Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SMS SPAM DETECTION USING ASSOCIATION RULE MINING BASED ON SMS STRUCTURAL FEATURES

[1]NOOR GHAZI M. JAMEEL

[1]Lecturer. Department of Computer Networks, Technical College of Informatics, Sulaimani Polytechnic University, Kurdistan Region, Iraq

E-mail: [1]noor.ghazi@spu.edu.iq

ABSTRACT

The popularity of using mobile phones has led to an increase in sending SMS messages. SMS messages are considered as a rapid way of communication due to its low cost and easy usage. As a result, SMS was target for many types of threats, one of these is spamming. SMS spam is unwanted message sent to many mobile phone users, and cause many problems like annoyance, consuming mobile network bandwidth and other real threats like scam, stealing personal information and installing malware. In this paper, a new system is proposed to detect spam SMSs using Apriori algorithm. This algorithm is used to generate association rules which applied to new SMSs to classify them into spam or legitimate. The system used structural features only instead of textual features or tokens. These features extracted from two publicly available datasets which consists of spam and legitimate SMSs. The rules are generated with different minimum support and minimum confidence values. The generated rules are applied to the test dataset then the rules which achieved higher accuracy are used in the proposed system. The aim of this work is presents a new and fast approach to detect spam SMS using structural features only and to find out if structural features are enough to detect spam SMSs instead of bag of words which depends on preprocessing and consists of many steps like parsing, tokenization, stop word removal and stemming. Good accuracy achieved with 97.65% using rules generated by Apriori association rule mining algorithm with minimum support 0.2 and minimum confidence 0.8 based on SMS structural features only.
Keywords: *SMS, SMS Spam, Association Rule Mining, Apriori Algorithm*

1. INTRODUCTION

Short Message Service (SMS) is one of the common communication methods for people through mobile devices or internet connected computers. Due to rapid increase in the number of mobile phone users around the world, this increase attracted spammers to send spam SMS or sometimes called junk SMS. Spam SMS is unwanted message sent in large quantities often for commercial or advertising purposes to many recipients [1], [2], [3]. According to Truecaller report in 2017, found that 56% of American adults claimed to receive SMS spam and texts at least one every month, with approximately one of every four, (23%) claiming six or more spam texts received a month. On average, Americans receive 8.4 spam text messages in an average month approximately 30% from 2015 [4].

Short messages often consist of only a few words, limited message length, symbols, specific linguistic with abbreviations which is called SMS language, therefore present a challenge to traditional bag-of-words based spam filters [5], [6]. The main problems with SMS spam are annoyance, waste for time and it can also be costly since some people pay to receive text messages. Another aspect is that spam SMS may lead to other types of threats such as viruses, fraud, man in the middle attack and message disclosure [7], [8]. Consequently, detection of spam messages corresponds to a binary text classification problem where the classes are defined as “spam” and “legitimate”. Vast amount of text classification studies used the bag-of-words model to represent text documents where the exact ordering of words, or terms, in the documents is ignored but the number of term occurrences is considered. Even if SMS spam filtering can be treated as conventional text classification task, the structure of spam messages can be significantly different than that of formal texts. Since the size of an SMS message is limited with just 160 characters, both the message length and number of terms have of great importance. Also, the usage of upper or lower case characters can be indicator of spam. Similarly, some non-alphanumeric characters (e.g.,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

"!", "$") and numeric characters (e.g., phone numbers) are commonly encountered in spam messages. Finally, URL links are usually observed in SMS spam as well [2]. The work in this paper presents a new approach to quickly detect spam SMS by considering all structural characteristics and extract them from SMS messages instead of bag of words. Then according to these features, rules are generated using Apriori association rule data mining algorithm. The rules are used to classify SMS into spam or legitimate, then the specificity, recall and accuracy of the system are computed with different values for minimum support and minimum confidences. The aim of this work is find out if structural features are enough to detect spam SMSs instead of bag of words, if so then the process of detecting spam SMS will be fast since extracting bag of words consists of many steps and it is a time consuming process. And according to the results the proposed system presented a good accuracy. This paper is organized as follows. Section 2, illustrates related works for filtering and classification spam and legitimate SMS. In Section 3, SMS concept and Apriori algorithm are explained. Section 4, the proposed system is described. In Section 5 the results and performance analysis of our suggested method is discussed. The last section addresses conclusions and future work.

2. RELATED WORKS

Recently, there are many researchers conducted studies for filtering, detection and classifications SMS messages. below are some of related works in this field:

