International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Recurrent Neural Networks for Spam E-mail Classification on an Agglutinative Language

Sahin Işik[1*], Zuhal Kurt[2], Yildiray Anagun[3], Kemal Ozkan[4]

Submitted: 18/10/2020 Accepted : 24/12/2020

Abstract: In this study, we have provided an alternative solution to spam and legitimate email classification problem. The different deep learning architectures are applied on two feature selection methods, including the Mutual Information (MI) and Weighted Mutual Information (WMI). Firstly, feature selection methods including WMI and MI are applied to reduce number of selected terms. Secondly, the feature vectors are constructed with concept of the bag-of-words (BoW) model. Finally, the performance of system is analyzed with using Artificial Neural Network (ANN), Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BILSTM) models. After experimental simulations, we have observed that there is a competition between detection results of using WMI and MI when commented with accuracy rates for the agglutinative language, namely Turkish. The experimental scores show that the LSTM and BILSTM give 100% accuracy scores when combined with MI or WMI, for spam and legitimate emails. However, for particular cross-validation, the performance WMI is higher than MI features in terms e-mail grouping. It turns out that WMI and MI with deep learning architectures seem more robust to spam email detection when considering the high detection scores.

Keywords: *RNN, Odds Ratio, Mutual Information, Spam E-mail, LSTM*

1. Introduction

The technology brings the valuable advantages to our life. The one of biggest advantage can be noted as email communication service. With email conversation, the wasted time is greatly in many public and private organizations since it is extremely fast in terms of shipping. By sharing documents, opinions and information, the companies become more organized to reach their staffs and customers. Moreover, with a lost cost, the companies can share their advertisements, brand the potential clients and boost the sales. Technically, a company can identify the pain point by taking messages from targeted clients. However, there are some weakness of email communication, which are reasoned by viruses, phishing and spam emails. Therefore, some security preventions are required in order to reduce the vulnerability of documents and information loss.

Technically, the emails can be categorized in two classes, as spam (malicious) or legitimate (normal) [1-3]. The spam emails are called as junk or unsolicited bulk emails. Moreover, the meaning of spam cannot be restricted to malicious definition. The spam definition varies with respect to target recipients. In an example, some legitimate emails can irrigate the staffs of an occupational group. Therefore, we have to emphasize that system for spam email detection must be developed based on the recipient's needs or the recipient must able to decide which emails are spam. The main aim of a developed system is reducing the stealing of privacy information, wasting of resources and reading time. In this context, a robust system is greatly needed when it comes to assign an email as spam or legitimate.

According to some statistics provided by Kaspersky [4], the percentage of spam emails during the first quarter of 2014 year was 66.34%, while this percentage was 69.6% in 2013 year. Similarly, the Kaspersky Lab was reported that top spam countries were Asian ones when the geographical distribution of spam by countries are evaluated. In general, the Trojan viruses were observed as malicious attachments in spam emails. These harmful attachments are aimed to steal the confidential data, which may be credit card numbers, industry specific data, employment information and trade secrets. Moreover, the identity theft attacks have oriented on social networks. Interestingly, the Kaspersky records that 130 million phishing attacks in second quartile of 2019 year [5], as it was observed that most targeted organizations were banks and payment systems.

Going through the literature, we can observe that a crowded set of study is performed on spam email detection by using statistical feature extraction/selection methods and classifiers. The general methodology of these studies is categorized into four stages; preprocessing, feature extraction/selection, and classification. In a study, the Artificial Neural Networks (ANN) and Bayesian Networks used for spam email classification. The performance was reported to 90% accuracy rate for the Turkish language [2]. In another study, the filter approaches (Document Frequency (DF), Mutual Information (MI) and Chi-Square (CHI2) and Information Gain (IG)) and the wrapper approaches (Genetic Algorithm Selection (GS)) were used together as a hybrid approach for text classification [6]. Moreover, a mobile framework was introduced for real-time classify spam emails. In the referred study, the CHI2 and IG-based feature selection procedure is carried out and highest overall accuracy rate is reported as 90.17%, for 747 spam and

[1] *Computer Eng., Eskisehir Osmangazi University, Eskisehir, TURKEY, ORCID ID : 0000-0003-1768-7104*
[2] *Computer Eng., Atılım University, Ankara, TURKEY ORCID ID : 0000-0003-1740-6982*
[3] *Computer Eng., Eskisehir Osmangazi University, Eskisehir, TURKEY, ORCID ID : 0000-0003-2737-2720*
[4] *Computer Eng., Eskisehir Osmangazi University, Eskisehir, TURKEY, ORCID ID : 0000-0003-2252-2128*
** Corresponding Author Email: sahini@ogu.edu.tr*

4,827 legitimate emails of English language [1]. Additionally, the IG, Gini Index (GI) and CHI2 were applied with ANN and Decision Tree (DT) for email classification on Turkish language [7]. Furthermore, , the contribution of feature vector dimension was analyzed by using MI for spam email detection on Turkish e-mails [8]. Recently, the deep learning-based approaches [9-12] have been performed with a purpose of text classification.

