PAPER

Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features

Hatoon S. ALSAGRI[†a)], ***Member and*** **Mourad YKHLEF**[††], ***Nonmember***

SUMMARY Social media channels, such as Facebook, Twitter, and Instagram, have altered our world forever. People are now increasingly connected than ever and reveal a sort of digital persona. Although social media certainly has several remarkable features, the demerits are undeniable as well. Recent studies have indicated a correlation between high usage of social media sites and increased depression. The present study aims to exploit machine learning techniques for detecting a probable depressed Twitter user based on both, his/her network behavior and tweets. For this purpose, we trained and tested classifiers to distinguish whether a user is depressed or not using features extracted from his/her activities in the network and tweets. The results showed that the more features are used, the higher are the accuracy and F-measure scores in detecting depressed users. This method is a data-driven, predictive approach for early detection of depression or other mental illnesses. This study's main contribution is the exploration part of the features and its impact on detecting the depression level.

key words: *social media analytics, depression detection, machine learning (ML), support vector machine (SVM), naive Bayes, decision tree, feature selection*

1. Introduction

Depression is a common mental illness and a leading cause of disability worldwide, which may cause suicides. Globally, more than 300 million people are estimated to suffer from depression every year*. Generally, depression is diagnosed through face-to-face clinical depression criteria. However, at early stages of depression, 70% of the patients would not consult doctors, which may take their condition to advance stages s [1].

Recently, there has been a movement to leverage social medial data for detecting, estimating, and tracking the changes in occurrence of a disease [2]. The ubiquity of social media provides a rich opportunity to enhance the data available to mental health clinicians and researchers, enabling a better-informed and -equipped mental health field [3]. In addition, contagious negative emotions in social networks adversely affect people, leading to depression and other mental illnesses. Mental illness is known as a major risk factor of suicide; almost 80% of those who attempt or die by suicide are known to have had some form of mental illness [4], [5]. Depression is considered as the most common mental illness [6],but because of its unrecognition or denial, it has remained undiagnosed or untreated [7]. The onset of a major depression can be prevented by early recognition of its symptoms and their treatment through timely intervention [7].

Many studies have detected physical and mental illnesses derived from social media's huge information, in precise some studies were dedicated for depression [3], [8], [9]. De Choudhury et al. [8] ffound that tweets posted by individuals with major depressive disorder, as well as their social media activity, can be utilized to classify and predict if they are suffering from depression or are likely to suffer in the future. Nadeem et al, Harman et al. and Benton et al focused on whether the users' tweets were depressive in nature [10]–[12], while Tsugawa et al. and Coopersmith et al. focused on user activity in Twitter [3], [13].

This study also aims to detect whether the user is depressed, from the nature of his\ her tweets and activity in the network. It can be further used to identify other mental illnesses and might even form an underlying infrastructure for new mechanisms that would help detect and limit depression diffusion in social networks.

This study exploits data collected from 111 user profiles and more than 300,000 tweets. Many classifier techniques are employed to identify the depression level, of which support vector machine (SVM)-linear shows the best results, with the accuracy reaching 82.5 and F-measure reaching 0.79.

The rest of the article is organized as follows. Section 2 illustrates the literature review. Section 3 presents the scientific background of the classifiers used in the experiments. Section 4 explains the methodology of the study and the features extracted and computed. Section 5 describes the experiments and discusses results. Finally, Sect. 6 outlines the conclusions of the study.

2. Literature View

With the gradual increase in social media usage and the extensive level of self-disclosure within such platforms, efforts to detect depression from Twitter data have increased [9], [14]. Park et al. [15] indicated that depressed

Manuscript received February 4, 2020.
Manuscript revised April 7, 2020.
Manuscript publicized April 24, 2020.

[†]The author is with the Dept. of Information Systems, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia.

[††]The author is with Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia.

a) E-mail: hsagri@imamu.edu.sa

DOI: 10.1587/transinf.2020EDP7023

*https://www.who.int/en/news-room/fact-sheets/detail/depression

Twitter users tend to post tweets containing negative emotional sentiments more than healthy users. In addition, De Choudhury et al. [8] found that depressive signals are noticeable in tweets posted by users with major depressive disorder [9], [10], [16].

