Spam email classification and sentiment analysis based on semantic similarity methods

Ulligaddala Srinivasarao* and Aakanksha Sharaff

Department of Computer Science and Engineering,
National Institute of Technology Raipur,
Chhattisgarh, India
Email: usrinivasarao.phd2018.cs@nitrr.ac.in
Email: asharaff.cs@nitrr.ac.in
*Corresponding author

Abstract: Electronic mail has widely been used for communication purposes, and the spam filter is required in the e-mail to save storage and protect from security issues. Various techniques based on NLP methods are used to increase spam detection efficiency. Spam detection cannot handle the unbalanced classes and lower efficiency due to irrelevant feature extraction in existing approaches. In this research, sentiment analysis-based semantic FE and hybrid FS techniques were used to increase the spam and non-spam detection efficiency in e-mail. The sentiment analysis is carried out in this proposed method with semantic feature extraction and hybrid FS. The sentiment analysis measures the polarity of the input text and used for e-mail spam classification. Different semantic similarity feature extraction methods are used in this proposed method. The TF-IDF, Information Gain, and Gini Index were used. The proposed semantic similarity and hybrid FS were evaluated with various classifiers. The experimental analysis shows that the Gini index FS technique, word2vec FE, and SVM classifier show the higher performance of 95.17% and RF with Gini index and word2vec methods has 93.3% accuracy in e-mail spam detection.

Keywords: artificial neural network; ANN; hybrid feature selection; HFS; semantic similarity; SVM; TF-IDF.

Reference to this paper should be made as follows: Srinivasarao, U. and Sharaff, A. (2023) 'Spam email classification and sentiment analysis based on semantic similarity methods', *Int. J. Computational Science and Engineering*, Vol. 26, No. 1, pp.65–77.

Biographical notes: Ulligaddala Srinivasarao is currently pursuing PhD in Department of Computer Science and Engineering at National Institute of Technology Raipur, Chhattisgarh, India. He received his Bachelor of Technology (BTech) and Master of Technology (MTech) in Department of Computer Science and Engineering from VR Siddhartha Engineering College Vijayawada, Andhra Pradesh, in 2015 and 2017 respectively. His areas of interest are data science, data mining, text mining, sentiment analysis, information retrieval, machine learning and deep learning.

Aakanksha Sharaff is working as an Assistant Professor in Department of Computer Science and Engineering at National Institute of Technology, Raipur Chhattisgarh, India. She completed her schooling, graduation and post-graduation with honours. She has teaching experience of more than nine years. She has published more than 51 research papers in reputed international journals and conferences. She has supervised 50 undergraduate and five postgraduate projects and supervising five PhD scholars. Her research areas focus mainly on data science, text analytics, sentiment analysis, information retrieval, soft computing, artificial intelligence, machine and deep learning.

1 Introduction

E-mail has become extremely popular for the communication purposes and lots of e-mails have been exchanged in daily basis. Emails have been used for various purposes, such as sending an important message within organisation/ inter-organisation, communication between countries, job recruitment processing and advertisements (Venkatraman et al., 2020). Recently, e-mail spam has become a problem for e-mail users that occupies more band width and e-mail memory and a number of e-mail spams have been increased daily for advertisement. Chikh and Chikhi (2019) proposed various methods have been proposed to reduce the problem of e-mail spam and save more memory for e-mail users. E-mail users receiving spam e-mail messages have the chance of exposing to security issues and inappropriate content. Furthermore, spam messages waste resources in terms of storage space, bandwidth and productivity (Faris et al., 2019). Therefore,

an efficient spam detection method is required to detect and block the spam to protect the user from the security issues and inappropriate content (Jain et al., 2019).

Recently, machine learning (ML) techniques in natural language processing (NLP) have become efficient in the performance of the spam based sentiment detection. ML model maps the e-mail features such as words, n-gram model, etc. to classify into spam/ham class. The hand-crafted rules (knowledge) can be used manually to extract e-mail features (Saidani et al., 2020). A ML method often provides higher efficiency of spam detection in e-mail and greater number of trifle feature or reiterative features may diminish the precision of e-mail spam detection. In feature selection (FS) methods, the potential is to improve the classification performance by selecting relevant subsets from the dataset (Ezpeleta et al., 2020). The pre-processing method can be used to provide proper information of the e-mail and remove the unwanted information to improve the learning process of the model (Dedeturk and Akay, 2020). Despite the presence of various NLP model for the spam detection in e-mail, still some challenges are needed to be improved for better classification performance such as high dimensional data and lack of training samples. In real-world applications, the spam detection model often encounters a large number of data, imbalance data, more computation time and more memory consumption (Asdaghi and Soleimani, 2019; Li et al., 2018; Sanghani and Kotecha, 2019). Nagwani and Sharaff (2017) proposed a SMS spam identification using two leave classifications and clustering approaches.

In this research, the sentiment based semantic similarity method is applied for e-mail spam detection. The pre-processing methods such as tokenisation, stop word removal, stemming and lemmatisation methods were used to represent the input data effectively. The various feature extraction (FE), FS and classification methods were used to evaluate the performance of the e-mail spam/ham detection. The analysis shows that the Gini index (GI), FS, word2vec FE and SVM classification methods have a high performance in e-mail spam detection.

The paper is organised as follows: Literature review is provided in Section 2, and proposed methodology explained in Section 3. The experimental results are provided in Section 4 and the conclusion of this paper is provided in Section 5.

2 Related works

In recent years, e-mail has become a major application in daily life, so there are many spam e-mails that occupy space in the e-mail. Therefore, the spam e-mail has to be filtered using sentiment analysis to effectively manage the space. Recent methods involved in applying the sentimental analysis to classify the spam e-mail are reviewed in the following section.

2.1 Spam identification

Shuaib et al. (2019), proposed whale optimisation algorithm to select salient features from the e-mail message and rotation forest (RF) algorithm to classify the spam/ham e-mails. The e-mail dataset were used to evaluate the RF algorithm with and without WOA FS method. Rodrigues and Chiplunkar (2019) proposed a Hybrid Lexicon-Naïve Bayesian Classifier (HL-NBC) method for sentiment analysis. Chikh and Chikhi, (2019) applied a combination of improved negative selection algorithm (NSA) and fruit fly optimisation (FFO) for spam e-mail detection methods. The spam benchmark dataset was used to estimate the performance and NSA-FFO method shows considerable performance. Sharaff et al., (2015) have compared filter approach which are chi-square (CS) and information gain (IG) with SVM classifier in classifying the spam and non-spam e-mail messages. Sharaff and Nagwani (2016) developed an e-mail thread identification using two clustering methods. In Sharaff and Srinivasarao (2020), they identified spam/ham e-mails using content and subject-based.

Faris et al. (2019) presented random weight network (RWN) and genetic algorithm (GA) method to detect the spam/ham e-mails. The most related features were analysed in the process to improve the automatic identification of spam e-mail. The three benchmark datasets were used to estimate the performance and GA-RWN method, which has higher performance of accuracy. Dedeturk and Akay (2020), proposed artificial bee colony FS method and logistic regression (LR) classification model for e-mail spam detection. The three datasets such as Enron, Turkish e-mail and CSDMC 2010 were used for evaluation and this model shows the considerable performance. The FS method with Naïve Bayes classifiers were evaluated in WEBSPAM-UK 2007 dataset and result shows higher efficiency. Li et al. (2018) applied sampling method and de-noising auto encoder in the input e-mail data and proposed deep belief network (DBN) for the e-mail spam detection. The DBN method was evaluated on WEBSPAM-UK 2007 dataset and shows the improvement in the performance. Sanghani and Kotecha (2019) applied TF-IDF method to select the features, learning model to update the classifiers and a selection rank weight to upgrade the new feature sets. Three e-mail databases were used to evaluate the performance and developed technique to reduce the false positive error. Diale et al. (2019) proposed FS method and applied in the random forest (RF) classifiers for the detection of e-mail spam. The analysis shows that the developed model has higher performance in detection of e-mail spam. Pashiri et al. (2020) proposed a sine-cosine algorithm with an artificial neural network (ANN) for e-mail spam/ham identification. The analysis shows higher performance in e-mail spam/ham detection.