The main objective of this study is investigating the discriminative capability of recurrent neural networks with feature selection methods. for e-mail classification over a special agglutinative language, namely Turkish. The Ods Ratio (OR) is used to weight the MI features, called as WMI. After obtaining the discriminative features, three well-known learning methods, which are ANN, LSTM [13] and BILSTM [14] are utilized on determined BoW histograms. To assess the performance of each individual classifier, the dimension of feature vector is changed as 50x5, 75x5, 100x5, 125x5 and 150x5 by using the new term weighting feature. The other important research of this study is that it is concentrated on a complex language as hard to find general rules to process due its characteristic property.

The rest of the paper is organized that Section 2 displays the feature extraction stage. The Section 3 reveals the feature selection stage. The utilized classifiers are explained in the Section 4. The findings of the experimental study are presented in Section 5. Finally, a conclusion is given in last section.

2. Term Selection (Feature Selection)

The key idea behind feature selection is representing data with effective size in terms efficiency for a learning model construction, the feature selection strategy is important to represent the data with effective size to extract the interesting pattern. There various benefits of using feature selections, which can be listed as improving the running of algorithm, enhancing performance, improving generalization, reducing the memory and hardware usage. Feature selection can be fulfilled based on the classifier dependent or independent way, as supervised or unsupervised, respectively. In case of supervised way, the wrapper, filter and hybrid approach are utilized to select the global or local optimum subsets of features. For text categorization, the different types of statistical feature selection methods are developed as Boolean Weighting (BW), Term Frequency Weighting (TF), TF-IDF Weighting, TFC Weighting and Entropy Weighting [15], TF-Chi square (TF-CHI), TF-Relevance Frequency (TF-RF) [16] and TF-Odds Ratio (TF-OR) [17]. As a preprocessing step, the html tags, numbers, date and special characters have been removed from text to reduce negative effects of not-meaningful terms. To prevent performance declination in text classification, the stemming approach is broadly applied in case of ignoring these terms. Similarly, we have considered the first five characters of a Turkish word as stem. This stemming technique is known as fixed prefix stemming approach. After stemming, we have considered the favored feature selection techniques, given as MI and OR. The discriminative feature subsets are obtained and trained with classifiers to yield compact models. The details of feature selection methods are explained with following subsections.

2.1. Odds Ratio (OR)

The Odds Ratio [18] deals with the absence and presence of two events in space. For a given population, the Odds ratio is a measure of absence of A event in case of the presence of B event. In context of odds ratio refers to the probability of two events are normalized after dividing probability of absence condition with presence condition. According to concept of the odds ratio, each term is obtained with the following formula [19, 20]:

$$OR = \log \frac{P(t \mid C_i)[1 - P(t \mid \bar{C}_i)]}{[1 - P(t \mid C_i)]\mathrm{P}(t \mid \bar{C}_i)} \qquad (1)$$

In given equation, $P(t \mid C_i)$ and $P(t \mid \bar{C}_i)$ denote the probabilities of the term *t* when the presence and absence of term C_i is given, respectively.

2.2. Mutual Information (MI)

The MI unearths the correlation between two variables. It measures the mutual dependency between two variables. The MI formula is given in Eq. (2-3) [3].

For text classification, the highest MI score of a word means that it is commonly observed in the class A and occasionally observed in the class B. We have performed the MI features for feature selection.

Moreover, the MI score of each term is weighted by OR value in order to increase the success rate of assigning an e-mail to spam or legitimate category. The proposed feature selection method is named by Weighted MI (WMI).

$$MI(W) = \sum_{(w \in \{0,1\}, c \in \{S,N\})} P(W = w, C = c) x log \frac{P(W = w, C = c)}{P(W = w,)P(C = c)} \qquad (2)$$

$$\begin{aligned} MI\ (W) = {} & P\,(W = 0, C = S)\ x\ log_2(P(W = 0,\ C = S) / (P(W = 0)\ x\ P(C = S))) + \\ & P\,(W = 1, C = S)\ x\ log_2(P(W = 1,\ C = S) / (P(W = 1)\ x\ P(C = S))) + \\ & P\,(W = 0, C = N)\ x\ log_2(P(W = 0,\ C = N) / (P(W = 0)\ x\ P(C = N))) + \\ & P\,(W = 1, C = N)\ x\ log_2(P(W = 1,\ C = N) / (P(W = 1)\ x\ P(C = N))). \end{aligned} \qquad (3)$$

In (2) and (3),

S and N refer to the spam and normal emails, respectively.

P (W=0, C=S): the probability of the word not to be included in spam e-mails

P (W=1, C=S): the probability of the word included in spam e-mails.

