Using Classification Techniques to SMS Spam Filter

Halah Hadi Mansoor, Shaimaa Hameed Shaker

Abstract*: SMS is service that uses mobile phone that allows the users to exchange textual content. Spamming can be defined as sending unwanted content to a group of people for various purposes such as fraud. SMS spam is one form of spamming in which unwanted messages are delivered to many clients by spammers. Therefore, it has become necessary to develop SMS spam detection system to keep up with the current development of message services. Where the aim of this work is developing spam filter for Arabic and English languages by using two filter to be able to detect spam sms efficiently. Content based method was used to build spam filter for English and Arabic languages. based on this method, there are a number of steps should be taken which are Read English and Arabic dataset, Preprocessing phase, Feature Extraction and Classification. The first step after reading the dataset for Arabic and English languages is preprocessing phase which is important step to get more accurate results. The next step is extracting the features from the body of each message. Eight features have been extracted from English messages and six features from Arabic messages. Then features of messages for English and Arabic languages are splitted into two set: training set and testing set. Training set are used to train the algorithms while the test set are used evaluate the performance of proposed Spam filter for the English and Arabic language. In proposed system two classifiers are used. Naive Bayes is used as first classifier and neural network as second classifier. The incoming messages are passed through naive Bayes classifier. If it is classified as ham then passes to second classifier to make sure if it is spam, otherwise it doesn't passes to second classifier. The results of the proposed system were acceptable with 97% accuracy is obtained for English language when using eight features and 80% from dataset for training .And 95% accuracy is obtained for Arabic language with six features and 70% from dataset for training.

Keywords : SMS, Spam Filter, Naïve Bayes, Neural Network, Back Propagation, spam Detection.

I. INTRODUCTION

In today's world, a large number of people use mobile phones especially after the advent of the Internet. One of the most important services used by the client is SMS .The short messaging service (SMS) is a two-way service to send text over wireless systems. The standard length of message in Arabic is 70 and in English it is 160, it contains alphabetic characters, numbers and special symbols [1].

Due to the low cost of SMS, it is became the most popular messaging service and also because of reliability of network has made message service an economic option for GSM subscribers [2].

Revised Manuscript Received on October 05, 2019.

***** Correspondence Author

Halah Hadi Mansoor[1], Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers and Informatics, Baghdad, Iraq[1]

Asst. Prof. Dr. Shaimaa Hameed Shaker[2], Computer Science, University of Technology, Baghdad, Iraq[2]

Many users are tired of receiving, and removing spam sms. In 2012, as one survey in Asia per day is about 30% of the sms are spam on mobile phone. Thus one of the important things to be considered is wasting user time and annoyed by reading and deleting spam messages and also these sms may be for fraud [3].

Depending on the sms log and calls finding a direct or indirect relationship between the sender and the receiver requires a lot of processing, leading to a non-singular record that may need to data analyzed. Hence the proposed system is mainly designed to allow textual communication between users without having to know direct or indirect communication between sender and receiver but depends on the content of the message. The main goal of the system is to develop the SMS service filter by analyzing the message body and checking the content to define if it is spam or not for receiver [3].

The aim of the work is to develop SMS spam filter system for Arabic and English language to suit today's messages requirements. Where it able to classify messages in accurate and fast form with acceptable complexity degree. This is done based on two filters. Naive Bayes as primary filter and then apply Neural Network algorithm as a second filter on ham messages only. This system must be able to classify datasets with more security and detect if there were spam messages that were classified as ham messages by first classifier. And finally get more secure and efficient system.

II. RELATED WORKS

- **Hedieh Sajedi et al. (2016),** In this paper, a survey of many machine learning and hybrid algorithms that used for discovering SMS spam messages which focused on accuracy criterion to compare between them .the sources of data is Original articles written English present in google-scholar.com, Sciencedirect.com, IEEE explorer, Search.com, and the ACM library. Study choice: articles that used hybrid approaches and machine learning to filter SMS spam. Many articles have been deduced by searching in a pre-defined series and the result was reviewed by one author and examined by the second. And the third author reviewed and modified the Preliminary paper. Outcome: a total of 44 articles were chosen relating to machine learning and hybrid methods that are used to detect spam messages. From these papers, they were able to extract 28 methods and algorithms, and then only 15 algorithms have been chosen and compared in one table based on their accuracy, strengths, and weaknesses in detecting spam messages of the Tiago dataset of spam message.