Thus far, different features have been used to detect depression from Twitter data. De Choudhury et al. [8] collected more than 2 M tweets from 476 users who were clinically diagnosed as depressed and had Twitter accounts. They used behavioral attributes related to social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications to build a classifier that provides estimates of the risk of depression. They leveraged these distinguishing attributes to build an SVM classifier that can predict the risk of depression with 70% classification accuracy. Tsugawa et al. [13] revealed that frequencies of word usage, along with topic modeling, are useful features for the prediction model. Using the radial kernel SVM classifier, they obtained 69% classification accuracy in predicting depression of 81 participants out of the 209 collected using a questionnaire. In addition, Reece et al. [9] extracted predictive features for measuring the effect, linguistic style, and context from users' tweets; built models using these features with supervised learning algorithms, and successfully discriminated between depressed and healthy contents. Their data were collected from 105 out of the 204 depressed users and the CESD scores relied on the identification of depressed users. The best classifier performance was obtained using a 1200-tree random forest classifier, increasing the precision to 0.866, compared to other study results. Nadeem et al. [10] utilized the bag-of-words approach for better depression detection, which uses word occurrence frequencies to quantify the content of a tweet measured on a document level. They employed four types of binary classifiers: linear SVM classifier, decision tree (DT), Naive Bayes (NB) algorithm, and logistic regressive approach, and found that NB outperformed other classifiers with an accuracy of 81% and precision of 0.86. They used a corpus of more than 2.5 M tweets gathered from the Shared Task organizers of CLPsych 2015, online, from users who indicated they were diagnosed as depressed (326) or with PTSD. On the other hand, Nadeem et al. [10], Coppersmith et al. [3], and Mowery et al. [2] considered sentiment analysis as a feature to detect depression from Twitter data. Jamil et al. [14] concluded that the use of sentiment analysis, along with percentage of depressed tweets, increases the precision and recall of detecting depression. The classifier was trained on 95 users who disclosed their own depression (which was equal to 5% of users participating in the study, while the remaining 95% were healthy users), using SVM, which provided a recall of 0.875 and precision of 0.775.

De Choudhury et al. [8] and Jamil et al. [14] benefited from depressed people tweets for extracting features that helped increase the detection accuracy. De Choudhury et al. [8] built a depression lexicon of terms that are likely to appear in postings from individuals discussing depression or its symptoms in online settings. In contrast, Jamil et al. [17] used the percentage of depressed tweets, along with self-indication of depression, to decide whether a user should be removed from the training set and found this feature to increase the model's accuracy.

3. Background

Classifiers

SVM, Naive Bayes (NB), and Decision Tree (DT) are some of the widely used algorithms in natural language processing tasks [17]. Of these, SVM-linear classifier demonstrates the best performance. As there is no one algorithm suited for all tasks, researchers tend to try various algorithms and enhance them for the problem of their interest [17].

1. **NB:**
   NB is based on the "Bayes' theorem" in probability. As a requirement of this theorem, NB can be applied only if the features are independent of each other [18].

   $$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

   It is generally used in machine learning owing to its ability to efficiently merge the evidence from several features [19]. Although the NB classifier is considered the most straightforward method in the machine learning field, it is still competitive with SVM [10].
2. **DT:**
   DTs classify instances by sorting them based on the feature values. Each node in a DT represents a feature, and each division represents a value that the node can undertake [20];
   A DT involves partitioning the data into subdivisions that contain occurrences with similar values, using an important aspect in a DT, called split selection, which intends to find an attribute and its related splitting function for each test node in a DT. Splits are evaluated by calculating entropy [21].
   A DT can structure complicated nonlinear decision borders [22]. The extensive trees will be structured, and then excluded to reduce the cost-complexity criterion. The resulting tree would be easily explicable and can provide perception into the data structure that is claimed to be a main advantage of tree algorithms [17]. DTs simply pose a series of carefully constructed questions to classify a task that makes them straightforward in nature, for which they are extensively employed within the machine learning field [10].
3. **SVM:**
   SVM is a supervised learning model that underlines two different classes in a high-dimensional space. It can adjust several features while balancing the excellent performance, to reduce the possibility of overfitting [23]. SVM is famous for its powerful capability, specifically when working on real-world data, which include a decisive theoretical basis and its insensitivity to high-dimensional data [24], [25]. SVM is a type of