Geler et al. (2021) proposed a new method to predict satisfaction level of the customers regarding the food and restaurants. Cao et al. (2021) proposed a sentiment classification scheme for fine-grained cross-domain using deep learning. Hesp et al. (2021) performed a study for

exploring evolutionary interfering of communication right from the start of evolutionary generation of core affect. Too and Rahim Abdullah (2020) presented a new approach to select the wrapper feature using atom search optimisation. Heinrich and Wermter (2018) proposed a neuron cognitively probable model for embodied multimodal language grounding. Further it illustrates in a natural interaction of a robotic agent. Devi et al. (2019) performed a thorough study over conventional learning methods to analyse the effects of class overlap and class imbalance. Xie et al. (2021) proposed a new strategy using two-player general-sum Markov game. It is developed for adaptive attack tolerance that is based on inference. Boden et al. (2017) presented seven high-level messages along with five ethical principles useful for robotics.

2.2 Semantic similarity based spam identification

Arif et al. (2018) applied rule based system of learning classifier system (LCS) for sentimental analysis to detect spam from SMS and e-mail, twitter message, and movie review. The encoding scheme has been developed based on the TF-IDF and sentiment lexicon to represent classifier. Liu and Lee (2018) used in trajectory clustering algorithm for sentiment analysis. The real email data have been used to evaluate the feasibility of the developed method. Bahgat et al. (2018) presented the email filtering method based on the semantic similarity method and WordNet ontology. The semantic and similarity measures reduce the extracted large number of textual features and also reduce the storage space, time complexity. The standard benchmark Enron dataset were used.

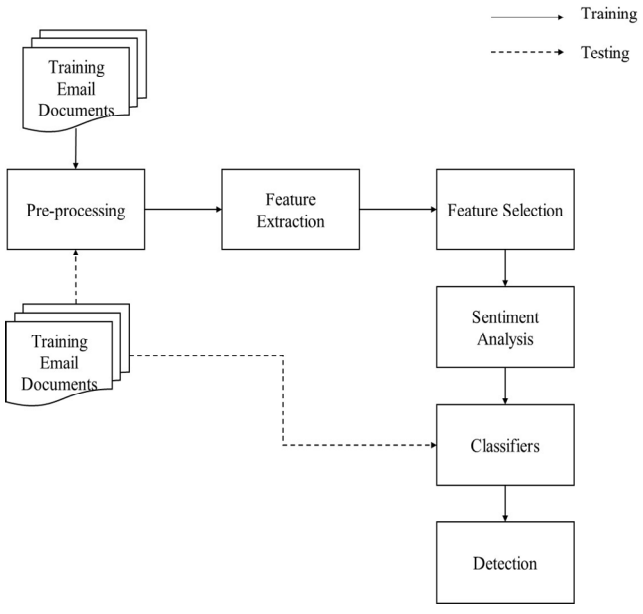
Barushka and Hajek (2018) proposed a TF-IDF, distribution based balancing algorithm, regularised deep multi-perceptron neural network (DMNN) model with rectified linear unit. The four benchmark databases were used to evaluate the performance of the proposed technique. The more complex features were captured by additional layers of neurons in the proposed method. Zhang et al. (2019) proposed quantum-inspired sentiment representation to represent the semantic and sentimental representation of the document. According to the subjective expression, namely adverbs and adjectives, sentimental phrases are extracted by QSR that are combined with the designed sentiment patterns.

Venkatraman et al. (2020) proposed a Naïve Bayes classifier with semantic similarity technique to analyse the ambiguity in the spam detection. The benchmark dataset such as Enron dataset, PU1, Ling-spam and Spam dataset were used to estimate the effectiveness of the Naïve Bayes with semantic similarity technique. The analysis shows Naïve Bayes with semantic similarity has higher performance in spam detection. Jain et al. (2019) applied semantic information based on WordNet and ConceptNet in CNN and LSTM for email spam detection. The semantic information increases the performance of the model and overfitting model affects the efficiency. Saidani et al. (2020) proposed two levels of semantic analysis for email spam/ham analysis. The first level involves in categorising the email and in second level, automatically extracted features were used for email spam detection. The analysis shows BoW and semantic information leads to improved results. de Mendizabal et al. (2020) applied NSGA-II with Naïve Bayes classifiers for the email spam detection. Mendez et al. (2019) applied IG, LDA and semantic similarity based FS method for the email spam detection. This model shows the high performance in the email spam detection.

3 Proposed method

Email has become the integral part of life and spam filter in the email is important to save storage space. Various techniques have been applied in the spam detection in email to increase the detection accuracy. In this research, the sentiment analysis based semantic similarity FE and hybrid FS based on the TF-IDF, IG, GI, Ambiguity measure and Distinguishing Feature selector were applied for email spam detection. Tokenisation, stop word removal, stemming and lemmatisation are used to filter the important information from the email data. The word 2 vectorisation, semantic similarity measure, word mover's distance (WMD) and local linear embedding are used in the FE method. The different classification methods such as OSLR, ANN, support vector machine, RF and decision tree (DT) are used to analysis the performance of the proposed method. The block diagram of the proposed semantic FE and the hybrid feature selection (HFS) method is shown in Figure 1.

Figure 1 The block diagram of sentiment analysis based semantic FE and HFS



3.1 Pre-processing techniques

Pre-processing techniques are commonly used in the NLP methods to reduce the redundant information from the dataset. The removal the unwanted information in email data helps to improve the classification performance and the

pre-processing techniques used in this research are Tokenisation, stop word removal, stemming and lemmatisation.

3.1.1 Tokenisation

Tokenisation method is involved in separating the composite text in the datasets into small tokens. The proposed model applies N-gram tokeniser to eliminate delimiters and word-spaces in the composite text.

3.1.2 Stop word removal

Stop words denote the most common words in a language such as 'of', 'is', 'the', and 'at'. Most researches in NLP consider stop word which affects the performance of the model and are removed from the input data before FE and FS process. The pre-compiled lists are the common method to remove stop words from the input data and it is used in this research.

3.1.3 Stemming and lemmatisation

Stemming and Lemmatisation is the normalisation method commonly used in NLP models to give a normalised form of input data. Stemming performs a basic form of approximation and doesn't replace the word. Lemmatisation method either removes the suffix or replaces the suffix completely from input data to form a lemma.

3.2 FE

FE methods such as WordNet, word2vector, smooth inverse frequency, cosine similarity (CS), Jensen – Shannon (JS) distance, WMD, local linear embedding and latent semantic index (LSA) were used in this proposed research.

3.2.1 WordNet semantic similarity

Ezzikouri et al. (2019) used WordNet semantic similarity measure between two terms based on the concept of each term described by a set of terms that can be used to measure its properties. WordNet uses the relationship with other similar terms in the hierarchical structure data. WordNet considers terms characteristics to measure similarity between different concepts, ignoring position and information on the taxonomy. WordNet can be measured using the equation (1).

$$Sim_{tvsk}(C1, C2) = \frac{|C1 \cap C2|}{|C1 \cap C2| + \alpha|C1 - C2| + (\alpha + 1)|C2 - C1|} \quad (1)$$

where two terms corresponding description is denoted as $C1$ and $C2$. The uncommon characteristics of relative importance are denoted as $\alpha \in [0, 1]$. The value of α increases with the similarity of terms and decreases the difference between the terms. The determination of α is based on observation and not necessarily a symmetric relation.

3.2.2 Word2 vector

Word2vec is used to express words based on the vector representation of a word, as shown in equation (2).

$$V = (v_1, v_2, v_3, \ldots, v_n) \quad (2)$$

where word space is denoted as V and vector of word space is denoted as v_1, v_2 of particular data.