P (W=0, C=N): the probability of the word not to be included in normal e-mails.

P (W=1, C=N): the probability of the word included in normal e-mails.

2.3. Weighted Mutual Information (WMI)

There are benefits of multiplying the OR with MI. As a special characteristic of the OR method, it measures the correlation between the presence or absence of term in class A and the presence or absence of same term in class B for binary text categorization problem. This fact allows us to explore the contribution of weighting the MI with OR. In most of studies, the performance of combining features of CHI2 and GI is investigated and taxonomy procedure is realized with a classifier such as SVM and K-NN. According to our knowledge, the weighting strategy is firstly analyzed in this study. The WMI can be represented with Eq. (4).

$$WMI(W) = OR(W) x MI(W) \qquad (4)$$

3. Feature Vector Construction

In general, the motivation behind feature extraction is improving performance of pattern classification as well as reducing the computational running time of the test and training stage. The main contributions of feature selection are efficiency and effectiveness in terms of memory usage and performance improvement. The dimension reduction can be fulfilled with different approach including subspace based feature projections approaches or simple heuristic thresholding technique. In order to project the higher dimensional data into lower one, a various algorithms have been proposed such as Principal Component Analysis (PCA) [21], Singular Value Decomposition (SVD) [22] and Independent Component Analysis (ICA) [23] or using a nonlinear method including Sammon's mapping [24]. These methods investigate linear correlation in feature space and eliminate redundant ones after projecting onto subspace orthonormal vectors. In case of feature extraction, we have represented each email sample with histogram, by applying the concept of the bag-of-words model. As a result, the raw email data is represented with histograms after feature selection stage. Thus, the efficiency and effectiveness of classification is improved after feature vector extraction.

4. Classifiers

In general, a text classification operation consists of the following stages: Preprocessing, Feature extraction/selection and Classification. In this study, a fixed stemming procedure is realized before term selection procedure. For this purpose, the only first five letters of a word are considered as stem. In case of feature selection, the MI and WMI methods is applied to reveal the more meaningful features. Finally, three learning methods including Artificial Neural Network (ANN), Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BILSTM) are performed in the classification stage. The parameter selection and details for each learning method is explained in the following subsections.

4.1. LSTM

LSTM is the modified type of Recurrent Neural Network family. As a one the most utilized deep learning model, the LSTM captures the trend behind the time series or sequence data. Depending on the task being processed, there are four different types of LSTM architectures, which are one-to-one, one-to-many, many-to-many and many-to-one. Since there is a binary classification case (spam and legitimate), we have handled the problem as a residual minimization. The processed LSTM is consisting from;

- 1 input layer (changes with respect to sequence size)
- 1 hidden layer with 100 hidden units
- 1 fully connected layer and
- 1 regression layer with MSE loss.

The details for parameter setting of LSTM is given as follows.

- **Optimizer:** Stochastic Gradient Descent with Momentum (SGDM)
- **Epoch:** 100
- **Mini Batch Size:** 128
- **Initial Learning Rate:** 0.001
- **Learn Rate Drop Period:** 50
- **Learn Rate Drop Rate Factor:** 0.1
- **Momentum:** 0.9
- **Learn Rate Schedule:** "piecewise"
- **Validation Frequency**: 50

In case of prediction stage, the threshold is 0.5 when it comes to assign the test label into spam or legitimate class. If the output value of network is less or equal to 0.5, the label of test sample is marked as spam, otherwise it is legitimate.

4.2. Artificial Neural Network

The ANN [25] is an intelligent system inspired from the biological neural networks for the pattern recognition, prediction, and data compression tasks. The main idea is adapting a system that have capabilities of neurons. From this perspective, a nonlinear system is modelled based on a set of nodes and connection between nodes by considering the particular trend stated on the data. The general stages behind an ANN algorithm can be simplified with following steps:

(1) Determination of a model for "best" representation of data.
(2) Specification of correct parameters for training stage.
(3) Then, some nonlinear operations are operated till network output matches the true target based on some error criteria.

To employ the ANN with the purpose of classification, we have preferred the Matlab implementation with two hidden layers having 20 neurons in each layer.

4.3. BILSTM

The BILSTM is generally performed when the learning problem is projected on sequences or series. Similarly, the BILSTM generate a learned model with the help of forgetting and remember gates, which are usually called by cells. The main difference between LSTM and BILSTM is that the LSTM learns the trend from beginning to end, while the BILSTM learns through two directions as from beginning to end and end to beginning.

Similarly, the applied BILSTM architecture is consisting from;

- 1 input layer (changes with respect to sequence size)
- 1 hidden layer with 100 hidden units
- 1 fully connected layer and
- 1 regression layer with MSE loss.

The details for parameter setting of LSTM is given as follows.