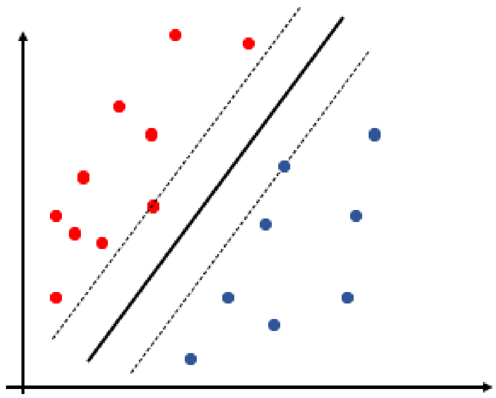


Fig. 1 SVM's maximum separating hyperplane

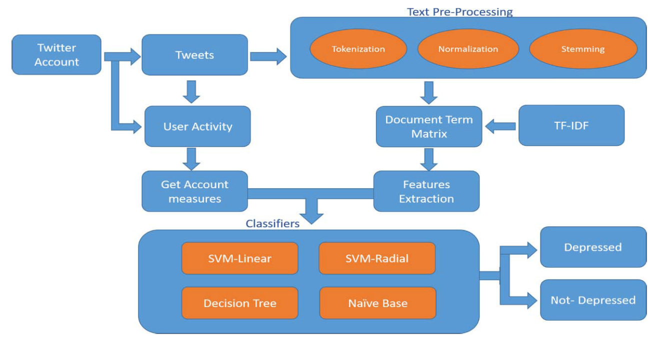


Fig. 2 DDACF classification model

algorithm with a set of labeled training examples for a binary class problem. The training algorithm in SVM creates a potential hyperplane, which divides the cases from the two classes. It escalates the distance between the divided hyperplane and the training examples closest to the hyperplane [26]. SVM can provide the prediction and determine which side of the hyperplane an object inclines [17]; as shown in Fig. 1.

Although SVM can tolerate the data outlier, it is computationally inefficient and sensitive to the kernel hyperparameters.

4. Methodology

A quantitative study is conducted to train and test various machine learning classifiers to determine whether a twitter account user is depressed, from tweets initiated by the user or his/her activities on Twitter.

Data preparation, feature extraction, and classification tasks are performed using various R packages and in R version 3.3 [27] using Rstudio IDE [28]. The classifiers are trained using 10-fold cross validation to avoid overfitting, and then tested on a held-out test set.

Figure 2 illustrates the depression detection using activity and content features (DDACF) classification model. First, all tweets for depressed and non-depressed accounts, as well as information of user account and activities such as number of followers, number of following, total number of posts, time of posts, number of mentions, and number of retweets, are retrieved. Next, all tweets of an account are assembled in one document.

Text preprocessing is applied to all documents. First, a corpus is created and tweets in each document are tokenized. Next, normalization is applied, where all characters are turned to lower case and punctuations, retweets, mentions, links, unrecognized emoji's, and symbols are removed. Usually, normalization includes removing stop words, such as first-person pronouns like "I," "me," and "you," but when removing stop words, we keep the first-person pronouns. Later, stemming is applied and a document term matrix (DTM) is created for each account [29]. The matrix indicates the frequency of words in each tweet, where each row indicates a document of tweets and each column indicates all words used in all accounts. TF-IDF is used to measure the words' weight.

Features applied on the DTM are then merged with account measures extracted from the social network and user activities. Results of the merge are then treated as independent variables in a classification algorithm to predict the dependent variable of an outcome of interest. Ultimately, we decide upon the DT, a linear and radial kernel support vector classifier, and an NB algorithm.

Feature Engineering

Feature engineering is referred to in machine learning as "the process of using domain knowledge of the data to create features that can be used by machine learning algorithms to find patterns" [17]. Features are generated to extract the information understood by a machine learning algorithm and might be useful for prediction [17].

Various features can be extracted from the activity histories and tweets of Twitter users. Features are extracted from the text after text preprocessing, when the text is in the desired format, as shown in Fig. 3. These features are computed for both the training and test sets. Table 1 lists the features and their possible values used for the classification model.

Previous studies have shown that first-person pronouns are useful predictors of depression. De Choudhury and Jamil [8], [17] indicated that the use of singular pronouns in comparison to second- and third-person pronouns is also an indicator of depression. Wang et al. [30] mentioned that depressed users lean more toward the use of first-person singular pronouns but less emoticons.

TF-IDF is also added to the classification model, and is used by Ramos, Resnik et al., and De Choudhury et al. [8], [31], [32] to rank words used by users with mental illness [17]. To reduce computation time, sparse vectors are removed from the matrix. Words having more than 95% zeros indicate low use in the user account, and as a result, are removed.

Inspired by Prieto et al. [33], information gain (IG) is added as a feature selector for the model. Prieto et al. [33] used IG to reduce features that improve the classification of depressed users, and reduced the time needed to generate

