International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-4, Issue-1, January 2018 Pages 94-98

Building Bi-lingual Anti-Spam SMS Filter

Heba Adel ,Dr. Maha A. Bayati

Abstract— **Short Messages Service (SMS) is one of the most popular telecommunication service packages that is used permanently due to its affordability and do not need the internet service. The growth of using SMS leads to the increase of SMS spam problem. So, SMS spam filter become a goal of many organizations to deal with those spams. This work proposes a spam classifications approach using "Naïve Bayesian" (NB) bi-lingual classifier. Based on the content; body of short messages, this classifier categorize input English/Arabic (E/A) messages as being Ham (legitimate) or Spam (unsolicited). As is the tradition, each message's body is represented by as set of features. These features are to be extracted from E/A SMS provided by certain datasets. The proposed filter was exterminated to measure it's efficiency under different settings of working permeates. For English SMS dataset, a total of 5574 SMS were considered; 70% for training and 30% for testing. For a total of 15-featuers, extracted from each SMS, an accuracy of 93% was achieved. For Arabic SMS, a total of 400 SMS were considered and under the same specifications for the English SMS, an accuracy 85% was reached. Using features selection, accuracy level was raised up to 95% for English SMS and 88% for Arabic SMS.**

Index Terms— **Bi-lingual , Anti-Spam, SMS**

I. Introduction

Mobile phone spam messages are form (unsolicited messages, especially advertising), directed at the text messaging . In other words SMS spam can be defined as unwanted or any junk text messages which is received on a mobile phone device. The SMS Spam problem is increasing daily with the increase in the use of text short messaging (SMS).The filtering mechanism available focuses firstly on email spam and it is one of the oldest problem but with the popularity of mobile phones, SMS spam is the one of the main issue these days [1]. The similarity of SMS spam filtering to email spam filtering indicate that certain technologies in email spam filter may be useful in struggle SMS spam. The content-based technologies used in email spam1 filtering that are candidate for SMS spam filtering contain both direct content filtering1 and collaborative content filtering techniques| [2]. This paper proposes system which classifies Bi-lingual SMS (E/A) in two categories spam or ham using Naïve Bayes (NB) algorithm.

Heba Adel ,Computer Science Dept., College of Science, Mustansiriyah University, Baghdad , Iraq
Dr. Maha A. Bayati, Computer Science Dept., College of Science, Mustansiriyah University, Baghdad-Iraq

II. RELATED WORK

[3]The researchers analyzed the possibility of using Bayesian filtering techniques used in blocking email spams in detecting and stopping mobile spams. Accordingly, two well-sized SMS spam test collections were built; one in English and the other was in Spanish. A number of messages were tested using these two techniques, which involve using representations and Machine Learning algorithms. Results have shown that Bayesian filtering techniques can effectively be transferred from email to SMS spams. [4], the researcher proposed an anti-spam technique based on the Artificial Immune System (AIS) for the purpose of filtering SMS spam messages. The proposed technique uses a set of input features t to spam detection model. The idea of this technique is based on classifying a message using trained dataset. The latter can be in the form of Phone Numbers, Spam Words, and Detectors. It uses a double collection of bulk SMS messages, ie., the Spam and Ham during the training process. Such a dataset can be built following a number of stages, such as: tokenizer, stop word filter, and training process. The study was experimental by nature; it was conducted on the iPhone Operating System (iOS). Results revealed that the proposed system could accurately classify the SMS spam and ham in comparison to the Naïve Bayesian algorithm. [5], a hybrid system was suggested for the purpose of classifying and detecting spam or ham SMS, using the Navïe Bayes classifier and an Apriori algorithm. The technique was characterized by being fully logic; its performance relied on the statistical character of the database. As a classifier, Navïe Bayes represents one of the most effective and significant1 learning algorithms in machine learning and data mining. It further represents one of the basic techniques invested in information retrieval. However, when user-specified minimum support and minimum confidence were used, a significant improvement and an effective accuracy, 98.7%, was noticed in comparison to the traditional Naive Bayes approach, which was 97.4% when conducted on UCI Data Repository. [6], an algorithm, called FIMESS (Filtering Mobile External SMS Spam) was suggested. It is characterized by its simple performance, and effective way of checking the message headers. Accordingly, the algorithm managed to SMS into its respective types, spam or ham. Another characteristic of this algorithm is that, FIMESS is able to invest the important information available in the SMS headers to be later used in SMS spam messages identification process. Away from the email metadata, which is said to be easily manipulated by the spammers, the SMS protocol helps supply useful information that can be invested in efficiently filtering SMS spams. The proposed scheme was examined on an Android platform, resulting in many encouraging results. [7], a group of researchers proposed a new method that involves a network for online SMS spam messages detection. This method

