Cross-Lingual Depression Detection for Twitter Users: A Comparative Sentiment Analysis of English and Arabic Tweets

Abdelmoniem Helmy (✉ abdelmoniem.hafez@cu.edu.eg)
Cairo University

Radwa Nassar
Cairo University

Nagy Ramdan
Cairo University

Research Article

Keywords: depression detection, social media analysis, machine learning models, sentiment Analysis, natural language processing. mental health monitoring

Posted Date: August 8th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3197428/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Depression, a common mental health issue, significantly disrupting an individual’s daily functioning and increasing premature mortality risk. The ubiquitous use of social media platforms for expressing sentiments and sharing daily activities provides a fertile ground for early detection of depression. This paper makes significant contributions in utilizing online platforms for depression detection. Firstly, we introduce five machine-learning models to detect depression in Arabic and English text from Twitter. For Arabic text, our optimal model achieved a high accuracy with an F1-score of 96.6% for binary classification of depressed and non-depressed tweets. For English text, excluding negations, the model accomplished an F1-score of 92% for binary classification and 88% for multi-classification (depressed, indifferent, happy). When considering negations, the model demonstrated a slightly lower performance with an 87% and 85% F1-score for binary and multi-classification respectively. Secondly, we present three unique corpora: one manually annotated Arabic corpus, and two automatically annotated English corpora—with and without negation. These corpora encompass a broad spectrum of emotional sentiments, enhancing the depth of our analysis. Lastly, the paper presents a novel web application for depression detection, implementing our refined models. This application enables the identification of depression symptoms in tweets and prediction of an individual’s depression trends, supporting both English and Arabic languages. This research represents a significant stride forward in mental health detection leveraging the widespread use of social media.

Highlights

- Social media platforms, particularly Twitter, offer a valuable tool for tracking mental health conditions, including depression, due to their widespread use and ability to capture real-time sentiments.
- Machine-learning models show high accuracy in detecting depression symptoms in English and Arabic tweets, with promising results for both binary and multi-classification tasks.
- A web application implementing the research models successfully detects depression and anxiety symptoms in live Twitter data, offering potential for real-time mental health monitoring and early intervention.

1. Introduction

Humans are getting more ambitious by nature these days, and they seek out every opportunity to advance professionally. Anxiety, depression, tension, annoyance, and dissatisfaction have grown so widespread that many people believe they are unavoidable in the workplace. According to the World Health Organization (WHO), Depression is a widespread ailment that affects 3.8 percent of the world's population, with 5.0 percent of adults and 5.7 percent of persons over sixty years old suffering from
depression [1]. In psychology, depression is a mood or emotional state characterized by feelings of poor self-worth or guilt, as well as a diminished ability to enjoy life [2]. Several of the following symptoms are common among people who are depressed: feelings of sadness, hopelessness, or pessimism; lowered self-esteem and increased self-depreciation; a decrease or loss of ability to enjoy everyday activities; decreased energy and vitality; slowness of thought or action; loss of appetite; and disturbed sleep or insomnia; thoughts of death or suicide [2]. In the worst-case scenario, depression can lead to suicide. Close to 800000 people die due to suicide every year. Suicide is the second leading cause of death in 15-29-year-olds [3]. The World Health Organization (WHO) estimates that one million individuals commit suicide each year, resulting in a mortality rate of 16 persons per 100,000, or one death every 40 seconds [4]. People regularly use social media platforms to express themselves and share their daily lives, considerably increasing the likelihood of using the internet as a rich resource for learning about the world and emphasizing one's own experiences [5, 6].

Social media profiles can provide information about people's mental health, such as whether or not they experience depression [7]. Twitter is the most preferred platform for sharing suicidal thoughts and some people write blogs on depression and anxiety [8]. There is a persistent need for automatic methods for analyzing the textual content which may contain obvious symptoms of depression or possibly suicide on social media in general and Twitter in particular in an attempt to assist who suffers from depression and mental health professionals in their various fields and their statistical research.

As a result, the objective of this paper is to examine users' tweets on the Twitter platform in order to predict depression severity. The data was gathered through Twitter API. While doing so, the paper looked into different text preprocessing techniques, various feature extraction methods such as term frequency-inverse document frequency and Bag of Word, as well as modeling techniques such as supervised classifiers such as Support Vector Machine, Random Forest, Logistic Regression, and Light Gradient Boosting Machine classifiers corresponding to each modality for improving the performance of automatic prediction and assisting in the early detection of depression.

The rest of the paper is structured as follows. Section II introduces most relevant comparable works. Section III provides a detailed description of the proposed methodology. Observations and outcomes are described in Section IV. Finally, Section V concludes the paper.

