Clustering as feature selection method in spam classification: uncovering sick-leave sellers

Mariam Elhussein and Samiha Brahimi
Computer Information Systems Department, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

Received 11 September 2021
Revised 28 October 2021
14 November 2021
Accepted 20 November 2021

Abstract

Purpose – This paper aims to propose a novel way of using textual clustering as a feature selection method. It is applied to identify the most important keywords in the profile classification. The method is demonstrated through the problem of sick-leave promoters on Twitter.

Design/methodology/approach – Four machine learning classifiers were used on a total of 35,578 tweets posted on Twitter. The data were manually labeled into two categories: promoter and nonpromoter. Classification performance was compared when the proposed clustering feature selection approach and the standard feature selection were applied.

Findings – Radom forest achieved the highest accuracy of 95.91% higher than similar work compared. Furthermore, using clustering as a feature selection method improved the Sensitivity of the model from 73.83% to 98.79%. Sensitivity (recall) is the most important measure of classifier performance when detecting promoters' accounts that have spam-like behavior.

Research limitations/implications – The method applied is novel, more testing is needed in other datasets before generalizing its results.

Practical implications – The model applied can be used by Saudi authorities to report on the accounts that sell sick-leaves online.

Originality/value – The research is proposing a new way textual clustering can be used in feature selection.

Keywords Profile classification, Supervised classification, Twitter, Clustering, Saudi Arabia

Paper type Research paper

1. Introduction

When not reporting to work, employees are expected to present proof if they claim to have had a medical condition. Sick leaves are documents provided by medical facilities issued by a doctor certifying that the person is suffering from a condition that allows them days off. Some employees and students abuse this allowance and issue documents illegally to have free day(s) off. In Saudi Arabia, employee absenteeism has been an issue for some time now. The government is combating the issuance of these documents by designing laws and regulations [1]. Despite these efforts, this type of documents is still being circulated. One mean of connecting to those who sell these documents is through Twitter. Promoters are accounts that sell these documents on social media. Since it is illegal, most of the accounts are either fake or pseudo accounts. Sick-leave promoters' tweeting behavior can be comparable to spamming behavior. Spammers tend to repeat the exact text multiple times within a short period of time [1].

Data availability: The data sets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-09-2021-0248

They use multiple hashtags, and they also capitalize on trending hashtags to gain exposure. For these reasons, promoters of sick-leave documents are treated in the same manner as spam accounts in this research.

Machine learning algorithms were employed to detect spam accounts on Twitter. Some of these approaches rely on features extracted from tweets, while others utilized textual content of tweets. This paper is attempting to uncover those who are involved in the sick-leave deception in Twitter. It contributes to this effort by:

(1) Analyzing a data set of 15,578 tweets downloaded from 2010 until January of 2021 and manually labels them. This is used as ground truth data.

(2) Using K-means clustering as feature selection approach to improve the performance of the classification model.

(3) Using the textual tweets of sick-leave promoters, construct a classification model by applying supervised learning algorithms. Four classifiers are used, including Decision Tree DT, Random Forest RF, Naïve Bayes NB, and Logistic Regression LR.

(4) Identifying the list of keywords that are most effective in revealing promoters.

This paper continues as follows: Related work of relevant research is presented, followed by proposed scheme, experiments and results, evaluation, and finally conclusion, implications and future work.

2. Related work

According to Twitter, profile detection is "the attempt to automatically infer the values of user attributes by leveraging observable information such as user behavior, network structure, and the linguistic content of the user's Twitter feed." [2]. Profile detection has been presented in many studies to distinguish the owner of the profile based on their interest and profile information. The detection efforts are mostly binary where the researchers want to identify whether or not the user is playing a certain role (male or female, bot or human, organization or individual, spam or not spam) [3].

Detecting spam accounts on Twitter can follow one or hybrid approaches of analysis based on: time-series analysis of tweets and interactions [4], features extracted from the user profile [5] and the text posted (tweets) [6, 7]. The first type applies time-series analysis to reveal trends. This can be the search of specific terms' count within a period of time such as in. The second type use features extracted from the user profile. Researchers investigate through profile interaction and the content of the tweet whether the account belongs to a human or a bot.