3.2.3 Smooth inverse frequency

Karipbayeva et al. (2019) used a smooth inverse frequency is the sentence embedding method and highly used in NLP due to its simplicity and competitive performance. Consider the context vector $C \in \mathbb{R}^d$, the word w probability is emitted in the context, using equation (3).

$$p(w|c) = \alpha p(w) + (1 - \alpha)\left(\frac{\exp((w, \tilde{c}))}{Z_{\tilde{c}}}\right) \quad (3)$$

with $\tilde{c} = \beta c_0 (1 - \beta) c, c_0 \perp c,$

where a scalar hyper parameter is denoted as $\alpha, \beta \in [0, 1]$, word embedding for ω is denoted as $w \in \mathbb{R}^d$, common discourse is denoted as $C_0 \in \mathbb{R}^d$, and the normalising constant is represented as $Z_{\tilde{c}} = \sum_{\omega \in w} \exp((\tilde{c}, w))$.

3.2.4 CS

CS is easy to interpret and simple to compute for sparse vector matrix, as it is widely used in information retrieval and text mining methods (Al-Anzi and AbuZeina, 2017). CS measures the cosine angle between two vectors, as shown in equation (4). The document with different totals of same composition is allowed to be treated identically and this makes this method popular for text analysis.

$$S_{cosine}(x, y) = \frac{x'y}{\|x\|\|y\|} \quad (4)$$

where $\|x\| = \sqrt{\sum_{i=1}^{l} x_i^2}$ and $\|y\| = \sqrt{\sum_{i=1}^{l} y_i^2}$ are the lengths of the vector x and y respectively.

3.2.5 Jensen – Shannon distance

The JS distance is based on KL distance and this is also indexed to measure the similarity of two probability distributions that helps to solve the a symmetry problem. The formula for JS (Xu et al., 2019), is shown in equation (5).

$$JS(P\|Q) = \frac{1}{2} KL\left(P\|\frac{P+Q}{2}\right) + KL\left(Q\|\frac{P+Q}{2}\right) \quad (5)$$

The *JS* value presents between 0 to. The KL denotes $D(P\|Q)$ and formula of *KL* is shown in equation (5).

$$D(P\|Q)=\sum_{x\in X}P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{6}$$

3.2.5 WMD

The WMD was introduced based on the earth mover's distance, which provide solution of transportation problem (Wu et al., 2018). WMD measure the distance between two text documents x, $y \in X$ considering the distance between the words. The number of distinct words in x and y is denoted as $|x|, |y|$. The normalised frequency vectors of each word in the documents x and y is denoted as $f_x \in \mathbb{R}^{|x|}$, $f_y \in \mathbb{R}^{|y|}$, respectively. The WMD distance between documents x and y is defined in equation (7).

$$WMD(x, y) \coloneqq \min_{F\in\mathbb{R}^{|x|\times|y|}}(C, F) \tag{7}$$

$$\text{s.t. } F1 = f_x,\ F^T1 = f_y$$

where transportation flow matrix is denoted as F, F_{ij}, represent the flow travelling amount from i^{th} word, x_i in x to j^{th} word y_j in y, and transportation cost is denoted as C as $C_{ij} \coloneqq dist(v_{x_i}, v_{y_j})$ is distance between two words evaluated in the word2vec embedding space. The Euclidean distance $dist(v_{x_i}, v_{y_j}) = \|v_{x_i} - v_{y_j}\|$ is popular choice and used in this research.

3.2.6 Local linear embedding

Linear Locally Embedding represents each data point based on a linear combination of k nearest neighbours (Zeng et al., 2020). The LLE can be expressed as follows in equation (8).

$$\min_w m\sum_{i=1}^{n}\left\|x_i^m-\sum_{j=1}^{k}w_{ij}^m x_{ij}^m\right\|_2^2,\ \sum_j^k w_{ji}^m=1 \tag{8}$$

The number of neighbours is denoted as k and a linear relationship weighting factor is denoted as w_{ij}^m. The neighbourhood sample x_i does not have sample x_j and this is set as $w_{ij} = 0$. The Lagrange multiplier method is denoted in equation (8) and LLE is measured based on formula in equations (9) and (10).

$$\min_F \sum_{m=1}^{2}\alpha_m tr(FA_mF^T) \tag{9}$$

$$A_m = (I-W_m)^T(I-W_m) \tag{10}$$

$$\text{s.t. } FF^T = nI$$

The LLE common subspace is denoted as F and matrix W_m with w_{ij}^m elements, m^{th} modal is represent as m to represent the text modality. The manifold structure controlling parameter is denoted as α_m to preserve m^{th} modality item.

3.2.7 LSA

The LSA method is developed for a text retrieval method. On the term-document matrix, LSA method measures singular value decomposition (SVD). New matrix is constructed to provide the original term-document matrix based on first maximal singular values, and respective singular vectors. The dimension of the new matrix is reduced by removing noise that helps to achieve excellent retrieval performance (Li et al., 2011).

The term document matrix is denoted as $A_{K\times N}$ [equation (11)] that related to K terms in N documents. Based on SVD, the matrix $A_{K\times N}$ is split into three matrices, such as

$$A_{K\times N} = S_{K\times n}S_{n\times n}(V_{N\times n})' \tag{11}$$

where the number of documents is denoted as , the number of terms is denoted as K, n = min(K, N), U and V have orthogonal columns, i.e. $UU^T = V^TV = I$, the singular values of $A_{K\times N}$, and the singular values are sorted in non-increasing order so that $\delta_i \geq \delta_j$ for $i < j$.

The truncated SVD of $A_{K\times N}$ is selected based on the first maximum of T singular values from matrix S and keeping the corresponding columns in U and V, as given in equation (12).

$$A'_{K\times N} = U_{K\times T}S_{T\times T}(V_{N\times T})' \tag{12}$$

In the least squares sense, the $A'_{K\times N}$ is the best approximation to $A_{K\times N}$ of any rank-T. The matrix $A_{K\times N}$ can be denoted in reduced dimension and latent semantic feature space is given in equation (13).

$$\bar{A}_{T\times N} = S_{T\times T}(V_{N\times T})' \tag{13}$$

where latent space dimensionality is denoted as T, and each column of $\bar{A}_{T\times N}$ corresponds to a latent semantic feature of each training dataset. The normalised projection features are denoted as $W(B) = [w(u_1, B), \ldots, w(u_k, B)]'$ and its latent semantic feature is denoted as in equation (14).ϕ

$$\phi(B) = (U_{K\times T})' \cdot W(B) \tag{14}$$

3.3 FS

FS methods such as TF-IDF, IG and GI were used in this research for e-mail Spam/ham detection.

3.3.1 Term frequency-inverse document frequency

Pashiri et al. (2020) used TF-IDF is developed from IDF with the heuristic intuition term occurs less frequently in the document that is a good discriminator and should be given more weight for the term. The TF-IDF term weighting formula is given in equation (15).

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{15}$$

where term weight is denoted as $W_{i,j}$ for term i in document j, the collected number of documents is denoted as , the term frequency is denoted as $tf_{i,j}$, and the document frequency is denoted as df_i.

3.3.2 IG

The IG is used in gene analysis that can be used to evaluate difference between conditional entropy, (Gao et al., 2017). The IG reduction of uncertainty is denoted as $g(Y, X)$, as shown in equation (16).

$$g(Y, X) = H(Y) - H(Y|X) \tag{16}$$

where Y dataset entropy is denoted as $H(Y)$ that measures the uncertainty involved in predicting random variable value. The conditional entropy is denoted as $H(Y|X)$ that represent known variable X uncertainty. The probability distribution is denoted as p. $H(Y)$ and $H(Y|X)$ can be measured in equations (17) and (18).

$$H(Y) = -\sum p(y) \log p(y) \tag{17}$$

$$H(Y|X) = \sum_{x \in X} p(x) H\left(Y|X = x\right) \tag{18}$$

3.3.3 GI

Data samples are denoted as S, the various class label attribute that denotes various classes of C_i, (i = 1; 2; …; m) (Manek et al., 2017). Based on the class labels attribute values, can be divided into m subsets (S_i, i =1; 2; …; m). If subset samples belongs to class C_i, and the number of samples in the subset is S_i, the GI is denoted as in equation (19).

$$GiniIndex(S) = 1 - \sum_{i=1}^{m} P_i^2 \tag{19}$$

where probability P_i of any sample C_i estimate by s_i / s. The GI initial form is used to measure 'impurity' attribute for classification. The GI equation is shown in equation (20).

$$GiniIndex(S) = \sum_{i=1}^{m} P_i^2 \tag{20}$$

The 'purity' of attribute for categorisation is shown in equation (20).