2. Related work

With the rise in social media usage and the high level of self-disclosure on these platforms, efforts to detect depression using Twitter data have grown [9,10]. Depressed Twitter users are more likely than healthy users to post tweets with negative emotional feelings [11]. In tweets written by users with major depressive disorder, depressive signals can be noticeable [12], [13], [14], [15].

Research has done by Salma Almouzini et al. [16] 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 TweetToSparseFeatureVector 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 [18] used two datasets for training and testing and three data sets for validation and four classifiers with BOW. First, the accuracy was 99% on the two datasets with very low accuracy reach to 0% on validation datasets due to over fitting. The author studied the impact of deleting the word “depression” and “diagnose” on the accuracy and the result reach 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 [19] 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. [20] The relevance of data pre-processing as a necessary step in developing a Machine Learning model and relying on how well the information or text data has been pre-processed 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 percent, the linear classifier was the most accurate.

In addition to Zunaira Jamil et al [49] whose 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. In order 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%, a precision of 70.83%, a recall of 85%, and F1 of 77.27%.
As well as Suyash Dabhane et al [24] whose studied the impact of implementing algorithms individually and implementing ensemble learners. The models have been training 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 over fitting was a common occurrence. As a result, to eliminate this issue of over fitting they trained their model using ensemble learning techniques and achieved an improved accuracy of roughly 87 percent.

Another work, Adedeji [50] 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-grams methods were used for feature extraction. Seven classifiers were applied to detect depression (classification and regression trees, linear discriminant analysis, C5.0, 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. [51] whose 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 percent, according to the findings. After that, ADA and SVM come in second and third, respectively, with 72 percent accuracy. When using BOW of characters as an NLP feature extraction strategy, ADA achieves this level of performance. When the TFIDF model was used to both words and characters, SVM performed the best. The researchers examined into 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 used unsupervised learning approach where there is no labeled data [52], [53], [54]. For example, Yang Zhenkai et al [54] surveyed 8,063 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 [51] 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.

And another group of work that focused on deep learning Recently Deep Learning techniques have proven great results in depression detection, especially with a huge corpus. Using deep learning to detect online depression is still limited work. Hereafter some works that utilized it for examples:

Mustafa R.U et al. [55] 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 has been used which classified the words into fourteen psychological attributes. Each word categorized by LIWC
was given a weight on the basis of 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, 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, 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.

And Orabi et al. [17] pointed to that 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 [21] 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 Twitter platform and the followers they have.

Also F. M. Shah et al. [56] Experimented with a different variety of feature sets included: 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 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 is also two works concern with other languages first, Xiaoxu Yao et al [22] collected more than 100 thousand chines Posts from Sina Weibo platform. They utilized the deep neural network Attention-Based Bidirectional Long Short-Term Memory model (Att-BLSTM) was employed 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 [23] collected more than 200 thousand public Norwegian posts from 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 authors’ proposed features + one hot encoder.

3. Proposed methodology

The paper suggested methodology will be introduced in this section. The assembled datasets are described and analyzed, the produced corpora are then revealed, text cleaning and pre-processing stages are then executed. And then, feature extraction methods are detailed. Finally, 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 Fetching

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 [22],[23],[24].

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. In order to build a corpus for study on 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 its 353 million users with daily access reach to 187 million users [25]. And what is worth to mention that Twitter is the most preferred platform for sharing suicide thoughts and some people write blogs on depression and anxiety [26].

Preparation of Arabic Data Set

Our Arabic dataset Arabic_Dep_tweets_10000 [58] 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 mix of Modern Standard Arabic (MSA), Egyptian dialect and Gulf dialects. Total of 57,391 Tweets have been collected divided into first, the data was obtained using a combination of words that express mental illness such as depression proclivity like

"""""""""""""""""""""""""""""""""""""""""
""""""""""""""""""""""""""""""""""""""
""""""""""""""""""""""""""""""""""""""
"""""""""""""""""""""""""""""""""""""
"""""""""""""""""""""""""""""""""""

And on the other hand combination of words that reflect happiness like

"""""""""""""""""""""""""""""""""""""""""
"""""""""""""""""""""""""""""""""""""""
""""""""""""""""""""""""""""""""""""
"""""""""""""""""""""""""""""""""""
""""""""""""""""""""""""""""""""""

in a total of 29,326 tweets.
“,”, “" 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.

Preparation of English Data Set

As mentioned above in 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.

DataSet I


DataSet II

In the second Eng_with_negation_57.000 [61] [62] 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_57000 dataset.

