Ain Shams Engineering Journal 9 (2018) 3259–3269

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Efficient email classification approach based on semantic methods

Eman M. Bahgat *, Sherine Rady, Walaa Gad, Ibrahim F. Moawad

Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

ARTICLE INFO

Article history:
Received 20 September 2017
Revised 3 March 2018
Accepted 11 June 2018
Available online 15 November 2018

Keywords:
Email classification
Spam
WordNet ontology
Semantic similarity
Features reduction

ABSTRACT

Emails have become one of the major applications in daily life. The continuous growth in the number of email users has led to a massive increase of unsolicited emails, which are also known as spam emails. Managing and classifying this huge number of emails is an important challenge. Most of the approaches introduced to solve this problem handled the high dimensionality of emails by using syntactic feature selection. In this paper, an efficient email filtering approach based on semantic methods is addressed. The proposed approach employs the WordNet ontology and applies different semantic based methods and similarity measures for reducing the huge number of extracted textual features, and hence the space and time complexities are reduced. Moreover, to get the minimal optimal features' set, feature dimensionality reduction has been integrated using feature selection techniques such as the Principal Component Analysis (PCA) and the Correlation Feature Selection (CFS). Experimental results on the standard benchmark Enron Dataset showed that the proposed semantic filtering approach combined with the feature selection achieves high computational performance at high space and time reduction rates. A comparative study for several classification algorithms indicated that the Logistic Regression achieves the highest accuracy compared to Naïve Bayes, Support Vector Machine, J48, Random Forest, and radial basis function networks. By integrating the CFS feature selection technique, the average recorded accuracy for the all used algorithms is above 90%, with more than 90% feature reduction. Besides, the conducted experiments showed that the proposed work has a highly significant performance with higher accuracy and less time compared to other related works.

1. Introduction

Electronic mails (Email) have become one of the most important and powerful communication ways in personal lives and business. Some users misuse the Emails by sending computer worms and spams which are unrequested information sent to the Email inboxes. The average spam Email messages sent every day have reached 54 billion messages based on statistics in 2014 [1]. Spam Emails cause an overload to the email servers, and consume network bandwidth and storage capacity. Therefore, Email filtering is a very important process to solve these problems. The filtering purpose is to identify and isolate the spam Emails.

Many mail server engines are using various authentication mechanisms to analyze Email content and categorize the Emails into white and black lists so; it can be optimized by users [2]. Using white and black lists, the new Email source is compared with a database to know if it is classified as spam before or not. On another side, an alternative approach filters Emails by extracting features from the Email body and using some classification methods, such as Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Neural Networks (NN). Most of the related works classify emails using the term occurrence in the email. Some few works additionally consider the semantic properties of the email text. Integrating semantic concepts and approaches for email classification is expected to add important benefits of enhancing the computational performance, in addition to the accuracy of classification.

In this paper, a novel approach to classify spam and ham Emails based on the Email body is presented. The approach employs various techniques that target the methods' complexities for enhancing the computational performances. The techniques used start

* Corresponding author.
E-mail addresses: eman.bahgat4@cis.asu.edu.eg (E.M. Bahgat), srady@cis.asu.edu.eg (S. Rady), walaagad@cis.asu.edu.eg (W. Gad), ibrahim_moawad@cis.asu.edu.eg (I.F. Moawad).

Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.asej.2018.06.001
2090-4479/© 2018 Production and hosting by Elsevier B.V. on behalf of Ain Shams University.

from pre-processing and cleaning the Email format, such as removing stop words and eliminating the irrelevant data, moving towards semantic modeling for compressing features, and finally employing the feature reduction and feature selection techniques. The main target is to preserve only the most important features. The given methods and techniques are integrated in a sequential process and proposed architecture.

In the semantic modeling, there is a need to exploit the semantic meaning within the Email to understand the context. Therefore, WordNet is employed as a semantic ontology [3]. In addition, different similarity measures have been presented using WordNet ontology for feature reduction. For further reduction for the dimensionality of textual features, feature reduction techniques are applied finally, and for that a comparison for two of the commonly used techniques is given: PCA (Principal Component Analysis) and CFS (Correlation Feature Selection). The experiments have been conducted using Enron E-mail dataset and different classifiers have been tested to classify emails into spam or ham.

To sum up, the proposed approach has a main contribution of employing the semantic relations and similarity measures of the WordNet ontology to reduce the Email feature semantically, which is an issue that have not been introduced thoroughly before. The additional integration of feature selection has contributed to presenting a methodology for preserving and classifying Email datasets at high dimensionality reduction rates.

The rest of the paper is organized as follows. Section 2 surveys the related work, while Section 3 introduces the proposed approach. The feature reduction methods are presented in depth in Section 4, which includes the semantic methods and the feature selection techniques. Section 5 reviews the classifier types utilized, and Section 6 presents the experimental evaluation and discusses the important key findings. Finally, the paper is concluded in Section 7 and illustrating the future work.

2. Related work

Supervised classification techniques have been applied extensively for Email filtering [4]. The pre-defined category labels (Ham or Spam) are assigned to documents based on the probability suggested by a training set of labeled documents. Some of the techniques used specifically for filtering spam Email are: Naive Bayes (NB) [5–7], Artificial Neural Net-works (ANN) [8], k-Nearest Neighbor (KNN) [6], Logistic Regression [7], C4.5 classifier [8,9], RBF Networks [2], Multi-Layer Perceptron (MLP) [8], AdaBoost [10,11], Support Vector Machine (SVM) [5,6,11], and Random Forest (RF) [11].

In [12], Sharma et al. proposed a technique based on ANN using Radial-Basis Function networks (RBF). In [13], an anti-spam filter named SENTINEL is applied. The filter extracts the natural language attributes from Email text that are related to writer stylometry. RF, SVM, and NB, and two meta-algorithms called ADABOOSTM1 and bootstrap aggregating (BAGGING) are used and evaluated using CSDMC2010 dataset.

Comparative analysis for different classification algorithms has been also done in several works. In [14] a spam classification approach is presented to extract features (terms) from Email body and additional readability features relative to the Email (e.g. word length and document length). The experiments have been conducted on four datasets; SpamAssassin, Enron-Spam, LingSpam, and CSDMC2010. Classification has been applied using NB, RF, SVM, Bagging and AdaBoost. They claimed that that the classifiers generated using meta-learning algorithms perform better than trees, functions, and probabilistic methods. In [15], another comparative analysis using four classifiers (NB, Logistic Regression, Neural Network, and RF) has been presented on Enron dataset, with significant performance been recorded for the RF classifier. Moreover, a [16] compared four classifiers; BayesNet, J48, SVM and LazyIBK. Their result showed that the BayesNet and J48 classifiers perform better than SVM. Some work also considered ensemble classifiers, such as in [41], where spam filtering is done using multiple classifiers.

The previously mentioned related works for Email classification do not take into consideration the problem of high dimensionality of and the associated complexity of filtering approach. Subsequently, this problem attracted some researchers to handle it in some work.

In [17], a content based spam filtering technique has been presented for classifying the spam and ham Emails, with feature selection techniques been applied, namely PCA (Principal Component Analysis) and CFS (Correlation Feature Selection). The presented approach has been tested on Enron corpus. The results show that using CFS saves the time for classifiers than PCA and SVM has the best prediction accuracy. In [18], a spam detection approach using RF classifier has been used on the Spambase Dataset, which allows feature selection and parameter optimization. In [19], feature selection algorithms based on similarity coefficients is applied on Spambase dataset to enhance the detection rate and improve the classification accuracy. In [20], an improved mutual information version is proposed combined with the word frequencies to calculate the correlation between Email features and their classes. The experiments were conducted on English corpus named PU1′s and Chinese corpus E-mail dataset.

