Quantifying Feature Importance for Detecting Depression using Random Forest

Hatoon AlSagri[1]
Information Systems Department[1]
College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic
University Riyadh, Saudi Arabia

Mourad Ykhlef[2]
Information Systems Department[1,2]
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

***Abstract*—Feature selection based on importance is a fundamental step in machine learning models because it serves as a vital technique to orient the use of variables to what is most efficient and effective for a given machine learning model. In this study, an explainable machine learning model based on Random forest, is built to address the problem of identification of depression level for Twitter users. This model reflects its transparency through calculating its feature importance. There are several techniques to quantify the importance of features. However, in this study, random forest is used as both a classifier, which has over-performing aspects over many classifiers such as decision trees, and a method for weighting the input features as their importance imply. In this study, the importance of features is measured using different techniques including random forest, and the results of these techniques are compared. Furthermore, feature importance uses the concept of weighting the input variables inside a complete system for recommending a solution for depressed persons. The experimental results confirm the superiority of random forest over other classifiers using three different methods for measuring the features importance. The accuracy of random forest classification reached 84.7%, and the importance of features increased the classifier accuracy to 84.9%.**

***Keywords*—Machine learning; random forest; feature selection; feature importance; depression; emotions; twitter**

I. Introduction

Depression is a leading cause of disability worldwide and a common mental illness. Globally, more than 300 million people are estimated to suffer from depression every year1.Face-to-face clinical diagnose is need to diagnose depression but 70% of the patients would not consult a doctor when they are at early stages of depression. This might cause patients to reach advance stages in their condition [1].

Several studies have reported that the diagnosis of mental illnesses has increased because of the use of social media platforms [2] [2], and these mental illnesses are one of the leading causes of disability and among the most of the devastating diseases that individuals suffer from worldwide according to the World Health Organization [3], [4], [5]. Therefore, to detect the users at risk for early referral to psychological assistance and treatment, machine learning algorithms have been employed.

Now-a-days, the data collected from millions of Internet of Things (IoT) devices, sensors, social media, etc. enables extremely enriched datasets. Although this is beneficial for machine learning researchers, this makes the data high dimensional it is quite common for datasets to have hundreds of features or more in most of the cases. Therefore, feature selection is an extremely vital process in the machine learning project lifecycle. Feature selections methods help reduce the dimensions without much loss of the total information. In addition, they help in understanding the features and their importance

Previously published papers have demonstrated that exploiting irrelevant features, along with the redundant ones, can impact the accuracy of classification significantly [2], [3], [4].Considering feature selection as a major step in any machine learning algorithm, it contains a step for measuring feature importance. Therefore, an effective feature selection technique that relies on computing the importance of features and remove irrelevant features those that may cause no impact or negative impact [5].

Prominent perspective to feature selection besides enhancing the accuracy, is weighting the features or in other words "feature importance." These weights could be exploited as weighting factors in further steps of recommending a remedy via recommendation techniques. Features' importance represents the statistical significance of each feature and to what extent it contributes to the model.

Random Forest (RF), among other machine learning algorithms, has been an excellent tool to learn feature representations [6], [7] because of its robust classification power and easily interpretable learning mechanism [8]. Features' importance can be estimated using different measures after being computed using RF. In this study, we apply RF as a classifier to detect depressed Twitter users with respect to features extracted from the users' Twitter content and activity. RF has proved to have an accuracy higher than those of the other classifiers namely decision tree (DT), Naïve Bayes (NB), and support vector machine (SVM) (kernel and linear) where they were implemented and tested and gained results of 82% for the SVM linear as the highest accuracy among the others. SVM, DT, and NB presented new features that increased the accuracy of identifying depressed users. By applying RF to the same data, we could find features' importance using three feature importance measures: overall, permutation, and tree interpreter feature importance measures. We were able to conclude that Tree interpreter feature importance measure proved to have the highest accuracy results when RF was recomputed after removing the least important features. In

addition, when the highest important features were removed, the accuracy of the classifier decreased significantly, proving the importance of these features.

