Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Website: Kjar.spu.edu.iq | Email: kjar@spu.edu.iq

Relevant SMS Spam Feature Selection Using Wrapper Approach and XGBoost Algorithm

Diyari Jalal Mussa
Information technology Department
Technical College of Informatics
Sulaimani Polytechnic University
Sulaimani, Iraq
Diyari.mussa@spu.edu.iq

Noor Ghazi M. Jameel
Computer Networks Department
Technical College of Informatics
Sulaimani Polytechnic University
Sulaimani, Iraq
Noor.ghazi@spu.edu.iq

Volume 4 - Issue 2
December 2019

DOI:
10.24017/science.2019.2.11

Received:
5 September 2019

Accepted:
15 November 2019

Abstract
In recent years with the widely usage of mobile devices, the problem of SMS Spam increased dramatically. Receiving those undesired messages continuously can cause frustration to users. And sometimes it can be harmful, by sending SMS messages containing fake web pages in order to steal users' confidential information. Besides spasm number of hazardous actions, there is limited number of spam filtering software. According to this paper, XGBoost algorithm used for handling SMS spam detection problem. Number of structural features was collected from previous studies. 15 structural features were extracted from Tiago's dataset, which is the most frequently used dataset by researchers. For selecting the optimal relevant features, two different types of wrapper feature selection algorithms were used in order to reduce and select best relevant features. The accuracy and performance obtained by the selected features via sequential backward selection method was better comparing to sequential forward selection method. The extracted nine optimal features can be a good representation of a spam SMS message. Additionally, the classification accuracy obtained by the proposed method using nine optimal features with XG Boost algorithm is 98.64 using 10-fold cross validation.

Keywords: SMS spam, wrapper methods, sequential feature selection, sequential forward selection, sequential backward selection, boosting classifier, extreme gradient boosting, XG Boost.

1. INTRODUCTION

Text messaging considered as one of the common transmission mechanisms between users of mobile devices. Texting has remarkably increased in recent years due to fast delivery, low-cost, and easily global reach. This expansion of text messaging has invited spammers to perform mobile spam message problem similar to spamming in e-mails. Essentially cell phone spam or SMS spam is whichever junk message or undesirable message reached to the user's mobile device as text via the short message service (SMS). SMS Spam usually sent automatically in bulk to randomly generated or selected mobile numbers. These types of texts can come in various forms – sometimes recipients were asked to respond to an email address or mobile number that does not lead to expected identity of the sender. In many cases, SMS spam may include free services, promotions, advertisements, etc., in order to manipulate recipient to provide personal information, defraud for economic purposes or political benefits[1][2]. This practice has recorded different stats in various countries. For instance a survey showed that 69% of mobile users in USA have received spam messages[3]. While in some countries of Asia such as India, Pakistan and China, spam messages are causing a disaster because more than 30% of the messages that users received were spam [2]. Fortunately, in recent years, governments around the world have presented serious involvement regarding the spam problem by objecting cost-effective, lawful and technical abilities for handling concerning issue. But still there are many countries suffering from this problem because of the absence of law and economic penalties. Technical measures to counter spam problem can be a convenient and powerful way to handle this issue. For instance, there are many applications for devices based on android operating systems and even IOS which can be used to filter spam texts while there are few who are familiar with such new techniques. In a different way the available filtering methods mostly concentrates on spamming of email, because it is considered as an earlier issue comparing to SMS spam[4]. While mobile devices are widely used around the world, spamming of SMS is one of the main problems nowadays. This issue has become common in peoples' everyday communication channels. It can be categorized as a threat, since spam contents could be fraudulent and lead users to harsh effects [5][6]. In addition, spamming could become a reason for an increase in internet crimes. Previously, spamming was used as a platform for advertisement and announcement of sale. While nowadays it became a tool for publishing lies, profanity; hence, determining an atmosphere for illegal activities. For instance, harmful attack that can be done through SMS spamming is SMiShing. It is dealing with deceiving target users into revealing confidential information. Considering SMS spam as a problem and potential threat led researchers to define many methods for detecting and filtering it. In this proposed method our intention is on detecting spam and normal SMS in an accurate way using a type of gradient boosting algorithm. Selecting most relevant features among structural attributes that are collected from previous studies using two different wrapper feature selection methods. The aim of this study is to compare among the relevant features via sequential wrapper selection methods, then choose the best relevant structural (binary and numeric) features that can make a distinction between a spam and a ham SMS and lead to better accuracy using XG Boost algorithm.