Other feature selection approaches were also proposed in [21]. Ref. [22] applied three sounding feature selection methods, Chi-square, Information Gain (IG) and Correlation Features Set (CFS), for the Email phishing classification problem. The OneRule, Part, Irep and decision trees (J48) have been used and the datasets in the experiments were collected from two sources: SpamAssassin and Nazario. They claimed that Part classification algorithm produced the best results, followed by JRip and then J48. In [37], two feature selection techniques have been used Chi-square and Information gain and tested with three classifiers Naïve Bayes, J48 and SVM on the Ling-Spam dataset. Another email filtering approach based on hybrid feature selection has been proposed in [38]. It combines term frequency and document frequency in the feature selection process, working on Naïve Bayes and SVM. The classifiers have been tested on four email datasets: PU1, SpamAssassin, Ling-spam, and Trec2007. 1000 documents are selected randomly from each dataset for computations. The number of features ranges from 200 to 1000 features.

Although the previous approaches handled the high dimensionality of emails by using different feature selection techniques, they didn't take into consideration the semantic meaning of the terms. Few works approached Emails filtering semantically for the sake of overcoming the ambiguity problems. They used the semantic relations only. In [23] a technique to categorize the E-mail based on semantic vector space model and WordNet has been introduced. The performance of their method has better results on small-scale Email training set. Three different classifiers have been tested on a collected 20-Newsgroups repository: SVM, KNN and Logistic Regression [23]. A model based on semantic feature is used to tackle the terms mismatch problem in [24]. SVM classifier has been used in the experiments. Similar work has been done in [25], applied on a public Chinese spam corpus, and in [26] using short text classification. In [39], natural language processing techniques have been used while clustering emails. Stylistic and semantic properties are considered in the clustering process. Regarding the semantic part, TF-IDF has been used for the most frequent terms. A two level model is proposed in [40] for data representation. The first one is to model the syntactic features using TF-IDF, while the second depends on semantic information using Wikipedia. Two datasets are used, Reuters-21578 and 20Newsgroups. Multi-layer SVM NN is applied to classify data. In such works, the semantic similarity techniques were not addressed.

In [27], an E-mail filtering approach based on classification tried to reduce the email features by applying stemming on the E-mail body. Five different classifiers have been applied on a subset of Enron dataset: Naïve Bayes, J48, SVM, Logistic Regression and Random Forest. Their proposed approach reduced the number of features by 43.5%. The SVM and Logistic regression classifiers produce the best accuracy result. In [28], a preliminary approach for feature reduction based on semantic is applied for filtering E-mails using the WordNet ontology. It aims to reduce the size of E-mail features. From their experiments, the reduction rate reached 36.5% and the Logistic regression has more accurate results compared to other classifiers.

The work of this paper tries to overcome the limitations addressed in the semantic approaches. The ambiguity problem in the context and the high dimensionality problem within the terms are handled through use of different semantic relations. As an enhancement for the high dimensionality problem, feature selection is a crucial step to get the most powerful and discriminatory features.

3. The proposed approach

The proposed approach includes several components for reducing the feature dimensionality to filter the E-mails into two classes: Ham and Spam. The proposed architecture is showed in Fig. 1.

It has two phases which are training and testing. The training phase consists of four main modules: Pre-processing; Feature Weighting; Feature Reduction and Classification. The testing phase consists of Pre-processing, Feature Weighting and Classification.

The most significant part in the architecture is a reduction module. This module consists of three proposed processes: Semantic discovery, Weight Function, and Feature Selection. The reduction module is an enhancement to a previous work introduced in [27,28].

Both the pre-processing and the feature weighting modules will be presented in Sections 3.1 and 3.2 respectively, while, Section 4 describes the feature reduction module in more detail.

3.1. Pre-processing

In the pre-processing module, Tokens are extracted and the irrelevant tokens such as numbers and symbols are removed. Tokens are extracted from both the body and subject line of the Email. After that, the stop words are eliminated. For more data cleaning, by consulting the WordNet as an English dictionary, only the meaningful words are considered. WordNet is interpreted as large lexical database for English Language, which groups English words into sets of synonyms called synsets [3].

In this module, stemming is not applied. This is relevant in order to preserve the meaning of the features. Instead of stemming, morphology is applied using WordNet. In morphology, all the features with the same root are considered as a token using.

Once an email document is pre-processed, the Email document 'd' can be represented by:

$$d_i = \langle (t_1, w_1), (t_2, w_2) \ldots, (t_n, w_n) \rangle$$

where, each term '*t*' is considered as a feature which has a corresponding weight '*w*' in a given *d*.

3.2. Feature weighting

This module calculates the weight of the extracted feature, where each term 't' is weighted by a weight 'w' using term frequency/inverse document frequency method (TF-IDF).

The Term frequency calculates the number of times the term 't' appears in the Email document 'd' as shown in the Eq. (1).

$$tf(t,d) = (f_d(t))/(max[f_d(t)]) \tag{1}$$

where $f_d(t)$ is frequency of term '*t*' in Email '*d*'.

The Inverse Document Frequency (IDF) estimates the importance of a given term is. It measures how rare a given term in the whole document, using equation (2):

$$IDF(t) = log(N/df_t) \tag{2}$$

where df_t is the no. of Emails with term '*t*', and '*N*' is the total number of Emails.

Finally, the TF-IDF is computed as the result of multiplication of Eqs. (1) and (2):

$$W = tf(t,d)IDF(t) \tag{3}$$

4. Feature reduction

The feature reduction module shown in architecture in Fig. 1 is the core process of the proposed approach. This module is responsible for reducing the Email extracted features in the previous steps by undergoing several different reduction techniques. It consists of three processes: Semantic-based reduction, Features weights updating and Feature selection. In this processing module, the synonyms of each feature are extracted, and then the extracted features are replaced with their synonym set concepts. Besides extracting the synonyms of each term, the hypernym/hyponym relations are considered by consulting the WordNet. Next, the similarity between words is measured using different semantic similarity measures. The Semantic similarity measures are calculated based on WordNet as well. This process is shown in Fig. 1 and begins by generating a set of reduced weighted features as an input for the next process "feature weights updating". A semantic weight is calculated for each term in the reduced feature set, and the outputs of this step is a set of updated feature weights.

Finally, feature selection process is applied on the current feature set to select the most discriminating and important features that assist the classification accuracy.

4.1. Semantic-based reduction

The main objective of this process is to reduce the feature size in each Email by employing different semantic techniques. After getting the weighted features from feature weighting module in the previous section, WordNet ontology is used as a lexical database to link English nouns, adjectives, verbs and adverbs to the sets of synonyms. Such synonyms are called synsets, which are linked together through semantic relations that determine word definitions [3].

In this process, synonyms set of each term in the Email are used to group the terms that have common synonyms. Furthermore, the hypernym/hyponym relationships among the noun synsets are considered. The hyponymy relation refers to "is a kind of" or "is a", where it links more general synsets to specific ones, while the hypernym relation represents the inverse of the hyponymy relation. This can help in merging the terms that have common parent or common children.

After applying semantic relations, different semantic similarity measures are applied in order to increase the reduction rate for the features. Fig. 2 shows the semantic-based reduction steps.