implies robust text signatures that are used for the identification of excessively sent SMS with a similar platform. This method was characterized as being robust against any slight modifications in SMS spam messages. It further implied utilizing a fast online algorithm that can be invested in a large number of carrier networks to detect spam activities before their delivery in large quantities.. This method does not save SMS contents; accordingly it maintains no privacy of mobile subscribers. [8], another work was conducted to examine the impact of some feature extraction and detection on filtering (SMS) spams using two different languages, namely Turkish and English. The process of filtering implied using some features originated from the bag-of-words (Bow) model. It further contained an ensemble of structural features (SF) to help solve the spam problem. Using information theoretic feature selection methods, the researchers were able to identify the distinctive Bow features. Various combinations of the Bow and SF1 were then fed used within the pattern classification algorithms to classify SMS messages. The filtering framework was then assessed using datasets from both Turkish and English, including the first publicly available Turkish SMS message. Comprehensively experimenting the respective datasets showed that the combinations of Bow and SFs provide a better classification performance on both of the datasets being analyzed. However, the impact of the feature selection methods being used was slightly different in each of these languages. [9], researchers proposed several solutions for the filtering and detecting SMS spams. They started by critically reviewing the available methods, challenges and future research recommendations on spam detection techniques, filtering and mitigation of mobile SMS spams. The highly famous techniques for SMS spam detection, filtering and mitigation were compared, shedding light on their datasets, findings and limitations Results showed that the majority of these studies developed a taxonomy to help solve the problem of SMS spams. Besides, those studies were based on the support vector machine and the Bayesian network when constructing SMS spam classifiers. [10], content-based filtering represented the highly invested technique in determining the type of the spams whether they are spam or ham. This was because such a type of process is characterized by being very dynamic and very challenging at the same time, and constantly changing when representing all information mathematically. . Naïve Bayes method changes the nature of a message using probability theory and support vector machine (SVM). These two classificatory methods are said to be efficient in different domains. As for Nepali SMS or Text classification have not yet considered comprehensively. Accordingly, it is highly recommended to examine their performance during the process of Spams classification. To that reason, the researchers used the Naïve Bayes and SVM-based classification techniques in the classification of Nepali SMS into Spam and non-Spam. Various texts were empirically analyzed to evaluate the classification accuracy of the methodologies being used. The result was SVM was 87.15% accurate whereas Naïve Bayes was 92.74% accurate, studies on SMS spam filtering and development were reviewed. Besides, a large amount of SMS data was collected and analyzed. The study represents the state of the art in SMS spam filtering; it reviewed various approaches in this regard using different datasets. Result revealed that the supervised learning algorithms are highly efficient in SMS spam classification; their accuracies reached up to 97%.

III. ANTI-SPAM TECHNIQUE

To cope with spam, several popular techniques can be used such as the following [11]:-

- **White and black listing:** The sender who is blacklisted is considered spammer, and his/her messages will be blocked. On the contrary, the messages that are sent from the sender's white list (e.g. the address book, contact list) represent legitimate and transferable.
- **Collaborative filtering**: This type of filtering is called social1 filtering. It filters the information based on people's recommendations. It is concept-based filter; that is, a message is tagged as a spam, it will be so to all other similar users where e is more than N=20 recipient) [12, 11].
- **Content-based filtering**: This is the highly invested approach where the spam features of each message are searched by words, as in: "free", "viagra", etc., or by the unfamiliar distribution of punctuation1 marks and capital letters, as in: in "BUY!!!!!!", etc. Content-based filtering represents one of the highly used approaches in detecting SMS spams. the process of detection is based on a set of attributes that help determine whether the message is ham or spam [11, 13].

Despite the fact that there are many approaches that can be used in spam filtering, content-based filtering, namely the Bayesian filtering, plays an important role in reducing spam messages [11]

IV. THE CONCEPT OF TEXT CLASSIFICATION

Use To more make things more clear about content-based filtering technique one needs to introduce the concept of text classification, it define one or more classes according to their contents. The text classification systems take care of all preprocessing tasks (tokenization, stop words removal and stemming). After preprocessing the text, classification systems proceeds by extracting features from that texts, and finally apply one of the machine learning algorithms to undergo the classification task. To improve the results of such systems one can use one or more features selection methods [14].