The third type is known as content analysis approach where tweet text is used to detect spam content. The analysis of text start by Bag-of-Words analysis, a popular approach to identify the k-top words in user groups [8]. Alternatively, studies use n-gram character features, unsupervised learning such as LDA and ensemble approach [9]. Content analysis of tweets also focuses on the fact that spammers on Twitter use malicious links. Therefore, the use of blacklist URLs is also another method applied [10].

Feature selection represents an important tool to balance the number of selected attributes to avoid overfitting the model (with too few attributes) and expensive computational time (with too many attributes). There are many methods for feature selection such as: wrapper methods [11], filter methods and unsupervised methods. Wrapper and filter methods are considered supervised approaches as they utilize the output to produce the best set of features. With textual data, unsupervised feature selection has been applied namely K-means clustering to select the best set of features with high-frequency words [12]. Four corpuses were experimented and test using three classifiers. SVM was found to have achieved better

performance when clustering was applied. Another approach involved selecting a list of features using *K*-means clustering and correlation analysis [13]. Using two textual data sets, NB showed improvement in accuracy. None of these studies applied *K*-means clustering as feature selection to classify profiles. Furthermore, all the text used consisted of lengthy documents and news databases. None of the texts used were short text (tweets). Table 1 summarizes these studies.

3. Proposed scheme

In this section, the proposed scheme is introduced, but first, an explanation of how data was collected.

3.1 Data collection

The data were retrieved by specifying a list of keywords that were identified using www.hashtagify.me. It is a tool that provides a list of relevant hashtags that are used frequently together. The keyword that was used to start finding the list was "سكليف", a transliteration of the word sick leave that is commonly used to refer to the document that is obtained. The retrieved list included keywords that are either the Arabic version of the word or variation with similar meaning. A total of nine words were used and a tweet is retrieved if it contained at least one of the nine. The location of the tweet was setup to be in Saudi Arabia as a condition for it to be selected.

Tweets between January 1, 2010, and January 8, 2021, were downloaded. The data have been manually labeled using two categories: promoter and nonpromoter. It was noticed from the data that people would write about sick leaves to either joke about needing a sick leave, promoting their sick-leave business, or ask for a sick leave. The majority of the tweets were humoring about needing a sick leave, very few were asking for one. For that reason, jokers and those who ask for sick leaves are considered as one category (nonpromoters) and the rest are promoters. Tweets were obtained and were ready for cleaning and preprocessing; 2, 413 tweets were identified as promoter tweets. The cleaning and preprocessing of the tweets included removing duplicates. Unifying the characters that contain such marks, like (إ, أ, آ, ٱ to be ا) and remove links and emojis. Table 1 details the list of attributes in the data set.

Research	How clustering was used	Contribution	Data set
Guru *et al.* (2019) [14]	TCR method to lower dimensionality	SVM classifier performed better	Reuters dataset
Chormunge and Jena (2018) [15]	Eliminate irrelevant features using feature clustering and cross correlation	NB accuracy improved	12 data sets of microarrays and texts
Malji *et al.* (2017) [16]	Improve processing time of feature selection	NB accuracy improved and less processing time	Two textual data sets
Nguyen *et al.* (2016) [13]	Remove irrelevant features using hybrid filter and clustering approach		Two news and medicine data set
Sheydaei *et al.* (2015) [17]	Cluster high-frequency keywords based on class labels	Better performance of multiple classification algorithms	Publicly available texts
Yang *et al.* (2014) [18]	Use the deviation of features from centroids as feature selection approach	Multiple classifiers showed better performance	Reuters, newsgroup, and webkb

Table 1. Summary of studies used clustering as feature selection

The tweet text goes through further preprocessing:

(1) Tokenization: Each tweet is converted to tokens. A token is any word that is preceded and followed by a space.