- Path based measure

Path Based Measures [32] are based on path lengths between two concepts. Three similarity measure versions are investigated

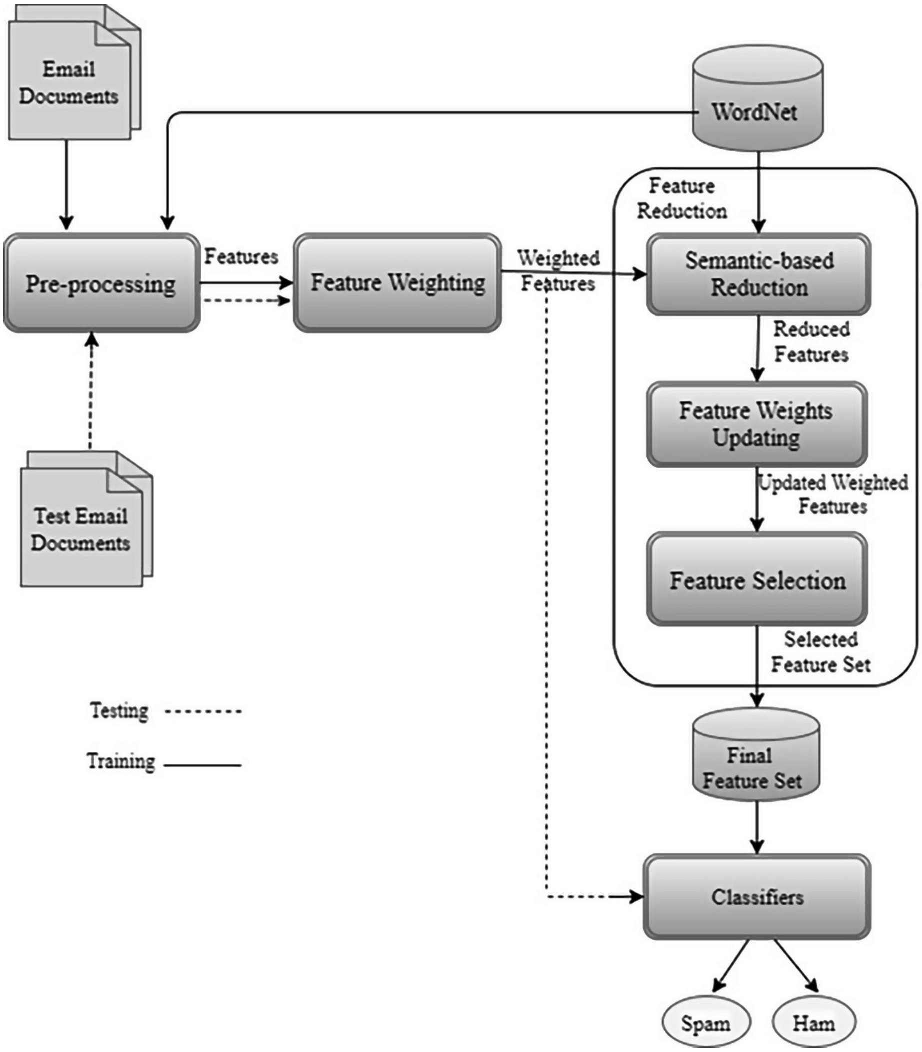


Fig. 1. The proposed system architecture.

for their performances which are Path Length measure, WUP measure, and LCH measure.

Path Length: calculates the semantic similarity of a pair of concepts by counting the number of nodes along the shortest path between the concepts in the 'is-a' hierarchies of WordNet. The path similarity score is inversely proportional to the number of nodes along the shortest path between the two words. Consequently, the equivalent measure equation is:

$$PATH(t_1, t_2) = 1/length(t_1, t_2) \tag{4}$$

where t_1 and t_2 are the two terms.

WUP measure: computes similarity by considering the depths of the two terms in the WordNet with the depth of the least common subsume (LCS) as shown the following equation:

$$WUP(t_1, t_2) = 2(depth(LCS(t_1, t_2)))/depth(t_1) + depth(t_1) \tag{5}$$

where lowest common subsume (LCS) is the most specific common ancestor among two synsets (t_1, t_2).

Leacock and Chodorow measure (LCH): finds the shortest path-length between two concepts, and then scales the resultant value by the max depth found in the "is–a" hierarchy in which they occur, according to the given equation:

$$LCH(t_1, t_2) = -log(length(t_1, t_2)/2Max(depth)) \tag{6}$$

- Information Content (IC) based:

provides a measure for the concept's specificity; The concept that occurs frequently is considered less specific and has a lower information content value, while the concept that rarely occurs is considered more specific and has a higher information content value [31]. One of the IC measures is the Resnik measure which calculates the information content of the least common subsume (LCS) of the two terms, using the following equation:

$$Resnik(t_1, t_2) = IC(LCS(t_1, t_2)) \tag{7}$$

The IC for a term (t) is defined by:

$$IC(t) = -logP(t) \tag{8}$$

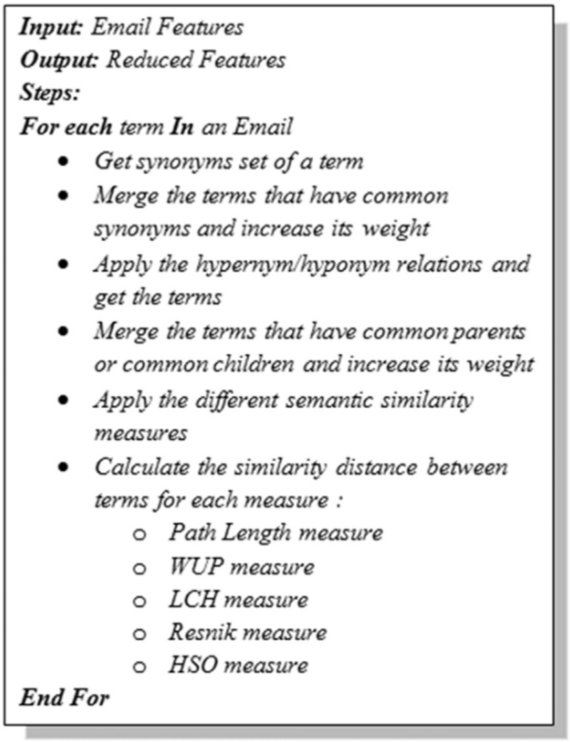
where $P(t)$ the probability of a term (t) in a given document containing (N) distinct terms.

$$P(t) = frequency(t)/N \tag{9}$$

- Relatedness measure:

Another type of relatedness measure has been applied which is HSO (Hirst and St-Onge) [33].

HSO measure calculates relatedness between terms by finding the path distance between the nodes, the number of changes in



```
Input: Email Features
Output: Reduced Features
Steps:
For each term In an Email
    • Get synonyms set of a term
    • Merge the terms that have common synonyms and increase its weight
    • Apply the hypernym/hyponym relations and get the terms
    • Merge the terms that have common parents or common children and increase its weight
    • Apply the different semantic similarity measures
    • Calculate the similarity distance between terms for each measure :
        o Path Length measure
        o WUP measure
        o LCH measure
        o Resnik measure
        o HSO measure
End For
```

Fig. 2. Semantic-based Reduction Steps.