Main Contributions of this paper can be summarized as:

- Applying RF to the RRACF model to classify depressed Twitter users more accurately
- Tree interpreter feature importance measure concludes best results of feature importance that has higher effect on classification accuracy.

The remainder of the paper is organized as follows. A literature review is provided in Section 2. Section 3 presents the background of RF and feature importance. Section 4 details the methodology used in this study. Section 5 describes the experiments and results. Finally, Section 6 outlines the conclusions of the study.

II. Literature Review

Efforts to detect mental illness and more specifically depression have increased gradually with the increase in social media usage [9], [10]. Guntuku et al. [11] indicated that tweets containing negative emotional sentiments are posted by depressed Twitter users more than by healthy users

Various studies have used different classifiers to detect depression and other mental illnesses. For clinical outcome prediction using gene expression data, Kong and Yu [8] presented a new classifier, where RF is integrated with deep neural network, and demonstrated that the accuracy is higher compared to those of the other classification models using simulation experiments.

Jotheeswaran and Koteeswaran [12] proved the efficiency of RF on a system developed for emotion detection, knowledge transformation, and predictive analysis using a Twitter dataset. From the experimental results, they concluded that the decision forest-based feature extraction increases the precision of classifier in contrast to decision tree-based feature selection [12].

Reece et al. [9] found that the computational analysis of Twitter data can be used to detect major changes in individual psychology. They extracted predictive features from users' tweets and built models with supervised learning classifiers using these features. The classifiers were trained to distinguish between depression and post-traumatic stress disorder (PTSD) in affected and healthy users.

The 1200tree RF outperformed other classifiers by exhibiting accuracy results higher than those of the classifiers used by Mitchell et al. [13] and Choudhury et al. [14] in depression classification reaching (0.866) and by Taubman-Ben-Ari et al. [15] and Nadeem [16] in PTSD classification reaching (0.934) [9].

Sau et al. [17]conducted a study to predict anxiety and depression among geriatric population and concluded that the RF algorithm delivers the best results with a predictive accuracy of 90% compared to the other machine learning classifiers.

TABLE I. Taxonomy of Random Forest Approaches for Feature Selection and Depression

Author	Technique	Data	Assessment
Mowery et al.,[18]	RF + DT	Twitter	Depression symptoms
Kong and Yu,[8]	RF + neural network	gene expression data	Feature representation for ranking gene importance
Reece et al.,[9]	RF	Twitter	Rank predictive features Depression and PTSD
Sau et al ,[17]	RF	geriatric patients evaluated for depression and anxiety	Predicting depression and anxiety

Mowery et al. [18] demonstrated that the machine learning algorithms used with Twitter data improved precision in detecting symptoms of depression compared to the use of keywords alone. Decision trees and RF resulted in a higher precision than that achieved by other machine learning algorithms.

Table I indicates approaches of RF for feature selection and depression. Mowery et al. [18] showed that feature representation increased classification of depressed people. Also, Kong & Yu [8] indicated that using RF to represent features fed to deep neural network increased the accuracy of the system.

Similarly, our study uses RF to find important features but using feature importance measures that up to our knowledge, has not been introduced for detection of depression.

III. Random Forest and Feature Importance Background

A. Random Forest (RF)

RF is an ensemble learning classification algorithm developed from multiple sub-decision trees [19]. The sub-decision trees are built using bagging and feature randomness to create an uncorrelated forest of trees that have a higher accuracy in prediction than that of any individual tree [20].Bagging is a common ensemble method that uses bootstrap sampling. Using the bagging procedure, a number of decision trees are generated from the original sample set through bootstrap sampling, and the features that are randomly selected from the original set are used for partitioning at each node [6], [21].Node is an elementary unit in any tree based algorithm. RF reduces the likelihood of over fitting generated during the use of single decision tree model [5]. In addition, the use of bootstrap sampling helps produce an optimal generalization ability and a higher accuracy classification model [19].