(2) Stop words and nonuseful words like pronouns and articles are removed [19].

(3) *N*-grams: *n*-grams are word sequences that are often co-occur. These can be two or more words. The data have been explored for up to 4-g.

(4) After that, term frequency-inverse document frequency (TF-IDF) approach has been applied to vectorize textual data. TF-IDF reflects the importance of a keyword in a document by giving high-frequent words more weight [20].

- For each term, frequencies are calculated. This is used to prune the word list and specify the list of words with the highest frequencies. The pruning condition is to keep words that were used more than 1,207 times (half the number of promoter tweets).
- Inverse document frequency (IDF) is calculated for each term. Each word is considered as a feature for each tweet and will have a weight. The formula for IDF is:

$$\text{IDF} = \log \frac{\text{Total number of tweetd}}{\text{number of tweets which have that word}}$$

At this stage, a list of eight words were identified. They represent the eight features along with their weights. Two of which are 2-g features. Table 2 shows the resulting wordlist.

The resulting data set includes these features along with the ID of each tweet. Figure 1 shows an example of the process that the tweet goes through during cleaning and preprocessing, and Figure 2 shows a sample from the resulting data set.

3.2 Classification techniques

Four classification algorithms are tested based on what was obtained from literature. NB, DT, RF and LR. NB is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. It assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [21].

DT is a supervised classifier where the data are continuously split according to a certain parameter. In DTs, each leaf is assigned to one class or its probability. Small variations in the training set result in different splits leading to a different DT. Thus, the error contribution due to variance is large [22].

RF consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become the model's prediction. Although RF has powerful properties, it is said to be less sensitive to the optimization of method parameters leading to a simpler training process [21].

Table 2.
List of attributes, their types and their description the data set

Attribute	Type	Sample data/description
From user	Text	1alilts1, oops3021, abotamaim
Tweet	Text	ليله_عمرك_علي_دفادف. #سكليف #عذر_طبي #اجازة_مرضية إجازات مرضية وأعذار طبية للموظفين والموظفات والطلبة والطالبات خاص whats 0590413491

Example

Original tweet: سكليف اعذار طبيه تقرير طبي إجازات مرضيه إشعار مراجعه اجازة مرافق مريض بالخاص حياك _____ التواصل بالخاص ارسل البيانات كامله وعدد الايام حتى يتم الرد عليك بأسرع وقت #الاتحاد_الاتفاق

Cleaned tweet: سكليف اعذار طبيه تقرير طبي اجازه مرضيه اشعار مراجعه اجازه مرافق مريض الخاص حياك التواصل الخاص ارسل البيانات كامله عدد الايام يتم الرد عليك اسرع وقت الاتحاد الاتفاق

Word list: "سكليف", "اعذار", "طبيه", "تقرير", "طبي", "اجازه", "مرضيه", "اشعار", "مراجعه", "مرافق", "اجازه", "مريض", "الخاص", "ارسل", "البيانات", "كامله", "عدد", "الايام", "يتم", "الرد", "عليك", "اسرع", "وقت", "الاتحاد", "الاتفاق"

2-grams: "سكليف اعذار", "اعذار طبيه", "طبيه تقرير", "تقرير طبي"..... "الاتحاد الاتفاق"

3-grams: "سكليف اعذار طبيه", "اعذار طبيه تقرير", "طبيه تقرير طبي"... "وقت الاتفاق الاتحاد"

4-grams: "سكليف اعذار طبيه تقرير", "اعذار طبيه تقرير طبي", "طبيه تقرير طبي اجازه".... "اسرع وقت الاتحاد الاتفاق"