3.4 Classifiers

Classifier uses the selected features from the FS method to classify the sentiment of the e-mail. Based on the sentiment analysis and selected features, the classifiers detect the spam e-mail from the input data.

3.4.1 DT

Tso and Yau (2007) proposed a DT model, a series of simple rules are applied to segment the data that are denoted in empirical tree. The set of rules perform the repetitive process of splitting the data for segmentation. The C5.0 is an improved version of C4.5 with differs as follows:

1 a nominal split have default branch-merging option

2 misclassification costs can be denoted

3 cross-validation and boosting are available

4 the rule set algorithm is improved.

The DT model has lower efficiency compared to neural networks for nonlinear data and is also affected by noisy data. The model is more suitable to predict categorical outcomes if sequential patterns and visible trends are available. The DT model has lower efficiency in time-series analysis.

3.4.2 Ordinary least squares regression (OLSR)

(Peng et al., 2019), An OLS is a linear approximation that reduces the sum of the squares of the distances between the observation points and the estimated points. The slope formula of ordinary least squares estimation is $\hat{\beta} = S_{XY}/S_{XX}$. Ordinary least squares is more suitable for the cases in which one of the two variables in equation (21).

$$\sum \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \tag{21}$$

3.4.3 ANN

ANN is a popular ML method that has been growing rapidly in recent years. ANN model has the capacity to handle nonlinear data and provide effective performance. Abid et al. (2020) developed a multi-layer neural architecture is a computation model. The ANN is inspired from human nervous system and the learning process of the ANN is based on pattern analysis of the network. ANN method is based on two processes, namely forward process and back propagation. In the activated network layer of forward process, the signals are processed in the forward direction, i.e., input to output. The error correction is the backward process based on bias term and connection weight. At each learning cycle, the back-propagation applies a gradient descent rule to minimise the network error. This method is repeated until the desired result is achieved. Many number of references related to neural networks with neural net model. The error value is used to weight the outputs and summed up in the output neuron. The input and output layer is explained as follows.

Input unit $o_1^1 = y$

Hidden units $o_1^2 = f(net_i), i = 1, \ldots, I \quad net_i = y.w_i^1 + b_i,$

where f is the sigmoid activation function.

Output unit $$N(y) = \sum_{i=1}^{I} \left(w_i^2 \cdot o_i^2 \right) = \sum_{i=1}^{I} \left(w_i^2 f \left(w_i^1 \cdot y + b_i \right) \right).$$

3.4.4 Support vector machine

The support vector machine is based on statistical learning method that uses the hyper plane to classify the data into various categories and hyper plane is developed from given dataset. The training feature dataset instances are labelled as $\{(x_i, y_i)\}, i = 1, 2, \ldots, N,$ where the number of instances is denoted as N, y_i is the class of instance x_i from input data. Hassonah et al. (2020) used an SVM, the maximum margin separating hyperplane is developed based on the closest points in high dimensional space. SVM computes the sum of distances between points of the hyper plane to closes points in high dimensional space to evaluate margin. The margin boundary function is computed as in equation (22).

$$Minimise W(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j k(x_i, x_j) - \sum_{i=1}^{N} \alpha_i \quad (22)$$

$$\forall_i : 0 \le \alpha_i \le C, \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0,$$

where α is a vector of N variables and soft margin parameter is denoted as C, $C > 0$. The SVM kernel function is denoted as $k(X_i, X_j)$. In this research, radial basis function (RBF) kernel is used, as in equation (23).

$$k(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2 \right), \gamma > 0 \quad (23)$$

where γ, r and d are kernel parameters.

3.4.5 RF

Izquierdo-Verdiguier and Zurita-Milla (2020) used a RF model which is the combination of DT and reduces the error in classification regression task based on bootstrap aggregation or bagging. The RF is fast and robust method to noise of the target data. The RF is to reduce the prediction error considered DTs within forest and the correlation among their predictions.

Focusing on one tree of the forest, let $P_i \in \mathbb{R}^{M_i \times N_i}$ where the i defines the i^{th} partition of samples (M_i) and features (N_i). Random samples are generated by selecting P_i from original data $\left(X \in \mathbb{R}^{M \times N} \right)$. At each node, subset feature N_i split the available samples (M_i).The best splitting feature and cut-off point are measured using GI. The samples having higher values compared to cut-off values are directed to the right node (v_R) or directed to the left node (v_L). Once several splits are performed, then samples moves from the root node (v_n) to the terminal nodes as a terminal leaves which supply the predictions of the samples. Forest provided ensemble prediction $\hat{Y} \in \mathbb{R}^{M \times 1}$ that obtained as the combination of individual trees results; typically using the majority vote rule for classification or the average for regression problems:

Classification $$\hat{Y}_i = \text{mode}\_\left(n = 1 \ldots N_{trees} \widehat{Y}_n \right)$$

Regression $$\hat{Y}_i = \frac{1}{N_{trees}} \sum_{n=1}^{N_{trees}} \widehat{Y}_n$$

where N_{trees} is the total number of trees used in the RF. The algorithm of sentiment based semantic similarity for spam detection is given as follows:

Algorithm of proposed method

Obtain input e-mail data
Perform tokenisation
    \\ Pre-processing
Perform stop word removal
Perform stemming and lemmatisation
Convert pre-processed word to vector using equation (2)
    \\ Feature extraction
Perform Gini index to select features based on equation (29)
    \\ Gini index feature selection
Apply selected features to SVM
Classify the input data using equation (22)
    \\ SVM classification

4 Experimental result

The e-mails are highly used for the communication purposes and spam mails are required to be removed to save storage. The various techniques were applied to increase the efficiency of spam/ham mail detection. In this research, semantic similarity FE and HFS methods were used to increase the efficiency of the spam/ham detection in the e-mail. The pre-processing method such as tokenisation, stop word removal, stemming and lemmatisation was used for effective representation of input e-mail data. The several classifiers such as OSLR, DT, ANN, SVM and RF were used to evaluate the performance of the method. The proposed method classifies the input e-mail data into three sentiment categories such as positive, negative and neutral. This section provides the detailed description about the performance of various methods.

- *Dataset*: (Metsis et al., 2006), Used an Enron-spam dataset to evaluate the performance of the proposed semantic similarity FE method in e-mail spam detection. There are totally 5,975 e-mails are present in the dataset with 4,672 ham e-mails and 1,303 spam e-mails.
- *Evaluation metrics*: The evaluation metrics such as accuracy, precision, recall and RMSE values were measured by the proposed semantic based similarity FE method. The formula for accuracy, precision, recall and RMSE were shown in equations (24)–(27).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

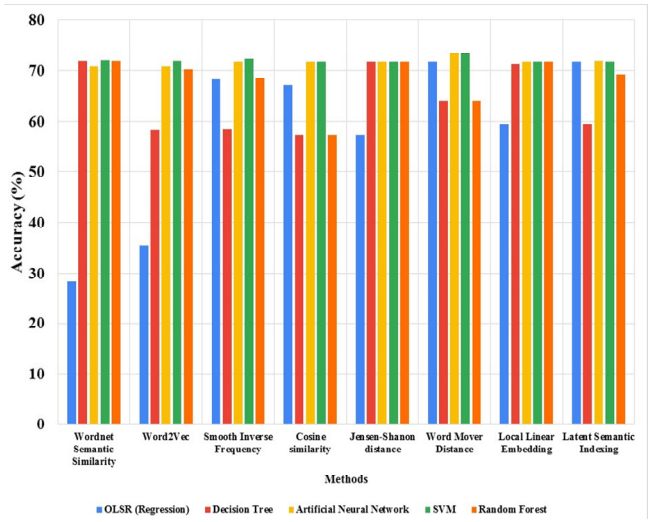
$$Recall = \frac{TP}{TP + FN} \tag{26}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(\widehat{y_i} - y_i\right)^2}{n}} \tag{27}$$

- *System requirement*: The proposed method is evaluated in the system consisting of Intel i5 processor, 8 GB of RAM and 2 GB graphics card. The proposed method is implemented using the python tool.