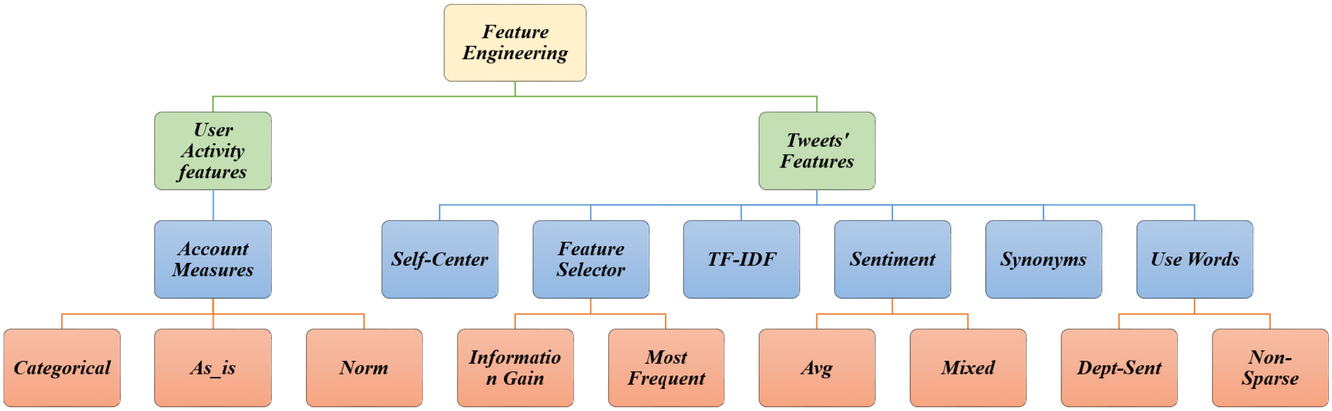


Fig. 3 Visualization of features used in the study

Table 1 Description of features and their possible values

Features	Possible Values	Description
Self-Center	T	Use first-person pronouns pronouns.
	F	Remove first-person pronouns when removing stop words.
TF-IDF	T	Determines the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document corpus.
	F	Word frequency use.
Feature Selector	Information Gain (IG)	-Measures the number of bits of information obtained for category prediction by determining the presence or absence of a term in a document. -Words are selected according to a higher IG.
	Most Frequent	Selects the most frequent words according to the word's higher frequency.
Sentiment	Avg.	For each user, sentiment is calculated for each tweet using sentence sentiment, and then, average of all tweets is calculated for the user.
	Mixed	Selects a higher sentiment for sentences that are negative or those that positive but have a hidden negative meaning
Use Words	Dept_Sent	Sentiment words, positive and negative ones,are extracted from depressed user's tweets.
	Non-Sparse	Words having more than 95% zeros are removed
Account Measures	As-is	User activities taken as they are (No. of posts,Avg of posts a day, Time of posts, No. of replies, No. of mentions, etc.).
	Norm	Activities are normalized, and average is calculated according to number of user posts.
	Categorical	Activities are categorized according to four quartiles (low, below average, average, and high).
Synonyms	T	Words in the matrix are grouped, and the frequency is summed according to their synonyms.
	F	Words are used as it is without reducing them according to synonyms.

the model. IG is used in machine learning as a term for goodness criterion.

Dept_Sent:

This feature, is inspired by De Choudhury et al. [8] concentrates depressed users' sentiment words. Sentiment words, positive and negative, are extracted from depressed users' tweets and put into files and all other words are removed for all users. While the exploited feature in this study, Dept_Sent, is inspired by De Choudhury [8], it is distinguished by the fact that it does not use static lexicon words for representing depression. Dept_Sent generalizes the depression lexicon and can be extended easily.

Account Measures:

Tsugawa et al. [13] showed that features obtained from user activities can be used to predict user depression with 69% accuracy. In addition, De Choudhury [8] used features obtained from the records of individual user activities on Twitter to identify depressed users. Tsugawa et al. and Del Vicario et al. [13], [34] indicated that the more a user is active, the higher is his/her tendency to express negative emotion when commenting, which will help indicate whether the user is depressed.

As a result, aggregated features are used in this paper to help detect depressed users on Twitter. The activities used for each user in this study are shown in Fig. 4.

Categorical:

For each user, activities will be categorized into four types (low, below average, average, and high), and they are defined using percentile values from quartile distribution (Q1, Q2, and Q3).

Synonyms:

This feature reduces the number of words in the matrix by finding similar words and adding frequencies of synonyms. Tsugawa et al. [13] used the bag-of-words approach to reduce the number of words and found that it helped increase the accuracy. Using WordNet [35], frequencies of synonyms are added; for example, if the depressed frequency is "10," sad is "5," and upset is "3," the frequencies will be added,

Fig. 4 User activity features extracted from user account

which will amount to "18," and put under one word. This will make the word stronger for detecting depression and reduce the number of words in the corpus, thus decreasing the computation time.