Rafique and Abulaish [9] used graph based learning to classify SMS into spam and legitimate. labeled messages are represented as a graph of words then the probability of occurrence and joint distribution among the tokens of spam and legitimate SMSs is computed. KL-Divergence is used to classify new SMSs based on the computed probabilities. Results show that the system achieved 98% detection rate with false alarm rate of less than 0.08 during the classification of spam SMS in 10-fold cross validation. Anchal and Sharma [10] proposed SMS spam filtering technique by applying principle component analysis algorithm and custom neural network with Independent Component analysis algorithm to classify SMS into spam and legitimate. Ishtiaq et al [11] used Navïe Bayes classifier and Apriori algorithm to classify SMS into spam and ham based on words in the SMS. UCI Data Repository was used as a dataset for training and testing and the achieved accuracy 98.7% compared to the traditional Navïe Bayes with accuracy 97.4%. Sin-Eon et al [12] proposed Frequency Ratio as a feature selection method to select important features then the performance of the system was compared with other methods like Naive Bayes, J-48 Decision Trees and Logistic regression. The results show that the accuracy achieved by Naive Bayes 94.70%, J-48 algorithm 94.82%, and Logistic algorithm 94.71%. Kamahazira and Mohd [13] studied the features extraction methods used in classifying spam and legitimate SMS. To find the best and relevant features to be used in this field. performance testing has been conducted to have a better perspective on the influence of spam words in recognizing spam messages. Tiago A. Almeida et al [14] proposed and evaluated a method to normalize and expand original short text messages in order to identify better attributes and enhance the classification performance. The proposed approach is based on lexicographic and semantic dictionaries along with some techniques for semantic analysis and context detection. Mohamed El Boujnouni [3] proposed a SMS spam filtering approach using an improved version of support vector domain description based on N-gram features for SMS classification into spam and ham SMS. For Feature selection, Information gain was used to select the most relevant features. The results show that the proposed approach that the SMS classification accuracy rate reached 95.13% in the training phase and 89.32% in the testing phase. Suleimana, and Al-Naymata [15] proposed SMS detection system based on using H2O platform instead of Weka to compare among different machine learning algorithms like random forest, deep learning and naïve bays. According to their results the best algorithm was the random forest which achieved to 96%, 86%, 91% and 0.977% in term of precision, recall, f-measure and accuracy respectively. Nagwani [16] used two stage process to identify and classify spam SMSs. In the first stage, SMS messages are classified into the two classes spam and non-spam using popular binary classifiers, and then at the second stage non-spam SMS messages are classified into the priority and normal SMS messages. Naïve Bayes (NB), Support Vector Machine (SVM), Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) are used to classify the SMS text message at different stages of classification. The results show that SVM algorithm performed better for filtering the spam messages and categorizing the priority messages. Choudhary and Jain [17] Studied spam SMS features and used ten features for SMS classification. Classification accuracy is tested

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

using WEKA tool and five machine learning algorithms: Naïve Bayes, Decision Table, Logistic Regression, J48, and Random Forest. According to the results, Random Forest algorithm achieved the best result with 96.1% as a true positive rate.

Most of previous works used either textual features which requires preprocessing and consists of many steps: parsing, tokenization, stop word removal and stemming in which it is a time consuming process. Others used combination between textual and structural features. In this work, just 6 structural features are used with two combined spam SMS datasets to generate association rules using Apriori association rule mining algorithm to detect Spam SMS messages.

3. CONCEPTS AND METHODS

In this section, the concepts and theoretical aspects are illustrated.

3.1 Short Message Service (SMS)

SMS stands for Short Message Service, and it is the commonly way for making a connection between the sender and the receiver through a mobile. The concept of sending and receiving SMS is shown in figure 1.

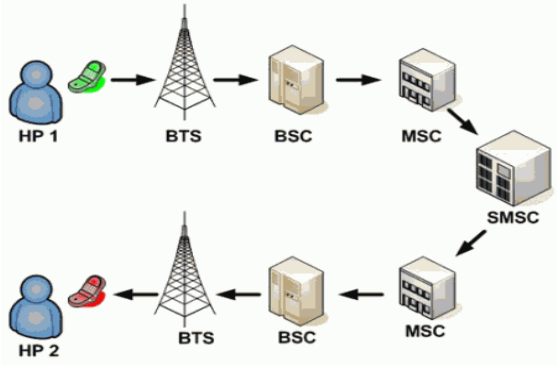

Figure 1: SMS Transmission

First the message is sent from the user to the base transceiver station (BTS) in which a wireless connection is established between the user and the network. The message is delivered to the receiver through the mobile switching center (MSC) which is accountable for routing voice calls and SMS, and short message service center (SMSC). The maximum message length in SMS in the beginning was 128 bytes carrying 146 characters using 7 bits for each character. This size enhanced to be 140 bytes as 160 characters for each SMS [18], [19].

3.2 Association Rule Mining

Association rule is one of the most research areas in data mining and machine learning. Association rule mining is a valuable tool that has been used widely in various areas [20]. Association rule uses two criteria, support and confidence. It identifies the relationships and rules generated by analyzing data for frequently used if/then patterns. Association rules are usually needed to satisfy a user-specified minimum support which is computed by equation (1) and a user specified minimum confidence which is computed by equation (2) [21].

Rule: X ==> Y

$$\text{Support} = \frac{frq(XY)}{N} \quad (1)$$

$$\text{Confidence} = \frac{frq(X,Y)}{frq(X)} \quad (2)$$

Where:
Support (s) = Fraction of transactions that contain both X and Y
Confidence(c)= Measures how often items in Y appear in transactions that contain X.
To produce association rules with specified minimum support and confidence. There are two stages:

- Stage 1: Generating item sets with the specified minimum support. The first stage proceeds by generating all one-item sets with the given minimum support and then using this to generate the two-item sets, three-item sets, and so on. Each operation involves a pass through the dataset to count the items in each set, and after the pass the surviving item sets are stored in a hash table—a standard data structure that allows elements stored in it to be found very quickly. From the one-item sets, candidate two-item sets are generated, and then a pass is made through the dataset, counting the support of each two-item set; at the end the candidate sets with less than minimum support are removed from the table. The candidate two-item sets are simply all of the one-item sets taken in pairs, because a two-item set cannot have the minimum support unless both its constituent one-item sets have the minimum support, too. This applies in general: A three-item set can only have the minimum support if all three of its two-item subsets have minimum support as well, and similarly for four-item sets.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Stage 2: From each item set determining the rules that have the specified minimum confidence. If only rules with a single test on the right side were sought, it would be simply a matter of considering each condition in turn as the consequent of the rule, deleting it from the item set, and dividing the support of the entire item set by the support of the resulting subset—obtained from the hash table—to yield the confidence of the corresponding rule.