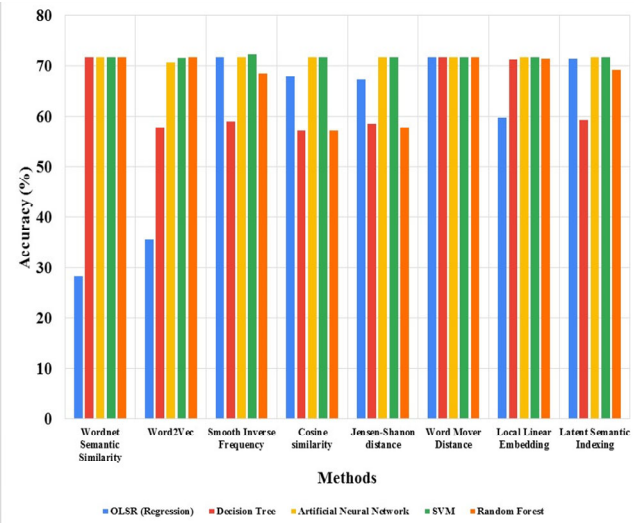
The accuracy of document frequency with various FS methods and classifiers in e-mail spam detection are shown in Figure 2. The SVM and RF classifiers have higher performance than other classifiers compared. The WMD with SVM and ANN classifier has higher performance in spam detection. The SVM classifier has stable performance in other FE method in e-mail spam detection. The SVM with WMD has accuracy of 73.42% and RF with WMD has 64.09% accuracy.

Figure 2 Accuracy of document frequency with various FE method (see online version for colours)



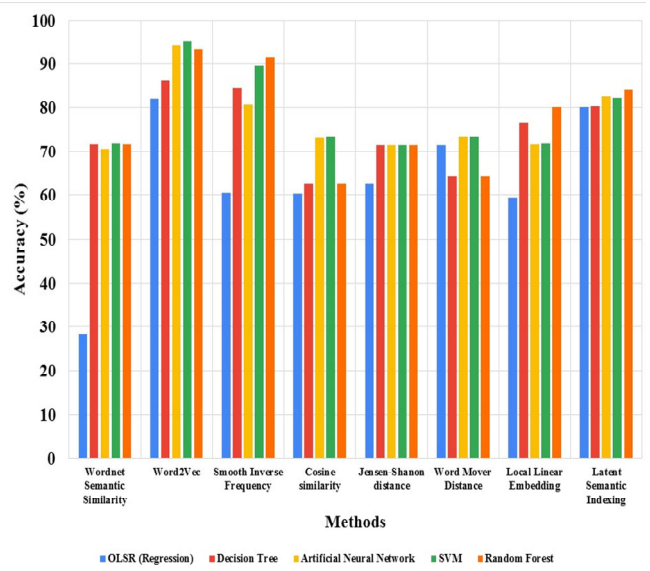
The accuracy of IG FS method with various FE and classifiers were shown in Figure 3. The SVM classifiers have higher performance with most of the FE method used in this research. The sentiment analysis, semantic FE and HFS effectively improves the performance of classification. The smooth inverse frequency FE method with SVM classifier has higher performance in IG FS method. The WMD provides considerable performance with various classifiers in e-mail spam detection. The SVM with smooth inverse frequency has 72.2% accuracy and RF with smooth inverse frequency has 68.46% accuracy.

Figure 3 Accuracy of IG FS with various FE (see online version for colours)



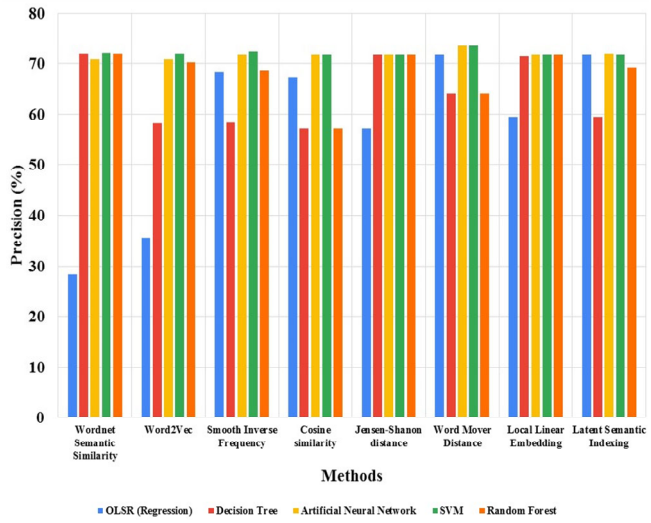
The accuracy of GI FS method with various FE and classifier were shown in Figure 4. The GI FS method has the higher performance compared to document frequency and IG. The word2vec FE method with SVM classifier provides the higher performance in e-mail spam detection. The RF with various FE method have considerable performance in e-mail spam detection. The SVM with word2vec method has an accuracy of 95.17% and RF with word2vec method has 93.3% accuracy.

Figure 4 Accuracy of GI FS method with various FE (see online version for colours)



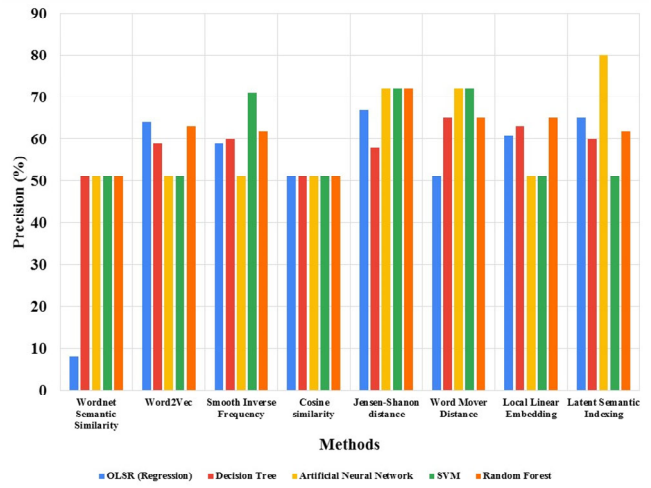
The precision value of document frequency with various FE method and classifiers are shown in Figure 5. The WMD with SVM and ANN classifier have higher precision value compared to the other FE method. The SVM provides the stable performance in various FE method and ANN method has considerable performance. The SVM with WMD has 73.42% precision and RF with WMD has 64.09% precision.

Figure 5 Precision of document frequency with various FE method (see online version for colours)



The precision of IG FS method with various FE and classifiers are shown in figure 6. The latent semantic indexing FE method with ANN classifier has higher precision value compared to other FE method. The SVM with WMD has considerable performance in e-mail spam detection. The ANN with the latent semantic indexing method has a precision value of 80% and SVM with the WMD method has 72% precision.

Figure 6 Precision of IG with various FE methods (see online version for colours)



The precision value of the GI with various FE and classifiers were shown in Figure 7. The word2vec FE with SVM has higher precision value compared to the other FE method. The smooth inverse frequency with RF has considerable performance in e-mail spam detection. The word2vec with the SVM method has a precision of 95% and the word2vec with ANN has 94% precision.

The recall value of document frequency with various FE method and classifiers were compared in Figure 8. The WMD with SVM and ANN have higher recall value than other FE method. The ANN and SVM classifiers have considerable performance with various FE in e-mail spam detection. The WMD with the SVM method has recall value of 73% and the WMD to RF has a 64% recall value.