The rest of this study is arranged as follows: in section 2 some relevant studies that are related to SMS spam detection are discussed, section 3 is about describing our proposed approach in details containing selection of the dataset, features extraction, features selection methods and the classification algorithm, section 4 presents results and other details about the tests outcomes, section 5 is comparing this study to another study in discussion and section 6 explains conclusion and a future study.

2. LITERATURE REVIEW

Nowadays there are different approaches for SMS spam filtration, while spamming is not limited to emails and web pages any longer. Various techniques of SMS spam detection lead

to availability of android applications to filter spam text messages, using classifications methods. In this section a review of most relevant studies related to this topic has done, with regard to some essential concepts such as dataset, feature selection and algorithms of machine learning. Yadav et al. [7] proposed SMS Assassin that deals with proposing a system based on mobile which is capable of spam detection using two methods; first one is using Bayesian algorithm and the second one is blacklisting. They used a dataset which was gathered in India, and it is a real world spam dataset collected in a span of two months which contains 4318 messages. Moreover, the features set that were used in their experiment contained 20 lightweight structural features. As for the dataset they split it into two part, 2000 messages for training and other half which contain 2318 messages determined for testing. The training data contained 2000 messages equally divided by two, 1000 messages for each ham and spam. The remaining data determined for testing via learning algorithms such as Support Vector Machine and Bayesian algorithm, the rest of the data were divided between ham (1195) and spam (1123). In their approach while the user receives a text via their cell phone, the proposed mobile system automatically detects the text without awareness of the user, and gets values of features and transmit them to the server in order to classify them. The message will be in spam folder if it is reported as spam. Uysal et al. [8] have examined the effect of various types of feature extraction and selection techniques on spam detection in two different languages: English and Turkish. The Turkish dataset were collected from volunteers which consists of 420 text messages of spam and 430 normal texts while the English corpus includes 425 spam texts with 450 determined for normal short texts. Moreover, the feature set consists of two types of features, the first one is bag-of-words (BoW) method, the second method is structural features (SF) with regard to problem of spam. Various features of BoW determined using methods of feature selection. Different sequence of SF and BoW were input into broadly applied pattern classification methods for detecting text messages that are spam or not. Then datasets of Turkish and English were used in order to evaluate detection models on the datasets. The results showed that the combination of both BoW and SF provides better outcome of classification most of the time rather than only BoW feature. Ahmed et al. [9] proposed a hybrid approach of SMS classification in order to filter ham or spam by using Apriori algorithm and Naïve Bayes classifier. The dataset collected from repository of UCI Machine learning under the name of "SMS Spam Collection Data Set", it includes 5574 text messages containing both spam and ham. Moreover the features that extracted were all word attributes. The proposed system for detecting spam messages has gained levels of accuracy that is competitive to state of art algorithm. Akbari et al. [10] presented a Boosting method for SMS spam detection and determined that boosting classifiers are good option among other classifiers when used dataset is unbalanced. They used Tiago's dataset, it consisted of 4,827 legitimate and 747 spam messages. Word attributes were used as features, removing unused ones leads to less number of features without affecting the accuracy, Gentle Boost was the used boosting classifier to classify the SMS messages. Zhang et al. [11] used ada boost algorithm as machine learning classifier in order to filter Chinese SMS spam messages. They introduced three types of filters that are weak and content-based for increasing final classification performance. According to the obtained outputs via Receiver Operating Characteristics (ROC) presented approach has some advantages; it has higher efficiency withless parameters comparing to already available spam detection models. Presented method's application is anticipated to filter most of the spam messages for mobile users. The proposed technique can be implemented on the text message classification with simple data processing and small number of training parameters. Choudhary et al. [12] proposed new approach that is able to filter and identify spam texts by applying learning classifier. According to their research, they studied the attribute of SMS spam messages comprehensively and they found 10 attributes while the dataset used in their study includes 2608 text message. They gathered publically available 2408 message from SMS Spam Corpus and they gathered 200 texts manually which contain 25 spam and 175 normal text messages. Their proposed technique which examined on multiple learning models and obtained topmost when Random Forest classification used, with scores of 96.5% recall and 1.02% false positive