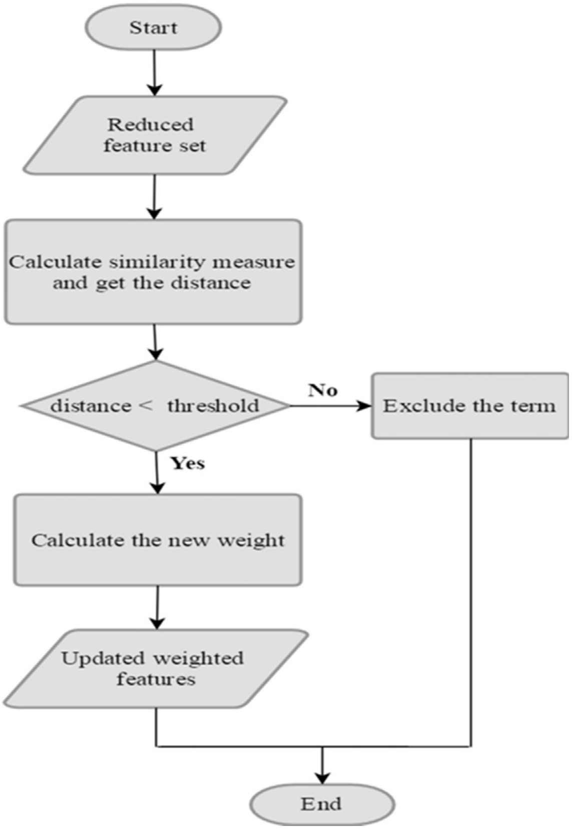


Fig. 3. The flowchart for feature weights update process.

direction of the path connecting two terms and the allowableness of the path. An Allowable Path is a path that does not diverge away from the meaning of the term. The HSO function is formulated as follows:

$$HSO(t_1, t_2) = C - PATH(t_1, t_2) - K \times dir \tag{10}$$

where *dir* is the number of changes of directions between two terms t_1 and t_2, and C, K are constants whose values are derived through experiments.

4.2. Feature weights update

After getting the reduced features set, a new weight is assigned to the terms based on the semantic similarity measure applied. After the semantic measure is calculated, a corresponding distance between two terms is generated. If this distance is less than a given threshold (indicating a similarity between the two terms), then the weight of this term is updated with a new weight value as calculated by:

$$F = \left(w_i, w_j\right) \times \left(1 - dist_{ij}\right) \tag{11}$$

where w_i and w_j are the weights (TF-IDF) of two terms i, j and dist_{ij} is the similarity distance between the two terms.

The threshold is normalized to be within the range [0.1, 1] for comparing through all the semantic similarity measures' performances. Fig. 3 outlines the flowchart for feature weights update process.

4.3. Feature selection

In addition to semantic based reduction, a second component is integrated in the reduction module for further reduction of the number of features to the optimal case, and hence, enhance the efficiency of the approach. Feature selection is a process that selects an optimal subset of relevant features. Two different feature selection techniques have been adopted in the proposed model: Principal Component Analysis (PCA) and Correlation Feature Selection (CFS). The choice of these selections is to be able to compare the results of presented approach with some of the related works.

Principal component analysis (PCA) is one of the ways to reduce the dimensionality of the data. PCA gets any important information from the data and introduces them in the form of orthogonal variables called principal components [17]. Correlation coefficient measures the interaction between features and estimates the correlation between a subset of features and class. Subsequently, it chooses the features that are highly correlated with the class features while having a low correlation with other attributes [30].

5. Classification

Supervised Classification is used to assign a class label to an unclassified tuple according to the learning process for instance set that is already classified. Different types of classifiers exist (probabilistic methods, tree-based, and neural network). The proposed approach is evaluated using six classifiers: Naïve Bayes, SVM, Bayesian Logistic Regression, J48, Random Forest, and RBF Network. These classifiers have been applied since they are the most commonly used in the literature of Email classification. Each classifier is depicted briefly.

The Naive Bayes algorithm is a simple conditional probabilistic classifier that uses Bayes' Theorem, which applies the Bayes' formula to measure conditional probability between two variables; in our case the features and the class [5,6].

Support vector machine (SVM) is a group of supervised classification algorithms. SVM creates a hyperplane, which separates the different types of data. In our case, we have two types of data spam or ham. It can effectively handle large number of features [6,11].

J48 is one of the famous decision tree-based algorithms. It creates a binary tree. At each node of the tree, it tries to select an effective attribute that split its set of instances into a group of subsets. J48 recursively visits each node in the tree and selects the optimal split until no more splits are available [9,10].

Logistic regression is used to predict the association between each form of the independent variable [8]. It groups these variables to calculate the probability that a specific event will happen.

Random Forest (RF) is an ensemble learning method for classification. It composed from a set of decision trees [11].

The radial-basis function networks (RBF) is a neural network technique. It illustrates curve-fitting problem, which is approximation in high dimensional features [12,30].

6. Approach evaluation

6.1. Dataset and development environment

The proposed approach has been evaluated on the standard benchmark Enron-Spam dataset [29]. It contains a large number of Emails. It consists of Emails from 6 Enron employees who have huge mail boxes. For experimentation, a subset of Enron corpus is extracted, forming 1000 Email documents and is distributed into 20% spam and 80% ham. Enron dataset has been selected since it contains a mix between personal and official emails, an issue that is missing in other email datasets.

The testing is conducted using 10-fold cross validation method; therefore, the data is split into ten subsets. For each iteration, a sample is considered as a testing set and remaining nine samples as a training set. Finally, the performance value is the average performance values for the ten iterations.

The experiments were tested on a Laptop with Intel core i7 processor, 8 GB memory and Windows 7 operating system. The WEKA tool [34] has been employed in the evaluation of the classifiers. It is a group of classification algorithms to perform data mining tasks.

6.2. Evaluation measures

The performance of classifiers is evaluated using several measures: Recall, Precision, F-score, Accuracy, Cohen's kappa and classification execution time. To compute these measures, the confusion matrix is calculated, which is composed of four items:

- True Positives (TP) these are the number of positive tuples that were correctly classified to that class.
- True Negatives (TN) these are the number of negative tuples that were correctly classified to that class.
- False Positives (FP) these refer to the number of negative tuples that were incorrectly classified as positive.
- False Negatives (FN) these refer to the number of positive tuples that were incorrectly classified as negative.

The Accuracy is the percentage of correct predictions to total predictions made:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{12}$$

The Precision represents the ratio of correctly predicted positive cases to the total predicted positive cases as represented by:

$$Precision = TP/(TP + FP) \tag{13}$$

The Recall indicates the ratio of correctly predicted positive examples as represented by:

$$Recall = TP/(TP + FN) \tag{14}$$

The F-score is the average weight between recall and precision; it is represented by:

$$F - score = (2 \times Precision \times Recall)/(Precision + Recall) \tag{15}$$

Cohen's kappa is a statistical measure that frequently tests the inter-rater agreement:

$$\kappa = P_o - P_e/1 - P_e \tag{16}$$

where P_o is the observed agreement and P_e is the chance agreement. Defined as:

$$p^e = 1/N^2 \sum_C n_{c1} n_{c2} \tag{17}$$

For category *C*, *N* is the number of items and n_{c1} is the number of times rater 1 predicted category *K*.

The values of these measures are in the range [0, 1], the time performance of the proposed approach is measured through the average classification execution time, which is measured by the average time needed to classify a given testing document in the testing phase and using the testing corpus.

6.3. Experimental results and discussion

Three main experiments have been conducted to test the feature reduction module as a core concept in this paper. The performance is recorded independently at each processing step of the proposed approach and architecture to show the effect of including such steps and modules. The first experiment is responsible for evaluating the several semantic similarity measures and types, to find out the most fitting semantic similarity measure as discussed in Section 4.1. The second experiment focuses on selecting the most discriminating and relevant features in the Email document as mentioned in Section 4.3. Finally, the third experiment is conducted for comparing the performance of the approach (accuracy and classification time) against other related work.

