al. (2020) [47]	No	Arabic	–	73 %
F. M. Shah et al. (2020) [52]	No	English	–	81 % F1
Mustafa R.U et al. (2020) [51]	No	English	No	91 %
Suyash Dabhane et al. (2021) [20]	No	English	–	87 %
Prof. S. J. Pachouly et al. (2021) [16]	No	English	–	87 %
Raymond Chiong et al. (2021) [14]	No	English	–	92 %
Safa, Bayat, and Moghtader (2021) [6]	No	English	–	91 %
Kour, H., & Gupta, M. K. (2022) [5].	No	English	–	94.2 %
Cha, Kim, and Park (2022) [8]	Yes	Korean	No	99 %
Khafaga et al. (2023) [62]	No	English	–	99.6 %
Our work	Yes	English & Arabic	Yes	87 % 96.6 %

- **Comparison with Recent Works:** In our study, we acknowledge the importance of benchmarking our methodology against more recent works in the field. We recognize that failing to compare our approach with these recent advancements may limit the comprehensive evaluation of our model's performance and its relevance in the current landscape of sentiment analysis and depression detection. Future research efforts should strive to include such comparative analyses to gain a deeper understanding of the state-of-the-art techniques.
- **Advanced Sentiment Analysis:** While our study primarily focused on linguistic and content-based features, integrating advanced sentiment analysis techniques could provide valuable insights into the polarity of tweets and further enhance the accuracy of depression

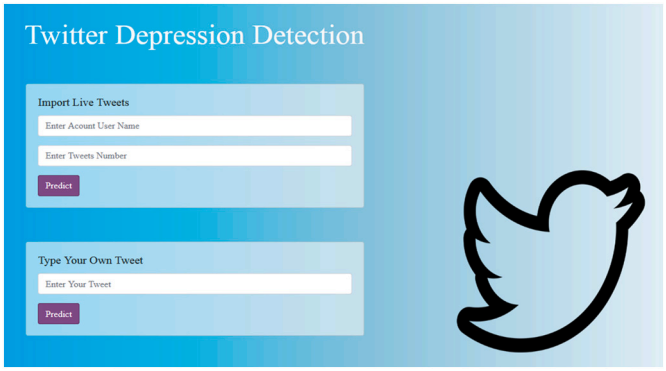


Fig. 15. English depression detection.

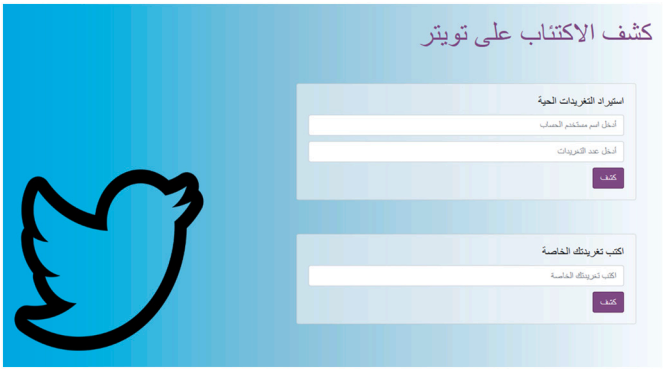


Fig. 16. Arabic depression detection page.

detection models. The value of sentiment analysis in understanding the emotional tone and polarity of tweets, which can provide additional insights into the mental state of individuals. The techniques can include subjectivity & tone, context & polarity, irony & sarcasm, comparisons, and emojis. Future iterations of our methodology should consider the inclusion of sentiment analysis as an additional feature.

- **Diversifying Feature Extraction Techniques:** Expanding the feature extraction mechanisms beyond Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (Tf-IDF) is another avenue for methodological improvement. Diversifying feature extraction techniques can potentially capture more nuanced linguistic patterns and contribute to more robust depression detection models. While our study focuses on using machine learning techniques, incorporating techniques like word embeddings or deep learning-based approaches may offer promising avenues for exploration.
- **Connecting with Recent Literature:** We acknowledge the importance of situating our research within the broader context of recent developments in mental disorder detection. Our study primarily focuses on a specific approach, and we acknowledge that it may not comprehensively cover the evolving landscape of depression detection. To bridge this gap, future research should seek to establish stronger connections with recent literature, especially those related to attentive relation networks and ensemble hybrid learning methods for automated depression detection.

By addressing these considerations in future research endeavors, we can advance the field of depression detection and contribute to the development of more accurate and comprehensive models.

8. Research limitations

The following limitations should be considered when interpreting the results of our study and in the design of future research in this area.