rate. Wang et al. [13] contributed a new categorized SMS spam dataset which named HIT SMS Spam Corpus. The new dataset consists of 13,078 messages which contain 4,204 spam and 8874 legitimate messages. They proposed a new technique for SMS spam detection that can reduce the feature sparse issue in the process of SMS spam classification. Furthermore, no preprocessing rules and non-linguistic attributes are used in their approach. Nivaashini et al [14] proposed SMS spam detection system using a deep learning algorithm. The features were extracted using Restricted Boltzmann Machine (RBM), which classifies the SMS into spam and ham by Deep Neural Network classifier they used SMS Spams dataset from UCI Machine Learning repository. Then the results are compared with other machine learning algorithms such as naïve Bayes, support vector machine and random forest. The proposed approaches achieved highest accuracy 98.18 compared with the other algorithms.

Most of related works used textual features as an indicator for SMS spam detection. Textural features extraction requires many preprocessing steps such as tokenization, stop words removal, stemming before classification and detection phase. Others used combination between textual and structural features. In this paper, just 9 optimal structural features which consists of binary and numeric features used for SMS spam detection by XGBoost algorithm. Using optimal structural features with XGboost algorithm is considered as an added point to this work since no preprocessing is required to extract these features. XGBoost is one of the new gradient boosting algorithms designed for speed and performance and deals with unbalanced dataset.

3. METHODS AND MATERIALS

The proposed approaches suggested detecting spam and hamming messages with XGboost algorithm. The input to XGboost algorithm is the most relevant features from wrapper feature selection algorithm. In this paper a comparison between two approaches of wrapper features selection are used: sequential forward and backward features selection algorithms. Relevant features from the two approaches are applied which considered input for the classifier to detect SMS spam messages. According to the accuracy of the classifier the best feature selection approach is used. The process starts with selecting a suitable dataset which is publically available dataset. Then, the collected features from previous studies were extracted from spam and ham messages to generate a vector of features. The decision in this proposed system is made after selecting the best relevant features from comparing two wrapper feature selection methods. Then the specified classifier which takes outcome of optimal selected attributes from feature selection methods, will be ready for filtering spam and ham SMS efficiently. The proposed approach is demonstrated in figure 1.