5. Experiment and Results

5.1 Evaluation Measures

Critique and cross validation of the feasibility of this automated predictions will be through standard accuracy (Acc), precision (P), recall (R), and F1 scores as well as through the confusion matrix (CM) and recipient operating classification curves (ROC), which are defined as follows.

$$Acc = \frac{truePositive + trueNegative}{truePositive + trueNegative + falsePositive + falseNegative}.$$

$$P = \frac{truePositives}{truePostives + falsePostives}$$

$$R = \frac{truePositives}{truePostives + falseNegatives}$$

The tradeoff between R (false negatives) and P (false positive) has been compromised by considering the F1-Measure.

$$F1 = \frac{2 \times P \times R}{P + R}$$

Hence, it is important to achieve both high recall and high precision.
The CM can be used to generate a point in the ROC space using metrics that are defined as follows [36]:

$$CM = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

The ROC assessment technique uses the TP rate and the FP rate, which are defined as follows [17]:

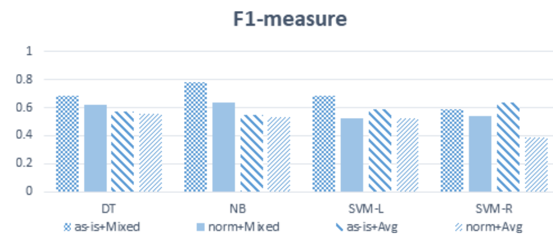


Fig. 5 F1 measures for different algorithms

$$TP\ rate = \frac{TruePositive}{ActualPositives}$$

$$FP\ rate = \frac{FalsePositive}{ActualNegatives}$$

5.2 Data Collection

We use a dataset for Twitter users who suffer from depression. The self-reports are collected by searching Twitter using a regular expression ("diagnosed with depression."). All Tweets are chosen to be in English. Candidate users are filtered manually. The manual labeling is done unanimously between two different psychologists. If any case has conflicts between them, it has been eliminated from final dataset. Later, all their recent tweets are continuously crawled using the Twitter Search API. Total number of 500 users were collected with more than 1M tweets, 334 users were classified as depressed. To ensure that users are disclosing their own depression and not talking about a friend or a family member, a human annotator reviews these tweets. For each user, up to 3000 of their most recent public tweets are included in the dataset, and each user is isolated from the others. Note that this 3000-tweet limit is derived from Twitter's archival polices [37]. Non-depressed users are collected randomly and checked manually to ensure they never post any tweet containing the character string "depress." In an effort to minimize noisy and unreliable data, users with fewer than five Twitter posts are excluded.

5.3 Results

The experiments are conducted on all possible combinations of feature values, using various classification algorithms (SVM with different kernels, DT, and NB). The expected labels for any training/testing sample are depressed/not depressed.

From previous studies, "first-person pronouns" and "TF-IDF" have been proven to be discriminant for depression identification. This finding has been proven during experiments run via measuring the correlation between features and class labels. For obtaining optimal classification results, various feature combinations are exploited. First, account measures (as-is and normalization) with sentiment features (mixed, average, and none of them) are used for training and testing. Figure 5 summarizes the F1 measures for different algorithms.

Figure 6 shows that using mixed sentiments with the

Table 2 Evaluation measures for different feature selectors

	InfoGain				MostFreq			
	Precision	**Recall**	**F1-measure**	**AUC**	**Precision**	**Recall**	**F1-measure**	**AUC**
DT	0.5625	0.4737	0.5143	0.5702	0.6364	0.3889	0.4828	0.6035
NB	0.8571	0.3529	0.5	0.6547	1	0.3	0.4615	0.65
SVM-L	0.6190	0.6842	0.65	0.6516	0.6	0.5217	0.558	0.5256
SVM-K	0.6087	0.7	0.6512	0.625	0.6190	0.6842	0.65	0.6516

Table 3 Evaluation measures for considering "Dept-Sent" feature

Classifiers	**Feature**	**Accuracy**	**Precision**	**Recall**	**F1-measure**	**AUC**
DT	Without Dept-Sent	75	0.7647	0.5652	0.65	0.6649
	With Dept-Sent	72.5	0.4736	0.6923	0.5625	0.6609
NB	Without Dept-Sent	65	0.7692	0.4545	0.5714	0.6439
	With Dept-Sent	67.5	0.5714	0.6	0.5854	0.575
SVM-L	Without Dept-Sent	75	0.6875	0.5	0.5789	0.6111
	With Dept-Sent	72.5	**0.681818**	**0.625**	**0.6521**	**0.5937**
SVM-R	Without Dept-Sent	75	0.6315	0.5714	0.6	0.6015
	With Dept-Sent	75	0.7333	0.55	0.6285	0.675