- **Optimizer:** Stochastic Gradient Descent with Momentum (SGDM)
- **Epoch:** 100
- **Mini Batch Size:** 128
- **Initial Learning Rate:** 0.001
- **Learn Rate Drop Period:** 50
- **Learn Rate Drop Rate Factor:** 0.1
- **Momentum:** 0.9
- **Learn Rate Schedule:** "piecewise"
- **Validation Frequency**: 50

Similarly, for the prediction case, the threshold is 0.5 when it comes to determining the test label into spam or legitimate class. If the predicted value is less or equal to 0.5, the label of test sample is marked as spam, otherwise it is legitimate.

5. Experimental Study

5.1. Database

In this study, Due to the lack of publicly available Turkish e-mail datasets, we have preferred a recently published dataset [3]. The referred dataset consists of 400 spams and 400 legitimate e-mails. To analyze the performance of proposed method, we have utilized the whole dataset.

The all of experiments given in this work were carried out on the same hardware (Intel core i5-3210M with 2.5 GHz CPU and 4 GB memory) with a software invoked on the MATLAB environment.

5.2. Studies on LSTM

Fig. 1 (a-b) demonstrates the results of LSTM obtained by MI and WMI features for 4-fold cross-validation. Results of Fig. 1 (b) indicates that performing the LSTM on WMI features provides the highest accuracy rates as 100% for legitimate email classification in case of all feature vector dimension. It generates the 100% detection rates for spam email classification over 3rd and 4th cross-fold validation for all feature dimensions.

Moreover, the performance of utilized LTSM is presented in Fig.

2 (a-b), when considering the MI features. In case of legitimate recognition, LSTM again gives the maximum recognition score as 100% when the feature vector size is 50, 75, 100, 125 or 150. One can say that the performance of WMI features is slightly better than the MI. For all cross-validations, the average results are higher than 99%.

5.3. Studies on Artificial Neural Network

The Fig. 3 (a-b) summarizes the classification rates returned from the ANN classifier when conducting WMI features in terms of spam and legitimate e-mails, respectively. By analyzing the results, it has been observed that ANN gives maximum rates as 93.25% and 98.00% in case spam and legitimate e-mails if the feature vector size is chosen as 100 and 125, respectively.

Moreover, the results obtained from MI features with ANN classifier is depicted on Fig. 4 (a-b) for spam and legitimate e-mails, respectively. At a glance, it should be emphasized that the performance of ANN on legitimate e-mails is superior to spam e-mails. While the successive results obtained by ANN on spam e-mails is maximum 95.50% with 150 dimension of feature set, but in case of legitimate ones, the accuracy rate appears as 98.75% with 50 dimension of feature set.

5.4. Studies on BILSTM

The performance of BILSTM has also investigated for spam and legitimate email detection. The utilized BILSTM with weighted mutual information gives 100% accuracy rates for all feature vector dimension and 4-fold cross validation. Given the results on Fig. 5 (a-b), one can observe that the BILSTM combined with WMI features shows the valuable results for both legitimate and spam email classification.

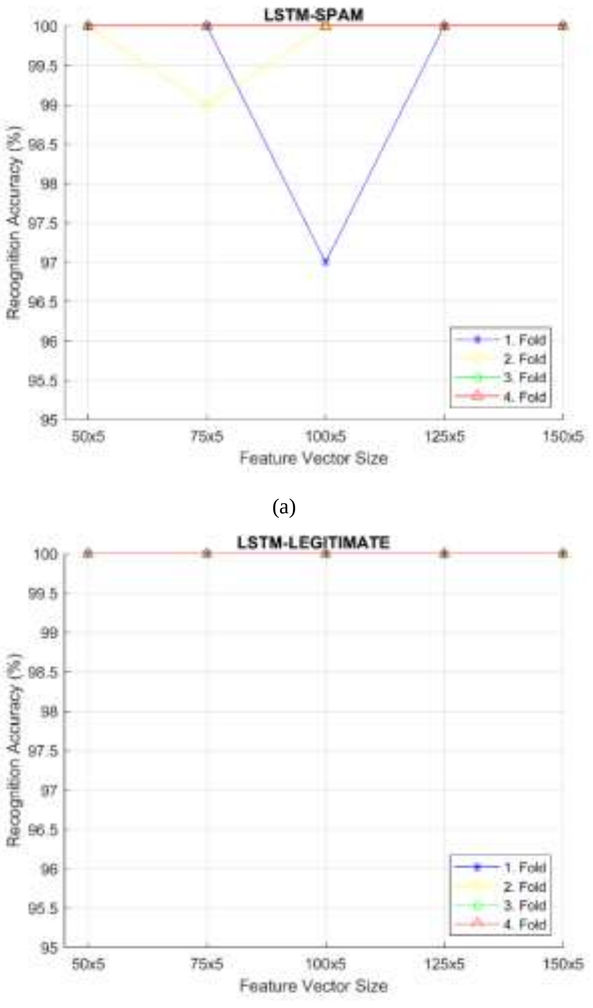
Furthermore, for MI features, the performance of BILSTM shows valuable detection results. Particularly on legitimate emails, the BILSTM reaches the 100% accuracy scores for all feature vector dimension. However, with increasing feature vector dimension, the performance of BILSTM considerably improved when spam email detection results are observed from Fig. 5(b).