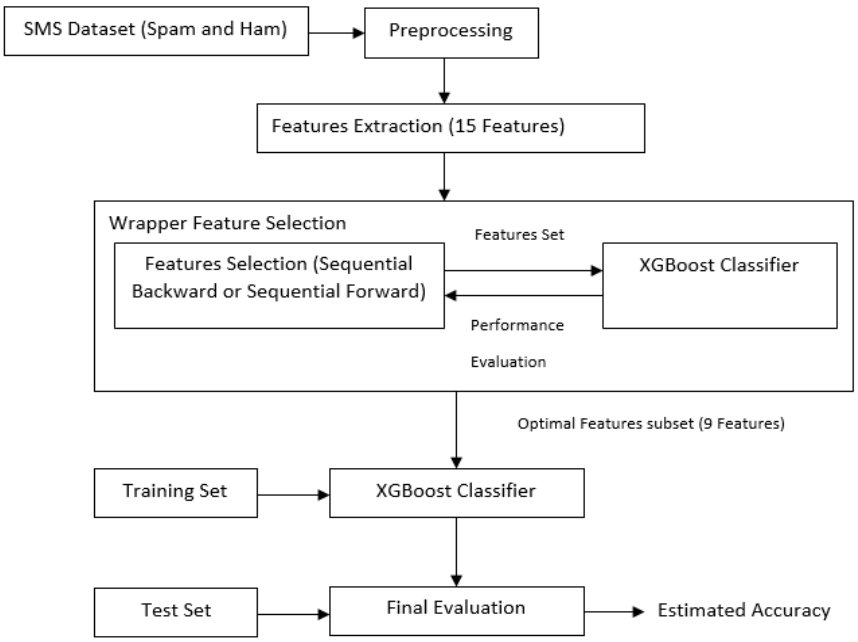

Figure 1: Proposed system architecture

3.1. Dataset loading

In spite of the fact that there are many e-mail datasets, or collections are available in order to be used by researchers [15], in this area of study there are few publically accessible SMS datasets. For this paper we chose the biggest publically existing spam corpus namely Tiago's dataset, which is widely used by researchers. It can be downloaded at UCI Repository freely under the name "SMS Spam Collection v.1"[16], It is imbalanced dataset, it made up of 5,574 messages which4,827 are normal texts and 747 are spam short messages. In figure 2, Statistics regarding chosen dataset is demonstrated.

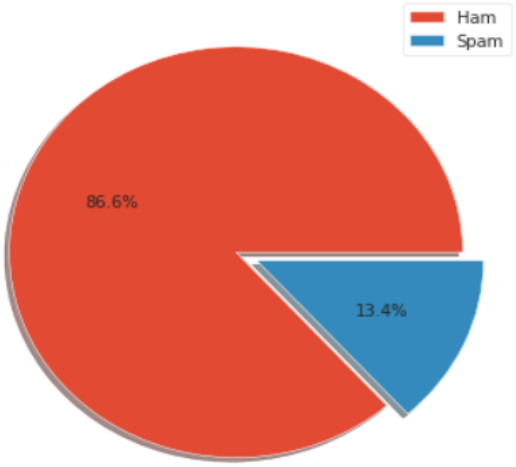

Figure 2: SMS spam collection statistics

3.2. Feature Extraction

Primarily SMS spam detection is same as e-mail spam filtering problem. SMS text message contains solely text constrained with 160 characters [17], On the other hand emails can

include hyperlinks, visuals, and files beside text [18]. Therefore; spam message filtration is considered as binary classification issue and labels are identified with "legitimate (ham)" or "spam". There are different ways implemented for features extraction, most of them based on using bag-of-words model [19] for extracting terms frequencies. While [12] used different approach, they studied spam messages attributes in detailbesideextracting10 structural variables, each feature will obtain value 0 or 1 according to their existence in the message. In this paper, the same approach was used but 15 features were extracted which consists of binary and numeric features.

The filtration of SMS spam deals with the task of normal text classification, but undoubtedly there are differences between spam structure and formal text. For instance, spam messages may contain URL [12] [7], where spammer perhaps deceive the receiver to click on the phishing web address. As the spammers use various sorts of symbols for representing their messages, for instance plus sign (+) represents services messages that are free. Math characters such as (+, -, /, >, <, ^) will be identified as structural feature. Besides using upper case words, lower case words can also represent as spam features. Studies also consider having phone number and special symbol such as (£, $, !, *, &, #, ~,') in a text message represents a distinctive indicator for spam message. The existence of dots (.) and emotions for instance (:) , ;) , :p , -_- , etc.), the presence of suspicious key words which spammers use, mostly should be a good spam features [12].

The 15 features which were extracted in this study are illustrated in table 1.