In each experiment, the performance of each classifier is measured using the different performance measures: Accuracy, Recall, Precision, F-score, and Time, and a comparison between the classifiers' performance is conducted. The different experiments and the results obtained are explained below.

6.3.1. Selecting the best similarity measure experiment

The different semantic similarity measures have been applied as explained in Section 4.1. This experiment has been conducted after using the three semantic relations, which are synonym, hypernym, and hyponym relations. The main goal of this experiment is to select the best similarity measure which can reduce the size of features at the most suitable threshold value, indicating the similarity between features. Different threshold values were studied to evaluate the accuracy at each value with the six different classifiers.

Figs. 4–8 show the accuracy of each similarity measure for the semantic discovery module. In Fig. 4, the WUP similarity measure

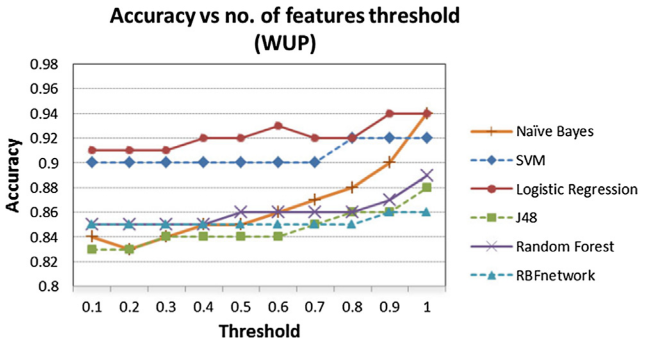


Fig. 4. Accuracy performance of different classifiers for WUP similarity measure versus number of features threshold.

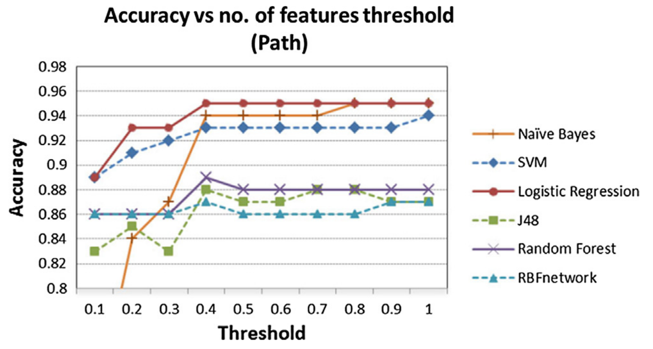


Fig. 5. Accuracy performance of different classifiers for Path similarity measure versus number of features threshold.

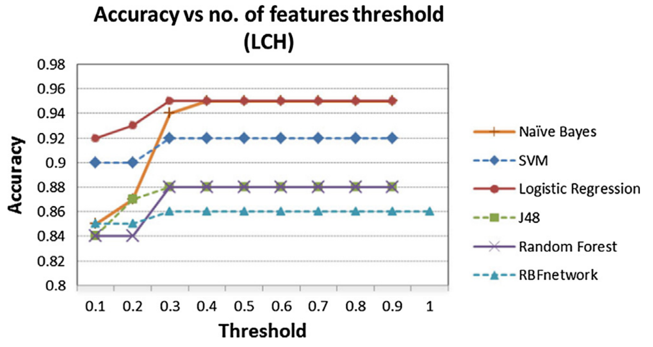


Fig. 6. Accuracy performance of different classifiers for LCH similarity measure versus number of features threshold.

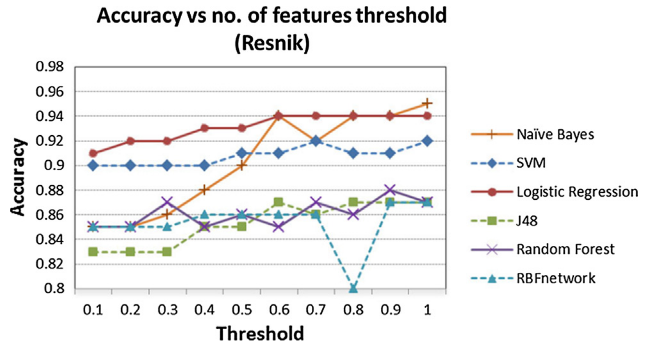


Fig. 7. Accuracy performance of different classifiers for Resnik similarity measure versus number of features threshold.

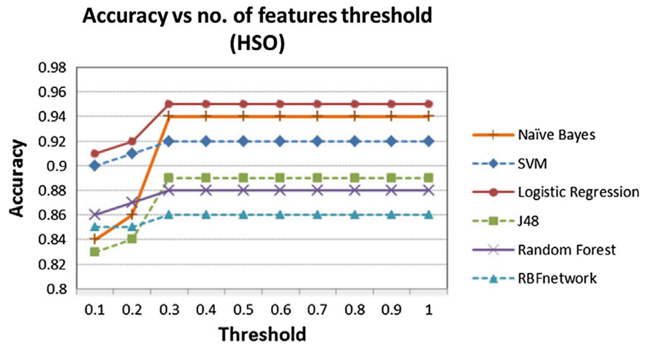


Fig. 8. Accuracy performance of different classifiers for HSO similarity measure versus number of features threshold.

has its best performance at threshold value 0.6 with less number of features. In Fig. 5, the Path measure records the best accuracy at threshold value 0.4. In Fig. 6, the LCH measure performs the best at 0.3. In Fig. 7, the Resnik measure proceeds a good performance at value 0.6. In Fig. 8, the HSO similarity measure performs best accuracy at 0.3. Hence, the best tuning parameter of the threshold was found to be within the range [0.3, 0.6], which is adopted for all the semantic similarity measures for common comparison. Obviously from all the Figs. 4–8, the LR has the highest and most stable accuracy results comparing with the other classifiers.

It is noted that the path similarity measure performs the best. This can be interpreted due to the distance between words in the spam Emails are most probably large and far from each other, since most of spam Emails are generated randomly containing words not related to each other, while in ham classification the distance is relatively shorter. The path similarity measure focuses on the hierarchical relation between two terms in the taxonomy, in our case the length between two terms in the wordnet. This can be an advantage while working on email documents. Otherwise, the other similarity measures worked based on LCS between two terms and the IC [42].

This conclusion is propagated in the next following experiments. In other words, it is adopted to be used for testing the effect of feature selection techniques. An illustrate example scenario was given in Fig. 9, which shows in sequential steps how the semantic reduction framework works. This example has been run on path similarity measure.

Table 1 shows the results of performances using path measure at the threshold value 0.4; where it recorded the best accuracy at this value. The Naïve Bayes had a precision value of 0.95 (the best value). Subsequently the Logistic Regression and SVM had recorded relatively similar results of 0.94 precision values. The RBF classifier had the minimal value of 0.86. With respect to Kappa measure, it is observed that the Naïve Bayes and Logistic Regression achieved the highest value of 0.83. The J48 and RBF recorded the least value of 0.6.

The Random Forest has the minimum execution time with 0.2 s, followed by Naïve Bayes with 0.3 s. The J48 classifier has the maximum exertion time with 5.0 s.

6.3.2. Selecting the most important feature set experiment

The main goal of this experiment is to test the effect of including PCA and CFS as mentioned in Section 4.3 to filter the most important feature set.