B. Features' Importance

Along with improving the accuracy that has been shown in a majority of RF studies, RF provides feature importance measures as one of its useful derivatives that has contributed to its popularity [22], [23], [24], [25], [26]. Feature importance measures are the Overall, Permutation, and Tree Interpreter feature importance [25], [26].

Overall feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node impurity should be decreased since we are going deeper into tree levels, so the impact of node can be

objectively quantified by the drop of impurity through the node. Gini impurity is calculated for each node where it is possible to calculate the node probability based on the number of samples that reach the node divided by the total number. In this case, higher values correspond to more important features. Overall feature importance starts by:

1) Calculating nodes importance n_j of node j for every decision tree, using the following equation:

$$n_j = W_j C_j - W_{left\ (j)} C_{left\ (j)} - W_{right(j)} C_{right(j)} \quad (1)$$

$where\ W_j$ is the node j reachability probability and C_j is Gini impurity of the node. The same is for the *right* node and *left* node children of node j.

2) The importance of each feature (F) in the tree is calculated using Eq. 2, where *m* is total number of nodes:

$$F(j) = \frac{n_j}{\sum_{i=1}^{m} n_i} \quad (2)$$

3) The importance for each feature in RF (collection of k Trees) is calculated using the following equation:

$$\text{Feature Importance } (i) = \frac{\sum_{j=1}^{m} F(j)}{k} \quad (3)$$

Permutation Features Importance: starts by training the baseline model and recording the score by evaluating the validation set or training set.

For all features in features set, do:

1) Re-shuffle one feature values in the evaluated dataset.
2) Re-pass the dataset to the model and re-calculate the metric for the modified dataset.
3) The feature importance is computed as the difference between the benchmark score and the score from the permuted dataset.

Tree Interpreter Features Importance: begins withtraining the baseline model and recording the score by evaluating the validation set or training set. For all features in features set do:

1) Drop the node (feature)
2) Re-pass the dataset to the model and re-calculate the metric for the modified dataset.
3) The feature importance is computed as the difference between the benchmark score and the score from the modified dataset.

IV. Methodology

In this study, we focus on identifying the importance of features that help detect depression from users' accounts including both tweets' content and activity. The model contains different hyper parameters such as number of trees, depth, validation set, etc. The optimal combination among these hyper parameters has been found through executing exhaustive grid search. The system depression detection using activity and content features -random forest (DDACF-RF) proposed is in Fig. 1.This system uses RF for classifying users' mental conditions and identifying the importance of features. Data preparation, feature extraction, and classification tasks are performed using various R packages, and in R version 3.3 [27], they are performed using Rstudio IDE [28]. The RF classifier is trained using 10-fold cross validation, each contains both training and testing set, to avoid over fitting, and it is then tested on a held-out test set Initially, all tweets from the accounts of depressed and non-depressed users are retrieved along with their information and activities such as number of followers, number of following, and total number of posts. Next, text preprocessing is applied to all the documents through tokenization, normalization, and stemming which is done through splitting words, removing punctuation and returning word to its stem. Then, a document term matrix (DTM), which designates the frequency of words in each tweet, is created for each account. The weights of words are measured using Term Frequency-Inverse Document Frequency(TF-IDF). The features applied on the DTM are then merged with the account measures extracted from the social network and user activities as illustrated in Fig. 3. Finally, the results of the merge are treated as independent variables in the RF classification algorithm to predict whether a user is depressed or not. Fig. 1 illustrates the DDACF-RF classification model.

Three different feature importance measures are applied to find the best importance measure to weigh the features. Diaz-Uriarte and Alvarez de Andres strategy is later applied to conclude the best feature importance measure among the three [29]. Dıaz-Uriarte and Alvarez de Andres strategy depended on computing RF, then removing 20% of the most important features, and then computing RF again [29], [30]. In this study, we removed the most and least important features and then recomputed RF in both cases. Using the three different feature importance measures discussed previously, the importance of features was calculated; then, the most and least important features were removed independently and RF was recomputed.