Figure 1.
An example of a tweet going through cleaning and preprocessing

Row No.	اجازه	اجازه_مرضيه	اعذار	اعذار_طبي	خاص	سكليف	طبي	مرضيه
27	0.427	0.224	0.244	0.252	0.618	0.081	0.455	0.223
28	0.427	0.224	0.244	0.252	0.618	0.081	0.455	0.223
29	0.427	0.224	0.244	0.252	0.618	0.081	0.455	0.223
30	0.427	0.224	0.244	0.252	0.618	0.081	0.455	0.223
31	0.427	0.224	0.244	0.252	0.618	0.081	0.455	0.223
32	0.427	0.224	0.244	0.252	0.618	0.081	0.455	0.223
33	0.475	0.498	0	0	0.458	0.090	0.253	0.495
34	0.475	0.498	0	0	0.458	0.090	0.253	0.495
35	0	0	0	0	0	1	0	0
36	0	0	0	0	0	1	0	0
37	0	0	0	0	0	1	0	0
38	0.394	0	0	0	0.761	0.298	0.420	0
39	0.366	0.383	0	0	0.705	0.277	0	0.381
40	0.442	0.463	0.504	0.261	0	0	0.235	0.461
41	0.297	0.312	0.339	0.351	0.574	0.225	0.317	0.310

Figure 2.
Snip of the data set after cleaning and preprocessing

LR is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes [21]. It is an extensively employed algorithm for classification in industry.

3.3 Clustering technique

Clustering is an unsupervised learning technique where similar instances are grouped together. *K*-means algorithm is one of the most common approaches to apply clustering. It has been applied to multiple problems such as recommender systems, image processing and

text mining [10]. Compared to other clustering algorithms, it is considered to be time efficient due to its linear complexity. It converges at O(J*K*m*N) with K clusters and J number of iterations, where m is the number of instances in the data set and N is the number of features [10]. In the problems where the number of clusters is unknown, multiple iterations of the algorithm are run in order to find the optimum value of K. Many approaches are used to find the best K including elbow approach, cross-validation and Silhouette approach [10]. In the current work, the number of K has already been set to 2.

The K-means algorithm starts by randomly selecting k instances (in this case two) as initial centroids of the clusters. After that, the distance between each of the remaining instances and the two centroids is calculated. The instance is assigned to a certain cluster if it is close to it. Once all instances are assigned, the mean of the distances between the instances and their centroid is calculated, and it becomes the new centroid. The process is repeated until the optimum clustering is reached using Eqn (1) [10], where μ_k is the mean for cluster k, N_k is the number of instances in the cluster k and x_i is one of the instances that belong to cluster k.

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N_k} x_i \tag{1}$$

3.5 Study model

The proposed model utilizes K-means clustering algorithm to identify features (terms) to be used in the classification efforts. This means that K-means clustering is applied to identify the terms that were useful in differentiating between the two clusters. After that, the list of terms is experimented with the four classification algorithms to see their performance compared to the standard feature selection approach. The clustering algorithm produces a list of features that are considered determinants in the clustering effort. They determine the similarity and dissimilarity between the instances and their centroids. The process is explained in the pseudocode showed in Figure 3.

4. Experiments and results

Promoters represent 16.8% of the data set. This means that the ratio of promoter to not-promoter is 1:5. This is showing an imbalance in the data set and needs to be considered when the classification algorithms are run in order to overcome any possible overfitting. Experiments are set using data set of ratio 1:1, 1:2 and 1:3.

4.1 Clustering analysis as feature selection

K-means algorithm was applied with $K = 2$. With topic modeling, the TFIDF operator was able to generate a list of 30,232 words. Nine words were used in clustering based on their

Classification based on Clustering

Input:
C: a collection of tweets
G: ground-truth data of labeled tweets
Output: Classification performance
1: C_TFIDF ← Generate TFIDF (C)
2: listOfFeatures ← Cluster (C_TFIDF, 2) //two clusters
3: updated_G ← Select_Features (G, listOfFeatures)
4: **Repeat until** all classifiers are tested
    Run Classifier (updated_G)
    Return performance

Figure 3.
Classification using clustering for feature selection

weight in the document. In Figure 4 the variation in term occurrences in each cluster shows which terms were most efficient in the clustering distinction. Words such as (تواصل connection), (خاص private), (تقارير reports) and (حكومي government) are appearing in instances belonging to cluster_0, while the word (سكليف sick leave) is more in cluster_1.