3.3 Apriori Algorithm

Apriori algorithm is one of the efficient algorithms of association rules mining. The algorithm finds frequent itemsets and used level-wise search. It generates the rules by two steps [22]:

1. Find all the frequent itemsets that satisfy minimum support.
2. Generate all the rules that satisfy minimum confidence using the frequent itemsets.

Itemset consists of set of items, the frequent itemset is an itemset that has support above minimum support. The frequent individual items in the database are identified and extended to larger and larger itemsets as long as those itemsets appear sufficiently in the database. After all frequent itemsets are found, it generates rules [21] [22].

4. THE PROPOSED SYSTEM

SMS contains some proprieties and characteristics, these proprieties are from text of the SMS which is treated as bag of words and some of these proprieties are structural features. From these characteristics, SMS(s) can be classified into spam and legitimate. In this paper, the proposed system classifies SMS into spam or legitimate using association rule data mining by Apriori algorithm based on structural features only. However, set of steps are performed: SMS collection, features extraction, convert features into binary, rules generation using Apriori data mining algorithm to classify SMS into spam and legitimate SMS. The overall process is described in general in the block diagram in figure 2 and discussed below:

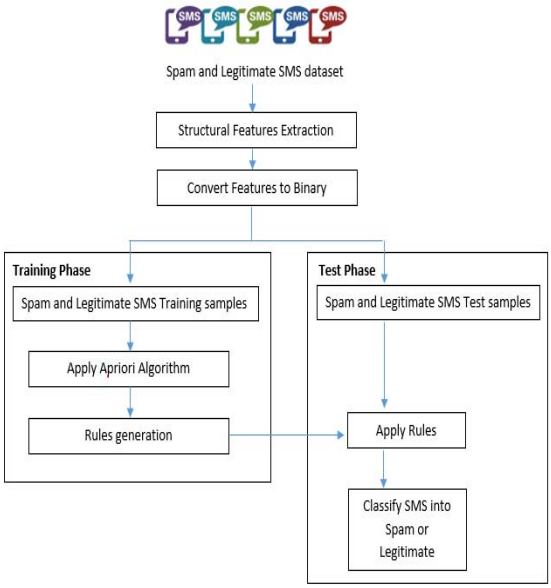

Figure 2: The Proposed System

4.1 SMS Collection

Two publicly available datasets are used in our proposed system: SMS Spam Collection v.1 UCI Machine learning repository which consists of 5574 SMSs of spam and legitimate (https://archive.ics.uci.edu/ml/datasets/sms+spam+collection), and data from DIT SMS Spam dataset which consists of 1353 spam SMS (http://www.dit.ie/computing/research/resources/smsdata/). Both datasets are combined together to be used for features extraction process. The dataset is divided randomly into two subsets: The first contains 60% of spam and legitimate SMSs and the second contains 40% of spam and legitimate SMSs. The proposed system works in two steps the training will be performed with the first subset and the testing will be performed with the second. Table 1 illustrate the number and percentage of spam and legitimate SMS used in the proposed system.

Table 1: Statistics of the SMS Dataset

Type of SMS	Number of messages	Percentage
Legitimate	4827	69.68%
Spam	2100	30.32%
Total	6927	100%

4.2 SMS Structural Features Extraction

Since SMS messages are different from normal text messages, there are limits in text message length, number of terms, uppercase characters, numbers, alphanumeric characters and the existence of URL are important indicators of spam SMS. In this paper only set of structural

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

features are extracted from spam and legitimate SMS dataset and these features are illustrated in table 2. The feature extraction process is developed by Visual Basic. Net programming language.

Table 2: Statistics of the SMS dataset [2]

No.	Feature	Description
SF1	Message Length	Number of all characters in the SMS
SF2	Number of Terms	Number of tokens or terms in the SMS
SF3	Uppercase Character Ratio	Number of uppercase characters divided by the message length.
SF4	Non alphanumeric Character Ratio	Number of Non alphanumeric characters divided by the message length.
SF5	Numeric Character Ratio	Number of numeric characters divided by the message length.
SF6	Existence of URL	Check if SMS contains URL or No

The proposed system focuses on specifying the relevant features that differentiate spam SMS from legitimate SMS. In order to specify the relevant features, certain statistical analysis and calculations were carried out on the SMS dataset with total of 2100 spam messages and total of 4827 legitimate messages. Since some of the structural features has continuous value and Apriori algorithm cannot handle these values so based on 6 heuristics, features (SF1, SF2, SF3, SF4, SF5) were defined. And according to these heuristics the features are converted to binary and stored as 10 features. The last feature (F6) is a binary value if the URL exists in the SMS the feature value is stored as 1 else its value is 0. Then these features are subjected to association rule mining to effectively determine the legitimate and spam SMS.