Table 1: All 15 binary and numeric features

Feature No.	Feature	Type	Reference
Feature1	mathematical symbols (+, -, /, >, <, ^)	Binary	[12]
Feature2	Presence of full URL	Binary	[12], [7]
Feature3	Dots	Binary	[12]
Feature4	Special symbol (£, $, !, *, &, #, ~,')	Binary	[12]
Feature5	Emotions (emoji or smiles)	Binary	[12], [7]
Feature6	Lowercase words	Binary	[12]
Feature7	Uppercase words	Binary	[12]
Feature8	Phone number	Binary	[12]
Feature9	Message Length	Numeric	[12]
Feature10	Count of "/" slash	Numeric	[7]
Feature11	Count of numeric words (numbers)	Numeric	[7]
Feature12	Count of alphanumeric words (ex: na18)	Numeric	[7]
Feature13	Number of words	Numeric	[7]
Feature14	Count of Spam words	Numeric	[7]
Feature15	Presence of Spam words	Binary	[12]

The extracted binary structural features are in binary format means if the feature exists in the message, the feature value will be one otherwise it will be zero in the feature vector.

3.3. Feature Selection

Variable selection or feature selection is one of the vital parts of spam filtering. It helps in improving performance and decrease models logical errors via eliminating unimportant attributes. There are different feature selection methodologies such as filter, wrappers and embedded. In this paper two types of wrapper methods based on sequential search algorithm are compared. Primarily, sequential selection considered as suboptimal techniques that attains series of feature subsets, for instance by increasing or eliminating worst or the best variable from the feature set [20]. These methods evaluate all possible sequence of the attributes and select the combination that produces the best outcome for particular machine learning algorithm. Here we concentrate on only two types: sequential forward selection and sequential backward selection.

- Sequential forward selection (SFS) has number of phases in order to be implemented, the first phase starts with classifiers performance evaluation with respect to each features.

The feature that achieves the best is chosen out of all attributes. In the second step, the first selected feature from first step is tried in combinations of all other attributes. The sequence of two attribute that yield the best algorithm performance is selected. The process goes on until particular number of attributes is selected.

- While sequential backward selection (SBS), is exact opposite of step forward feature selection. The first step of its implementation, one attribute is removed in round-robin fashion from the attribute set also the classifiers performance is evaluated. The set of attributes that obtains best performance is maintained. In the second step, repeatedly one feature is removed in a round-robin fashion and the performance of all the sequence of features excluding the two features is assessed. This process goes on until the particular number of features remains in the dataset. In figure 3 wrapper feature selection is demonstrated in general.

In this study 15(binary and numeric) features were extracted. Two Wrapper feature selection algorithms are used to select an optimal feature subset. A comparison between the two algorithms: sequential forward selection and backward selection been applied depending on extreme gradient boosting algorithm (XGBoost) as classifier. For measuring the classifier accuracy, 10-fold cross validation approach in which the dataset is divided into 10 fold, in each iteration one fold is used for testing and the others are used for training the model.

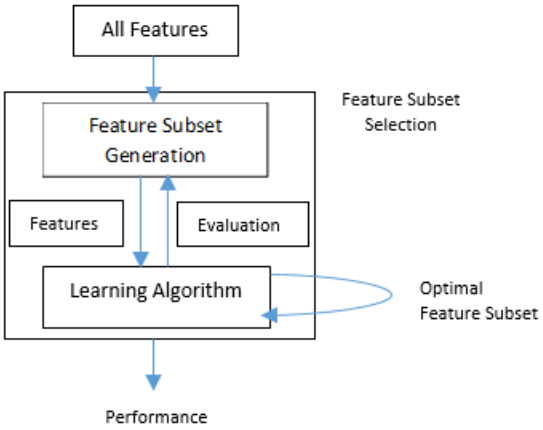

Figure 3: Wrapper method for feature selection [21]

3.4. Classification

In this part, an extreme gradient boosting (XGBoost) is used in the proposed approach as a classifier for SMS spam detection and very good performance and accuracy is obtained by using this algorithm. Primarily XGBoost have been selected as the classifier because it has a very good result when it comes to handling imbalanced dataset [20]. XGBoost is a library of distributed gradient boosting developed as vastly flexible, convenient and efficient. XGBoost grant tree boosting in parallel that handle various problems of data science rapidly and in accurate manner [22].