A. Data Collection

This study dataset concentrates on Twitter users who suffer from depression. Using a regular expression ("diagnosed with depression"), self-reported tweets are collected from Twitter. All Tweets are chosen to be in English and gathered in May-July 2018. Candidate users are filtered manually. These tweets are then processed by a human annotator to certify that the users are revealing their own depression and not talking about someone else. 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. 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 [31]. Non-depressed users are collected randomly and checked manually to ensure that they have never posted any tweet containing the character string "depress". In an effort to minimize the noisy and unreliable data, users with fewer than five Twitter posts are excluded.

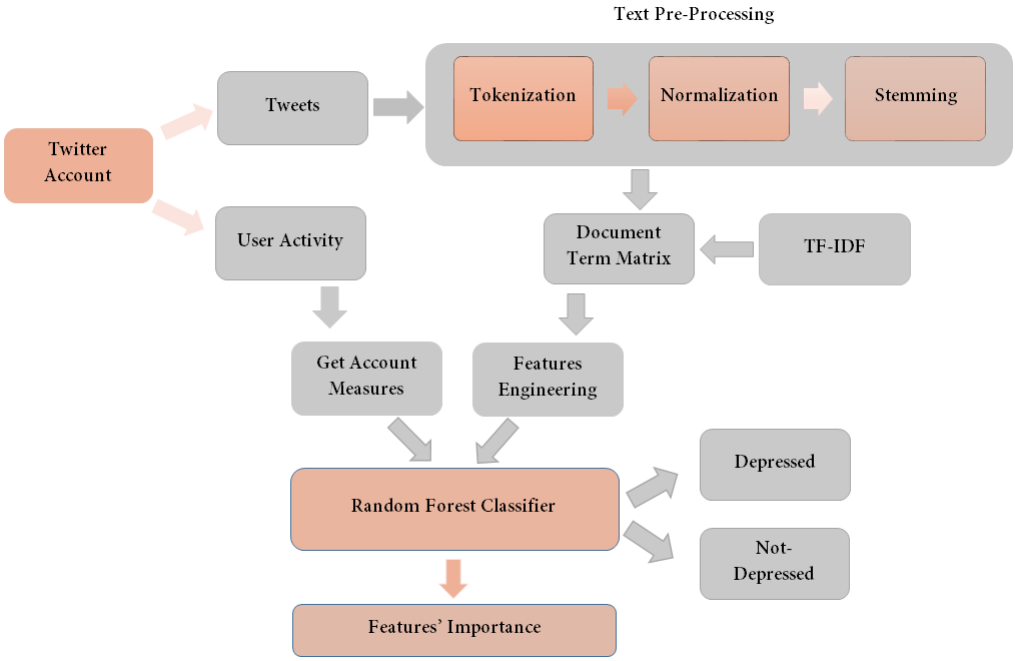

Fig. 1. DDACF-RF Classification Model

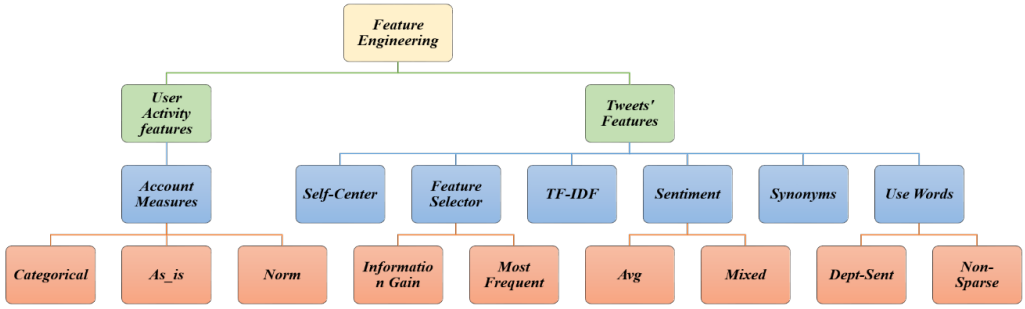

Fig. 2. Visualization of Features used in the Study

B. Feature Engineering

In machine learning, feature engineering is referred to as "the process of using domain knowledge of the data to create features that can be used by machine learning algorithms to find patterns" [10]. The information that are recognized by machine learning and might be beneficial for prediction are extracted by generating the features [10].