Heuristic 1: Message Length

SMS is a limited length text. It consists of uppercase, lowercase letters, numbers, non-alphanumeric characters. For each SMS in the dataset, SMS message length is extracted and examined. For the input data set (2100 spam SMSs and 4827 legitimate SMSs), message length is analyzed for spam and legitimate SMSs. The distribution of the message length for spam SMS is shown in figure 3 and the average length of the message (l) in spam SMS is found to be greater than or equal 136 characters. The distribution of the message length for legitimate SMS is shown in figure 4 and the average length of the message length (l) in legitimate SMS is found to be 72 characters. Therefore, the heuristic is defined as:

H1 = If length>= l → Spam SMS

Else Legitimate SMS

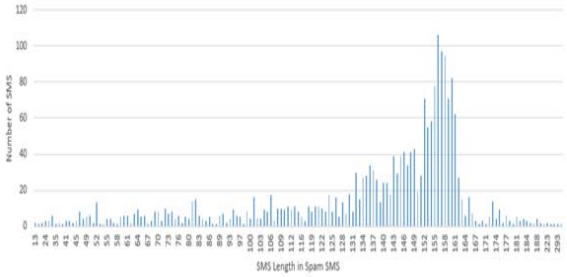

Figure 3: Message Length in Spam SMS

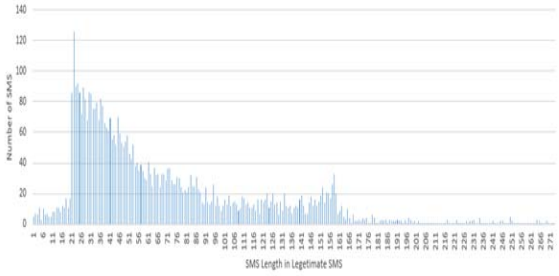

Figure 4: Message Length in Legitimate SMS

Heuristic 2: Number of Terms

In this work, the number of terms in SMS is considered as a feature to identify spam and legitimate SMS. The number of terms (μ) examines in spam and legitimate SMSs. For the input data set (2100 spam SMSs and 4827 legitimate SMSs), the number of terms is analyzed for spam and legitimate SMSs. The distribution of number of terms in the spam SMSs and legitimate SMSs are analyzed and shown in figure 5 and figure 6 respectively. The result shows that the average number of terms (μ) in spam SMS is found to be greater than 23 and the average number of terms (μ) in legitimate SMS is found to be 14.

H2 =If (number of terms) > μ → Spam SMS

Else Legitimate SMS

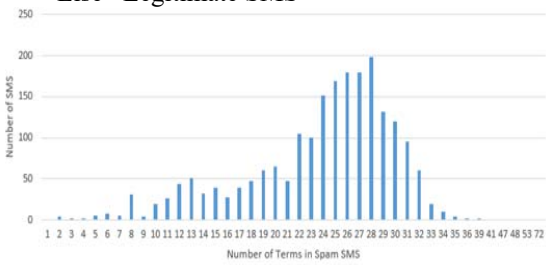

Figure 5: Number of Terms in Spam SMS

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

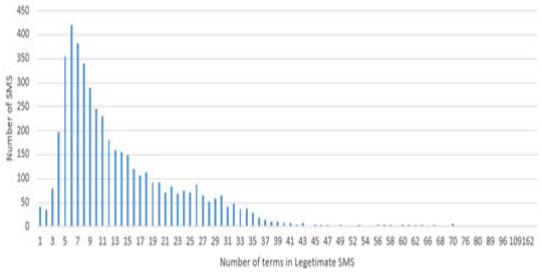

Figure 6: Number of Terms in Legitimate SMS

Heuristic 3: Uppercase Character Ratio

SMS contains uppercase and lowercase characters, computing uppercase character ratio considered as one of the features for identifying spam and legitimate SMSs. For the input dataset (2100 spam SMSs and 4827 legitimate SMSs), the uppercase character ratio is analyzed for spam and legitimate SMSs and are plotted in figure 7 and figure 8 respectively. The Result shows that the uppercase character ratio (α) in spam SMS is greater than 0.1 and the uppercase character ratio (α) in legitimate SMS is 0.05. Therefore, the heuristic is defined as:

H3= If (uppercase character ratio) <α → legitimate SMS
Else Spam SMS

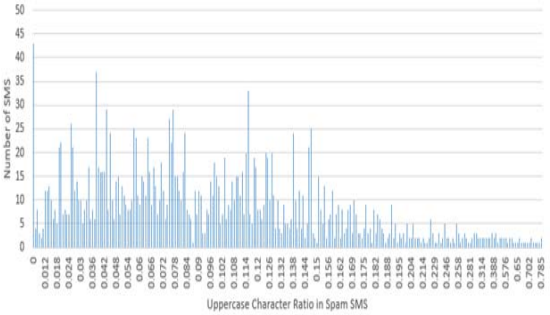

Figure 7: Uppercase Character Ratio in Spam SMS

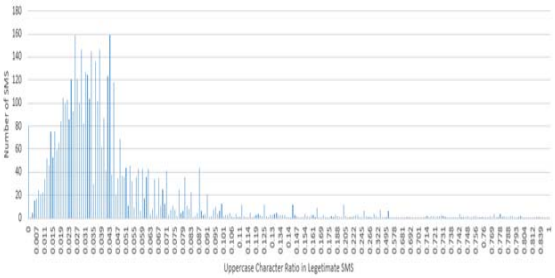

Figure 8: Uppercase Character Ratio in Legitimate SMS

Heuristic 4: Non Alphanumeric Character Ratio

SMS contains non-alphanumeric characters, computing non-alphanumeric character ratio considered as one of the features for identifying spam and legitimate SMSs. The distribution of non-alphanumeric character ratio for spam SMS is plotted in figure 9. The average value for non-alphanumeric ratio (∂) in the spam SMS is found to be greater than 0.04. The distribution of non-alphanumeric ratio for legitimate SMS is plotted in figure 10. The average value for non-alphanumeric character ratio (∂) in the legitimate SMS is found to be 0.068. the heuristic defined as follows:

H4= If (non-alphanumeric ratio) >∂→ Spam SMS
Else Legitimate SMS

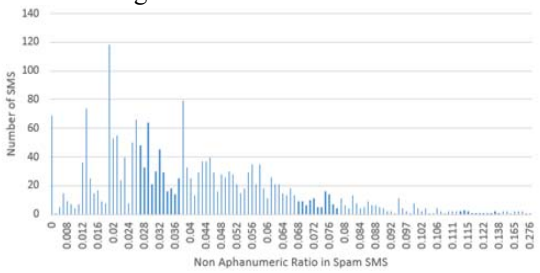

Figure 9: Non Alphanumeric Ratio in Spam SMS

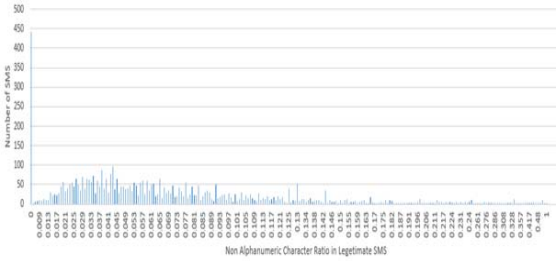

Figure 10: Non Alphanumeric Ratio in Legitimate SMS

Heuristic 5: Numeric Character Ratio

Numeric characters ratio is computed from the SMS. The distribution of numeric character ratio for spam SMS is plotted in Figure 11. The mean value for numeric character ratio (Ω) in the spam SMS is found to be greater than or equal 0.01. The distribution of numeric ratio for legitimate SMS is plotted in figure 12. The mean value for numeric character ratio (Ω) in the legitimate SMS is found to be 0.004. the heuristic defined as follows:

H5= If (numeric character ratio) >= Ω → Spam SMS
Else, Legitimate SMS

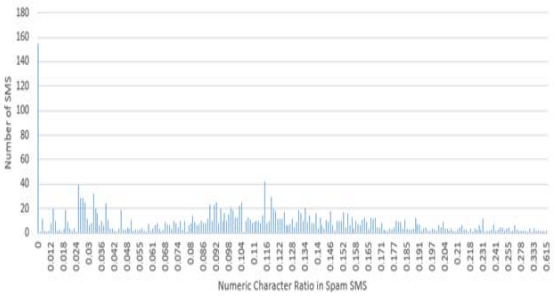

Figure 11: Numeric Character Ratio in Spam SMS

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

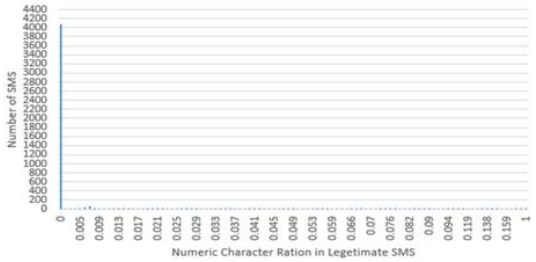

Figure 12: Numeric Character Ratio in Legitimate SMS

Heuristic 6: Existence of URL

In the dataset, the existence of URL in SMS is examined. It was found that 9.5 % of spam SMS contained URL. And 0.04% of the legitimate SMS contained URL. Figure 13 shows the number of URLs in spam SMSs and figure 14 shows the number of URLs in legitimate SMSs Therefore, the heuristic is defined as

H6= If (URL in SMS) → Spam SMS

Else, Legitimate SMS

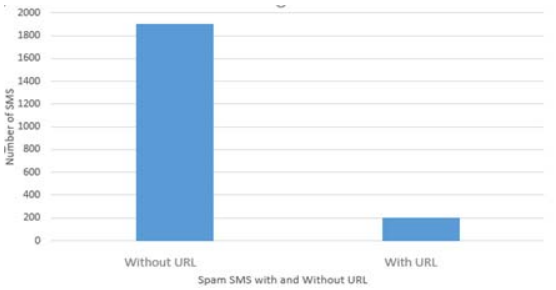

Figure 13: Number of SMSs With and Without URLs in Spam Dataset

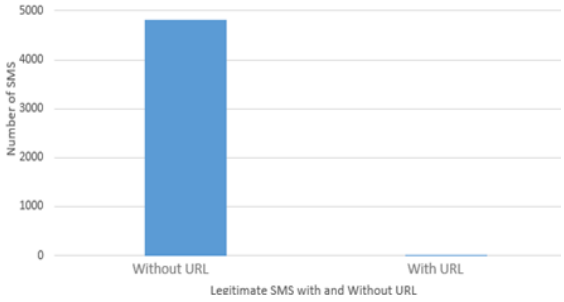

Figure 14: Number of SMSs With and Without URLs in Legitimate Dataset

4.3 Training Phase

The dataset is divided into training and testing samples. The number of training samples is 4200 legitimate and spam SMSs. Which are used to generate the rules for classifying SMSs using Apriori association rule mining algorithm.