- **Inadequate Arabic Corpora**: Our study's findings are constrained by the limited availability of Arabic corpora suitable for sentiment analysis and depression detection. This limitation impacts the depth and breadth of our analysis, restricting the generalizability of our results to a broader Arabic-speaking population.
- **Limited Representation of Depressive Users**: We primarily relied on Twitter users who openly discussed experiences and emotions that may be connected to depression. This method of identifying users with depressive tendencies is indirect and potentially unreliable. There is no way to ascertain the accuracy of self-disclosures or whether these individuals may have other mental health conditions without professional evaluations. Furthermore, Twitter users may use terms or expressions not included in our list of keywords, potentially leading to the omission of relevant data.
- **Possibility of Misdiagnosis**: Our study's reliance on self-disclosed information from Twitter users may not align with clinical diagnoses of depression. It is crucial to acknowledge that self-disclosed mental health information on social media platforms may not always reflect accurate diagnoses, potentially impacting the precision of our findings.
- **Limited Keyword List**: The keyword list we employed for identifying depression-related content on Twitter may not encompass all expressions and phrases associated with depression. Consequently, there is a possibility that relevant data were excluded, which could affect the comprehensiveness of our analysis.
- **Generalizability**: The findings of our study may not extend seamlessly to other social media platforms or languages. Characteristics, behaviors, and linguistic expressions may differ across platforms and cultures, influencing the applicability of our models and results.

Model Generalization: While our machine learning models have shown promising performance within the scope of this study, their effectiveness in different contexts or with distinct datasets may vary. Generalization and model validation in diverse scenarios should be undertaken for more robust conclusions.

9. Conclusion

In this study, we have presented a comprehensive analysis of depression detection using machine learning techniques applied to Twitter data. Our research has made significant contributions to the existing body of knowledge in several ways.

First, we addressed the critical need for diverse and language-specific depression corpora by introducing a manually labeled Arabic depression corpus and two new automatically labeled English depression corpora. These resources have the potential to enhance depression detection research in Arabic and English languages.

Second, we conducted extensive experiments to evaluate various text preprocessing techniques, feature extraction methods, and supervised classifiers. Our findings shed light on the effectiveness of different combinations of these techniques in accurately predicting depression severity, thus providing valuable insights for future researchers and practitioners in the field.

Third, our study demonstrates the utility of machine learning in mental health prediction, despite the growing interest in deep learning and multimodal analysis. While recent trends suggest promising avenues for improving accuracy, machine learning models remain a viable and interpretable approach for early depression detection.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Brown C, Fields B, Matthysee T. Advances in mental health in South Africa. Lancet Psychiatry 2019;6(3):177–8.
[2] Cha J, Kim S, Park E. A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. Artif Intell Med 2022; 127:104325. https://doi.org/10.1016/j.artmed.2022.104325.
[3] Chen L, Wang Y. Big data analytics on social networks for real-time depression detection. Front Psychol 2021;12:688382. https://doi.org/10.3389/fpsyt.2021.688382.
[4] Johnson RB, Jones PQ, Smith AM. Mental health trends in the digital age: a critical review of current research. J Ment Health Technol 2018;1(1):15–30.
[5] Kour H, Gupta MK. An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. J Artif Intell Med 2022;54:102612. https://doi.org/10.1016/j.iaim.2022.102612.
[6] Safa R, Bayat P, Moghtader L. Automatic detection of depression symptoms in Twitter using multimodal analysis. Comput Hum Behav 2021;124:106937. https://doi.org/10.1016/j.chb.2021.106937.
[7] Smith J, Jones L. Detecting depression on social media. Digit Health 2020;6. https://doi.org/10.1177/2055207620952804 [2055207620952804].
[8] Cha J, Kim S, Park E. A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. 2022 [*Journal Name*, Volume(Issue), Page numbers. DOI or URL].
[9] Smith JA, Brown LK. Recent trends in depression detection: deep learning and multimodal analysis. J Ment Health Res 2023;45(2):123–37.
[10] Brown AR, Johnson MS. Machine learning techniques for mental health detection: a comprehensive review. Int J Artif Intell Med 2022;35(4):567–83.
[11] Johnson SP, Patel RN. Comparative analysis of text preprocessing techniques and supervised classifiers in mental health prediction. J Artif Intell Healthc 2021;15(3): 312–28.
[12] Almouzini Salma, Khemakhem Maher, Alageel Asem. Detecting Arabic depressed users from Twitter data. Procedia Comput Sci 2019:1–9 [ELSVIER].
[13] A. H. B. P. O. M. H, Orabi I. Deep learning for depression detection of Twitter users. In: Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic; 2018. https://doi.org/10.18653/v1/w18-0609.
[14] Chiong Raymond, Budhi Gregorius Satia, Dhakal Sandeep, Chiong Fabian. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Elsvier; 2021. p. 1–12.