Among the suggested methods is DCA algorithm, the large cellular network method and graph-based KNN are three most accurate in filtering unwanted messages of Tiago dataset. Furthermore it, hybrid methods were discussed in this paper [4].

- **Sheetal Ashokrao Sable and Prof. P.N. Kalavadekar (2016),** In this paper, researchers used hybrid system to classify SMS for detecting spam or ham, by using diverse algorithms such as Naïve Bayes classifier and Apriori Algorithm. So there are a bunch of important steps in classification SMS that should be done, such as collect SMS dataset, selecting of features, creation of vector, filtering process and updating system where Two types of SMS classification are enlisted as Black and White. Naïve Bayes can be considered as one of the most effective and important learning algorithms for data mining and machine learning as it has been considered as the basic method for retrieval the information [5].
- **Shafi'i Muhammad Abdulhamid et al. (2017),** The researchers focused on presenting and reviewing the methods currently available to deal with the problem of detecting SMS spam and filtering and mitigation of mobile SMS spams, as well as reviewing challenges and future research directions and providing recommendations for future researchers in this field. So, the related researches are analyzed and reviewed .The most common techniques for spam discovery, filtering and reduction are compared, including the used datasets, their results, constraints, challenges and future trends of research. This review is designed to assist experts to identify open areas that need further improvement. Besides, those studies were based on the construction of the SMS spam classifier on SVM and Bayesian network [6].
- **Heba Adel, Dr. Maha A. Bayati 2018**, in this paper, Naïve Bayesian" (NB) was proposed for spam classification system and to build bi-lingual classifier. Based on content based method, the classifier can classify input Arabic messages as being legitimate or unsolicited. The proposed filter was evaluated to measure its efficiency. For Arabic sms dataset a total of 400 SMS were splitting to 70% for training and 30% for testing .with 15 features were extracted from each sms, an accuracy 85% was achieved [7].

III. SPAM FILTERING METHODS

There are two basic methods for detecting the SMS spams [8].

1. **Collaboration-based method:** This method is based on user feedback and shared user experience. Collaborative filtering is a way of making automatic predictions (filtering) about user interests by collecting preferences or tasting information from multiple users (collaboration).
2. **Content-based method:** This method focuses on analyzing the textual content of messages. This is more common due to the difficulty of accessing data related to usage and user experience.

IV. THE PROPOSED SYSTEM DESIGN

Now we introduce the Methodology of work and the designs that are used to build SMS Spam Filter classification application. Spam filter is application that is used to detect unwanted message and not allow it to save in user's inbox. Python was used to develop sms spam classification program and machines learning algorithm such as naive Bayes and neural network to learn my system.to build spam filter there are steps should be taken as shown in figure 1.