4.3.1 Apply apriori algorithm

The mechanism to classify SMS is generated using Apriori algorithm in which the different heuristics are used to convert SMS structural features to binary to be used by the Apriori algorithm. The generated rules are used to check the SMS before it sent to the user's mobile. Rules are extracted from the SMSs and collected over 4200 SMSs. Applying Apriori algorithm and rules generation were performed using WEKA 3.8 data mining tool which consists of different machine learning algorithms.

4.3.2 Rules generation

Rules are generated in the form of "if–then" statements depending on the features values. the hidden relationship among the features is found by association rule mining. In the proposed work, the frequently occurring features in the training dataset are detected by association rule mining to discover the associations and correlations among features in the dataset to gain insight into which features are frequently appear together in spam SMS. Each feature has a Boolean variable representing the presence or absence of that feature. Each SMS can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed that reflect features that are frequently associated together. These patterns can be represented in the form of association rules. All the features in the dataset are binary attributes with the class of each SMS and the Apriori algorithm used for identifying the recurring patterns. Different values for minimum support and minimum confidences are used to generate the rules and the rules which achieved best accuracy on the test samples were selected to be used in our system. Figure (15) shows pseudocode for the Apriori algorithm and its related procedures, to discover frequent itemsets for mining Boolean association rules. Step 1 of Apriori finds the frequent 1-itemsets, L1. In steps 2 through 10, Lk−1 is used to generate candidates Ck to find Lk for k≥2. The apriori_gen procedure generates the candidates and then uses the Apriori property to eliminate those having a subset that is not frequent (step 3). Once all of the candidates have been generated, the database is scanned (step 4). For each transaction, a subset function is used to find all subsets of the transaction that are candidates (step 5), and the count for each of these candidates is accumulated (steps 6 and 7). Finally, all the candidates satisfying the minimum support (step 9) form the set of frequent itemsets, L (step 11). A procedure can then be called to generate association rules from the frequent itemsets. The apriori_gen procedure performs two kinds of actions, namely, join and prune. In the join component, Lk−1 is joined with Lk−1 to generate potential candidates (steps 1–4). The prune component (steps 5–7) employs the Apriori property to remove candidates that have a

Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

subset that is not frequent. The test for infrequent subsets is shown in procedure has infrequent subset [23].

Input:

- D, a data set,
- min_sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

```
(1)  L1 = find_frequent_1-itemsets(D);
(2)  for (k = 2; Lk−1 ≠ φ; k++) {
(3)      Ck = apriori_gen(Lk−1);
(4)      for each transaction t ∈ D { // scan D for counts
(5)          Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)          for each candidate c ∈ Ct
(7)              c.count++;
(8)      }
(9)      Lk = {c ∈ Ck|c.count ≥ min_sup}
(10) }
(11) return L = ∪k Lk;

procedure apriori_gen(Lk−1:frequent (k − 1)-itemsets)
(1)  for each itemset l1 ∈ Lk−1
(2)      for each itemset l2 ∈ Lk−1
(3)          if (l1[1] = l2[1]) ∧ (l1[2] = l2[2])
                 ∧... ∧ (l1[k − 2] = l2[k − 2]) ∧ (l1[k − 1] < l2[k − 1]) then {
(4)              c = l1 ⋈ l2; // join step: generate candidates
(5)              if has_infrequent_subset(c, Lk−1) then
(6)                  delete c; // prune step: remove unfruitful candidate
(7)              else add c to Ck;
(8)          }
(9)  return Ck;

procedure has_infrequent_subset(c: candidate k-itemset;
         Lk−1: frequent (k − 1)-itemsets); // use prior knowledge
(1)  for each (k − 1)-subset s of c
(2)      if s ∉ Lk−1 then
(3)          return TRUE;
(4)  return FALSE;
```

Figure 15: Apriori Algorithm [23]

Table 3 shows the rules generated for spam SMS and provide higher accuracy for classifying spam and legitimate SMSs achieved by using minimum support value 0.2 and minimum confidence 0.8. Figure 16 shows the number of rules generated for spam SMS training set with different minimum support and minimum confidence values.

Table 3: Generated Rules by Apriori Algorithm for Spam SMS

Rule No.	Rule	Confidence
1	message-length >= 136, numeric-ratio> 0.01 ==> class= spam	0.94
2	message-length>= 136, number-terms> 23, numeric-ratio> 0.01 ==> class= spam	0.94
3	number-terms> 23, numeric-ratio> 0.01==> class=spam	0.91
4	non-alphanumeric> 0.04, numeric-ratio> 0.01==> class=spam	0.89
5	uppercase-ratio> 0.05 ==> class=spam	0.86
6	numeric-ratio> 0.01==> class=spam	0.85
7	number-terms> 23, non-alphanumeric> 0.04 ==> class=spam	0.83

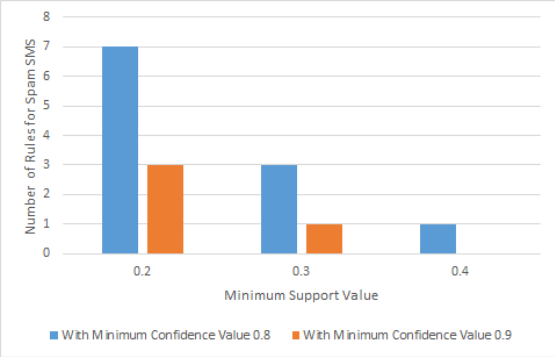

Figure 16: Number of rules for spam SMS training dataset

4.4 Testing Phase

For Test Phase, 2727 legitimate and spam SMSs are used as test dataset. By applying the generated rules from training phase on test dataset. specificity, recall and accuracy are computed.