The performance of *K*-means clustering is evaluated using the average distance within centroid. The larger the number, the better, and in this case, it is 0.584. Another number to look at is the average within centroid distance within each cluster. For cluster_0, it is 0, and for cluster_1, it is 0.821.

Table 4 presents the list of most influential terms in Arabic along with their translation and the distances between them and the centroid of each cluster. The difference shows the extent at which the term belongs to a certain cluster. If a certain tweet contains one of the terms, it supports its assignment to the cluster centroid closest to it. In the table, five terms have high absolute difference values ranging from 0.108 to 0.503. The other terms are showing very low difference. The top five terms are used with the classification algorithms as features.

In Table 5, RF is showing the best performance based on most of the measures.

4 2 Experimenting with different ratios

The four selected classification algorithms were conducted using the 10-fold cross-validation technique. The results are shown for the original data and the data after sampling to treat the imbalance (see Table 6). The highest accuracy has been achieved using RF with the original

Figure 4. Distribution of term occurrences based on clusters

Word	Translation	Total frequency	Tweet frequency
سكليف	Sickleave (transliteration)	10,465	9,350
اجازة	Vacation	6,432	4,622
طبي	Medical	6,266	4,294
مرضية	Sick	5,794	4,405
اجازة_مرضية	Sick-leave	5,720	4,374
اعذار	Excuse	5,381	3,940
اعذار_طبي	Medical excuse	4,872	3,772
خاص	Private (inbox)	1,974	1,611

Table 3. wordlist, their translation, frequencies and the number of tweets they appeared in

ratio reaching for up to 94.92% with the highest specificity of 98.57%. Other significant results were achieved under the ratio 1:1 where the DT achieved 95.9% precision. RF also achieved the highest *f*-measure of 89.11% under 1:1 ratio. NB achieved the highest recall of 88.73% under 1:1 ratio. The ratios 1:2 and 1:3 did not achieve any significant results.

To further attempt improving the results, correlation between attributes is calculated to be applied in backward elimination. The process of elimination starts by including all attributes and eliminating the least significant attribute and then runs the classifier. The process continues until the best performance is reached. For the purpose of this analysis, backward elimination is applied with only RF since it achieved the best results. The result of backward elimination with RF improved the accuracy, recall and *f*-measure. However, it slightly reduced the specificity. RF using four features managed to reach to 95.01% accuracy, 90.81%

Table 4.
The terms and their centroid values based on each cluster and their absolute difference.

Term	Translation	cluster_0	cluster_1	Absolute difference
حكومي	Government	0.001	0.504	0.503
خاص	Private	0.007	0.461	0.454
تواصل	Contact	0.007	0.302	0.295
تقارير	Reports	0.033	0.141	0.108
سكليف	Sickleave	0.235	0.021	0.214
طبي	Medical	0.121	0.123	0.002
مرضية	sick	0.124	0.113	0.011
اعذار	Excuse	0.136	0.11	0.026
إجازة	Vacation	0.134	0.108	0.026

Table 5.
Classification results with clustering as feature selection

Algorithm	Accuracy	Precision	Recall	*f*-measure	Specificity
NB	91.61	94.85	95.08	94.96	74.43
DT	95.07	95.24	*99.02*	97.09	75.50
RF	*95.91*	*96.38*	98.79	*97.57*	*81.64*
LR	90.18	92.53	95.96	94.21	61.58

Table 6.
Comparing classification performance using different ratios without feature selection