The activity histories and tweets of Twitter users are used to extract various features reaching more than 150 thousands input features (raw features) containing words from tweets and account activities Fig. 3. This information undergoes preprocessing Fig. 1 before the engineered features are obtained, and once the engineered features are obtained Fig. 4, they are computed for both the training and test sets. Fig. 4 shows the features obtained from the tweets and activities of user accounts. These features are used as the variables for the classification model. Table II lists the features and their possible values used for the classification model, Where T (true) and F (false) for possible values indicate the use of this feature or not. For example, when the possible value for TF-IDF is T meaning TF-IDF is used for the experiment and if it's F that means word frequency is used instead.

C. Self-Center

Previous studies have shown that first-person pronouns are useful predictors of depression. De Choudhury and Jamil [32], [10] indicated that the use of singular pronouns in comparison to second- and third-person pronouns is also an indicator of depression. Thus, we skip removing the first-person pronouns with other stop words in the normalization step in the proposed classification model to increase the efficiency of the classification algorithm.

Fig. 3. User Activity Features Extracted from User Account

TABLE II. DESCRIPTION OF FEATURES AND THEIR POSSIBLE VALUES

Features	*Possible Values*	*Description*
Self-Center	T	Use first-person pronouns.
	F	Remove all stop words associated with the first-person pronouns.
TF-IDF	T	Determine the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document corpus.
	F	Use word frequency.
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 the higher IG.
	Most Frequent	Selects the most frequent words according to the words' higher frequency.
Sentiment	Avg	For each user, sentiment is calculated for each tweet using sentence sentiment, and then, the average of all tweets is calculated.
	Mixed	Selects a higher sentiment for sentences that are negative or positive with a hidden negative indication.
Use-Words	Dept-Sent	Sentiment words—positive and negative words—extracted from depressed user's tweets.
	Non-Sparse	Words with zeros more than 95% are removed.
Account Measures	As-is	User activities are taken as they are (number of posts, average number of posts a day, time of posts, number of replies, number of mentions, etc.).
	Norm	Activities are normalized, and average is calculated according to the number of user posts.
	Categorical	Activities are categorized according to 4 quartiles (low, below average, average, and high).
Synonyms	T	Words in the matrix are grouped, and the frequency is added based on their synonyms.
	F	Words are used as they are without reducing them based on their synonyms.

D. Feature Selector

For selecting features there are two possible values, either Most-frequent which select the most frequent words according to the words' higher frequency or Information gain. Inspired by Prieto et al. [21], information gain (IG) is added as a feature selector for the model. Prieto et al. [21] used IG to reduce features that improve the classification of depressed users, and reduced the time needed to generate the model. IG is used in machine learning as a term for goodness criterion.

E. Sentiment

Sentence sentiment is used for each tweet in the user's account, then the average of all tweets' sentiment is calculated and this is the Avg feature. Mixed feature calculates sentence sentiment for sentences that are either negative or positive but have hidden negative indication.

F. Use Words

This feature has two possible values, either non-sparse meaning non-sparse words are used and sparse words having more than 95% zeros are removed, or Dept_Sent. Depression Sentiment (Dept_Sent) is a feature, inspired by De Choudhury et al. [32], concentrates on depressed users' sentiment words. From tweets crawled for this study, sentiment words, positive and negative, are extracted from depressed users' tweets and put into files and all other words are removed for all users. The exploited feature in this study, Dept_Sent, 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.

G. Account Measures

Tsugawa et al. [31] showed that features obtained from user activities can be used to predict user depression with 69% accuracy. In addition, De Choudhury [32] used features obtained from the records of individual user activities on Twitter to identify depressed users. Tsugawa et al. and Del Vicario et al. [31], [33] 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. Activities extracted from each user account such as retweets, mentions, etc. used in this study are shown in Fig. 2.