Naive Bayesian (NB)

It represents one of the highly famous machine learning algorithms. It is based on Bayes' theorem together with some independent assumptions between predictors [5]

- **The Bayes theorem:**
- Bayes theorem helps finds the probability of a hypothesis where the event Y gives the observed training data, as shown in the following equation:

$$P\ (Y|X) = \frac{P(X|Y)P(X)}{P(X)}$$

…………..Equ.1

This simple formula has been invested practically in many applications. This is because it easily finds the probabilities, P(X | Y), P(Y), P(X) when required. This theorem is the basics of Bayesian statistics probability of a new event by examining its earlier probability estimates that are derived empirically from the data. The following section explains the various ways the Bayesian statistics conducts the statistical analyses [15]

NB CLASSIFIERS

This classifier is one of the supervised text learning algorithms that is used in the process of spam filtering. The NB classifier represents each of the patterns X (SMS), as vector of feature values. Here "($f_1, f_2, \cdots \cdots f_n$) "is the feature vector of X, and C = (C_1, C_2) is the given to two classes. It is necessary to solve the probabilities (P_1, P_2) of X belong in to class C1 and C2, and the class corresponding to max(P_1, P_2) would be the desired one. So the problem can be defined as following equation :

$$C_X = \arg max \frac{P(f1, f2,,,, fn)|C_j)P(C_j)}{P(X)} \text{ .... (Equ.2)}$$

P(X) is usually regarded as a constant and has no any impact on the solving of maximum value, so (Equ.2) is1 equal to (Equ.3):

$$C_X = \arg\max P(f1, f2,,,, fn)|C_j)P(C_j) \text{.......... (Equ.3)}$$

The Naïve Bayesian Model to simplify the calculation of (Equ.3), that is, the features in vector X are independent of each other, so (Equ.3) can be further defined as follows :

$$C_X = \arg\max \prod_i^n P(x_i | C_j)P(C_j) \text{ .... (Equ.4)}$$

Note that Equ.11 includes two types of parameters, $P(C_j)$ and $P(x_i | C_j)$, there parameter are defined as in (Equ.5) and (Equ.6) , respectively. The two parameters can be derived from the training data [26].

$$P(C_j) = \frac{\text{the number of comments in } C_j}{\text{the number of all comments}} \text{ .... (Equ.5)}$$

$$P(x_i | C_j) = \frac{\text{the number of feature } x_i \text{ appearing in } C_j}{\text{the total number of all features appearing in } C_j} \text{...(Equ.6)}$$

Building Bi-lingual SMS filter:

This part presents the design of the proposed system for filtering English as well as Arabic short messages .System's workflows the following basic:

- **Reading SMS (English and Arabic Datasets)**

Considering the UCI machine learnings repository dataset, The "SMS Spam Collection dataset" is used by this system. It is a set of English SMS in .XML file format. It contains a total of 5,574 partitioned into 4,827 ham SMS and 747 spam ones.

This dataset is first manipulated by saving it in a data file where each SMS is separated by an "ID" and is identified by a corresponding text body and H/S label.

As about Arabic SMS, no resource is found to provide for Arabic SMS dataset. Hence, Arabic dataset was collected manually in a way or another, and was set to act upon as for English dataset.

- **Preprocessing phase**

To facilitate for filtering SMS, it is necessary to preprocess messages thought the following steps:

i. Tokenization: - It breaks the body down into words, hence cleaning up all white space.

ii. Stop words removal: - the stop words (English/Arabic) are some sort common words, those provide no useful information to help deciding the class of some SMS.SO it prefer to remove form text SMS.

iii. Stemming:- Unlike emails, SMS cannot be suitably undergone stemming for two reasons :

1- Short messages are almost written in local languages where abbreviation are frequently used. For example:

"U dun say so early ... U c already then say......"

2- No BOW is used here, only a limited number of spam words (List in table (3) and table (4) in Appendix) is considered, hence it is not worthier do stemming.

- **Extraction Feature phase**

In this phase extract fifteen features are extracted from the body of each SMS. It is worth to mention that these features were hardly determined suite both English and Arabic text bodies due to lack of recourses concerning Arabic messages. Table below illustrate the fifteen features that is used.

No	Feature name	Description
F1	Message length	Number of all characters
F2	Number of words	Number of words obtained using alphanumeric tokenization
F3	Uppercase character Ratio	Number of uppercase characters normalized by the message length
F4	Non-alphanumeric character ratio	Number of non-alphanumeric characters normalized by the message length
F5	Numeric character Ratio	Number of numeric characters normalized by the message length
F6	Presence of URL	Presence of "http" and/or "www" Terms
F7	Spam words	The number of spam words
F8	Abbreviations	Number of abbreviations
F9	Number of Non-alphanumer	Number of non-alphanumeric characters

	ic character	
F10	Uppercase words	Number of uppercase words
F11	Uppercase words ratio	Number of uppercase words normalized by the message length
F12	Words ratio	Number of words obtained using alphanumeric tokenization normalized by the message length
F13	Country	Number of Countries
F14	Digits Ratio	Number of digits normalized by the message length
F15	Trade Markets	Number of trade Markets

- **Normalization phase**

Just after extracting fifteen features from each SMS body, time to apply normalization to reduce the variance of values between those.