3.5. Evaluation metrics

For determining XGBoost's performance regarding spam filtering problem, a number of possible results is dealt with. These results based on outcome obtained from the confusion matrix which used for evaluating the performance classification model. The evaluation metrics used in this study contains recall (true positive rate), specificity, precision and accuracy. While the confusion matrix includes true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Moreover, for evaluating spam filtering system, those standard metrics should be considered.

- True Positives (TP) – refers to the case in which the classifier predicted that the message is spam and actually is spam.
- True Negative (TN) – the classifier predicted that the message is ham and actually is ham.
- False Positive (FP) – the classifier predicted the message is spam, but actually it is not.
- False Negative (FN) – in this case the classifier predicted that the message is ham, but actually is spam.
- Accuracy – refers to the ability to differentiate the messages correctly.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Recall – refers to the percentage of classified truly spam messages in spam detection problem.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

- Specificity – in spam filtration problem true negative rate refers to the percentage of classified truly normal text messages.

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

- Precision –determines spam messages measurement that are truly identified by specified classifier besides demonstrating the definite propriety.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

4. RESULTS

In this section, sequence of experiments implemented on Tiago's dataset aiming to find the best combination of features. At first we collected features on the basis of normal messages and spam short messages behavior, there after derived 15attributesof the dataset for creating the vector of features. After features extraction from SMS Spam corpus, XGBoost algorithm is implemented to detect SMS spam messages. Python programming is used for features extraction and classification beside the usage of 10-fold cross validation for training and testing. The main objective is to select the best among 15 features using two different feature selection methods based on wrapper approach. Different outcome obtained when sequential forward selection and sequential backward selection been applied. Table 2 shows the output of the optimal feature subset regarding to each method.

Table 2: The best 9 selected features regarding SBS and SFS

Selected features by SBS		**Selected features by SFS**	
Feature No.	**Feature**	**Feature No.**	**Feature**
Feature2	Presence of full URL	Feature2	Presence of full URL
Feature4	Special symbol (£, $, !, *, &, #, ~,')	Feature5	Emotions (emoji or smiles)
Feature6	Lowercase words	Feature8	Phone number
Feature7	Uppercase words	Feature10	Count of "/" slash
Feature8	Phone number	Feature11	Count of numeric words (numbers)
Feature9	Message Length	Feature12	Count of alphanumeric words (ex: na18)
Feature11	Count of numeric words (numbers)	Feature13	Number of words

Feature13	Number of words	Feature14	Count of Spam words
Feature14	Counting Spam words	Feature15	Presence of Spam words

Those selected features scored almost the best accuracy in each separate test regarding each wrapper method. For instance, 9out of 15 features led to the best accuracy when SBS implemented. In figure 4 and 5 accuracy result is demonstrated with selected number of features regarding each SBS and SFS.