Three different possible values for this feature (As_is, Norm, Categorical). As_is uses user's activities as it is while Norm uses the activities after calculating the average according to the number of user's posts. Categorical is a new feature that has been introduced uniquely in this study. It relies on categorizing activities of each user into four types (low, below average, average, and high), whose delimiters are defined using percentile values from quartile distribution (Q1, Q2, and Q3).

H. Synonyms

Tsugawa et al. [31] used the bag-of-words approach to reduce the number of words and found that it helped increase

TABLE III. DISTRIBUTION OF POSSIBLE VALUES OF EACH FEATURE

Feature	*Possible Value (v)*	*Distribution p(v)*
Self-Center	T	0.54
	F	0.45
TF-IDF	T	0.53
	F	0.46
Feature Selector	Information gain (IG)	0.33
	Most Frequent	0.33
	None	0.34
Sentiment	Avg	0.33
	Mixed	0.33
	None	0.34
Use-Words	Dept-Sent	0.45
	Non-Sparse	0.54
Account Measures	As-is	0.34
	Norm	0.33
	Categorical	0.33
Synonyms	T	0.09
	F	0.91

the accuracy. This feature reduces the number of words in the matrix by finding similar words and adding frequencies of synonyms, using Word Net.

This will make the word stronger for detecting depression and reduce the number of words in the corpus, thus decreasing the computation time.

Tree based methods have been picked for this study due to the categorical nature of the features Table III.

TABLE IV. RESULTS OF RANDOM FOREST CLASSIFICATION MODEL EXPERIMENTS

Features	*Accuracy %*	*Precision*	*Recall*	*f-measure*	*RF tree*
Initial features	67.8	0.36	0.615	0.457	2000
Dept-Sent	69.5	0.38	0.615	0.470	2000
Dept-Sent +Categ	72.9	0.45	0.642	0.529	2000
Dept-Sent +Categ +Synonyms	84.7	0.52	0.9	0.667	2000

TABLE V. COMPARISON OF RF RESULTS WITH THOSE OF OTHER CLASSIFIERS

Classifier	*Accuracy %*	*Precision*	*Recall*	*f-measure*
DT	77.5	0.65	0.59	0.619
NB	80	0.65	0.81	0.723
SVM-L	82.5	0.74	0.85	0.791
SVM-R	77.5	0.71	0.63	0.667
RF	84.7	0.53	0.9	0.667

Metrics	**Formula**
Accuracy	$Acc = \frac{truepositives + truenegatives}{truepositives + truenegatives + falsepositives + falsenegatives}$
Precision	$P = \frac{truepositives}{truepostives + falsepostives}$
Recall	$R = \frac{truepositives}{truepositives + falsenegatives}$
F-measure	$F1 = \frac{2 * P * R}{P + R}$

V. EXPERIMENTAL RESULTS

A. Results

This study was conducted on all possible combinations of feature values, using RF classifier. The expected labels for any training/testing sample are depressed/not depressed. Feature importance, used to find the features that mostly help increase the classification accuracy and determine the user's mental condition, was an important result of the study.

Feature engineering used with NB, DT, and SVM used for detecting depressed users proved that utilizing a rich, diverse, and discriminating feature set that contains both tweet text and behavioral trends of different users helped increase the classification accuracy.

For that this study follows the same experimental steps and proves that the conclusion evaluation metrics increased when new features were added.

The experiments are compared with respect to four metrics, namely accuracy, precision, recall, and f-measure. All the experiments used "first-person pronouns" and "TF-IDF" that have already been proven to discriminant for depression identification [32]. In addition, considering "InfoGain" as the feature selector and "mixed" as the sentiment feature, the results were obtained for the first experiment. Further, in the second experiment "Dept-Sent" was added as feature along with features from first experiment. "Categorical" as the account measures feature was added in the third experiment, and "synonyms" in the last experiment. As a result, we observed an increase in all evaluation measures, where the accuracy reached 84.7% and recall was 0.9 as shown in Table IV.