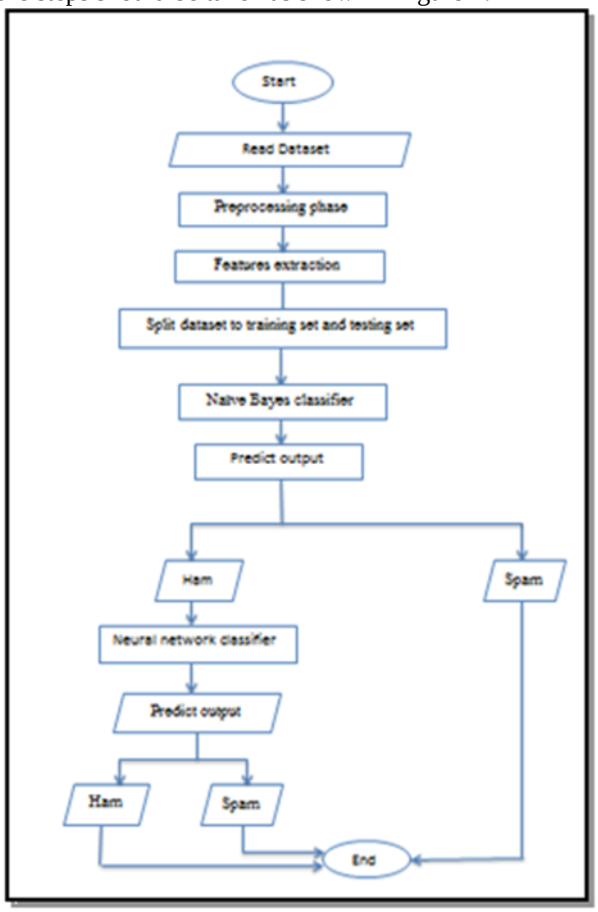

Figure1 flowchart illustrated the design of proposal system

A. Read Data Set (English and Arabic Datasets)

The first step in the system is a load the dataset and stores it in a csv file format. Two dataset were used which is English and Arabic dataset. UCI machine learnings repository dataset are used for English which contains 5,574 short messages divided into 4, 827 ham SMS and 747 spam ones. As for the Arabic dataset there are no sources available to download it so it is created with translate some of English SMS and download some message from Zain Official Website. Read_csv function was used to read two dataset and then made ham or spam label for each SMS.

B. Preprocessing phase

In spam filtering, preprocessing of text information is very important and critical. According to previous literature on spam filtering for email or sms, the tokenization step may be the most significant in the process of message analysis and learning a classifier on them. Main purpose of pre-processing text data is to remove data that does not provide useful information regarding the class of the document. Most of the data cleaning steps that are widely used in preprocessing tasks is removing the stop words [9].

Algorithm (1) Preprocessing

```
Input: body of SMS
       Stop words      // list of English and Arabic stop words
Output: sms in form of Token      //SMS with prepressing phase
 Begin:
  For  each SMS body do:
  If sms [word]! = 'space' then    //remove white space
   If sms [word] doesn't exist in stop words list then
                                          //remove stops word
       Data [token] =sms [word]
    Else:
         Counter++
  End for
End
```

C. Features Extraction phase

The process of extracting features from messages is a very important on which the accuracy of the machine learning algorithms depends. Where accurate analysis of sms dataset and extracted more accurate features Leads to increase the probability detection of spam message. Short messages have Specific limit length and contain only text without any attachments file or graphics. While in email have no limited for text length and also contains attachments or graphics. There are two types of sms messages namely the "ham" message and the "spam" message. Spam and ham can be distinguished using many features.

Extracting useful and well features from each message that help in filter SMS messages efficiently. Moreover, properties of sms spam messages are studied in depth and detect some useful features that lead to effective sms spam detection. Table-I illustrates features that are extracted from sms and used it in our spam filter system. Where eight features are extracted from English sms and six features are extracted from Arabic sms.

D. Training and Testing Set

Two dataset are used one for English language and the other for Arabic language that saved in csv file. English dataset were splitting into two part which are 80% for training and 20% for testing but Arabic dataset were splitting into 70% training and 30% testing as illustrated in table-II. Training set use to train algorithm in training phase but testing set used to evaluate the performance

Table-I list of Features for English and Arabic dataset

No	Feature Name	Description
F1	Message length (for Arabic & English)	Number of all characters
F2	Uppercase letters	Number of uppercase characters
F3	Non-alphanumeric Character (for Arabic & English)	Number of non-alphanumeric characters
F4	Numeric character (for Arabic & English)	Number of numeric characters
F5	Presence of URL (for Arabic & English)	Presence of "http" and/or "www" Terms
F6	Spam words (for Arabic & English)	The number of spam words (the list of spam words in Appendix at table (3) and table (4))
F7	Uppercase words	Number of uppercase words
F8	Phone Number (for Arabic & English)	Phone numbers keys for group of Countries