Table 4 Evaluation measures for considering "Categorical" feature

Classifiers	**Feature**	**Accuracy**	**Precision**	**Recall**	**F1-measure**	**AUC**
DT	With Dept-Sent	72.5	0.4736	0.6923	0.5625	0.6609
	Dept-Sent and Categ	70	0.6315	0.6315	0.6315	0.6491
NB	With Dept-Sent	67.5	0.5714	0.6	0.5853	0.575
	Dept-Sent and Categ	75	0.6842	0.7222	0.7027	0.7247
SVM-L	With Dept-Sent	72.5	0.6818	0.625	0.6521	0.5937
	Dept-Sent and Categ	72.5	0.6667	0.6363	0.6511	0.6237
SVM-R	With Dept-Sent	75	0.7333	0.55	0.6285	0.675
	Dept-Sent and Categ	77.5	**0.7368**	**0.7778**	**0.7567**	**0.7752**

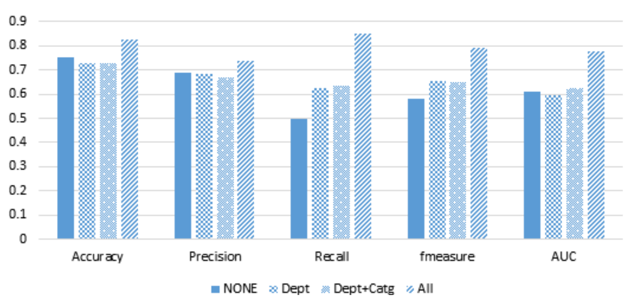


Fig. 6 Results of applying all features to SVM-L

account measures as they are (as-is) produces a higher F1 measure for all the exploited classification algorithms. The F-measure was 0.784 for the NB and 0.68 for both DT and SVM-L.

In addition, training and testing are conducted to reach out the optimal feature selector (InfoGain and Most-Freq). The experimental results show that InfoGain provides higher f-measure results for all classification models, except for the SVM radial kernel, as shown in Table 2.

Another experiment considers the "DeptSent" feature for the obtained data, and the results show that the accuracy and f-measure increase for all classifiers, except for DT. The F measure with the DT classifier decreases due to a decrease in precision; meanwhile, the recall increases from 0.56 to 0.69 (less false negatives). In addition, the F measure increases from 0.57 to 0.65 with SVM-L, and shows a slight up-lift with NB and SVM-R, as indicated in Table 3.

Table 3 shows the trade-off between recall and precision for each model. SVM shows lower variance between the two measurements.

The next step in the experiment is considering the "Categoral" feature, which is a newly derived feature that has not been considered previously. Exploiting this feature shows a noticeable increase in the F measure with all classification models, except for SVM-L, which shows no change but a slight increase in the recall, as shown in Table 4. These results support De Choudhury [8], as he concluded that social media activity provides useful signals that can be utilized to classify whether an individual is suffering or will suffer from depression.

The results support the same conclusion: although the dataset becomes richer, the detection algorithm becomes more stable and the trade-off between precision and recall becomes narrower. Moreover, SVM outperforms other detection algorithms with better accuracy.

Besides the other features, the synonyms feature is also applied to the training and testing sets and the classification models show an increase in the F measure, except for SVM-R. Moreover, DT shows a slight decrease but large increase in accuracy. SVM-L and NB show an increase that reaches 82% accuracy with SVM-L and 80% with NB. The F measure is also increased, reaching 0.79 in SVM-L, as listed in Table 5.

After conducting many experiments with variations in the exploited features, the results emphasize that enriching the model with discriminant features yields better results.

Table 5 Evaluation measures for considering "Synonyms" feature

Classifiers	Feature	Accuracy	Precision	Recall	F1-measure	AUC
DT	+ Categ	70	0.6315	0.6315	0.6315	0.6491
	Categ+ Synom	77.5	0.65	0.5909	0.6190	0.6010
NB	+ Categ	75	0.6842	0.7222	0.7027	0.7247
	Categ + Synom	80	0.6538	0.8095	0.7234	0.6679
SVM-L	+ Categ	72.5	0.6667	0.6363	0.6511	0.6237
	Categ+ Synom	82.5	**0.7391**	**0.85**	**0.7906**	**0.775**
SVM-R	+ Categ	77.5	0.7368	0.7778	0.7567	0.7752
	Categ + Synom	77.5	0.7058	0.6315	0.6667	0.6967

Table 6 Comparative analysis between the proposed model and previously proposed models

	(De Choudry 2013) [8]	***(Nadeem 2016)*** [10]	***(Reece 2016)*** [9]	***Proposed work***
Dataset	2M tweets 476 users (questionnaire)	2.5M tweets (326 depressed) (self-diagnosed)	105 depressed users out of 204 (CESD scores)	67 depressed users out of 111 (self-diagnosed) and more than 300,000 tweets
ML models exploited	SVM	NB	1200-Random Forest	Linear SVM
Used Features	- LIWC - Sentiment analysis - Social engagement - Language - Social network	- BOW - Sentiment analysis	- LIWC - Sentiment analysis - Time series - LabMT	-LIWC -Sentiment analysis -Social activity -Synonyms
Obtained Accuracy	Accuracy: 70%	Accuracy: 81% Precision: 0.86	Precision: 0.866	Accuracy:82.5% Recall: 0.85

Moreover, the SVM-linear classifier shows best results and invariant behavior, despite its extensive performance complexity.