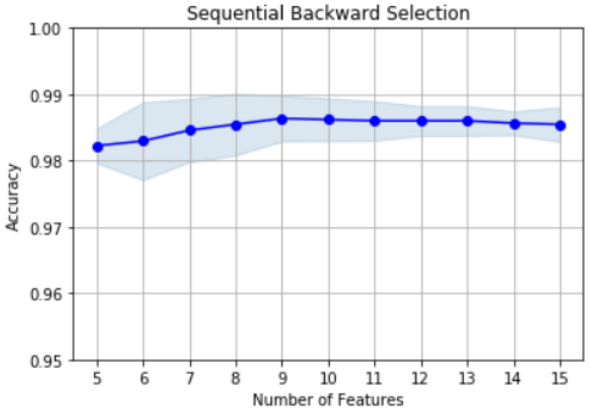

Figure 4: SBS accuracy regarding feature number

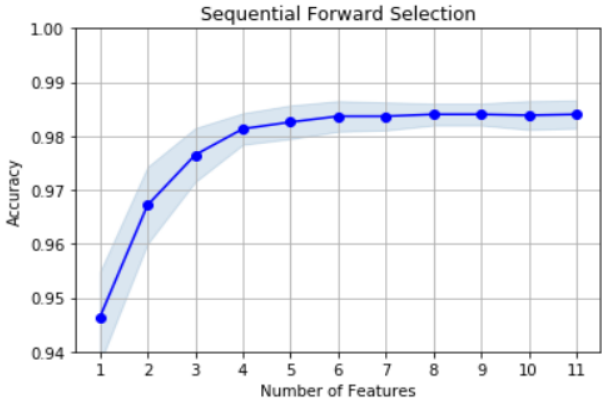

Figure 5: SFS accuracy regarding feature number

When optimal 9 feature accuracy and their evaluation metrics compared to each other as demonstrated in table 3, both of them is almost close to each other in terms of accuracy. But there are some differences that can be pointed out, which makes the results obtained by SBS better than SFS.

Table 3: Comparison between the evaluation metrics of SBS and SFS

Accuracy and Evaluation metrics	Sequential backward selection	Sequential forward selection
Accuracy	98.64	98.40
Recall	0.9224	0.909
Specificity	0.9963	0.9957
Precision	0.9745	0.97

5. DISCUSSION

Depending on the results achieved by XGBoost classification algorithm, it obtained that the performance and accuracy can be considered very well in this context based on using structural features which needs less time for features extraction phase compared the textual features or words attributes. As a sort of gradient boosting algorithm can be a good competitor to other algorithms used in this area. A comparative study is done with [10] in which different types of boosting algorithms used words attributes and Tiago's dataset. Results of the comparison is demonstrated in table 4 which shows the proposed approach with structural features gave better accuracy compared with other boosting algorithms with words attributes. We should also note that for comparison among methods there are some aspects such as simplicity, storage and accuracy that should not be neglected.

Table 4: Comparison between the proposed approach and other boosting algorithms in [10]

	Features Type	**Accuracy**	**TP**	**TN**	**FP**	**FN**
XGBoost (Proposed method)	9Structural Features	98.64	689	4809	18	58
GentleBoost[10]	Textual Features (32 words)	98.30	677	4800	25	70
TotalBoost [10]		98.17	682	4788	37	65
LogitBoost [10]		98.24	664	4810	15	83
LPBoost [10]		97.22	637	4780	45	110
AdaBoost [10]		93.95	416	4819	6	331
RobustBoost [10]		91.04	253	4820	5	494
RusBoost [10]		91.04	253	4820	5	494

6. CONCLUSION

Nowadays as text messaging usage is increasing, the problem of spam message has increased. This paper proposed using extreme gradient boosting algorithm (XGBoost) for SMS spam filtration and detection. To select the optimal feature subset from 15 structural features; two wrapper feature selection algorithms are used. A comparison between two different algorithms of wrapper feature selection methods has been applied. According to the results the sequential backward algorithm selected most 9 relevant structural features in terms of accuracy comparing to sequential forward selection. XGBoost gave %98.64 of accuracy when used for handling this imbalanced dataset. For future study, we need to find or create even better spam and ham features in order to handle spam message filtration problems and obtain better accuracy, and also try to find more spam corpus in different languages in order to create one spam filtration system that is compatible to different languages.

REFERENCE

[1] T. A. Almeida, J. M. G. Hidalgo and T. P. Silva, "Towards SMS Spam Filtering: Results under a New Dataset," International Journal of Information Security Science, vol. 2, no. 1, pp. 1-18, 2013.