Table-II training and testing set groups

Number of Sample	Dataset type	phase
4457	English set	Training
1115	English set	Testing
350	Arabic set	Training
150	Arabic set	Testing

E. Classifier

Classification is procedure of predicting and discovering the category of given data points. Categories are sometimes called labels /goals or Classes. The predictive modeling of a classification is the task of approximating a mapping function (f) from the input variables (X) to separated output variables (y). As example that can be defined as a classification problem is detecting of spam message and this classification is binary because there are only two classes which is spam and ham. Classifier needed some training dataset to understand how given input variables relate to the class. In sms spam filter case, used known spam and ham sms dataset as the training data and when the classifier is trained accurately, it can be used to detect an unknown sms. There are a lot of machine learning classification algorithms available now but it is not possible to infer which one is better than the other [10].

The following algorithms were used for experiments.

1. Naïve Bayes Classifier

It is the easiest probabilistic classifiers which are lean on Bayes theorem with strong naïve independence assumption. Naive Bayes as a probabilistic model is very simple and shows good performance under conditions where the occurring words are independent of each other. The function of the classifier is to classify the Arabic and English SMS into two classes, which is spam or ham depending on the extracted features. Naive Bayes classifier passes into two stages which are training and testing phase.

A. Training phase

At this stage the following is calculated

a) Spam SMS class and Ham SMS class probabilities of complete number of sms sample are calculating using the following Equation.

$$P(C_j) = \frac{\text{the number of SMS in } C_j}{\text{the number of all SMS}}$$

Where C_j = type of class which is either spam or ham

b) The P(Ci) probability of certain sample SMS(X) which is Belong to one of the two classes are Calculated, this is done by calculating the probability of appearance of individual feature x_j in either class using the following Equation

$$P(x_i|C_j) = \frac{\text{the number of feature } x_i \text{ appearing in } C_j}{\text{the total number of all features appearing in } C_j}$$

B. Testing phase

This is the phase to classify unknown SMS (in the dataset) into Spam or Ham message. This phase uses the result from training phase with entries of new SMS messages to classify it. The testing phase is

done on the new message set as follows.

a. for each sms sample X calculate the probabilities (If feature F is present in the training then we will take the probabilities for both classes)
b. if the specific value for some features x does not exists in training phase then Applying the "near range" mechanism .as shown in algorithm 3.2
c. For each sample X, by using "Bayes theorem" calculate the posterior probabilities. By applying the following equation

$$C_X = \arg\max \sum_{i}^{n} P(x_i | C_j) P(C_j)$$

d. Then depended on the result from c, decided the class of X with largest probability for it.

Algorithm (2) Nearest Range

Input: Feature F, Training Set
Output: ham and spam probabilities for input Feature F
Begin:
For all Pi (F) in training set for Ci //where Pi (F) = ham or Spam probability for all Features and Ci= type of class
Find two nearest Features
Calculate average of two features by applying
Pi(F)=(Pi(F1)+Pi(F2))/2
Return Pi(F)
End for
End

2. Neural Network Algorithm

The Proposed system uses 5,574 SMS for English set and 500 for Arabic set. Each one of them implement as a multilayer pass forward neural networks with back propagation learning algorithm. Each neural network has one hidden layer. The number of nodes in input layer depends on number of features that are extracted from the message. And the number of nodes in output layer depends on number of output classes. While the number of nodes in hidden layer can be determined by computing the average number of nodes in input and output layers or can by trial and error. The datasets are implemented as neural networks with back propagation learning algorithm. The activation function that is used in back propagation learning algorithm is the sigmoid function.

A. Training phase

In training phase will be used the pattern mode with back propagation learning algorithm to train the network on each sms in system. The training phase with pack propagation method that used in the proposed classification system is illustrated in algorithm 3

In training phase, neural network algorithm used the same training sets that are used in naive Bayes to train network.