Fig. 10 shows that the CFS technique achieves higher accuracy with SVM and RBF Network, and the same accuracy with J48 and Random Forest. Nevertheless, the proposed semantic technique only without applying the feature selection (the experiment in Section 6.3.1) recorded better performance with Naïve Bayes and Logistic regression. This can be because Naïve Bayes and logistic regression can perform well with respect to accuracy when handling high dimensional features [35]. From Fig. 10, it is observed that the Naïve Bayes and SVM didn't perform well in comparison to CFS when using the PCA technique [36].

6.3.3. Comparison to related works

Comparative experiments have been conducted to compare the proposed methods to the related work presented in [14,17], and which use the same Enron dataset. Figs. 11 and 12 show such comparison, where the accuracy and execution time are illustrated respectively, for the common classifiers used in experimentation. It is worth mentioning that comparison is conducted while considering that the related work worked on the same size of Enron dataset, though the samples might be different.

In Fig. 11, it is shown that the proposed approach has better accuracy measure with NB and SVM. The achieved accuracy for

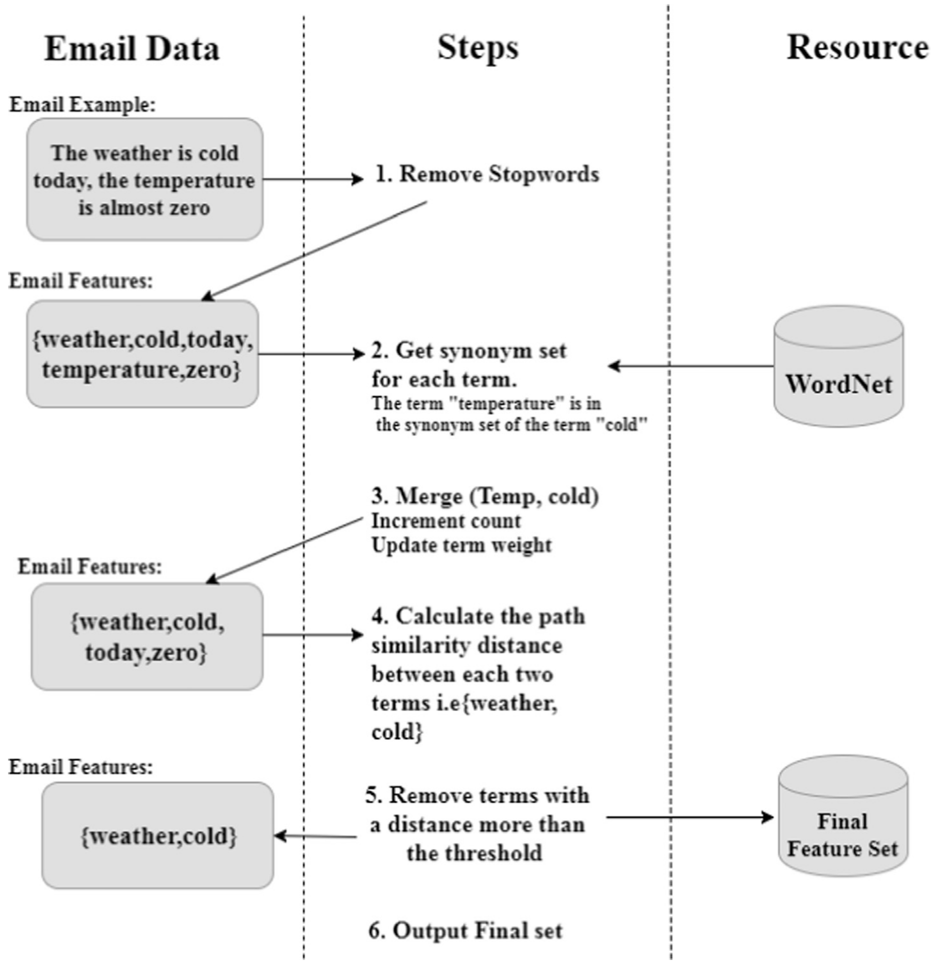

Email Data
Steps
Resource
Email Example:
The weather is cold today, the temperature is almost zero
1. Remove Stopwords
Email Features:
{weather,cold,today, temperature,zero}
2. Get synonym set for each term.
The term "temperature" is in the synonym set of the term "cold"
WordNet
3. Merge (Temp, cold)
Increment count
Update term weight
Email Features:
{weather,cold, today,zero}
4. Calculate the path similarity distance between each two terms i.e{weather, cold}
Email Features:
{weather,cold}
5. Remove terms with a distance more than the threshold
Final Feature Set
6. Output Final set

Fig. 9. Illustrative example for semantic reduction framework (using path similarity).

Table 1
Performance of path similarity measure at threshold 0.4.

Evaluation Criteria	Classifiers					
	NB	SVM	Logistic Regression	J48	Random Forest	RBF Network
Accuracy	0.94	0.93	0.95	0.88	0.89	0.86
Precision	**0.95**	**0.93**	**0.94**	0.88	0.88	0.86
Recall	0.94	0.93	0.95	0.88	0.89	0.86
F-Score	0.94	0.93	0.95	0.88	0.88	0.86
Kappa	0.83	0.78	0.82	0.6	0.61	0.6
Time (s)	**0.3**	2.1	0.6	5.0	**0.2**	3.0

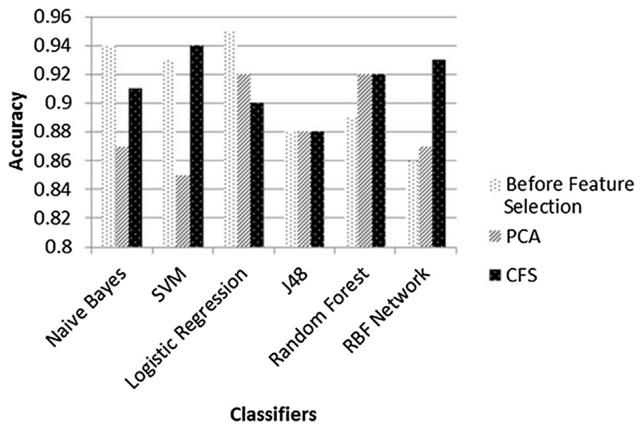


Fig. 10. Comparing accuracy performance versus the two feature selection techniques used (PCA and CFS).

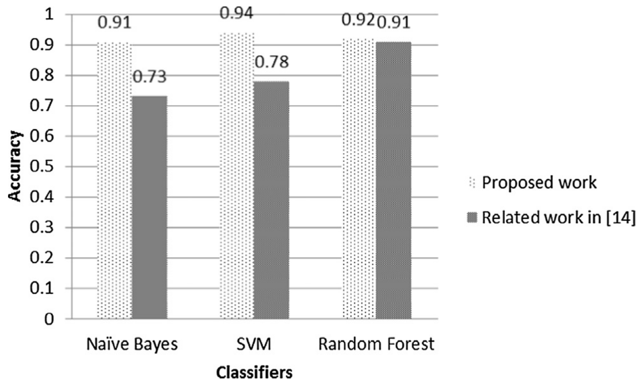


Fig. 11. Comparing the accuracy performance after using CFS feature selection technique against related work in [14].

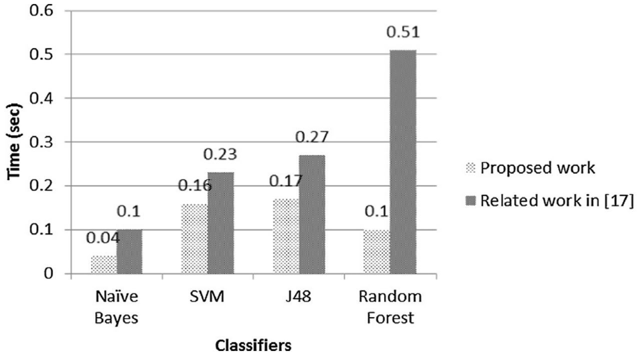


Fig. 12. Comparing the time after using CFS feature selection technique against related work in [17].