Figure 6 shows the overall impact of including "Synonyms" in the feature set (SVM-linear), which increases the frequency of words used by depressed people and gives the words strength when fed to the classification model. It also reduces the number of words in the corpus, which decreases the computation time.

The main contribution of this study lies in exploiting the concept of merging both tweets' text besides the account historical activities. Both types of features enable the ML models to monitor the change in the mental and psychological states of the user. Table 6 lists a comparative analysis between the proposed model and the previously proposed ones. From the previous analysis, we conclude that the proposed model outperforms the previously proposed ones in terms of accuracy, due to the diversity and richness of its feature set.

6. Conclusion

This paper defines a binary classification problem as identifying whether a person is depressed, based on his tweets and Twitter profile activity. Different machine learning algorithms are exploited and different feature datasets are explored. Many preprocessing steps are performed, including data preparation and aligning, data labeling, and feature extraction and selection. The SVM model has achieved optimal accuracy metric combinations; it converts an extremely non-linear classification problem into a linearly separable problem. Although the DT model is comprehensive and follows understandable steps, it can fail if exposed to brand-new data. This study can be considered as a step toward building a complete social media-based platform for analyzing and predicting mental and psychological issues and recommending solutions for these users. The main contribution of this study lies in exploiting a rich, diverse, and discriminating feature set that contains both tweet text and behavioral trends of different users. This study can be extended in the future by considering more ML models that are highly unlikely to over-fit the used data and find a more dependable way to measure the features' impact.

Acknowledgments

The authors would also like to thank the Deanship of scientific research for funding and supporting this research through initiative of DSR Graduate Students Research Support (GSR) at King Saud University.

They would also like to thank Dr. Fathy Mostafa from Ain Shams University Specialized Hospital and Dr. Afnan Alwabili from King Faisal Takhassusi Hospital for their greatfull assistance in certifying that Twitter users in the dataset are practicing some depression symptoms.

References

[1] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," IJCAI, pp.3838–3844, 2017.

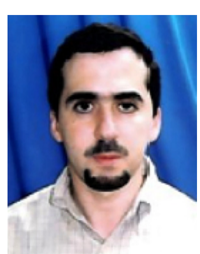
[2] D. Mowery, C. Bryan, and M. Conway, "Feature studies to inform the classification of depressive symptoms from twitter data for population health," arXiv preprint arXiv:1701.08229, 2017.

[3] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, pp.51–60, 2014.

[4] S.J. Stack, "Mental illness and suicide," The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society, pp.1618–1623,