Algorithm 3.4 Back Propagation

Input: Features of sms //eight input matrix for English language And six input matrix for Arabic language
Output: Trained data with appropriate weights
Begin:
Step 1: initialize a random small weights on all links between layers in the network
Step 2: Keep doing epochs: for every pattern (input feature matrix with target) from training set do:

a) Calculate the actual output by using sigmoid function and applying an equation

$$y(k) = F\left(\sum_{i=0}^{m} w_i(k).x_i(k) + b\right)$$

Where:
xi (k) is input value in discrete time k where **i** goes from 0 to m
wi (k) is weight value in discrete time k where i goes from 0 to m
b :- is bias
F: - is transfer function
y (k) is output value in discrete time k

b) Compute the difference between the determined and the target values of the output layer by equation

$$\delta_j = y_j(1 - y_j)(d_j - y_j)$$

Where d_j is the target output of node ***j*** and ***y j*** is the actual output

c) Compute the error for all nodes in the hidden layer by applying equation

$$\delta_j = X_j(1 - X_j)\sum \delta_{jm} w_{jk}$$

Where, k is over all nodes in the layers above node j.

d) Then update weights between output layer and hidden layer and between hidden layer and input layer by applying equation

$$w_{ij}(t+1) = w_{ij}(t) + \eta\, \delta_j X_j$$

Where:
Wij(t) is the weight from hidden node i or from an input to node j at time t,
Xi is either the output node i or is an input,
η is a gain term
δ_j is an error term for node j

Step 3: repeat from step 2 until errors become small than mse_{min} or an epoch complete
Step 4: Save weights.
End.

B. Testing phase

In proposed system, test phase used only the messages that are predicted by naive Bayes as ham messages. As we shown in testing flowchart enter the testing pattern to network to make prediction. We use the weights that are stored from training phase and only apply the equation.

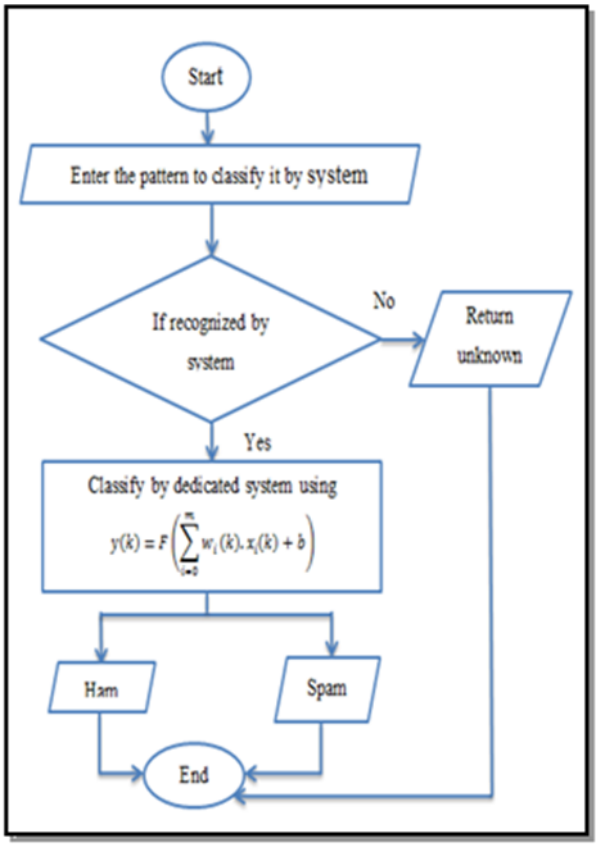

Figure (2) Testing Phase flowchart for neural network algorithm

V. EXPERIMENTS AND EVALUATION OF RESULTS

Here we will present the evaluation of the performance of the proposed approach for solving the SMS Spam Problem. The criteria used to measure the performance of the system (naive Bayesian with neural network spam classifier) were: accuracy, recall, precision, false positive and false negative.