Figure 7 Precision value of GI with various FE methods (see online version for colours)

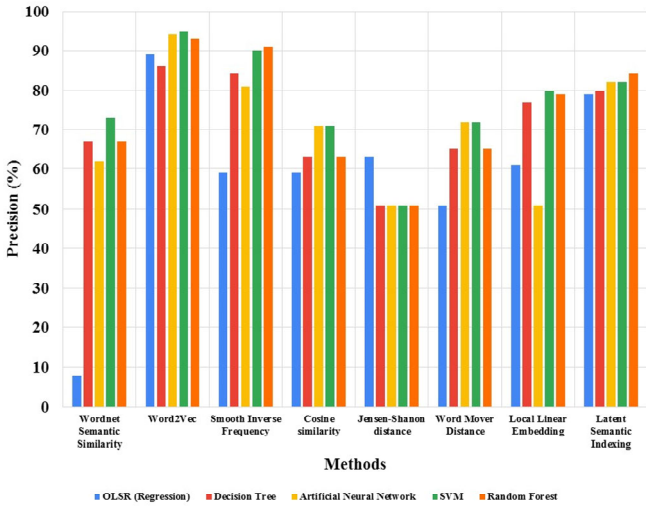
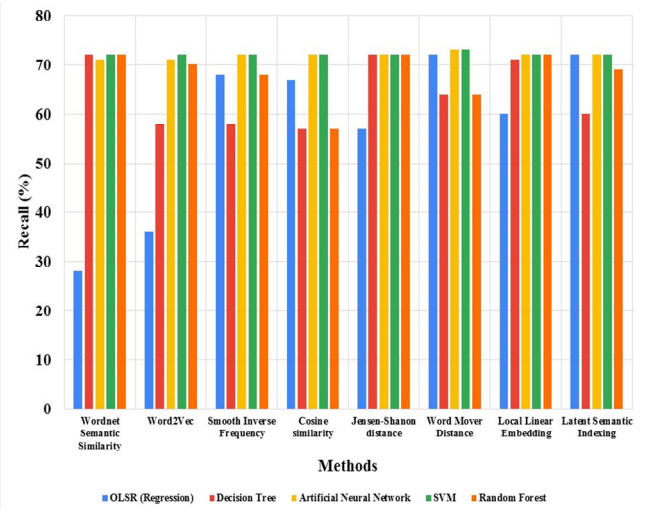


Figure 8 Recall value of document frequency with various FE (see online version for colours)



The recall value of IG with various FE and classifiers were compared in Figure 9. The WMD with SVM and ANN have higher recall value compared to the other FE method. The WMD with SVM has recall value of 73% and the WMD to RF has a 64% recall value.

The recall value of GI with various FE and classifiers were compared in figure 10. The word2vec with SVM has higher recall value compared to other FE methods. The semantic FE and HFS method effectively improves the Spam classification performance. The word2vec with the SVM method has the higher recall value compared to document frequency and IG and FS methods. The word2vec with the SVM method has recall value of 95% and word2vec with ANN method has a 94% recall value.

Figure 9 Recall value of IG with various FE (see online version for colours)

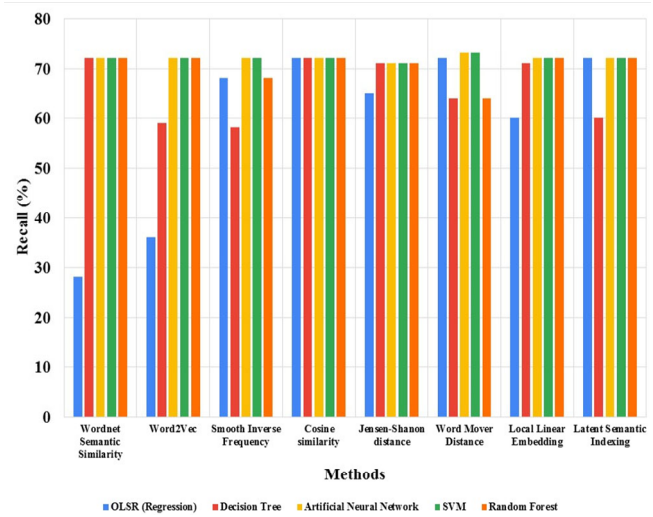


Figure 10 The recall value of GI with various FE methods (see online version for colours)

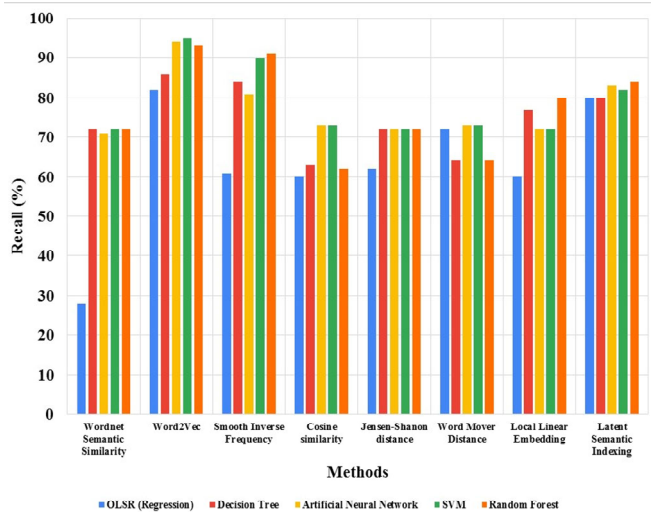
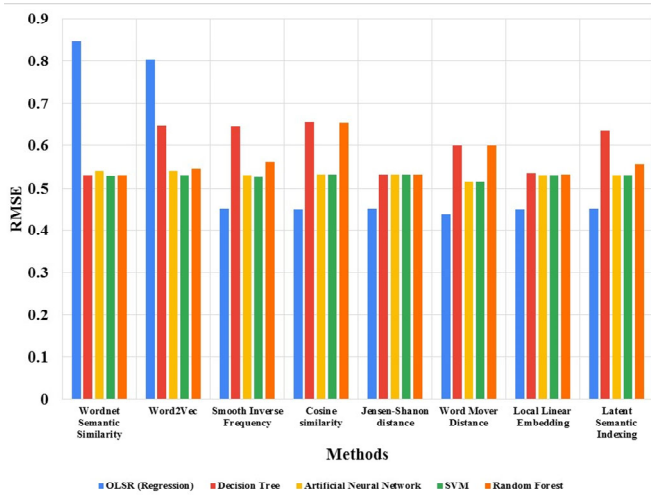


Figure 11 RMSE value of document frequency with various FE methods (see online version for colours)



The RMSE value of document frequency with various FE and classifiers were compared in Figures 11–13. This shows that sentiment analysis improves the performance of Spam classification in e-mail text. The OLSR method has a lower error value with some FE methods such as smooth inverse frequency, CS, WMD, etc. The SVM classifier has considerable lower RMSE value compared to other classifier compared with this research.

The accuracy of the 10-fold cross validation method with various FE, FS and classifier were shown in Figure 14. The GI FS, Jensen-Shanon FE method and SVM classifier method have high performance in the 10-fold cross validation analysis. In word2vec FE method with SVM classifier provides the second higher performance in e-mail spam detection. The RF with various FE method have considerable performance in e-mail spam detection. The SVM with word2vec method has an accuracy of 95.17% and RF with word2vec method has 94.2% accuracy. This shows that the proposed method effectively improves the performance of Spam classification due to sentiment analysis. The sentiment analysis, semantic FE and HFS improves the performance of spam classification.