5. RESULTS

The algorithm generated different number of rules and the confusion matrix for different minimum support and minimum confidence is computed which is shown in table 4. Based on the test results, the true positives TP (number of spam SMS messages identified as spam), the false positives FP (number of legitimate SMS messages identified as spam), true negatives TN (number of legitimate SMS messages identifies as legitimate), and the false negatives FN (number of spam SMS messages identified as legitimate) are computed. Also different precision, recall and accuracy values using different valued for minimum support and minimum confidence are computed and the result is shown in table 5. The rules which were generated by Apriori algorithm with minimum support value 0.2 and minimum confidence value 0.8 achieved higher accuracy with 97.65% and they are selected to be used in the proposed system to classify spam and legitimate SMSs. the following metrics were deduced: precision, recall and accuracy in equations (3), (4), and (5) respectively. Figure 17 shows the accuracy values when the minimum confidence is 0.9. Figure 18 shows the accuracy values when the minimum confidence is 0.8.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{Tp}{Tp+FN} \quad (4)$$

$$Accuracy = \frac{Tp+TN}{Tp+TN+FP+FN} \quad (5)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: The Confusion Matrix

Minimum Support	Minimum Confidence	FP	FN	TP	TN
0.2	0.9	0.016	0.301	0.7	0.984
0.2	0.8	0.006	0.041	0.959	0.994
0.3	0.9	0.024	0.4	0.601	0.976
0.3	0.8	0.016	0.053	0.947	0.984
0.4	0.9	0.117	0.043	0.957	0.883
0.4	0.8	0.117	0.043	0.957	0.883

Table 5: The Values of Precision, Recall and Accuracy

Min. Support	Min. Confidence	Precisi-on	Recall	Accuracy
0.2	0.9	0.9776	0.6995	0.8418
0.2	0.8	0.9942	0.9586	0.9765
0.3	0.9	0.9613	0.6005	0.7882
0.3	0.8	0.9834	0.9471	0.9656
0.4	0.9	0.891	0.9567	0.9198
0.4	0.8	0.891	0.9567	0.9198

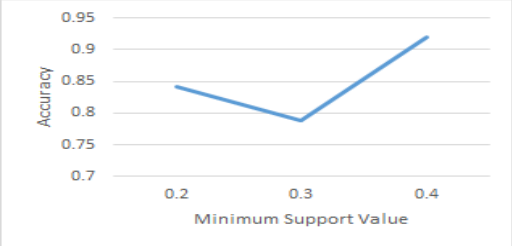

Figure 17: Accuracy with Minimum Confidence 0.9

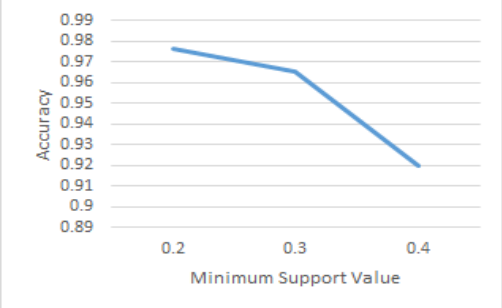

Figure 18: Accuracy with Minimum Confidence 0.8

6. DISCUSSION

In order to demonstrate the effectiveness of the proposed system, the system result was compared with some of the previous work in this field. The results show that using just 6 structural features achieved good accuracy which is 97.65% using rules generated by Apriori association rule mining algorithm with minimum support 0.2 and minimum confidence 0.8. the importance of using just structural features was to reduce features extraction time compared to the use of bag of works extracted from SMS messages. Most researchers used bag of words extracted from SMS messages such as in [3], [9], [11], [12], [15] and [17] achieved accuracy between 68% and 98% depends on the algorithm which they used. So according the results from the previous works compared to the work in this paper, the proposed system achieved good results with less number of features and less processing time for feature extraction process using Apriori association rule mining algorithm.

7. CONCLUSION

In this paper, a new system is proposed to detect and classify SMSs into to spam and legitimate using Apriori association rule algorithm which is used to discover the patterns among the SMS structural features. Two publicly available datasets are used. the proposed scheme is evaluated with different values for minimum support and minimum confidence which are used to generate the rules. The rules are tested on test dataset. the accuracy was computed. The rules which achieved higher accuracy where selected to be used on the proposed system. Higher accuracy 97.65% achieved using rules generated with minimum support 0.2 and minimum confidence 0.8 and less number of features. According to the results, this approach achieved good accuracy based on extracting only six structural features from the SMS instead of textural features which requires many preprocessing steps for tokens extraction in which require more time for feature extraction step to detect spam SMS. At this point and based on the results, structural features are considered enough to detect spam SMS messages with low false positive rate. For Future work other association rules algorithms will be used and compared with the results of Apriori algorithm since apriori algorithm is considered slow and requires many database scans. The proposed system as a future work will also be applied to detect spam messages sent through Instant Messaging applications (e.g., WhatsApp, Viber...etc.), because they have similar properties and different techniques and algorithms will be used.