- **Accuracy:** Performance is often measured with regard to accuracy. The accuracy can be defined as

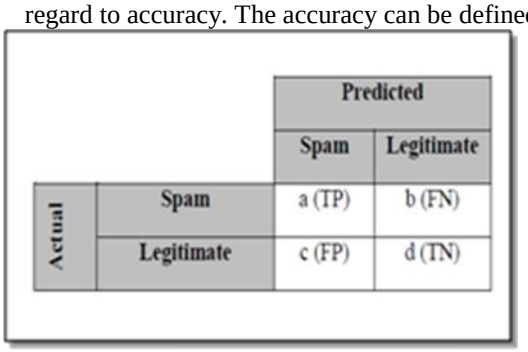

Figure 3 Confusion matrix

the proportion of the total number of SMS that were classified correctly, and is calculated as shown in following equation

$$Accuracy = \frac{a+d}{a+b+c+d}$$

- **Precision:** this can be defined as the proportion of the predicted positive cases that were correct, in other words it measures the degree to which the spam classified SMS are truly spams. It is calculated using the following equation

$$\text{Precision} = \frac{a}{a+c}$$

- **Recall:** This can be defined as the proportion of positive cases that were correctly identified, in other words it measures the percentage of SMS spam that the system addresses correctly. It is calculated using the following equation

$$\text{Recall} = \frac{a}{a+b}$$

- False negative (FN) is defined as the number of positives cases that were incorrectly classified as negative, which occurs when, the spam sms is classified as legitimate.
- False positive (FP) is defined as the number of negatives cases that were incorrectly classified as positive

1. Naïve Bayes Performance

In sms spam filter system, the Naive Bayes was used as the first filter and the input to classifier algorithm was the eight features for the English language and six features for the Arabic language that mentioned in table-I .The result as shown in table III and table IV for English dataset and table V and table VI for Arabic dataset.

Table-III Confusion matrix of Naïve Bayes for English SMS

	Pred: ham	Pred: spam
True ham	941	22
True spam	46	106

Table-IV Evaluation measures of naïve Bayes for English SMS

	Accuracy	Precision	Recall	support
ham	93.9%	0.95	0.98	963
spam		0.83	0.70	152

Table-V Confusion matrix of Naïve Bayes for Arabic SMS

	Pred: ham	Pred: spam
True ham	84	0
True spam	11	55

Table-VI Evaluation measures of naïve Bayes for Arabic SMS

	Accuracy	Precision	Recall	support
ham	92.6%	0.88	1.00	84
spam		1.00	0.83	66

2. System Performance

Neural network is used to improve the performance of the naive algorithm and increase security of system. After obtained the naive Bayes results as first filter then taken only sms that was predicted as ham messages from first filter and pass them to neural network as a second filter to obtain more secure prediction system. And the result as shown in table VII and table VIII for English dataset, Table IX and Table X for Arabic dataset .As we can see in the table VII, the ratio of FN (spam sms is classified as ham) messages decreased to more than half

Table-VII Confusion matrix of System for English SMS

	Pred: ham	Pred: spam
True ham	946	25
True spam	12	132

Table-VIII Evaluation measures of System for English SMS

	Accuracy	Precision	Recall	support
ham	97%	0.99	0.97	971
spam		0.84	0.92	144

Table-IX Confusion matrix of System for Arabic SMS

	Pred: ham	Pred: spam
True ham	83	1
True spam	6	60

Table-X Evaluation measures of System for Arabic SMS

	Accuracy	Precision	Recall	support
ham	95%	0.93	0.99	84
spam		0.98	0.91	66

VI. CONCLUSION

To classify sms into Spam or Ham, NB classifier needs to be trained, and then tested. The size of training and testing datasets affect the performance of the proposed classifier. It has been recognized when split dataset to 80% for training and 20% for testing that give the best results when it is splitted to 70% for training and 30% for testing as shown in Table –XI and Table-XII for Naïve Bayes