3.2 Data Cleaning and Pre-processing

Arabic Data Sets

Text cleaning is our first mission, this stage including 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 its 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... ] since they used in both depression and non-depression cases. In addition to Tashkeel removal [] and character Normalization also is an important pre-processing process for example [ to be ], [ to be ], [ to be ] So that the algorithm does not treat the words that contain diacritics or not normalized as another words. Next Tokenization which converting the normal text strings into list of tokens. Finally, the stemming to return the words to its original roots for example ( – ), ( – )
## English Data Sets

As mentioned in Arabic dataset our English text cleaning also including: 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 lower case 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 to 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 their 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 [29].

Finally, lemmatization, it is the process of assembling various inflected forms of a word, or lemma. It converts a number of 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 [29]. 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

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 in order to interpret them and generate reliable outputs [57].

**Arabic_Dep_10000 Annotation**

Each tweet is manually annotated into 1 for “Depressed” class or 0 for “Non depressed” class. After remove 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 40000-Egyptian-tweets corpus [27] 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 10000 Arabic tweets.
English Datasets Annotation

After corpus fetching, cleaning and preprocessing stages, the turn of 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. [21] in depression detection. The number of tweets for each class for binary and multi classifications for both Eng-Dep-60000 and Eng_with_negation_57.000 dataset is as mentioned in (Table 1), (Table 2), (Table 3) and (Table 4).

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>17,149 Tweet</td>
</tr>
<tr>
<td>Indifferent</td>
<td>23,152 Tweet</td>
</tr>
<tr>
<td>Happy</td>
<td>19,871 Tweet</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>21,802 Tweet</td>
</tr>
<tr>
<td>Non_dep</td>
<td>38,369 Tweet</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>24,662 Tweet</td>
</tr>
<tr>
<td>Indifferent</td>
<td>11,923 Tweet</td>
</tr>
<tr>
<td>Happy</td>
<td>20,806 Tweet</td>
</tr>
</tbody>
</table>

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

### 3.4 Feature Extraction

Feature extraction techniques are a necessary step for the text classification process. In order 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 [30], [31]. The paper experimented with Term Frequency–Inverse Document Frequency Tf-Idf and Bag of Word BOW techniques.

**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 in order to achieve words' weights [32].

**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 [33].

The total numbers of features for our datasets were 13,950 for Eng_without_negation_60.000 dataset, 12,647 features for Eng_with_negation_57.000 dataset, and 22,012 features for Arabic_Dep_10000 dataset. After using all features that were extracted; selectpercentile method which Select features according to a percentile of the highest scores was used for feature selection to keep only 10% of features.

### 3.5 Data Resampling

As shown in (Table 1) and (Table 2) 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 creating additional instances for the minority class using Synthetic Minority Oversampling Technique (SMOTE) method [44]. Second, undersampling which in turn randomly removes instances from the majority class using OneSidedSelection method [45]. finally combined sampling which merges the two previous resampling methods using SMOTETomek [46].
3.6 Machine learning supervised algorithms

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 ratio 70:30 and using 5-fold CV also with ratio 80:20 to avoid over fitting. The selected ML algorithms are described in the next subsections.

**Light Gradient Boosting Machine (Lgbm) [34],[35],[36]**

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 a number of 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).

**Random Forest**

The random forest algorithm is an extension of the bagging method which is belongs to ensemble learning techniques. It makes an uncorrelated forest of decision trees using bagging and feature randomization and combines together the results of various decision trees to get a single outcome [37].

**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 [38]. 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 [39] so the largest the distance, the lower the error generated by classifier [41]. 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 [40]. 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 [42].
Logistic Regression

The probability of a dependent categorical variable is predicted by logistic regression. The dependent's binary variables have yes/no codes. 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 $x$ [43].

4. Experimental Results

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 hold out and 5-fold CV methods. The following subsections detailed each dataset result. All classifiers were evaluated by five performance measures as shown below [19], [47]:

1. Confusion Matrix: A table called the confusion matrix is used to describe how well a classifier performs on data whose real values are known. Confusion Matrix-related terms include:
   - True Positives (TP) these are successfully predicted positive values, which means that both the actual and predicted classes are yes.
   - True Negatives (TN) are correctly predicted negative values, meaning that the actual class value and the predicted class value are both no.
   - False Positives (FP) are when the predicted class is yes but the actual class is no
   - False Negatives (FN) are when the predicted class is no when the actual class is yes.

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.

   $$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

3. Precision: The percentage of positively classified data that is genuinely positive is what this metric measures.

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

4. Recall: It measures the percentage of positive values that the algorithm successfully identified.

   $$\text{Recall} = \frac{TP}{TP+FN}$$
F1-Score: It is the harmony of recall and precision. It is, in other words, the harmonic mean of recall and precision.

\[
F1-Score = \frac{2\times P \times R}{P + R}
\]

Another assessment metric was added to the proposed method to better examine its effectiveness it is the Receiver Operating Characteristic (ROC) chart. The true positive rate (TPR) VS 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) [48].