5.5. Evaluation of the System

The performance of average detection results is given in Table 1, 2, 3 and 4. Specifically for ANN, one can witness that there is a declining trend in performance with increasing feature vector size. This can be attributed to the curse of dimensionality problem as the curse of dimensionality [26-29] arises when the feature vector size becomes bigger than a specific size. It is an inevitable problem in case of high dimensional data. However, for LSTM and BILSTM, we can note that the bigger dimension does not affect the performance.

As another interesting point, the all of the classifiers present superior results in case of legitimate e-mail classification when compared with the spam ones. This property can be attributed to a discriminative characteristic of selected legitimate and spam words.

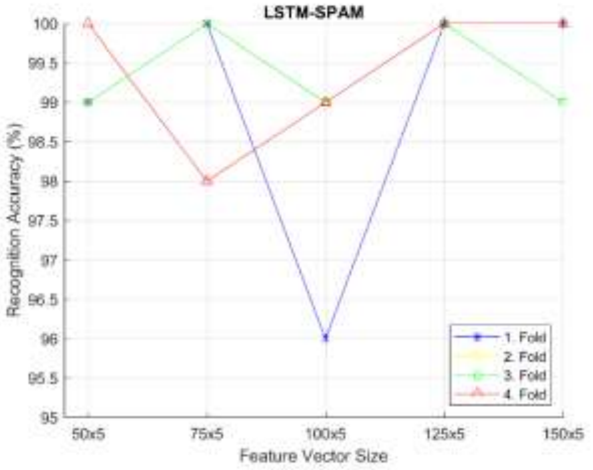
To compare the performance of WMI and MI features combined with different classifiers, the average accuracy results for each classifier have been given in Table (1-2) and Table (3-4) in terms of spam and legitimate email detection. Upon inspecting the Table 1 and 2, one can emphasize that the highest average performances are 100% and 100.00%, when performing the BILSTM with WMI features, for spam and legitimate emails respectively. However, the performance of ANN is slightly lower than the LSTM and BILSTM. This performance limitation can be explained with lower discriminative capability of ANN.



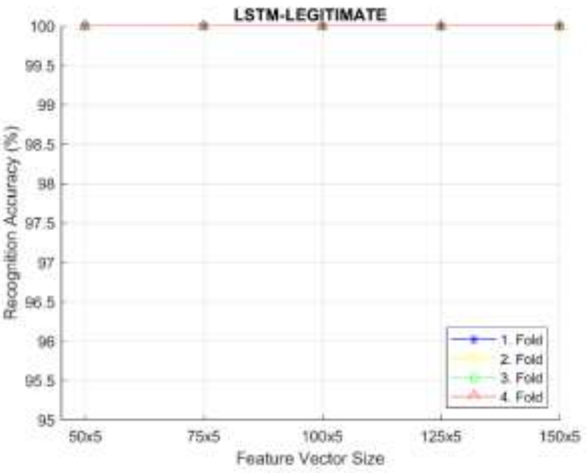
(a)

(b)

Fig. 1. Performance of LSTM on WMI.

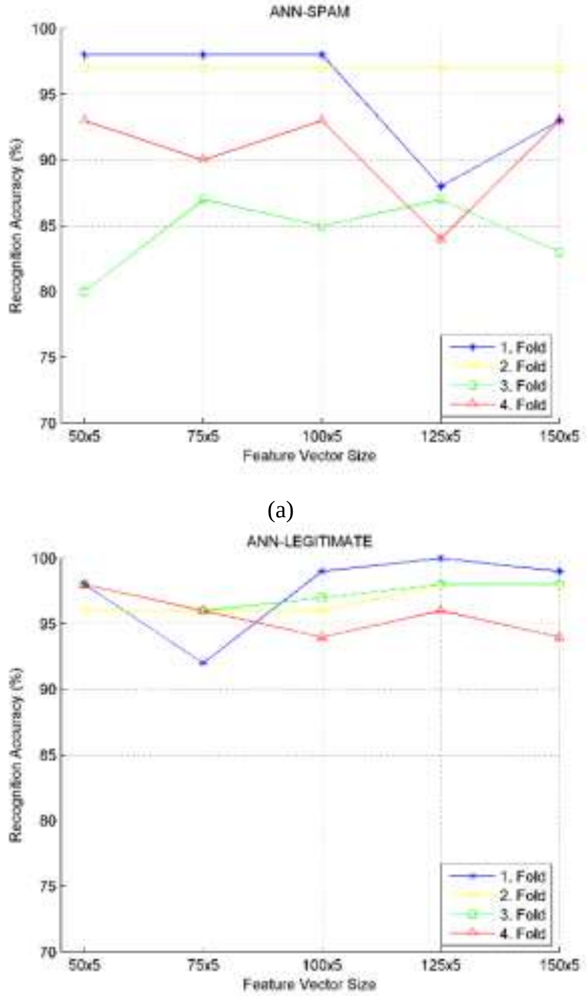


(a)