Algorithm	Ratio	Accuracy	Precision	Recall	*f*-measure	Specificity
NB	1:1	75.09	69.9	*88.73*	78.09	61.46
DT	1:1	88	*95.9*	80.52	87.52	96.56
RF	1:1	89.83	95.77	83.34	*89.11*	96.31
LR	1:1	87.26	93.4	80.19	86.28	94.32
NB	1:2	80.56	66.29	85.04	74.48	78.33
DT	1:2	91.41	95.67	77.74	85.76	98.24
RF	1:2	92.13	94.42	81.18	87.29	97.6
LR	1:2	89.45	91.92	74.93	82.53	96.71
NB	1:3	90.85	85.24	76.71	80.73	95.57
DT	1:3	93.15	93.82	77.75	85.01	98.29
RF	1:3	93.16	93.91	77.7	85.02	98.31
LR	1:3	91.35	89.7	73.89	81	97.17
NB	Original	92.66	79.69	75.54	77.54	96.12
DT	Original	94.77	90.86	76.62	83.09	98.43
RF	Original	*94.92*	91.59	73.83	83.52	*98.57*
LR	Original	93.48	85.82	73.31	79.04	97.56

precision, 78.24% recall, 84.01% *f*-measure and 98.39% specificity. The four attributes that were used were (اجازه vacation, اعذار excuses, خاص private messaging, سكليف sick-leave (transliteration)). The rest of the algorithms showed similar improvement. Figure 5 shows the comparison between two approaches.

5. Evaluation

5.1 Evaluation criteria

In the literature dealing with spam detection, some standard metrics are used. These include accuracy, precision, sensitivity, *f*-measure and specificity. Accuracy is the total ratio of correctly predicted as promoter to the total cases (Eqn 2). Sensitivity, also known as recall or true positive rate, reflects the percentage of the positively predicted as promoter to those predicted positive (Eqn 3). Specificity is the measure of instances that were correctly predicted as not-promoter (Eqn 4). Precision is the percentage of instances that were correctly predicted as promoters to the percentage of positively and negatively predicted (Eqn 5). Finally, the *f*-measure is calculated as a harmonic mean for precision and sensitivity (Eqn 6).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2}$$

$$\text{Sensitivity (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$F - \text{measure} = 2 \times \frac{\text{Precision.Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

5.2 Evaluating the model

As RF achieved the best accuracy, a discussion of the other measures is also important to reflect on the model's performance. RF showed also highest precision, specificity and *f*-measure; however, it achieved low sensitivity. This means that the model is likely to generate false-negatives 26.17% of the times. On the other hand, the model is able to correctly identify an instance to be not-promoter 98.57% of the time. These results were based on the original data set. The results improved when applying feature selection using backward elimination. However, sensitivity remained low at 78.24%. When applying clustering as

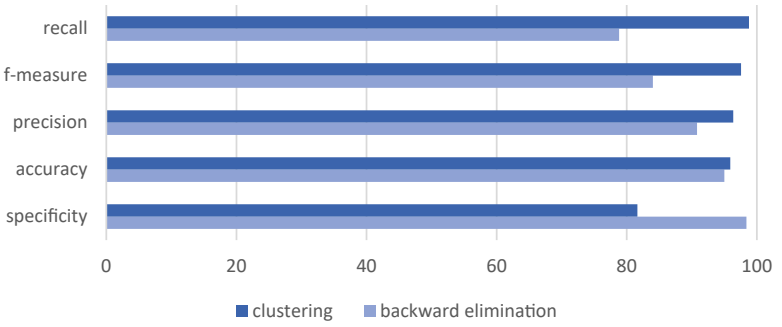

Figure 5. Comparing the performance of RF using backward elimination and clustering and feature selection approach

feature selection, sensitivity improved significantly. Other measures also improved including accuracy, precision and *f*-measure. It is also visible that a decline happened in specificity from 98.39% to 81.64%. This means that the model's ability to detect tweets belonging to not-promoter is less.

Figure 6 compares the performance of RF without feature selection, with backward elimination and using features identified by clustering.

The difference between features selected using backward elimination and the ones selected by clustering is in the number of features and terms included. Backward elimination produced four terms including اجازه vacation, اعذار excuses, خاص private messaging and سكليف sick leave. While clustering produced five keywords including حكومي government, خاص private, تواصل contact, تقارير reports and سكليف sick leave. It is noticed that سكليف sick leave was used in both approaches. This can be attributed to the fact that the word has been mentioned 10,465 times. On the other hand, clustering introduced a new set of keywords including حكومي government which refers to the type of sick leave issued from a government hospital. تواصل contact and خاص private which refers to requesting those interested in buying sick leaves to communicate using private messaging or WhatsApp. Also, تقارير reports which refers to the sick-leave documents.