This section divided into first, 1. English corpora experiments which contain the binary and multi results for each corpus. Second, 2. Arabic corpus experiments results and, finally, 3. The web application.

### 4.1 English corpora experiments

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

Preliminary experiments were conducted first which includes: ML classifiers were applied after selecting 10% of the original features of the imbalanced two corpora by using chi square method. 10% of the features have been used since it provides 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 oneside selection, and SMOTE+ Tomek for combined sampling in order to balance the original corpora.

### Eng_without_negation_60.000 corpus results

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.

**Binary classification experiments**

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>Tf-Idf</td>
<td>38,369</td>
<td>21,802</td>
</tr>
<tr>
<td>Non_dep</td>
<td>38,369</td>
<td>33,123</td>
<td>38,369</td>
</tr>
<tr>
<td>Depressed</td>
<td>BOW</td>
<td>38,369</td>
<td>21,802</td>
</tr>
<tr>
<td>Non_dep</td>
<td>38,369</td>
<td>36,167</td>
<td>38,277</td>
</tr>
</tbody>
</table>

Tf-Idf results

In Table 6, the average of f1 score for all models using first, feature selection (imbalanced), second over sampling, third under sampling and finally combined sampling with Tf-Idf.

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 over and combined sampling were 4% with the RBF-SVM as a best model, whereas, under-sampling archived slight improvement for only RF, L-SVM, and RBF-SVM with 1%.

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Imbalanced corpus</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>L gbm</td>
<td>86%</td>
<td>89%</td>
<td>86%</td>
<td>89%</td>
</tr>
<tr>
<td>RF</td>
<td>86%</td>
<td>90%</td>
<td>87%</td>
<td>90%</td>
</tr>
<tr>
<td>L-svm</td>
<td>87%</td>
<td>90%</td>
<td>88%</td>
<td>90%</td>
</tr>
<tr>
<td>Rbf -svm</td>
<td>88%</td>
<td>92%</td>
<td>89%</td>
<td>92%</td>
</tr>
<tr>
<td>LR</td>
<td>85%</td>
<td>87%</td>
<td>85%</td>
<td>87%</td>
</tr>
</tbody>
</table>

We found that under sampling technique OneSidedSelection (OSS) does not fully balance the dataset in this experiment as shown in table 1. Since it eliminate many instances from the majority class were, including redundant examples and ambiguous examples. It means that there has 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.

The 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 much costs either for individuals or society and even can save a person life. On the other hand, if a person predicted as willing to have a depression but it is not, will not has 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 depressed class.

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

<table>
<thead>
<tr>
<th>Actual</th>
<th>Depressed</th>
<th>Non_dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Depressed</td>
<td>7202</td>
<td>739</td>
</tr>
<tr>
<td>Non_dep</td>
<td>399</td>
<td>6937</td>
</tr>
</tbody>
</table>

Figure 5 displays a receiver operating characteristic (ROC) chart for binary experiment with 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 figure 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.

*BOW results*

Table 8 illustrates the BOW method experiments with resampling techniques especially over and combined sampling which achieved a simple enhancement when compared with Tf-Idf.
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 depressed class with BOW in this binary experiment.

A receiver operating characteristic (ROC) chart for binary experiments with combined resampling technique 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.

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Imbalanced corpus</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>L gbm</td>
<td>86%</td>
<td>87%</td>
<td>86%</td>
<td>87%</td>
</tr>
<tr>
<td>RF</td>
<td>86%</td>
<td>88%</td>
<td>86%</td>
<td>88%</td>
</tr>
<tr>
<td>L-svm</td>
<td><strong>88%</strong></td>
<td><strong>89%</strong></td>
<td><strong>89%</strong></td>
<td><strong>89%</strong></td>
</tr>
<tr>
<td>Rbf-svm</td>
<td>87%</td>
<td>88%</td>
<td>88%</td>
<td>88%</td>
</tr>
<tr>
<td>LR</td>
<td>87%</td>
<td>89%</td>
<td>88%</td>
<td>89%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Actual</th>
<th>Depressed</th>
<th>Non_dep</th>
<th>P=</th>
<th>R=</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depressed</td>
<td>6845</td>
<td>726</td>
<td><strong>P=90.4%</strong></td>
<td><strong>R= 89.1%</strong></td>
</tr>
<tr>
<td>Non_dep</td>
<td>829</td>
<td>6948</td>
<td><strong>P=89.3%</strong></td>
<td><strong>R= 90.5%</strong></td>
</tr>
</tbody>
</table>

Multi classification experiments

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.