(b)

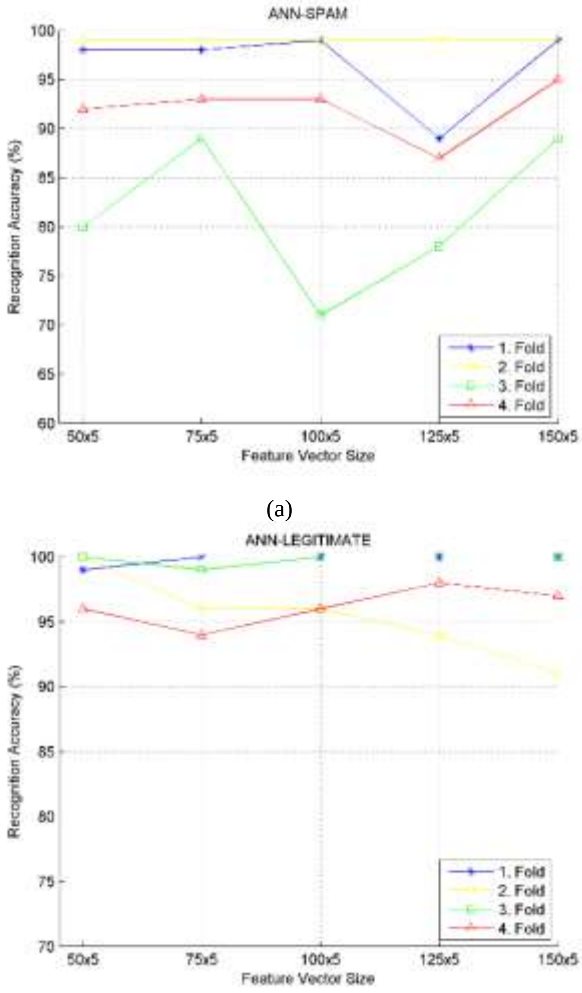
Fig. 2. Performance of LSTM on MI.



(a)

(b)

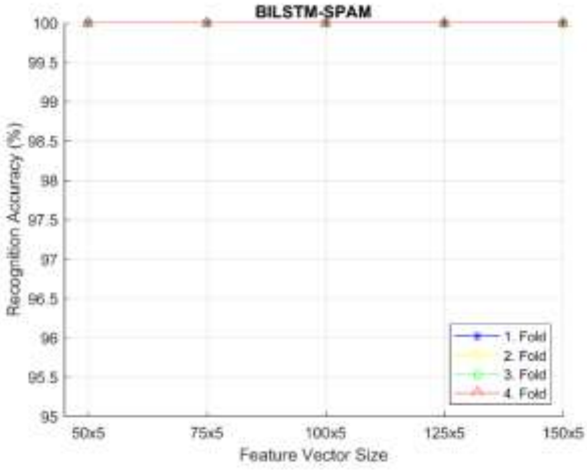
Fig. 3. Performance of ANN on WMI



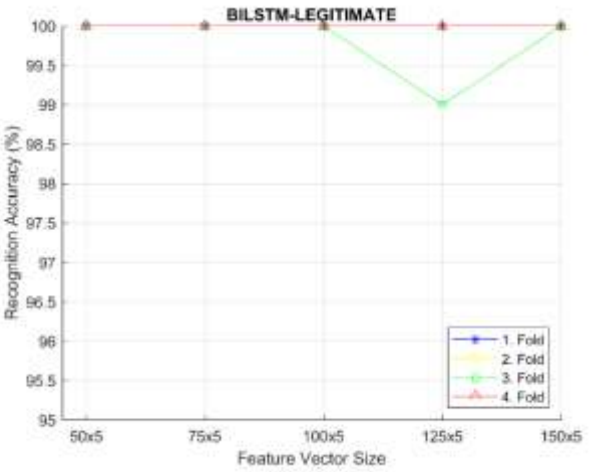
(a)

(b)

Fig 4. Performance of ANN on MI.