5.3 Comparing with related work

The focus is on studies that used textual analysis of tweet content to classify spam/nonspam accounts.

All of the studies in Table 7 used Twitter data from either publicly available data sets or data downloaded from Twitter. This study showed the highest performance compared to the previous studies. In fact, recall improvement is considered the most significant contribution as it shows the sensitivity of the model in detecting promoters. According to [7], the majority of studies of spam detection rely on recall as a performance measure.

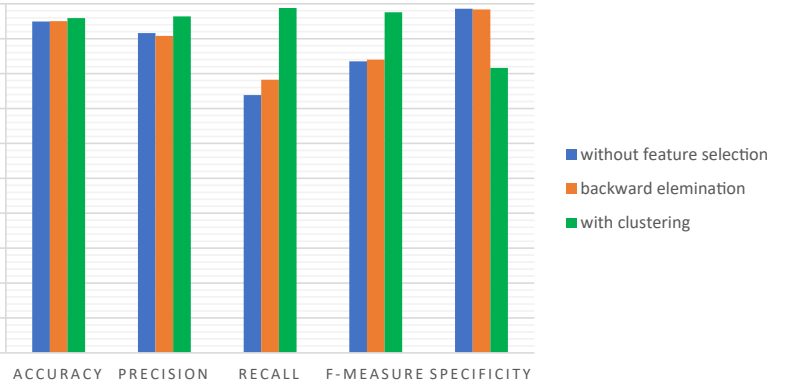

Figure 6.
Comparing the performance of RF: without feature selection, with backward elimination and with clustering

Research	FS method	Classifier	Accuracy	Recall	*F*-measure
Al-Azani *et al.* (2019) [23]	Skip-grams	SVM	87.33%	87.33%	87.33%
Ashour *et al.* (2019) [24]	*N*-grams	RF	–	78.4%	78.36%
Afzal and Mehmood (2016) [25]	Information gain	NB	95.42%	–	–
This study	*K*-means clustering	RF	95.91%	98.79%	97.57%

Table 7.
Studies used tweet content for spam classification

6. Conclusion, implications and future directions

This work is dealing with the problem of Twitter accounts that sell undeserved sick leaves in Saudi Arabia. The model proposed utilizes *K*-means clustering as feature selection approach to identify the most important keywords in determining each cluster. The resulting features are tested with four classification algorithms. When comparing the performance of these algorithms without *K*-means clustering, it was found that clustering improved the classification performance of all the algorithms. Most importantly, the sensitivity of the classification model improved. The study also identified a list of keywords that can be used as determinants in the classification of sick-leave promoters.

The major implications of this issue can be directly influencing the efforts of Saudi Arabia to identify the accounts that are engaged in illegally selling sick-leave documents. Detecting and reporting these accounts to Twitter means that the mean of communication between those seeking the service and those promoting it is broken. The authors understand that other platforms maybe utilized; however, it is considered as contributing to the other efforts to combat these actions. Future directions can be in investigating other platforms to compare the behavior of promoters across platforms. Technically, future work can involve experimenting with ensemble machine learning techniques and testing the model with other standard databases for spam detection.

Note

1. A newspaper article explaining the punishment of issuing undeserved sick leaves in Saudi Arabia: https://www.okaz.com.sa/article/905731