**Table 10.** Eng\_without\_negation\_60.000 tweets counts for each class after data resampling in the multi classification experiment

<table>
<thead>
<tr>
<th>Class</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Depressed</strong></td>
<td>Tf-Idf</td>
<td>23,152</td>
<td>17,149</td>
</tr>
<tr>
<td><strong>Indifferent</strong></td>
<td></td>
<td>23,152</td>
<td>18,394</td>
</tr>
<tr>
<td><strong>Happy</strong></td>
<td></td>
<td>23,152</td>
<td>15,655</td>
</tr>
<tr>
<td><strong>Depressed</strong></td>
<td>BOW</td>
<td>23,152</td>
<td>17,149</td>
</tr>
<tr>
<td><strong>Indifferent</strong></td>
<td></td>
<td>23,152</td>
<td>21,087</td>
</tr>
<tr>
<td><strong>Happy</strong></td>
<td></td>
<td>23,152</td>
<td>14,877</td>
</tr>
</tbody>
</table>

*Tf-Idf results*

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.

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 by 91.2% recall for depressed class.

**Table 11.** Eng\_without\_negation\_60.000 F1-scores average of multi-classification for all classifiers using Tf-Idf

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Imbalanced corpus</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lgbm</td>
<td>82%</td>
<td>82%</td>
<td>80%</td>
<td>82%</td>
</tr>
<tr>
<td>RF</td>
<td>83%</td>
<td>84%</td>
<td>82%</td>
<td>84%</td>
</tr>
<tr>
<td>L- svm</td>
<td>88%</td>
<td><strong>88%</strong></td>
<td>87%</td>
<td>88%</td>
</tr>
<tr>
<td>Rbf-svm</td>
<td>84%</td>
<td>85%</td>
<td>84%</td>
<td>85%</td>
</tr>
<tr>
<td>LR</td>
<td>82%</td>
<td>83%</td>
<td>81%</td>
<td>83%</td>
</tr>
</tbody>
</table>
Table 12. Confusion matrix of RBF-SVM using Tf-Idf and combined resampling of Eng_without_negation_60.000 in multi classification experiment

<table>
<thead>
<tr>
<th></th>
<th>Depressed</th>
<th>Indifferent</th>
<th>Happy</th>
<th>P=</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depressed</td>
<td>4226</td>
<td>368</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>91.1%</td>
</tr>
<tr>
<td>Indifferent</td>
<td>355</td>
<td>3791</td>
<td>401</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83%</td>
</tr>
<tr>
<td>Happy</td>
<td>49</td>
<td>371</td>
<td>4084</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>91%</td>
</tr>
</tbody>
</table>

R=91.2%  R=84%  R=90.2%

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

In the following graph the model has three curves represent each class since 2 for “depressed” class, 1 for “indifferent” class and 0 for “happy” class using 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.

BOW results

With BOW method in 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 were not achieved any enhancement percentage. Undersampling achieved a decrease for 4 models; 2% with Lgbm and 1% with RF, L-SVM, and LR while RBF-SVM Didn't achieves any improvement with under sampling. The best classifier with applying BOW was L-SVM with 88% F1 score.

Table 13. Eng_without_negation_60.000 F1-scores average of multi-classification for all classifiers using BOW
Table 14 illustrates the confusion matrices of multi classification experiment with recall and precision of L-SVM with BOW which achieved the highest percentages by 87.4% recall for depressed class.

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

<table>
<thead>
<tr>
<th>Actual</th>
<th>Depressed</th>
<th>Indifferent</th>
<th>Happy</th>
<th>p=</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>Depressed</td>
<td>4051</td>
<td>475</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Indifferent</td>
<td>493</td>
<td>3758</td>
<td>492</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>87</td>
<td>397</td>
<td>4063</td>
</tr>
</tbody>
</table>

| R=87.4% | R=81.1% | R=88.3% |

Fig 8: ROC curves for L-SVM that 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.

Eng_with_negation_57.000 corpus results

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 classifications.

Binary classification experiments
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.

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>31,871</td>
<td>25,520</td>
<td>31,871</td>
</tr>
<tr>
<td>Non dep</td>
<td>31,871</td>
<td>25,924</td>
<td>30,968</td>
</tr>
<tr>
<td>Depressed</td>
<td>31,871</td>
<td>25,520</td>
<td>31,871</td>
</tr>
<tr>
<td>Non dep</td>
<td>31,871</td>
<td>28,591</td>
<td>31,803</td>
</tr>
</tbody>
</table>

**Tf-Idf results**

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 for all classifiers. In this experiment, RF achieved the highest F1 score with 87.9% but L-SVM and LR were the lowest with 79% F1 score according to combined sampling results. Table 17 illustrates the confusion matrixes for RF which achieved the highest recall by 91.1% for depressed class.