Figure 12 RMSE of IG with various FE methods (see online version for colours)

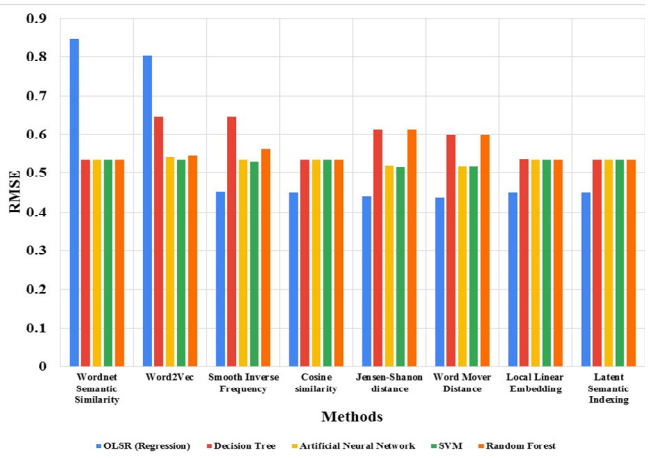


Figure 13 RMSE of GI with various FE methods (see online version for colours)

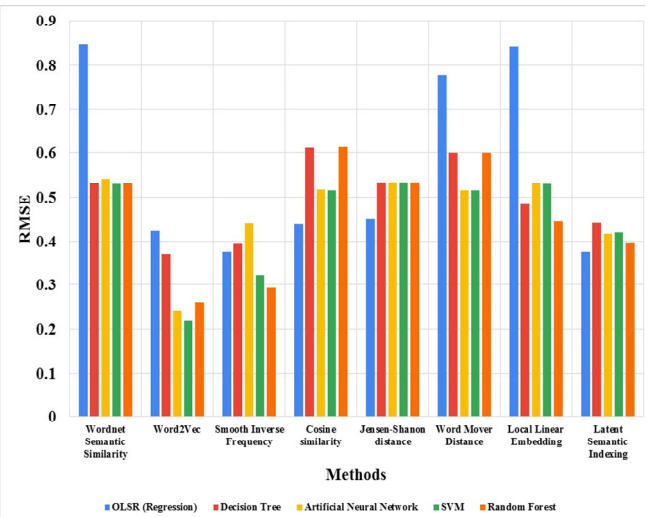
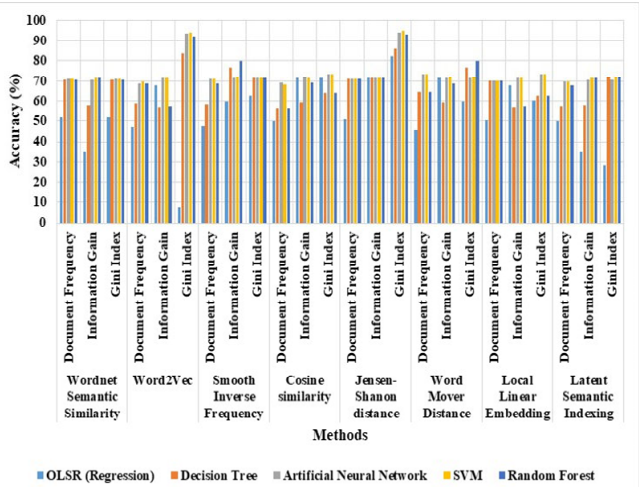


Figure 14 Accuracy of 10-fold cross validation (see online version for colours)



The analysis shows that the GI FS method, word2vec FE and SVM classifier show the high performance in detecting the spam in e-mail. The WMD with document frequency and IG with SVM classifier have considerable performance in e-mail spam detection. The SVM classifier shows the higher performance in the analysis, and ANN and RF classifier shows considerable performance in e-mail spam detection.

5 Summary and conclusions

E-mail is widely used for communication purposes and spam detection is required in the e-mail to save storage. The existing methods involved in the spam detection have the limitation of irrelevant FS and failed to handle the unbalance classes. In this paper, the sentiment analysis based semantic FE and the HFS method were used to increase the efficiency of the spam detection in e-mail. This research involves in applying the sentiment analysis is one of the features along with semantic FE and hybrid FS method. The sentiment analysis measures the polarity of input e-mail text to improve the efficiency of spam classification. The word 2 vectorisation, semantic similarity, WMD, and local linear embedding were the FE methods and the TF-IDF, IG, GI, ambiguity measure and distinguish feature selector were the FS methods use in this research. The classifiers classify the input e-mail data into three sentiment categories and spam detection is performed based on sentiment analysis.

The experimental analysis shows that the GI FS method and word2vec FE method with SVM have the higher performance than other compared method. The GI-word2vec-SVM method has 95.17% accuracy and GI-word2vec-RF method has 93.3% accuracy. The proposed semantic similarity and hybrid FS effectively improves the e-mail spam classification due to relevant FS. The proposed method has advantage of considering semantic and hybrid FS method and existing method has limitation of irrelevant FS in e-mail spam classification. The proposed method is capable to apply in the e-mail services to effectively classify the ham and spam. Furthermore, the proposed method is useful to understand the polarity of the input text and capable to apply in messing services also. The proposed method suffers from the limitations of high computation complexity and deep learning method is possible solution to apply instead of FE and FS to overcome this limitation. The future work of this research involves applying deep learning method to analyse the relevant information to improve the spam detection method.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

The authors would like to thank the National Institute of Technology Raipur, India for providing infrastructure and facilities to carry out this research work.

References

Abid, F., Li, C. and Alam, M. (2020) 'Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks', *Computer Communications*, Vol. 157, pp.102–115, doi.org/10.1016/j.comcom.2020.04.002.

Al-Anzi, F.S. and AbuZeina, D. (2017) 'Toward an enhanced Arabic text classification using cosine similarity and latent semantic indexing', *Journal of King Saud University – Computer and Information Sciences*, Vol. 29, No. 2, pp.189–195.

Arif, M.H., Li, J., Iqbal, M. and Liu, K. (2018) 'Sentiment analysis and spam detection in short informal text using learning classifier systems', *Soft Computing*, Vol. 22, No. 21, pp.7281–7291.

Asdaghi, F. and Soleimani, A. (2019) 'An effective feature selection method for web spam detection', *Knowledge-Based Systems*, Vol. 166, pp.198–206, doi.org/10.1016/j.knosys.2018.12.026.

Bahgat, E.M., Rady, S., Gad, W. and Moawad, I.F. (2018) 'Efficient email classification approach based on semantic methods', *Ain Shams Engineering Journal*, Vol. 9, No. 4, pp.3259–3269.

Barushka, A. and Hajek, P. (2018) 'Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks', *Applied Intelligence*, Vol. 48, No. 10, pp.3538–3556.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S. and Winfield, A. (2017) 'Principles of robotics: regulating robots in the real world', *Connection Science*, Vol. 29, No. 2, pp.124–129.

Cao, Z., Zhou, Y., Yang, A. and Peng, S. (2021) 'Deep transfer learning mechanism for fine-grained cross-domain sentiment classification', *Connection Science*, Vol. 33, No. 4, pp.911–928, doi.org/10.1080/09540091.2021.1912711.

Chikh, R. and Chikhi, S. (2019) 'Clustered negative selection algorithm and fruit fly optimization for email spam detection', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 10, No. 1, pp.143–152.

de Mendizabal, I.V., Basto-Fernandes, V., Ezpeleta, E., Méndez, J.R. and Zurutuza, U. (2020) 'SDRS: a new lossless dimensionality reduction for text corpora', *Information Processing and Management*, Vol. 57, No. 4, p.102249.

Dedeturk, B.K. and Akay, B. (2020) 'Spam filtering using a logistic regression model trained by an artificial bee colony algorithm', *Applied Soft Computing*, Vol. 91, No. 4, p.106229, doi.org/10.1016/j.asoc.2020.106229.

Devi, D., Biswas, S.K. and Purkayastha, B. (2019) 'Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique', *Connection Science*, Vol. 31, No. 2, pp.105–142.

Diale, M., Celik, T. and van der Walt, C. (2019) 'Unsupervised feature learning for spam email filtering', *Computers and Electrical Engineering*, Vol. 74, pp.89–104, doi.org/10.1016/j.compeleceng.2019.01.004.