References

1. Inuwa-Dutse I, Liptrott M, Korkontzelos I. Detection of spam-posting accounts on Twitter. Neurocomputing. 2018; 315: 496-511.
2. Mbarek A, Jamoussi S, Charfi A, Ben Hamadou A. Suicidal profiles detection in twitter. WEBIST 2019 - Proceedings of the 15th International Conference on Web Information Systems and Technologies, 2019: 289-96.
3. Li L, Song Z, Zhang X, Fox EA. A hybrid model for role-related user classification on Twitter. 2018; 1. arXiv:1811.10202.
4. De Bie T, Lijffijt J, Mesnage C, Santos-Rodriguez R. Detecting trends in twitter time series. IEEE International Workshop on Machine Learning for Signal Processing, MLSP, 2016; 1-6. 2016-November.
5. Sowmya P, Chatterjee M. Detection of fake and clone accounts in twitter using classification and distance measure algorithms. Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020, 2020; 67-70.
6. Rao S, Verma AK, Bhatia T. A review on social spam detection: challenges, open issues, and future directions. Expert Syst Appl. 2021; 186.
7. Abkenar SB, Kashani MH, Akbari M, Mahdipour E. Twitter spam detection: a systematic review. 2020; 2: 1-18.
8. Cichosz P. A case study in text mining of discussion forum posts: classification with bag of words and global vectors. Int J Appl Math Comput Sci, 2018; 28(4): 787-801.
9. Raghuram MA, Akshay K, Chandrasekaran K. Efficient user profiling in twitter social network using traditional classifiers. Adv Intell Sys Comp, 2016; 385: 399-411.
10. Adewole KS, Han T, Wu W, Song H, Sangaiah AK. Twitter spam account detection based on clustering and classification methods. J Supercomput. 2020; 76(7): 4802-37.
11. Pintas JT, Fernandes LAF, Garcia ACB. Feature selection methods for text classification: a systematic literature review. Artif Intell Rev. 2021; 54: 6149-6200.

12. Edward E. Comparing methods of text categorization. Uppsala: Uppsala Universitet; 2018.
13. Nam LNH, Quoc HB. A combined approach for filter feature selection in document classification. Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI. 2016; 2016(January): 317-24.
14. Guru DS, Swarnalatha K, Kumar NV, Anami BS. Effective technique to reduce the dimension of text data. Int J Comput Vis Image Process. 2019; 10(1): 67-85.
15. Chormunge S, Jena S. Correlation based feature selection with clustering for high dimensional data. J Electr Syst Inf Technol. 2018; 5(3): 542-49.
16. Malji P, Sakhare S. Significance of entropy correlation coefficient over symmetric uncertainty on FAST clustering feature selection algorithm. Proceedings of 2017 11th International Conference on Intelligent Systems and Control, ISCO. 2017: 457-63.
17. Sheydaei N, Saraee M, Shahgholian A. A novel feature selection method for text classification using association rules and clustering. J Inf Sci. 2015; 41(1): 3-15.
18. Yang J, Liu Z, Qu Z, Wang J. Feature selection method based on crossed centroid for text categorization. 2014 IEEE/ACIS 15th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2014 - Proceedings, 2014.
19. Bahassine S, Madani A, Al-Sarem M, Kissi M. Feature selection using an improved Chi-square for Arabic text classification. J King Saud Univ - Comput Inf Sci. 2020; 32(2): 225-31.
20. Huilgol P. Quick introduction to bag-of-words (BoW) and TF-IDF for creating features from text, Analytics Vidyha, 2020, [Online], available at: https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/.
21. Pranckevičius T, Marcinkevičius V, Comparison of naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Balt J Mod Comput. 2017; 5(2): 221.
22. Dev VA, Eden MR. Gradient boosted decision trees for lithology classification. Computer Aided Chem Eng, 2019; 47: 113-118.
23. Al-Azani S, El-Alfy ESM. Detection of Arabic spam tweets using word embedding and machine learning. 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2018, 2018.
24. Ashour M, Salama C, El-Kharashi MW. Detecting spam tweets using character N-gram features. Proceedings - 2018 13th International Conference on Computer Engineering and Systems, ICCES 2018, 2019; 190-95.
25. Afzal H, Mehmood K. Spam filtering of bi-lingual tweets using machine learning. International Conference on Advanced Communication Technology, ICACT, 2016; 2016-March: 710-14.

Corresponding author
Mariam Elhussein can be contacted at: maelhussein@iau.edu.sa