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Imbalanced</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lgbm</td>
<td>80%</td>
<td>83%</td>
<td>80%</td>
<td>83%</td>
</tr>
<tr>
<td>RF</td>
<td><strong>82%</strong></td>
<td><strong>86%</strong></td>
<td><strong>82%</strong></td>
<td><strong>87%</strong></td>
</tr>
<tr>
<td>L-svm</td>
<td>74%</td>
<td>78%</td>
<td>73%</td>
<td>79%</td>
</tr>
<tr>
<td>Rbf -svm</td>
<td>82%</td>
<td>85%</td>
<td>82%</td>
<td>86%</td>
</tr>
<tr>
<td>LR</td>
<td>74%</td>
<td>79%</td>
<td>74%</td>
<td>79%</td>
</tr>
</tbody>
</table>

**Table 17.** Confusion matrix of RF using Tf-Idf and combined resampling of Eng_with_negation_57.000 binary classification experiment
Fig. 9 illustrates that the AUCs for all classifiers are close to 1 but the AUC 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 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.

**BOW results**

It is clear from Table 18 that Over and combined sampling achieved 4% improvement with Lgbm, 6% with L-SVM and LR, 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.

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

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Imbalanced</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lgbm</td>
<td>79%</td>
<td>83%</td>
<td>81%</td>
<td>83%</td>
</tr>
<tr>
<td>RF</td>
<td>83%</td>
<td>86%</td>
<td>83%</td>
<td>85%</td>
</tr>
<tr>
<td>L-svm</td>
<td>73%</td>
<td>79%</td>
<td>72%</td>
<td>79%</td>
</tr>
<tr>
<td>Rbf-svm</td>
<td><strong>84%</strong></td>
<td><strong>86%</strong></td>
<td><strong>83%</strong></td>
<td><strong>86%</strong></td>
</tr>
<tr>
<td>LR</td>
<td>73%</td>
<td>79%</td>
<td>73%</td>
<td>79%</td>
</tr>
</tbody>
</table>

Table 19. Confusion matrix of RBF-SVM using BOW and combined sampling of Eng_with_negation_57.000 binary classification experiment
<table>
<thead>
<tr>
<th></th>
<th>Depressed</th>
<th>Non_dep</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depressed</td>
<td>5812</td>
<td>1018</td>
<td>P=85%</td>
</tr>
<tr>
<td>Non_dep</td>
<td>561</td>
<td>5076</td>
<td>P=90%</td>
</tr>
</tbody>
</table>

\[ R = 91.1\% \quad R = 83.2\% \]

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.

**Multi classification**

Preliminary experiments were applied on the original imbalanced status. In addition to resampled three classes. Table 20 illustrates the count of tweets for the three classes after all resampling techniques.

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Depressed</strong></td>
<td>24,662</td>
<td>17,669</td>
<td>24,662</td>
</tr>
<tr>
<td><strong>Indifferent</strong></td>
<td>24,662</td>
<td>11,923</td>
<td>24,662</td>
</tr>
<tr>
<td><strong>Happy</strong></td>
<td>24,662</td>
<td>14,128</td>
<td>23,884</td>
</tr>
<tr>
<td><strong>Depressed</strong></td>
<td>24,662</td>
<td>22,559</td>
<td>24,662</td>
</tr>
<tr>
<td><strong>Indifferent</strong></td>
<td>24,662</td>
<td>16,571</td>
<td>24,662</td>
</tr>
<tr>
<td><strong>Happy</strong></td>
<td>24,662</td>
<td>11,923</td>
<td>24,585</td>
</tr>
</tbody>
</table>

**Tf-Idf results**

As illustrated from table 21 that still both over and combined sampling achieves the F1 score improvement of all models while using Tf-Idf. 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% with combined resampling.

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

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Imbalanced corpus</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lgbm</td>
<td>77%</td>
<td>77%</td>
<td>72%</td>
<td>79%</td>
</tr>
<tr>
<td>RF</td>
<td><strong>83%</strong></td>
<td><strong>83%</strong></td>
<td>78%</td>
<td><strong>85%</strong></td>
</tr>
<tr>
<td>L-svm</td>
<td>73%</td>
<td>73%</td>
<td>71%</td>
<td>77%</td>
</tr>
<tr>
<td>Rbf- svm</td>
<td>77%</td>
<td>77%</td>
<td>77%</td>
<td>84%</td>
</tr>
<tr>
<td>LR</td>
<td>70%</td>
<td>70%</td>
<td>68%</td>
<td>73%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Actual</th>
<th>Depressed</th>
<th>Indifferent</th>
<th>Happy</th>
<th>P=83.4%</th>
<th>R=80.3%</th>
<th>R=93.9%</th>
<th>R=81.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>Depressed</td>
<td>3962</td>
<td>127</td>
<td>661</td>
<td>P=83.4%</td>
<td>R=80.3%</td>
<td>R=93.9%</td>
</tr>
<tr>
<td>Indifferent</td>
<td>278</td>
<td>4654</td>
<td>215</td>
<td>P=90.4%</td>
<td>R=93.9%</td>
<td>R=81.5%</td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>693</td>
<td>171</td>
<td>3884</td>
<td>P=81.8%</td>
<td>R=93.9%</td>
<td>R=81.5%</td>
<td></td>
</tr>
</tbody>
</table>

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.