Table V reveals the increase in accuracy and recall obtained using RF when compared to the other classifiers used in our previous work[34].

After training the RF, the importance of features was found to have a significant impact on the outcome values. Feature importance measures help find each feature's importance as a measure by which the accuracy is decreased when that feature is removed and vice versa—by which the accuracy is increased when that feature is included. Fig. 4 shows the most important

TABLE VI. EFFECT OF THE REMOVAL OF MOST AND LEAST IMPORTANT FEATURES ON THE CLASSIFIER RESULTS

Features Importance Measure	*Removing most important features*	*Removing least important features*
Overall	78.1	83.81
Permutation	77.85	83.29
Tree Interpreter	72.9	84.908

features developed from the model.

We can conclude that the account measures (retweets, hash tags, ...) and the words extracted from users' contents have great significance on the detection of depressed users as their importance indicates. It's noticeable that the number of retweets appears with high importance in identification of the depression level. However, it can be explained through that as much time the user is online on Twitter or having more interactions reflects how much he is disconnected in reality. The three different feature importance measures resulted in different outcomes. These measures were then validated to obtain the best way of calculating feature's importance using Diaz-Uriarte and Alvarez de Andres strategy [29], [30].

The five most and least important features were removed independently and RF was recomputed. For example, the overall feature importance measure was used to compute RF. From the outcomes of overall feature importance, the five most important features were removed and RF was recomputed, and then, the five least important features were removed and RF was recomputed. Same strategy was repeated for the permutation and tree interpreter feature importance measures. On applying the strategy to the sample with an accuracy of (84.7), we found that the tree interpreter feature importance exhibited the highest accuracy (84.908) when the five least important features were removed and exhibited the least accuracy (72.9) when the five most important features were removed. Results of all feature importance measures are summarized in Table VI. shows the increase in accuracy of the classifier results after removing the least important features, which are less significant to the model. In addition, it shows the decrease in accuracy of the classifier results when the five most important features, demonstrating the importance of these features to the model efficiency, are removed. From the results, we can observe that the tree interpreter importance measure exhibits the highest results.

Table VI results show that the higher the decrease in accuracy reveals that the features removed were more important. It shows that tree interpreter found the most important features which caused more decrease in the accuracy than the overall and permutation importance measures. Also, removing the least important features increase the accuracy showing that tree importance measure was able to find the least important features that needed to be removed to get better classification accuracy.

The increase in accuracy was very small which was sufficient for our study to find and validate the best and most representative feature importance measure aiming to find a quantitative method to weight features. This method is needed in future work and will be employed to find a remedy for depressed Twitter users.

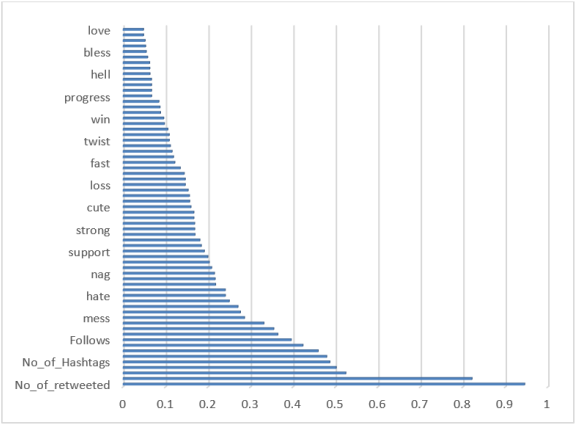

Fig. 4. Importance of Features

VI. CONCLUSION

RF has proven to be an efficient classifier with respect to DT, NB, SVM-L, and SVM-R. In addition, it offers feature importance as an average gain achieved during forest construction. The feature importance revealed the features that do not add value to the classifier's performance and helped increase the accuracy. The uniqueness of this study was indicated in the different importance measures used, where the tree interpreter importance measure outperformed the other importance measures. The application of importance measures to the features extracted from both tweets and activities of user accounts helped classify the depressed users in the dataset more accurately.