Ezpeleta, E., Velez de Mendizabal, I., Hidalgo, J.M.G. and Zurutuza, U. (2020) 'Novel email spam detection method using sentiment analysis and personality recognition', *Logic Journal of the IGPL*, Vol. 28, No. 1, pp.83–94.

Ezzikouri, H., Madani, Y., Erritali, M. and Oukessou, M. (2019) 'A new approach for calculating semantic similarity between words using WordNet and set theory', *Procedia Computer Science*, Vol. 151, pp.1261–1265, doi.org/10.1016/j.procs.2019.04.182.

Faris, H., Ala'M, A.Z., Heidari, A.A., Aljarah, I., Mafarja, M., Hassonah, M.A. and Fujita, H. (2019) 'An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks', *Information Fusion*, Vol. 48, pp.67–83, doi.org/10.1016/j.inffus.2018.08.0020

Gao, L., Ye, M., Lu, X. and Huang, D. (2017) 'Hybrid method based on information gain and support vector machine for gene selection in cancer classification', *Genomics, Proteomics and Bioinformatics*, Vol. 15, No. 6, pp.389–395.

Geler, Z., Savić, M., Bratić, B., Kurbalija, V., Ivanović, M. and Dai, W. (2021) 'Sentiment prediction based on analysis of customers assessments in food serving businesses', *Connection Science*, Vol. 33, No. 3, pp.1–19, doi.org/10.1080/09540091.2020.1870436.

Hassonah, M.A., Al-Sayyed, R., Rodan, A., Ala'M., A.Z., Aljarah, I. and Faris, H. (2020) 'An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter', *Knowledge-Based Systems*, Vol. 192, p.105353, doi.org/10.1016/j.knosys.2019.105353.

Heinrich, S. and Wermter, S. (2018) 'Interactive natural language acquisition in a multi-modal recurrent neural architecture', *Connection Science*, Vol. 30, No. 1, pp.99–133.

Hesp, C., Heerebout, B.T. and Phaf, R.H. (2021) 'Evolutionary computation for bottom-up hypothesis generation on emotion and communication', *Connection Science*, Vol. 33, No. 2, pp.296–320.

Izquierdo-Verdiguier, E. and Zurita-Milla, R. (2020) 'An evaluation of guided regularized rand om forest for classification and regression tasks in remote sensing', *International Journal of Applied Earth Observation and Geoinformation*, Vol. 88, p.102051, doi.org/10.1016/j.jag.2020.102051.

Jain, G., Sharma, M. and Agarwal, B. (2019) 'Spam detection in social media using convolutional and long short term memory neural network', *Annals of Mathematics and Artificial Intelligence*, Vol. 85, No. 1, pp.21–44.

Karipbayeva, A., Sorokina, A. and Assylbekov, Z. (2019) *A Critique of the Smooth Inverse Frequency Sentence Embeddings*, arXiv: 1909.13494.

Li, D.X., Peng, J.Y., Li, Z. and Bu, Q. (2011) 'LSA based multi-instance learning algorithm for image retrieval', *Signal Processing*, Vol. 91, No. 8, pp.1993–2000.

Li, Y., Nie, X. and Huang, R. (2018) 'Web spam classification method based on deep belief networks', *Expert Systems with Applications*, Vol. 96, pp.261–270, doi.org/10.1016/j.eswa.2017.12.016.

Liu, S. and Lee, I. (2018) 'Discovering sentiment sequence within email data through trajectory representation', *Expert Systems with Applications*, Vol. 99, pp.1–11, doi.org/10.1016/j.eswa.2018.01.026.

Manek, A.S., Shenoy, P.D., Mohan, M.C. and Venugopal, K.R. (2017) 'Aspect term extraction for sentiment analysis in large movie reviews using Gini index feature selection method and SVM classifier', *World Wide Web*, Vol. 20, No. 2, pp.135–154.

Mendez, J.R., Cotos-Yanez, T.R. and Ruano-Ordas, D. (2019) 'A new semantic-based feature selection method for spam filtering', *Applied Soft Computing*, Vol. 76, pp.89–104, doi.org/10.1016/j.asoc.2018.12.008.

Metsis, V., Androutsopoulos, I. and Paliouras, G. (2006) 'Spam filtering with Naive Bayes-which naive bayes?', in *CEAS*, July, Vol. 17, pp.28–69.

Nagwani, N.K. and Sharaff, A. (2017) 'SMS spam filtering and thread identification using bi-level text classification and clustering techniques', *Journal of Information Science*, Vol. 43, No. 1, pp.75–87.

Pashiri, R.T., Rostami, Y. and Mahrami, M. (2020) 'Spam detection through feature selection using artificial neural network and sine–cosine algorithm', *Mathematical Sciences*, Vol. 14, No. 3, pp.193–199.

Peng, Z., Guan, L., Liao, Y. and Lian, S. (2019) 'Estimating total leaf chlorophyll content of gannan navel orange leaves using hyperspectral data based on partial least squares regression', *IEEE Access*, Vol. 7, pp.155540–155551, doi:10.1109/ACCESS.2019.2949866.

Rodrigues, A.P. and Chiplunkar, N.N. (2019) 'A new big data approach for topic classification and sentiment analysis of Twitter data', *Evolutionary Intelligence*, pp.1–11.

Saidani, N., Adi, K. and Allili, M.S. (2020) 'A semantic-based classification approach for an enhanced spam detection', *Computers and Security*, Vol. 94, p.101716, doi.org/10.1016/j.cose.2020.101716.

Sanghani, G. and Kotecha, K. (2019) 'Incremental personalized e-mail spam filter using novel TFDCR feature selection with dynamic feature update', *Expert Systems with Applications*, Vol. 115, pp.287–299, doi.org/10.1016/j.eswa.2018.07.049.

Sharaff, A. and Nagwani, N.K. (2016) 'Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques', *Journal of Information Science*, Vol. 42, No. 2, pp.200–212.

Sharaff, A. and Srinivasarao, U. (2020) 'Towards classification of email through selection of informative features', in *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, IEEE, January, pp.316–320.

Sharaff, A., Nagwani, N.K. and Swami, K. (2015) 'Impact of feature selection technique on email classification', *Int. J. Knowl. Eng.*, Vol. 1, No. 1, pp.59–63.

Shuaib, M., Adebayo, O.S., Osho, O., Idris, I., Alhassan, J.K. and Rana, N. (2019) 'Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification', *SN Applied Sciences*, Vol. 1, No. 5, p.390.

Too, J. and Rahim Abdullah, A. (2020) 'Binary atom search optimisation approaches for feature selection', *Connection Science*, Vol. 32, No. 4, pp.406–430.

Tso, G.K. and Yau, K.K. (2007) 'Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks', *Energy*, Vol. 32, No. 9, pp.1761–1768.

Venkatraman, S., Surendiran, B. and Kumar, P.A.R. (2020) 'Spam e-mail classification for the internet of things environment using semantic similarity approach', *The Journal of Supercomputing*, Vol. 76, No. 2, pp.756–776.

Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y. and Witbrock, M.J. (2018) *Word Mover's Embedding: from Word2vec to Document Embedding*, arXiv: 1811.01713.

Xie, Y.X., Ji, L.X., Li, L.S., Guo, Z. and Baker, T. (2021) 'An adaptive defense mechanism to prevent advanced persistent threats', *Connection Science*, Vol. 33, No. 2, pp.359–379.

Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C. and Yao, H. (2019) 'Research on topic detection and tracking for online news texts', *IEEE Access*, Vol. 7, pp.58407–58418, doi: 10.1109/ACCESS.2019.2914097.

Zeng, H., Zhang, H. and Zhu, L. (2020) 'Label consistent locally linear embedding based cross-modal hashing', *Information Processing and Management*, 57, No. 6, pp.102136.

Zhang, Y., Song, D., Zhang, P., Li, X. and Wang, P. (2019) 'A quantum-inspired sentiment representation model for Twitter sentiment analysis', *Applied Intelligence*, Vol. 49, No. 8, pp.3093–3108.