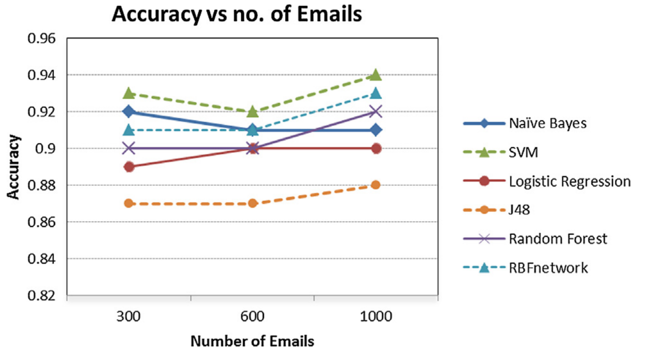


Fig. 13. Accuracy performance of different classifiers using different sizes of datasets.

NB is 91% versus 0.73% for related work, which is significantly better. Moreover, SVM has an accuracy of 0.94 while in the related work it is 0.78 which is also higher. For the Random Forest, it has lightly higher accuracy than the related work.

In Fig. 12, it is also noticed that the proposed approach has less time for building the model for all classifiers than the related work [17]. These conclusions are while considering that the related work worked on the same size of Enron dataset and the same feature reduction techniques. The executed time for NB is 0.04 versus 0.1 for related work and for the SVM is 0.16 versus 0.23 for the related work. The J48 classifier has execution time of 0.17 while in the related work it is 0.27. The Random Forest classifier has execution time of 0.1 s versus 0.51 for the related work, which is five times faster.

Table 2 summarizes different methods used from the proposed approach and architecture, where several performance measures are compared. The first column shows the results of the using the classifiers only. The second column shows the results after applying the semantic relations and the path similarity measure.

When applying feature selection (CFS) + semantic relations and similarity measures, the SVM and RBF Network had the best accuracy value of about 0.94. Their execution time decreased by 92% and 70% respectively in feature selection phase due to feature reduction.

The number of features when using classifiers only was 8975. This number decreased by 38.4%, reaching 5526 features when using the synonyms. By integrating semantic methods, the number of features is reduced to 2978 features forming a reduction percentage of 66.8%, hence a significant high reduction rate is achieved. Eventually, using an additional feature selection step, the number of features reached 70 only, forming the highest reduction percentage of value equal to 97%. It is shown that the Naïve Bayes and Logistic Regression classifiers have the best (least) execution time compared to other classifiers.

Fig. 13 presented the performance of the proposed approach on different dataset sizes (300, 600, and 1000). In this figure, the behavior of different dataset sizes has been studied. It shows that almost all the classifiers have better accuracy when applied on large datasets.

7. Conclusion and future work

In this work, an approach for E-mail filtering is introduced, targeting both accuracy and complexity performance enhancements. It is based on introducing semantic modeling to solve the problem of high dimensionality of features by considering the semantic properties of words. The semantic modeling makes use of semantic relations and semantic similarity measures to compress features in their dimensional space. Feature selection reduction techniques have been moreover for further reduction to achieve optimal feature compression. A set of different classifiers have been studied to test their performances to segregate Emails as spam or ham experiments on the Enron dataset. It has been shown from experiments that the path similarity measure performs the best. Introducing CFS as a technique for feature selection enhanced the accuracy of some classifiers compared to employing the semantic similarity only. Classifiers like Random Forest and RBF Network managed to reach accuracy values 92% and 93% respectively. The

Table 2
Comparing the accuracy value and time of the proposed approach at 3 different experimental phases.

Comparison measures	classifiers	Classifier only	Classifier + Semantic relations and similarity measures	Feature selection (CFS) + Semantic relations and similarity measures
Accuracy	Naïve Bayes	0.92	0.94	0.91
	SVM	0.94	0.93	**0.94**
	Logistic regression	0.81	0.95	0.9
	J48	0.84	0.88	0.88
	Random forest	0.89	0.89	0.92
	RBF Network	0.84	0.86	**0.93**
Time (sec)	Naïve Bayes	1.02	0.3	0.04
	SVM	7.1	2.1	**0.16**
	Logistic regression	3.8	0.6	0.04
	J48	15.5	5.0	0.17
	Random forest	0.5	0.2	0.1
	RBF Network	24.7	3.0	**0.9**
Number of features		8975	2978	70
Reduction percentage %		–	66.8%	97.6%

presented approach also maintained the accuracy of other classifiers such as SVM and J48. Combined semantic and feature selecting models managed to reduce the total size of features to 70 features without loss in the classification accuracy, subsequently, with lot of savings in time for the classifiers.

In the future work, ensemble classification can be tested. In addition, the behavior of the proposed semantic approach will be studied on different datasets.