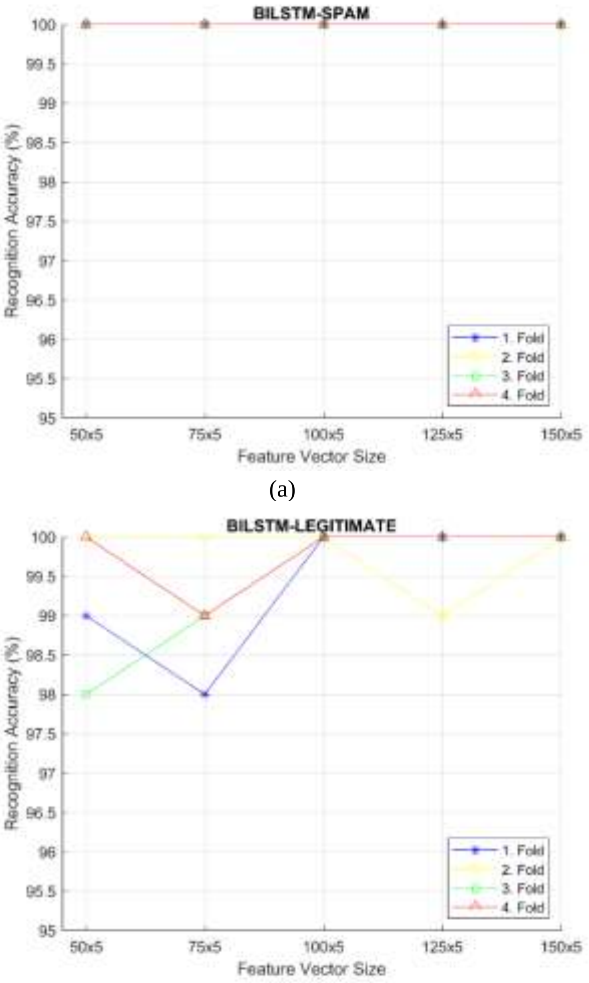


(a)



(b)

Fig. 5. Performance of BILSTM on WMI.



(a)

(b)

Fig. 6. Performance of BILSTM on MI.

After analyzing the results of MI features given in Tables 3 and 4, we can observe that the highest scores are determined with BILSTM and LSTM for spam and legitimate email detection, respectively. One can say that WMI features give higher results than MI features.

Table 1. Overall performance with WMI on spams.

	SPAM				
SIZE	50x5	75x5	100x5	125x5	150x5
ANN	92.00	**93.00**	93.25	89.00	91.50
LSTM	**100.00**	99.75	99.25	**100.00**	**100.00**
BILSTM	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AVERAGE	97.33	97.58	**97.50**	96.33	97.17

Table 2. Overall performance with WMI on legitimates.

	LEGITIMATE				
SIZE	50x5	75x5	100x5	125x5	150x5
ANN	97.50	95.00	96.50	**98.00**	97.25
LSTM	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
BILSTM	**100.00**	**100.00**	**100.00**	99.75	**100.00**
AVERAGE	99.17	98.33	98.83	**99.25**	99.08

Table 3. Overall performance with MI on spams.

	SPAM				
SIZE	50x5	75x5	100x5	125x5	150x5
ANN	92.25	94.75	90.50	88.25	**95.50**
LSTM	99.25	99.50	98.25	**100.00**	99.75
BILSTM	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
AVERAGE	97.17	98.08	96.25	96.08	**98.42**

Table 4. Overall performance with MI on legitimates.

	LEGITIMATE				
SIZE	50x5	75x5	100x5	125x5	150x5
ANN	**98.75**	97.25	98.00	98.00	97.00
LSTM	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
BILSTM	99.25	99.00	**100.00**	99.75	**100.00**
AVERAGE	**99.33**	98.75	99.33	99.25	99.00

This study shows that using WMI or MI with LSTM or BILSTM classifier gives highest recognition rates, which is given as 100% for both spam and legitimate emails. Comparing to literature, the combination of CHI2 and ANN has yielded a 91% accuracy rate for spam emails and 97% for legitimate emails [7]. Moreover, the performance of MI and ANN was reported to 91.08% and highest detection score of MI and DT was accounted for 87.67% when the feature vectors dimension are selected as 150x5 and 75x5, respectively [8]. Additionally, the running time for feature extraction is roughly about 0.001 seconds for MI and WMI.

Moreover, we have compared the performance of the proposed method with some studies that are used the same dataset. In a study on filtering Turkish SMS messages [3], the highest Micro-F1 score was reported as 0.98, which was obtained by performing Support Vector Machine (SVM) classifier on Structural Features (SF) and 50% of Bag of Word (BOW) features chosen by CHI2 feature selection method. The SF refers to the number of terms obtained using alphanumeric tokenization. Again, the 0.95 Micro-F1 score was determined by using K-NN classifier with a fusion set of SF and 50% of BoW features selected by GI. In another study [1], the valuable classification score, as 90.17%, was determined after applying the binary classification model on top-10 of features selected by CHI2. Comparing with our proposed deep learning models, the BILSTM gives superior scores, which is reported as 100% for spam emails.