[15] Hemanthkumar M, Latha A. Depression detection with sentiment analysis of tweets. Int Res J Eng Technol May 2019:1197–201.
[16] Pachouly ProfSJ, Raut Gargee, Bute Kshama, Tambe Rushikesh, Bhavsar Shruti. Depression detection on social media network (Twitter) using sentiment analysis. Int Res J Eng Technol 2021:1834–9.
[17] Razak CSA, Zulkarnain MA, Hamid SHA, Anuar NB, Jali MZ, Meon H. Tweep: a system development to detect depression in twitter posts. In: Alfred R, Lim Y, Haviluddin H, On C, editors. Computational science and technology. Lecture notes in electrical engineering; 2020. p. 543–52.
[18] Yao Xiaoxu, Guang Yu, Tang Jingyun, Zhang Jialing. Extracting depressive symptoms and their associations from an online depression community. Comput Hum Behav 2021:1–10.
[19] Uddin MZ, Dysthe KK, et al. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Comput & Applic 2021:1–24 [Springer].
[20] Dabhane Suyash, Chawan Prof Pramila M. Depression detection on social media using machine learning techniques. IJSRD 2021:1–5.
[21] Backlinko. Twitter users [Accessed: 31 oct 2022, [online] Avilable], https://backlinko.com/twitter-users.
[22] Verma B, Gupta S, Goel L. A survey on sentiment analysis for depression detection|. In: Komanapalli VLN, Sivakumaran N, Hampannavar S, editors. Advances in automation, signal processing, instrumentation, and control. Lecture notes in electrical engineering, vol. 700. Singapore: Springer; 2021.
[23] Kora Rania, Mohammed Ammar. Corpus on Arabic Egyptian tweets. Harvard Dataverse; 2019.
[25] https://marcobonzanini.com/2015/01/26/stemming-lemmatisation-and-pos-tagging-with-python-and-nltk/.
[26] AlSagri Hatoon S, Ykhlef Mourad. Machine learning-based approach for depression detection in twitter using content and activity features. IEICE Trans Inf Syst 2020; 103(8):1–16.
[27] Salau AO, Jain S. Feature extraction: a survey of the types, techniques, applications. In: 2019 international conference on signal processing and communication (ICSC); 2019. p. 158–64. https://doi.org/10.1109/ICSC45622.2019.8938371.
[28] Kim SW, Gil JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum Cent Comput Inf Sci 2019;9:30.
[29] Zheng Alice, Casari Amanda. Feature Engineering for Machine Learning. 2018. p. 42–5.
[30] Schapire RE. The boosting approach to machine learning: an overview. In: Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B, editors. Nonlinear estimation and classification. Lecture notes in statistics. vol. 171. New York, NY: Springer; 2003.
[31] Debasmita Mishra, Bighnaraj Naik, Janmenjoy Nayak, Alireza Souri, Pandit Byomakesha Dash, Vimal S. Light gradient boosting machine with optimized hyperparameters for identification of malicious access in IoT network. Digit Commun Netw 2022;3:125–37.
[32] Wentao Cai, Ruihua Wei, Lihong Xu, Xiaotao Ding. A method for modelling greenhouse temperature using gradient boost decision tree. Inf Proces Agric 2022; 9:343–54.
[33] Alessia Sarica, Aldo Quattrone. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Front Aging Neurosci 2017;9:1–12.
[34] Tripathy A, Agrawal A, Rath SK. Classi_cation of sentimental reviews using machine learning techniques. Procedia Comput Sci Jan. 2015;57 [pp. 821_829].
[35] Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Appl Jul./Aug. 2008;13(4):18–28.
[36] Apostolidis-Afentoulis Vasileios. SVM Classification with linear and RBF kernels. 2015. https://doi.org/10.13140/RG.2.1.3351.4083.
[38] Liu Yuxihayden. Python Machine Learning by Example. 2nd ed. 2019. p. 137–51.
[39] Sasikala P, Lourdusamy Mary Immaculate Sheela. Sentiment analysis and prediction of online reviews with empty ratings. Int J Appl Eng Res 2018: 11525–31.
[40] Chawla Nitesh, Bowyer Kevin, Hall Lawrence, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res (JAIR) 2002;16:321–57. https://doi.org/10.1613/jair.953.
[41] Kubat M, Matwin S. Addressing the course of imbalanced trainingsets: one sided selection.proceedings of the 14 th international conference on machine learning, ICML'97, Morga Kaufmann. 1997. p. 179–86.
[42] Zeng M, Zou B, Wei F, Liu X, Wang L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In: 2016 IEEE international conference of online analysis and computing science (ICOACS); 2016. p. 225–8. https://doi.org/10.1109/ICOACS.2016.7563084.