Results of this study prove the benefit of feature importance in obtaining the best solution for depressed people and for mentally ill people in general. In future study, feature importance can be used to obtain the values of features that increase the efficiency of any model. In addition, the least important features that decrease the productivity and increase the time elapsed to obtain desired results can be eliminated in early stages of the study.

ACKNOWLEDGMENTS

The authors would 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 Specialist Hospital for their grateful assistance in certifying that Twitter users in the dataset are practicing.

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." in *IJCAI*, 2017, pp. 3838–3844.

[2] D. Chutia, D. K. Bhattacharyya, J. Sarma, and P. N. L. Raju, "An effective ensemble classification framework using random forests and a correlation based feature selection technique," *Transactions in GIS*, vol. 21, no. 6, pp. 1165–1178, 2017.

[3] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 Science and Information Conference*. IEEE, 2014, pp. 372–378.

[4] A. Chinnaswamy and R. Srinivasan, "Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data," in *Innovations in bio-inspired computing and applications*. Springer, 2016, pp. 229–239.

[5] S. Nakariyakul, "High-dimensional hybrid feature selection using interaction information-guided search," *Knowledge-Based Systems*, vol. 145, pp. 59–66, 2018.

[6] A. Tang and J. T. Foong, "A qualitative evaluation of random forest feature learning," in *Recent Advances on Soft Computing and Data Mining*. Springer, 2014, pp. 359–368.

[7] C. Vens and F. Costa, "Random forest based feature induction," in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 744–753.

[8] Y. Kong and T. Yu, "A deep neural network model using random forest to extract feature representation for gene expression data classification," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.

[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, pp. 1–11, 2017.

[10] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression to detect at-risk users," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, 2017, pp. 32–40.

[11] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.

[12] J. Jotheeswaran and S. Koteeswaran, "Feature selection using random forest method for sentiment analysis," *Indian Journal of Science and Technology*, vol. 9, no. 3, pp. 1–7, 2016.

[13] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," *The Lancet*, vol. 374, no. 9690, pp. 609–619, 2009.

[14] N. Choudhury and S. A. Begum, "A survey on case-based reasoning in medicine," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 8, pp. 136–144, 2016.

[15] O. Taubman-Ben-Ari, J. Rabinowitz, D. Feldman, and R. Vaturi, "Post-traumatic stress disorder in primary-care settings: prevalence and physicians' detection," *Psychological medicine*, vol. 31, no. 3, pp. 555–560, 2001.

[16] M. Nadeem, "Identifying depression on twitter," *arXiv preprint arXiv:1607.07384*, 2016.

[17] A. Sau and I. Bhakta, "Predicting anxiety and depression in elderly patients using machine learning technology," *Healthcare Technology Letters*, vol. 4, no. 6, pp. 238–243, 2017.

[18] D. L. Mowery, Y. A. Park, C. Bryan, and M. Conway, "Towards automatically classifying depressive symptoms from twitter data for population health," in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 2016, pp. 182–191.

[19] H. Cai, Y. Chen, J. Han, X. Zhang, and B. Hu, "Study on feature selection methods for depression detection using three-electrode eeg data," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 10, no. 3, pp. 558–565, 2018.

[20] T. Yiu, "Understanding random forest," *Medium*, 2012.

[21] 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, 2014.

[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[24] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[25] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures," in *Workshop on Statistical Modelling of Complex Systems*. Citeseer, 2006.

[26] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[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] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[30] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[31] S. Tsugawa, Y. Mogi, Y. Kikuchi, F. Kishino, K. Fujita, Y. Itoh, and H. Ohsaki, "On estimating depressive tendency of twitter users from their tweet data," in *IEEE Virtual Reality*, vol. 2, 2013, pp. 29–32.

[32] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[33] 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.

[34] H. AlSagri and M. Ykhlef, "Machine learning-based approach for depression detection in twitter using content and activity features," *IEICE TRANSACTIONS on Information and Systems*, vol. E103-D, no. 07, 2020.