REFRENCES:

[1] "Short Message Service Security", The Government of the Hong Kong Special Administrative Region, 2008.

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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ELEKTRONIKA IR ELEKTROTECHNIKA Vol. 19, No. 5, 3013.

[3] Mohamed El Boujnouni, "SMS Spam Filtering Using N-gram Method, Information Gain Metric and an Improved Version of SVDD Classifier", *Journal of Engineering Science and Technology Review*, Vol. 10, No. 1, 2017, pp. 131- 137.

[4] https://blog.truecaller.com/2017/04/19/truecaller-us-spam-report-2017/
Truecaller Insights Special Report: An Estimated 22.1M Americans Lost $9.5B in Phone Scams Last Year, 2017

[5] Gordon V. Cormack, José María Gómez, and Enrique Puertas Sánz, "Spam Filtering for Short Messages", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* Lisbon, Portugal, November 06 - 10, 2007, pp. 313-320.

[6] Iosif Androulidakis, Vasileios Vlachos, and Alexandros Papanikolaou, "FIMESS: FIltering Mobile External SMS Spam", Proceedings *of the 6th Balkan Conference in Informatics*, Thessaloniki, Greece, September 19 - 21, 2013, pp. 221-227.

[7] Tiago A. Almeida, José María Gómez, and Akebo Yamakami, "Contributions to the Study of SMS Spam Filtering: New Collection and Results", *Proceedings of the 11th ACM symposium on Document engineering*, Mountain View, California, USA, September 19 - 22, 2011, pp. 259-262.

[8] M. Taufiq Nuruzzaman, Changmoo Lee, and Deokjai Choi, "Independent and Personal SMS Spam Filtering", *IEEE International Conference on Computer and Information Technology*, Pafos, Cyprus, 31 Aug.-2 Sept. 2011, pp. 429–435.

[9] Muhammad Zubair Rafique and Muhammad Abulaish, "Graph-Based Learning Model for Detection of SMS Spam on Smart Phones", *Wireless Communications and Mobile Computing Conference (IWCMC)*, Limassol, Cyprus , Aug 27-31, 2012.

[10] Anchal and Abhilash Sharma, "SMS Spam Detection Scheme Using Custom Neural Network with ICA", *Global Journal of Advanced Engineering Technologies*, Vol. 3, No. 3, 2014, pp. 269-273.

[11] Ishtiaq Ahmed, Donghai Guan, and Tae Choong 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.

[12] Sin-Eon Kim, Jung-Tae Jo, and Sang-Hyun Choi, "SMS Spam Filtering Using Keyword Frequency Ratio", *International Journal of Security and Its Applications*, Vol. 9, No. 1, 2015, pp. 329-336.

[13] Kamahazira Zainal, and Mohd Zalisham Jali, "A Review of Feature Extraction Optimization in SMS Spam Messages Classification", *SCDS 2016, CCIS 652*, 2016, pp. 158-170.

[14] Tiago A. Almeida, Tiago P. Silva, Igor Santos, and José M. Gómez Hidalgo, "Text Normalization and Semantic Indexing to Enhance Instant Messaging and SMS Spam Filtering", *Knowledge-Based Systems*, 2016, 108: 25-32.

[15] Dima Suleimana, and Ghazi Al-Naymata, "SMS Spam Detection using H2O Framework", *the 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, Procedia Computer Science 113*, 2017, pp. 154–161.

[16] Naresh Kumar Nagwani , " A Bi-Level Text Classification Approach for SMS Spam Filtering and Identifying Priority Messages ", *The International Arab Journal of Information Technology*, Vol. 14, No. 4, July 2017.

[17] Neelam Choudhary, and Ankit Kumar Jain "Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique", *ICAICR 2017, CCIS 712*, Springer Nature Singapore Pte Ltd, 2017, pp. 18–30.

[18] L. Novak and M. Svensson, "MMS- Building on the success of SMS", *Ericsson Rev*, Vol. 78, No. 3, 2001, pp. 102-109.

[19] F. Hillebrand, F. Trosby, K. Holley and I. Harris, "Short Message Service (SMS), the Creation of Personal Global Text Messaging", *John Wiley and Sons Publication*, 2010.

[20] Alaa Al Deen Mustafa Nofal, and Sulieman Bani-Ahmad, "Classification Based on Association-Rule Mining Techniques: A General Survey and Empirical Comparative Evaluation", *Ubiquitous Computing and Communication Journal*, Vol. 5, No. 3, 2011.

[21] Trupti A. Kumbhare, and Santosh V. Chobe "An Overview of Association Rule Mining Algorithms", *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 1, 2014, pp. 927-930.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[22] Bing Liu, Yiming Ma, and Ching-Kian Wong, "Classification Using Association Rules: Weaknesses and Enhancements", *Data Mining for Scientific and Engineering Applications*, Vol. 2, 2001 pp. 591-605.

[23] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining Concepts and Techniques", Third Edition", Elsevier, 2012.