*BOW results*
When the BOW method has been used with resampling techniques especially over and combined sampling achieved remarkable enhancement. The best classifiers with applying BOW were RF and RBF-SVM with 84% F1 score with over and combined sampling as shown in table 23.

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.

**Table 22.** Eng_with_negation_57.000 F1- scores averages of multi classification for all classifiers using BOW

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Imbalanced corpus</th>
<th>Over sampling</th>
<th>Under sampling</th>
<th>Combined sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lgbm</td>
<td>76%</td>
<td>79%</td>
<td>75%</td>
<td>79%</td>
</tr>
<tr>
<td>RF</td>
<td><strong>81%</strong></td>
<td><strong>84%</strong></td>
<td><strong>79%</strong></td>
<td><strong>84%</strong></td>
</tr>
<tr>
<td>L-svm</td>
<td>74%</td>
<td>76%</td>
<td>72%</td>
<td>77%</td>
</tr>
<tr>
<td>Rbf-svm</td>
<td><strong>81%</strong></td>
<td><strong>84%</strong></td>
<td><strong>79%</strong></td>
<td><strong>84%</strong></td>
</tr>
<tr>
<td>LR</td>
<td>71%</td>
<td>74%</td>
<td>69%</td>
<td>74%</td>
</tr>
</tbody>
</table>

**Table 23.** Confusion matrix of RBF-SVM using BOW and combined resampling of Eng_with_negation_57.000 in multi classification experiment

<table>
<thead>
<tr>
<th>Actual</th>
<th>Depressed</th>
<th>Indifferent</th>
<th>Happy</th>
<th>P=80.2%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>Depressed</td>
<td>4005</td>
<td>193</td>
<td>792</td>
</tr>
<tr>
<td></td>
<td>Indifferent</td>
<td>138</td>
<td>4543</td>
<td>211</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>789</td>
<td>197</td>
<td>3914</td>
</tr>
</tbody>
</table>

R=81.2% R=92% R=79.6%
Fig. 12. ROC curves for RF that applied with combined sampling and BOW. It achieved the best performance in differentiating between classes among the other classifiers.

4.2 Arabic corpus experiments

The experiments of the Arabic_Dep_tweets_10000 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.

Tf-Idf results

As shown from Table 24, 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 24. All ML classifiers’ f1-scores averages of Arabic_Dep_tweets_10000 experiments with Tf-Idf

<table>
<thead>
<tr>
<th>Classifier</th>
<th>All features</th>
<th>Feature selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lgbm</td>
<td>96.1%</td>
<td>96.2%</td>
</tr>
<tr>
<td>RF</td>
<td>95.6%</td>
<td>96%</td>
</tr>
<tr>
<td>L-svm</td>
<td>96.4%</td>
<td>96.2%</td>
</tr>
<tr>
<td>Rbf-svm</td>
<td><strong>96.6%</strong></td>
<td><strong>96.5%</strong></td>
</tr>
<tr>
<td>LR</td>
<td>96.3%</td>
<td>96.2%</td>
</tr>
</tbody>
</table>

Table 25 illustrates the confusion matrixes for RBF-SVM which achieved the highest recall for depresses 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.

Table 25. Confusion matrix of RBF-SVM using Tf-Idf of Arabic_Dep_tweets_10000 using all features
<table>
<thead>
<tr>
<th>Actual</th>
<th>Depressed</th>
<th>Non_depred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>Depressed</td>
<td>1485</td>
</tr>
<tr>
<td></td>
<td>Non_depred</td>
<td>78</td>
</tr>
</tbody>
</table>

**BOW results**

As cleared from table 26, 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.

**Table 26 All ML classifiers’ f1-scores average of Arabic_Dep_tweets_10000 experiments with BOW**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>All features</th>
<th>Feature selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lgbm</td>
<td>96.3%</td>
<td>96.2%</td>
</tr>
<tr>
<td>RF</td>
<td>95.7%</td>
<td>96.1%</td>
</tr>
<tr>
<td>L-svm</td>
<td>95.9%</td>
<td>96.1%</td>
</tr>
<tr>
<td>Rbf -svm</td>
<td>96.2%</td>
<td>96.1%</td>
</tr>
<tr>
<td>LR</td>
<td>96.4%</td>
<td>96.3%</td>
</tr>
</tbody>
</table>

Table 27 illustrates the confusion matrixes for and LR which achieved the highest recall for depresses 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.