[2] S. J. Delany, M. Buckley and D. Greene, "SMS spam filtering: Methods and data," Expert Systems with Applications, vol. 39, no. 10, pp. 9899-9908, 2012.

[3] X. Hu and F. Yan, "Sampling of Mass SMS Filtering Algorithm Based on Frequent Time-domain Area," in Third International Conference on Knowledge Discovery and Data Mining, Phuket, Thailand, 2010.

[4] D. Puniškis , R. Laurutis and R. Dirmeikis , "An Artificial Neural Nets for Spam e-mail Recognition," Elektronika ir Elektrotechnika, vol. 69, no. 5, p. 73 –76, 2006.

[5] W. L. Huang, Y. Liu, Z. Q. Zhong and Z. M. Shen, "Complex Network Based SMS Filtering Algorithm," Acta Automatica Sinica, vol. 7, no. 35, p. 990–996, 2009.

[6] X. Zheng , C. Liu and Z. Yu, "Chinese short messages service spam filtering based on logistic regression," Journal of Heilongjiang Institute of Technology, vol. 4, no. 24, p. 36–39, 2010.

[7] K. Yadav, P. Kumaraguru, A. Goyal, A. Gupta and V. Naik, "SMSAssassin: crowdsourcing driven mobile-based system for SMS spam filtering," in Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, Phoenix, Arizona, 2011.

[8] A. K. Uysal, S. Gunal, S. Ergin and E. S. Gunal, "The Impact of Feature Extraction and Selection on SMS Spam

Filtering," Elektronika ir Elektrotechnika, vol. 19, no. 5, p. 2013, 2013.

[9] I. Ahmed, D. Guan and T. C. Chung , "SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset," International Journal of Machine Learning and Computing, vol. 4, no. 2, 2014.

[10] F. Akbari and H. Sajedi, "SMS Spam Detection using Selected Text Features and Boosting Classifiers," in 7th Conference on Information and Knowledge Technology (IKT), Urmia, Iran, 2015.

[11] X. Zhang , G. Xiong, Y. Hu, F. Zhu, X. Dong and T. R. Nyberg, "A Method of SMS Spam Filtering Based on AdaBoost Algorithm," in 12th World Congress on Intelligent Control and Automation, Guilin, China, 2016.

[12] N. Choudhary and A. k. Jain, "Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique," Advanced Informatics for Computing Research, 2017.

[13] J. Ma, Y. Zhang, Z. Wang and B. Chen, "A New Fine-grain SMS Corpus and Its Corresponding Classifier Using Probabilistic Topic Model," KSII Transactions on Internet and Information Systems, vol. 12, no. 2, 2018.

[14] M. Nivaashini, R.S.Soundariya, A.Kodieswari, and P.Thangaraj,"SMS Spam Detection using Deep Neural Network", International Journal of Pure and Applied Mathematics, Volume 119 No. 18, pp. 2425-2436 , 2018,

[15] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras and C. D. Spyropoulos , "An Evaluation of Naive Bayesian Anti-Spam Filtering," in Proc. of the Workshop on Machine Learning in the New Information Age, pp. 9-17, 2000.

[16] SMS Spam Collection v.1 dataset, https://archive.ics.uci.edu/ml/datasets/sms+spam+collection.

[17] "Technical realization of the Short Message Service (SMS). Point-to-Point," ETSI, GSM 03.40, 1992.

[18] S. Günal, S. Ergin, M. B. Gülmezoğlu and Ö. N. Gerek, "On Feature Extraction for Spam E-Mail Detection," in Lecture Notes in Computer Science, , pp. 635-642, 2006.

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

[20] P. Su, Y. Liu and X. Song, "Research on Intrusion Detection Method Based on Improved Smote and XGBoost," in Proceedings of the 8th International Conference on Communication and Network Security, Qingdao, China, 2018.

[21] B. Kumari and T. Swarnkar, "Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review," International Journal of Computer Science and Information Technologies, Vol. 2 (3),pp. 1048-1053, 2011.

[22] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016.