2014.
[5] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, “Discovering shifts to suicidal ideation from mental health content in social media,” Proceedings of the 2016 CHI conference on human factors in computing systems, pp.2098–2110, ACM, 2016.
[6] W.H. Organization et al., “Depression and other common mental disorders: global health estimates,” Tech. Rep., World Health Organization, 2017.
[7] A.H. Yazdavar, H.S. Al-Olimat, T. Banerjee, K. Thirunarayan, and A.P. Sheth, “Analyzing clinical depressive symptoms in twitter,” 2016.
[8] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, “Predicting depression via social media,” Seventh International AAAI Conference on Weblogs and Social Media, 2013.
[9] A.G. Reece, A.J. Reagan, K.L. Lix, P.S. Dodds, C.M. Danforth, and E.J. Langer, “Forecasting the onset and course of mental illness with twitter data,” Scientific reports, vol.7, no.1, p.13006, 2017.
[10] M. Nadeem, “Identifying depression on twitter,” arXiv preprint arXiv:1607.07384, 2016.
[11] G. Harman and M.H. Dredze, “Measuring post traumatic stress disorder in twitter,” ICWSM, 2014.
[12] A. Benton, M. Mitchell, and D. Hovy, “Multi-task learning for mental health using social media text,” arXiv preprint arXiv:1712.03538, 2017.
[13] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, “Recognizing depression from twitter activity,” Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp.3187–3196, ACM, 2015.
[14] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, “Monitoring tweets for depression to detect at-risk users,” Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality, Vancouver, BC, pp.32–40, Association for Computational Linguistics, Aug. 2017.
[15] G. Park, H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, M. Kosinski, D.J. Stillwell, L.H. Ungar, and M.E.P. Seligman, “Automatic personality assessment through social media language,” Journal of personality and social psychology, vol.108, no.6, pp.934–952, 2015.
[16] G. Coppersmith, C. Harman, and M. Dredze, “Measuring post traumatic stress disorder in twitter,” Eighth International AAAI Conference on Weblogs and Social Media, 2014.
[17] Z. Jamil, Monitoring tweets for depression to detect at-risk users, Ph.D. thesis, Université d’Ottawa/University of Ottawa, 2017.
[18] A. Khan, B. Baharudin, L.H. Lee, and K. Khan, “A review of machine learning algorithms for text-documents classification,” Journal of advances in information technology, vol.1, no.1, pp.4–20, 2010.
[19] C.D. Manning, C.D. Manning, and H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
[20] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[21] V. Podgorelec and M. Zorman, “Decision tree learning,” Encyclopedia of Complexity and Systems Science, pp.1–28, 2015.
[22] A.R. Webb, Statistical Pattern Recognition, John Wiley & Sons, 2003.
[23] D.L. Mowery, A. Park, C. Bryan, and M. Conway, “Towards automatically classifying depressive symptoms from twitter data for population health,” Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES), pp.182–191, 2016.
[24] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, “Top 10 algorithms in data mining,” Knowledge and information systems, vol.14, no.1, pp.1–37, 2008.
[25] M.J. Patel, A. Khalaf, and H.J. Aizenstein, “Studying depression using imaging and machine learning methods,” NeuroImage: Clinical, vol.10, pp.115–123, 2016.
[26] N. Cristianini and J. Shawe-Taylor, “An introduction to support vector machines and other kernel-based learning methods,” Cambridge university press, 2000.
[27] R.C. Team et al., “R: A language and environment for statistical computing,” 2013.
[28] R. Team et al., “Rstudio: integrated development for r. rstudio,” Inc., Boston, MA, vol.639, p.640, 2015.
[29] B. Grün and K. Hornik, “topicmodels: An r package for fitting topic models,” Journal of Statistical Software, vol.40, no.13, pp.1–30, 2011.
[30] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, “A depression detection model based on sentiment analysis in micro-blog social network,” Pacific-Asia Conference on Knowledge Discovery and Data Mining, vol.7867, pp.201–213, Springer, 2013.
[31] J. Ramos et al., “Using tf-idf to determine word relevance in document queries,” Proceedings of the First Instructional Conference on Machine Learning, pp.133–142, Piscataway, NJ, 2003.
[32] P. Resnik, W. Armstrong, L. Claudino, and T. Nguyen, “The university of maryland clpsych 2015 shared task system,” Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp.54–60, 2015.
[33] V.M. Prieto, S. Matos, M. Alvarez, F. Cacheda, and J.L. Oliveira, “Twitter: a good place to detect health conditions,” PloS one, vol.9, no.1, p.e86191, 2014.
[34] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, “Echo chambers: Emotional contagion and group polarization on facebook,” Scientific reports, vol.6, p.37825, 2016.
[35] G.A. Miller, WordNet: An Electronic Lexical Database, MIT Press, 1998.
[36] C. Ferri, J. Hernández-Orallo, and R. Modroiu, “An experimental comparison of performance measures for classification,” Pattern Recognition Letters, vol.30, no.1, pp.27–38, 2009.
[37] S. Tsugawa, Y. Mogi, Y. Kikuchi, F. Kishino, K. Fujita, Y. Itoh, and H. Ohsaki, “On estimating depressive tendencies of twitter users utilizing their tweet data,” 2013 IEEE Virtual Reality (VR), pp.1–4, IEEE, 2013.

Hatoon S. Alsagri received the Masters degree in information systems from the Department of Information Systems, College of Computer and Information Sciences, King Saud University, where she is currently pursuing the Ph.D. degree in information systems. She is currently a Lecturer with the Department of Information Systems, College of Computer and Information Sciences, Al-Imam Mohammad bin Saud Islamic University. During her graduate studies, she has had the opportunity to participate in various conferences. Her main research interests include Data Mining, Social Analysis and Information Diffusion.



Mourad Ykhlef is a full professor in the Department of Information Systems, KSU, Saudi Arabia. He got his Ph.D. in Computer Science from university Bordeaux 1 (France). His M.Sc. degree was in Artificial Intelligence from Paris 13 (France). He got B. Eng in Computer Science from Constantine University. His main research interests include Data Mining and Bioinspired computing.