**Table 27. Confusion matrix of LR using BOW of Arabic_Dep_tweets_10000 using all features**
4.3 Twittpy web 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 through check the status of the specific Twitter user that its user name 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.

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

In live tweets prediction, after entering the user name and the count of tweets and click “predict” on the English page or “ ” on the Arabic page. The models will run and the tweets will display 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.

5. Conclusion

More than two-thirds of suicides each year are caused by depression, which is the most common mental illness. Unfortunately, a lot of cases go untreated due to self-denial or a failure to recognize them. Social media posts can be a useful tool for tracking a variety of mental health conditions, including depression, given the exponential rise in social media usage.

The most commonly utilized languages on social media are English and Arabic. The lack of adequate Arabic corpora makes it difficult to use machine learning techniques in depression detection for the Arabic language. Thus, the paper introduced a new manually labeled Arabic depression corpus of 10,000
tweets, and two new automatically labeled English depression corpora, one without negation of 60,172 tweets and the other with negation of 57,392 tweets.

The best results in our experiments were first; with Arabic_Dep_tweets_10000, RBF-SVM with Tf-Idf achieved the highest f1-score by 96.6%. Second, in the preliminary experiments with Eng_without_negation_60.000 corpus, RBF-SVM with Tf-Idf achieved the highest f1 score of 88% in binary classification and 87% in multi-classification.

Third, with Eng_with_negation_57.000 corpus preliminary experiments, RF with Tf-Idf over-performed the other models with an 82% f1 score in binary and 83% in multi-classification. Additionally, this work applied data resampling techniques to enhance the results of our imbalanced two English corpora. Particularly, we used SMOTE for over-sampling, Oneside selection for under-sampling, and SMOTE+Tomek for combined sampling.

After all experiments, it can be concluded that Tf-Idf has excelled BOW in most of the experiments of all three corpora.

An additional observation is that the combined sampling technique on the two English corpora has provided further improvements with an f1 score of 92% in binary classification and 88% in multi-classification by RBF-SVM with Eng_without_negation_60.000 corpus, and 87% in binary classification and 85% in multi-classification by RF with Eng_with_negation_57.000. As opposed to under-sampling which made no progress with the performance of the classifiers. Since it is not fully balance our two corpora.

Finally, we tested our best models with new live Twitter data through a Twittpy web application. It detects tweets that contain depression and anxiety symptoms in both English and Arabic languages. Also, it enables us to check specific Twitter users’ tweets as live tweets or even check any individual tweet. It displays users’ depression percentages and the status of each tweet that has been retrieved.

As a future work, applying deep learning techniques for depression detection on our corpora and then contrast the outcomes. Other kinds of mental disorders such as Post-traumatic stress disorder (PTSD) can be investigated.

**Declarations**

**Data Availability**

References


Figure 1

Depression detection proposed methodology

Figures
Figure 2

Arabic_Dep_tweets_10000 dataset word cloud
Figure 3

Eng_without_negation_60.000 dataset word cloud
Figure 4

Eng_with_negation_57.000 dataset word cloud
Figure 5

all classifiers ROC curves and AUCs for Eng_without_negation_60.000 binary classification with combined resampling using Tf-Idf
Figure 6

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

Figure 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

Figure 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

Figure 9

All classifiers ROC curves and AUCs for Eng_with_negation_57.000 binary classification with combined resampling using Tf-Idf
Figure 10

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

Figure 11

ROC curves and AUCs for RF classifier with combined sampling and Tf-Idf for multi classification of Eng_with_negation_57.000 corpus
Figure 12

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

Figure 13

ROC curves and AUCs for Arabic_Dep_tweets_10000 using Tf-Idf with all features
Figure 14

ROC curves and AUCs for Arabic_Dep_tweets_10000 using BOW with all features

Twitter Depression Detection

Import Live Tweets
- Enter Account User Name
- Enter Tweets Number
- Predict

Type Your Own Tweet
- Enter Your Tweet
- Predict
Figure 15

English depression detection page

Figure 16

Arabic depression detection page

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ArabicDepression10.000Tweets.xlsx
- BinaryLabeledDepressionEnglish57000Tweetnegationwords.xlsx
- MultiLabeledDepressionEnglish57000Tweetnegationwords.xlsx
- MultiLabeledDepressionEnglish60000Tweet.xlsx