[43] Ramachandran A, Gadwe A, Poddar D, Satavalekar S, Sahu S. Performance evaluation of different machine learning techniques using twitter data for identification of suicidal intent. In: 2020 international conference on electronics and sustainable communication systems (ICESC); 2020. p. 223–7. https://doi.org/10.1109/ICESC48915.2020.9155747.
[44] Mustafa RU, Ashraf N, Ahmed FS, Ferzund J, Shahzad B, Gelbukh A. A multiclass depression detection in social media based on sentiment analysis. In: Latifi S, editor. 17th international conference on information technology–new generations (ITNG 2020). Advances in intelligent systems and computing. vol 1134. Cham: Springer; 2020.
[45] Jamil Zunaira, Inkpen Diana, Buddhitha Prasadith, Kenton White. Monitoring tweets for depression to detect at-risk users. Association for Computational Linguistics; 2017. p. 32–40.
[46] Adegoke Adedeji. Detection of depression among Nigerians using machine learning techniques MSc research project data analytics. National College of Ireland; 2019. p. 1–31.
[47] Alghamdi NS, Hosni Mahmoud HA, Abraham A, Alanazi SA, García-Hernández L. Predicting depression symptoms in an Arabic psychological forum. In: IEEE access; 2020. p. 57317–34.
[48] Bansal Ruchi, Singh Jagdeep, Kaur Ranjodh. Machine learning and its applications: a review. 2020. p. 1392–8.
[49] Sah Shagan. Machine learning: a review of learning types, preprints. 2020. p. 1–8.
[50] Yang Zhenkai, Chen Chuansheng, Li Hanwen, Yao Li, Zhao Xiaojie. Unsupervised classifications of depression levels based on machine learning algorithms perform well as compared to traditional norm-based classifications. Front Psychol 2020: 1–9.
[51] Jain S, Narayan SP, Dewang RK, Bhartiya U, Meena N, Kumar V. A machine learning-based depression analysis and suicidal ideation detection system using questionnaires and twitter. In: 2019 IEEE students conference on engineering and systems (SCES); 2019. p. 1–6.
[52] Shah FM, et al. Early depression detection from social network using deep learning techniques. In: 2020 IEEE region 10 symposium (TENSYMP); 2020. p. 823–6.
[53] Stefania Russo, et al. The value of human data annotation for machine learning based anomaly detection in environmental systems. Water Res 2021;206:1–10 [ELSEVER].
[54] Nassar Radwa, Helmy AbdelMoniem, Ramadan Nagy. Depression corpus of Arabic tweets. Harvard Dataverse 2022. https://doi.org/10.7910/DVN/YHMYEQ.
[55] Nassar Radwa, Helmy AbdelMoniem, Ramadan Nagy. Binary labeled depression Corpus of 60,000 English tweets. Harvard Dataverse 2022. https://doi.org/10.7910/DVN/Z4M0HC.
[56] Nassar Radwa, Helmy AbdelMoniem, Ramadan Nagy. Multi labeled depression Corpus of 60,000 English tweets. Harvard Dataverse 2022. https://doi.org/10.7910/DVN/KXB9G4.
[57] Nassar Radwa, Helmy AbdelMoniem, Ramadan Nagy. Multi labeled depression Corpus of 57,000 English tweets. Harvard Dataverse 2022. https://doi.org/10.7910/DVN/CMDF4U.
[58] Nassar Radwa, Helmy AbdelMoniem, Ramadan Nagy. Binary labeled depression Corpus of 57000 English tweets. Harvard Dataverse 2022. https://doi.org/10.7910/DVN/RY1THB.
[61] Salas-Zárate R, Alor-Hernández G, Salas-Zárate MDP, Paredes-Valverde MA, Bustos-López M, Sánchez-Cervantes JL. Detecting depression signs on social media: a systematic literature review. Healthcare 2022;10:291. https://doi.org/10.3390/healthcare10020291.
[62] Khafaga DS, Auvdaiappan M, Deepa K, Abouhawwash M, Karim FK. Deep learning for depression detection using twitter data. Intell Autom Soft Comput 2023;36(2): 1301–13.
[63] Lin Chenhao, Hu Pengwei, Hui Su, Li Shaochun, Mei Jing, Zhou Jie, et al. SenseMood: depression detection on social media. In: Proceedings of the 2020 international conference on multimedia retrieval (ICMR '20). New York, NY, USA: Association for Computing Machinery; 2020. p. 407–11. https://doi.org/10.1145/3372278.3391932.
[64] Angskun J, Tipprasert S, Angskun T. Big data analytics on social networks for real-time depression detection. J Big Data 2022;9:69. https://doi.org/10.1186/s40537-022-00622-2.
[65] De Choudhury M, Sharma S, Logar T, Eekhout W. Mental health discourse on reddit: self-disclosure, social support, and anonymity. Trans ACM 2020;4(4):1–26.
[66] World Health Organization. Depression. https://www.who.int/news-room/fact-sheets/detail/depression; 2021.