6. Conclusion

In the given study, we have inspected the performance of proposed feature selection method WMI and MI after combined with LSTM and BILSTM to determine the label of email samples as spam or legitimate. Additionally, the performance of WMI and MI has been compared under different feature vector dimensions. After conducting simulations with ANN, LSTM, and BILSTM, the results indicate that using WMI and MI combined with LSTM or BILSTM achieves a 100% accuracy rate. As future work, the combination of MI with different feature selection methods including GI, CHI2, and IG can be analyzed for spam and legitimate email classification.

References

[1] A. K. Uysal, S. Gunal, S. Ergin, and E. S. Gunal, "A novel framework for SMS spam filtering," in 2012 International Symposium on Innovations in Intelligent Systems and Applications, 2012, pp. 1-4.

[2] L. Özgür, T. Güngör, and F. Gürgen, "Adaptive anti-spam filtering for agglutinative languages: a special case for Turkish," Pattern Recognition Letters, vol. 25, no. 16, pp. 1819-1831, 2004.

[3] A. Uysal, S. Gunal, S. Ergin, and E. Sora Gunal, "The Impact of Feature Extraction and Selection on SMS Spam Filtering," Elektronika ir Elektrotechnika, vol. 19, no. 5, pp. 67-72, 2012.

[4] usa.kaspersky.com. (2020). Spam and Phishing Statistics Report Q1-2014, [Online]. Available: https://usa.kaspersky.com/resource-center/threats/spam-statistics-report-q1-2014, Accessed: Dec. 11, 2020.

[5] itgovarnance.eu. (2020). Kaspersky records 130 million phishing attacks in Q2 2019, [Online]. Available: https://www.itgovernance.eu/, Accessed: Dec. 11, 2020.

[6] S. Gunal, "Hybrid feature selection for text classification," Turkish Journal of Electrical Engineering Computer Sciences, vol. 20, no. 2, pp. 1296-1311, 2012.

[7] S. Ergin and S. Isik, "The assessment of feature selection methods on agglutinative language for spam email detection: A special case for Turkish," in Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on, 2014, pp. 122-125.

[8] S. Ergin and S. Isik, "The investigation on the effect of feature vector dimension for spam email detection with a new framework," in Information Systems and Technologies (CISTI), 2014 9th Iberian Conference on, 2014, pp. 1-4.

[9] W. Fang, H. Luo, S. Xu, P. E. Love, Z. Lu, and C. Ye, "Automated text classification of near-misses from safety reports: An improved deep learning approach," Advanced Engineering Informatics, vol. 44, p. 101060, 2020.

[10] A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," Information Processing Management, vol. 57, no. 1, p. 102121, 2020.

[11] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion," Information Processing Management, vol. 56, no. 4, pp. 1245-1259, 2019.

[12] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS spam," Future Generation Computer Systems, vol. 102, pp. 524-533, 2020.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[14] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, 1997.

[15] N. Chirawichitchai, P. Sa-nguansat, and P. Meesad, "A Comparative Study on Feature Weight in Thai Document Categorization Framework," in IICS, 2010, pp. 257-266.

[16] M. Lan, C.-L. Tan, H.-B. Low, and S.-Y. Sung, "A comprehensive comparative study on term weighting schemes for text categorization with support vector machines," in Special interest tracks and posters of the 14th international conference on World Wide Web, 2005, pp. 1032-1033.

[17] Z.-H. Deng, S.-W. Tang, D.-Q. Yang, M. Z. L.-Y. Li, and K.-Q. Xie, "A comparative study on feature weight in text categorization," in Advanced Web Technologies and Applications: Springer, 2004, pp. 588-597.

[18] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," Expert Systems with Applications, vol. 36, no. 3, pp. 5432-5435, 2009.

[19] D. Mladenic, "Machine Learning on non-homogeneous, distributed text data," Ljubljana, Slovenia, Faculty of Computer and Information Science, University of Ljubljana, Diss, vol. 3, no. 3.1, p. 2, 1998.

[20] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 80-89, 2004.

[21] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on, 1991, pp. 586-591.

[22] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in A practical approach to microarray data analysis: Springer, 2003, pp. 91-109.

[23] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," Neural Networks, IEEE Transactions on, vol. 13, no. 6, pp. 1450-1464, 2002.

[24] J. W. Sammon, "A nonlinear mapping for data structure analysis," IEEE Transactions on computers, vol. 18, no. 5, pp. 401-409, 1969.

[25] C. M. Bishop, "Neural networks for pattern recognition," 1995.

[26] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998, pp. 604-613: ACM.

[27] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 19, no. 2, pp. 153-158, 1997.

[28] I. V. Oseledets and E. E. Tyrtyshnikov, "Breaking the curse of dimensionality, or how to use SVD in many dimensions," SIAM Journal on Scientific Computing, vol. 31, no. 5, pp. 3744-3759, 2009.

[29] D. Zongker and A. Jain, "Algorithms for feature selection: An evaluation," in Pattern Recognition, 1996., Proceedings of the 13th International Conference on, 1996, vol. 2, pp. 18-22.