Table- XI Confusion matrix of naïve Bayes with 70% from dataset for training

	Pred: ham	Pred: spam
True ham	1444	20
True spam	116	92

Table- XII evaluation measures of naïve Bayes with 70% from dataset for training

	Accuracy	Precision	Recall	support
ham	92%	0.93	0.99	1464
spam		0.82	0.44	208

Table-XIII Confusion matrix of System with 70% from dataset for training

	Pred: ham	Pred: spam
True ham	1430	34
True spam	31	177

Table-XIV evaluation measures of System with 70% from dataset for training

	Accuracy	Precision	Recall	support
ham	96%	0.98	0.98	1464
spam		0.84	0.85	208

Classifiers performance is enhancing with a bigger training sample size Thus the result in English dataset is best than the result in Arabic dataset as shown in table II. Eight features was used that mentioned in the table I for English dataset. And the results were shown in Table VIII. but when three features were deleted which are (special character, numeric character and upper word) and were used only five remain features in system, noticed that the algorithm's accuracy was reduced to 93% as shown in Table-XV

Table-XV performance of System with five features only

	Accuracy	Precision	Recall	support
ham	93%	0.95	0.96	1450
spam		0.75	0.70	222

So we concluded that the number and accurated of features extracted form each message have great impact on the accuracy of the system. aslo We conclude from our work that when two algorithms (naive and neural) are used, this increases the security of the system and provides better accuracy from using only Naive Bayes . Where the accuracy of the Naive Bayes algorithm was 93.9% for english dataset and 92.6% for arabic dataset . but When using Neural Network as second classifier the peformance of system was improved to 97% for english dataset and 95% for Arabic dataset

REFERENCES

1. Shirani-Mehr, H.J.u.h.c.s.e.p.S.a.r.-S.p.: 'SMS spam detection using machine learning approach', 2013
2. Zhu, Y., Tan, Y.J.I.T.o.I.F., and Security: 'A local-concentration-based feature extraction approach for spam filtering', 2010, 6, (2), pp.486-497
3. N. G. M. J. J. o. T. JAMEEL and A. I. Technology, "SMS SPAM DETECTION USING ASSOCIATION RULE MINING BASED ON SMS STRUCTURAL FEATURES," vol. 96, no. 12, 2018.
4. H. Sajedi, G. Z. Parast, and F. J. M. L. R. Akbari, "Sms spam filtering using machine learning techniques: A survey," vol. 1, no. 1, pp. 1-14, 2016.
5. S. Sable, P. J. I. J. o. I. i. E. R. Kalavadekar, and Technology, "SMS Classification Based on Naïve Bayes Classifier and Semi-supervised Learning," vol. 3, no. 7, 2016.
6. M. A. Shafi'I *et al.*, "A review on mobile SMS spam filtering techniques," vol. 5, pp. 15650-15666, 2017.
7. H. Adel, M. A. J. I. J. o. N. T. Bayati, and Research, "Building Bi-lingual Anti-Spam SMS Filter," vol. 4, no. 1.
8. G. Sethi and V. J. I. J. C. S. I. T. Bhootna, "SMS spam filtering application using Android," vol. 5, no. 3, pp. 4624-4626, 2014
9. S. Vijayarani, M. J. Ilamathi, M. J. I. J. o. C. S. Nithya, and C. Networks, "Preprocessing techniques for text mining-an overview," vol. 5, no. 1, pp. 7-16, 2015
10. G. Giacinto and F. J. P. R. Roli, "Dynamic classifier selection based on multiple classifier behaviour," vol. 34, no. 9, pp. 1879-1881, 2001

AUTHORS PROFILE

Halah Hadi Mansoor, received the B.Sc. degree in computer science from Al-Nahrain University, Baghdad, in 2012, and the High diploma degree in Web site technology from Iraqi Commission for Computers and Informatics, Informatics Institute for Postgraduate Studies, Baghdad, in 2016, she is currently pursuing the M.Sc. Degree in computer science from Iraqi Commission for Computers and Informatics, Informatics Institute for Postgraduate Studies, Baghdad. She was work as lecturer in university of Karbala for three years from 2013-2015.Her researches interests includes websites technology, AI using machine learning and spam filtering. She has published article relevance to his research interest Titled "Design and Implementation Dynamic Website for Electronic Library