References

[1] Internet Threats Trend Report. Cyberoam A SOPHOS Campany; 2014.
[2] Del Castillo M, Dolores, Ignacio Serrano J. An interactive hybrid system for identifying and filtering unsolicited e-mail. Intelligent Data Engineering and Automated Learning–IDEAL. Springer, Berlin Heidelberg; 2006. p. 779–88.
[3] Hristea FT. Semantic wordnet-based feature selection. In: The Naïve Bayes model for unsupervised word sense disambiguation. Heidelberg: Springer, Berlin; 2013. p. 17–33.
[4] Khan Aurangzeb et al. A review of machine learning algorithms for text-documents classification. J Adv Inform Technol 2010;1(1):4–20.
[5] Islam MS, Al Mahmud A, Islam MR. Machine learning approaches for modeling spammer behavior. Information Retrieval Technology. Heidelberg: Springer Berlin; 2010. p. 251–60.
[6] Blanzieri E, Bryl A. A survey of learning-based techniques of email spam filtering. Tech. rep. DIT-06-056, University of Trento, Information Engineering and Computer Science Department; 2008.
[7] Mitchell T. Generative and discriminative classifiers: naive Bayes and logistic regression. Manuscript available at http://www.cs.cm.edu/~tom/NewChapters.html; 2005.
[8] Renuka DK, Hamsapriya T, Chakkaravarthi MR, Surya PL. Spam classification based on supervised learning using machine learning techniques. In: International conference on process automation, control and computing (PACC). IEEE; 2011. p. 1–7.
[9] Shi L, Wang Q, Ma X, Weng M, Qiao H. Spam email classification using decision tree ensemble. J Comput Inform Syst 2012;8(3):949–56.
[10] Islam M, Zhou W. Architecture of adaptive spam filtering based on machine learning algorithms. ICA3PP, LNCS 4494. Berlin Heidelberg: Springer; 2007. p. 458–69.
[11] Islam Rafiqul, Yang Xiang. Email classification using data reduction method. In: Proceedings of the 5th international ICST conference on communications and networking in China. IEEE; 2010. p. 1–5.
[12] Sharma Reena, Kaur Gurjot. E-mail spam detection using SVM and RBF. Int J Modern Educat Comput Sci (IJMECS) 2016;8(4):57.
[13] Shams Rushdi, Mercer Robert E. Supervised classification of spam emails with natural language stylometry. Neural Comput Appl 2016;27(8):2315–31.
[14] Shams Reza, Mercer Robert E. Classifying spam emails using text and readability features. 13th international conference on data mining (ICDM). IEEE; 2013.
[15] More S, Kulkarni S. Data mining with machine learning applied for email deception. International conference on optical imaging sensor and security. IEEE; 2013.
[16] Sharaff A, Nagwani NK, Dhadse A. Comparative study of classification algorithms for spam email detection. In: Emerging research in computing, information, communication and applications. India: Springer; 2016. p. 237–44.
[17] Sharma Amit Kumar, Yadav Renuka. Spam mails filtering using different classifiers with feature selection and reduction technique. 2015 Fifth international conference on communication systems and network technologies (CSNT). IEEE; 2015.
[18] Lee SM, Kim DS, Kim JH, Park JS. Spam detection using feature selection and parameters optimization. In: International conference on complex, intelligent and software intensive systems. IEEE; 2010. p. 883–8.
[19] Abdelrahim AA, Elhadi AAE, Ibrahim H, Elmisbah N. Feature selection and similarity coefficient based method for email spam filtering. International conference on computing, electrical and electronics engineering (ICCEEE). IEEE; 2013.
[20] Ting Liang, Qingsong Yu. Spam feature selection based on the improved mutual information algorithm. Fourth international conference on multimedia information networking and security (MINES). IEEE; 2012.
[21] Wang R, Youssef AM, Elhakeem AK. On some feature selection strategies for spam filter design. In: CCECE'06. Canadian conference on electrical and computer engineering. IEEE; 2006. p. 2186–9.
[22] Issa Qabajeh, Thabtah Fadi. An experimental study for assessing email classification attributes using feature selection methods. 3rd International conference on advanced computer science applications and technologies (ACSAT), 2014. IEEE; 2014.
[23] Lu Z, Ding J. An efficient semantic VSM based email categorization method. International conference on computer application and system modeling. IEEE; 2010. V11-525.
[24] Yoo S, Gates D, Levin L, Fung S, Agarwal S, Freed M. Using semantic features to improve task identification in email messages. In: Natural language and information systems. Springer Berlin Heidelberg; 2008. p. 355–7.
[25] Wei Hu, Jinglong Du, Xing Yongkang. Spam filtering by semantics-based text classification. Eighth international conference on advanced computational intelligence (ICACI), 2016. IEEE; 2016.
[26] Tang HJ, Yan DF, Yuan TIAN. Semantic dictionary based method for short text classification. J China Univ Posts Telecommun 2013;20:15–9.
[27] Bahgat EM, Rady S, Gad W. An E-mail filtering approach using classification techniques. In: The 1st international conference on advanced intelligent system and informatics, Beni Suef, Egypt. Springer International Publishing; 2016. p. 321–31.
[28] Bahgat Eman M, Moawad Ibrahim F. Semantic-based feature reduction approach for e-mail classification. International conference on advanced intelligent systems and informatics. Springer International Publishing; 2016.
[29] Enron-Spam datasets. CSMINING group [accessed July 7, 2016] http://csmining.org/index.php/enron-spam-datasets.html.
[30] Karegowda Asha Gowda, Manjunath AS, Jayaram MA. Comparative study of attribute selection using gain ratio and correlation based feature selection. Int J Inform Technol Knowledge Manage 2010;2(2):271–7.
[31] Pedersen Ted, Patwardhan Siddharth, Michelizzi Jason. WordNet:: Similarity: measuring the relatedness of concepts." Demonstration papers at HLT-NAACL 2004. Association for Computational Linguistics; 2004.
[32] Kolhatkar Varada. An extended analysis of a method of all words sense disambiguation Diss. University of Minnesota; 2009.
[33] Slimani Thabet. Description and evaluation of semantic similarity measures approaches. arXiv preprint arXiv:1310.8059; 2013.
[34] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newslett 2009;11(1):10–8.
[35] Lim Noha et al. Classification of high-dimensional data with ensemble of logistic regression models. J Biopharm Stat 2009;20(1):160–71.
[36] Janecek Andreas, et al. On the relationship between feature selection and classification accuracy. New Challenges for Feature Selection in Data Mining and Knowledge Discovery; 2008.
[37] Sharaff A, Nagwani NK, Swami K. Impact of feature selection technique on email classification.
[38] Liu Y, Wang Y, Feng L, Zhu X. Term frequency com-bined hybrid feature selection method for spam filtering. Pattern Anal Appl 2016;19(2):369–83.
[39] Halder S, Tiwari R, Sprague A. Information extractionfrom spam emails using stylistic and semantic features to identify spammers. In: IEEE international conference on information reuse and integration (IRI), 2011. IEEE; 2011. p. 104–7.
[40] Suganya S, Gomathi C, et al. Syntax and semantics based efficient text classification framework. Int J Comput Appl 2013;65(15).
[41] Fdez-Glez J, Ruano-Ordas D, Méndez JR, Fdez-Riverola F, Laza R, Pavón R. A dynamic model for integrating simpleweb spam classification techniques. Expert Syst Appl 2015;42(21):7969–78.
[42] McInnes BT, Pedersen T. Evaluating measures of seman-tic similarity and relatedness to disambiguate terms in biomedicaltext. J Biomed Inform 2013;46(6):1116–24.

Eman Mohamed Bahgat received the BSc degree in Computer and Information Sciences from Ain Shams University, Cairo, Egypt, in 2011. In 2012, she joined the department of Information systems, Ain Shams University, as a Teaching Assistant. Her current research interest includes Email Classification based on semantic methods.

Sherine Rady is an Associate Professor at the Information Systems department of Faculty of Computer and Information Sciences - Ain Shams University in Cairo, Egypt. She got her B.Sc in Electrical Engineering (Computer and Systems) from the Faculty of Engineering - Ain Shams University. She acquired her M.Sc. in Computer and Information Sciences from Ain Shams University in 2003 and her Ph.D. from University of Mannheim in Germany in 2012. – Dr. Sherine Rady is a DAAD and JICA Alumni and her current research interests are Data Mining, Computer Vision, Medical Informatics and Information Retrieval.

Walaa Gad received the BSc and the MSc degrees in computers and information sciences in 2000 and 2005 respectively, from Ain Shams University, Cairo, Egypt. The master was about designing and planning a network model in the presence of obstacles using clustering around medoids techniques. She was a Ph. D student in

the Pattern and Machine Intelligence (PAMI) Group, Faculty of Electrical and Computer Engineering, University of Waterloo, Canada. She received her Ph.D in Computers and Information Sciences in 2010. The dissertation title is "Text Clustering Based on Semantic Measures". The work was done jointly between Faculty of Computers and Information Sciences, Ain Shams University and University of Waterloo in Canada. She is currently an Associate professor in faculty of computers and information sciences. She is the author of several publications. Her current research interests include data and knowledge mining, semantic web and machine learning, data warehouse and big data analytics.

Ibrahim Fathy Moawad is a full professor of Artificial Intelligence at Ain Shams University, Egypt. He also is the director of the Electronic and Knowledge Service Center and the Egyptian Universities Network, Supreme Council of Universities, Egypt. He got his Ph.D., M.Sc and B.Sc in 2005, 2000 and 1995 respectively. He has published more than 60 academic articles. He is a referee for many international journals and conferences. He is the author of the "A Framework for Multi-Agent Diagnosis System: Argumentation-based Negotiation Technique for Belief Conflict Resolution" book. He is the principle investigator of Ain Shams university for BioDialog project under the DAAD framework "Hochschuldialog mit der islamischen Welt", a collaborative project with Assuit university - Egypt, Safex university – Tunisia, and Jena university - Germany.