- **Features Selection phase**

Features selection

To improve classification accuracy, one should use one or more features selection methods

This approach statistically allocate a scoring to each feature to be later ranked accordingly. Then, the features will either be selected to be saved or ignored from the dataset. Figure (2-7) illustrates the Filter model in question. Various methods have been used for Filter selection, Figure (2-7) depicts this adopted model. Cases of the most commonly used ones include the following: [17] [18]

A.Term Frequency(TF) :- This method simply calculate the number each features appeared in a given text. Being in department of certain class, TF may be calculated over the entire test set as well. Selecting frequent terms will improve the chances that the features will be presented in future test cases (Equ.14) present the formula used to find TF [19].

$$\text{TF}=\sum_{1}^{C}\frac{H}{N}*F+\frac{S}{N}*F \quad \text{........ (Equ.7)}$$

Where:" N is the number of all data, H is the number "of ham, S is the number of Spam, F is the number that appears in certain class.

Information Gain (IG):- The basic idea behind this method is to find out how well each single features separates the given data set. Entropy of an information is used to measure the suspicion of a features in the dataset [20].However the entropy of Y is :

$$\text{H(Y)} = -\sum_{y\epsilon Y} p(Y)\log_2 p(Y) \quad \text{.........(Equ.8)}$$

Where p(Y) is the density function of the marginal probability for the variable "Y", which is a random number. If the values of "Y" obtained in the training data set "S" were divided according to the second feature "X" values, and the entropy of "Y" with regard to "X" had divisions that were less than "Y" prior to partitioning entropy, then there exists a relationship between "Y" and "X" features. Accordingly, the resulting entropy of "Y" after noticing "X" is

$$\text{H(Y|X)} = -\sum_{x\epsilon X} p(X)\sum_{y\epsilon Y} p(Y|X)\log_2 p(Y|X) \quad \text{.......(Equ.9)}$$

Where p(Y|X) represents the "y" that was given the conditional probability of "x". Using Entropy as impurity criterion for the training set "S", one can examine the measure that gives extra information about "Y" provided by "X". Thus, such a measure indicates the amount of decrease of the entropy of "Y". Such a measure is also called IG, as illustrated in the following equation:

$$\text{IG} = \text{H(Y)} - \text{H(Y|X)} = \text{H(X)} - \text{H(X|Y)} \quad \text{.........(Equ.10)}$$

Where: IG refers to the symmetrical measure, which gains the information about "Y" once the latter is observed to be equal to "X", and the reverse is true. The criterion weakness of IG is biased towards the features with more values regardless whether they values are informative or not [21].

C."Gain Ratio"(GR):- This is an adjustment of IG that reduces its bias. GR takes the number and size of section into account when choosing a features. It corrects the IG by taking the intrinsic information of a divided into account. Intrinsic information is the entropy of distribution of instances into sections (i.e. how much info do one needs to tell which section an instance belongs to). Value of features reduction as intrinsic information gets larger [22] . present the formula used to find GR of Certain features.

$$\text{GR(feature)} = \frac{\textit{Gain (featuer)}}{\textbf{intrinsic info.(featuer)}} \quad \text{.......(Equ.11)}$$

- **Classification phase**

Naïve Bayesian Classifier

This is a classification technique belong to the family of probability algorithms. It is based on Bay's theorem of conditional independently and is taking advantage of probability theory to predict the category of certain sample. In this scene, this work present an NB text classifier whose aim is to categorize E/A SMS as by Spam or Ham, on the basis of E/A dataset of sample SMS.

NB classification goes through the two phase of training and testing where operate on the feature vector, the list of feature gained upon stepping theory the feature engineering process.

1. Training phase :

This phase attempts to calculate the following probabilities:

a. Calculate the probability of Spam SMS class and Ham SMS class to the total number of SMS sample, using (Equ.5),

$$P\,(C_i) = \frac{|C_i|}{N}$$

Where: C_i = class type which is either spam or ham, N = total number of SMS.

a. Calculate the P (C_i) probability of certain sample SMS (X) being in either of two classes .This is done in terms of calculating probability of occurrence of individual feature x_j in either class , using (Equ.13), as shown below:

$$P\,(X|C_i) = \prod_{j=1}^{n} P(x_j|C_i)$$

Where x_j the element of feature (X) that is found (n) times in spam or ham, C_i is class type which is either spam or ham.

2. Testing phase :

Using the probabilities gained from the training phase, testing phase operation on SMS sample in the testing set as follows:

a. Calculate probabilities of each SMS sample X (in terms of each value for each features x) for both class, on the basis of probabilities from training phase.

b. If certain value for some features x , does not show , for X , during the training phase, then set that value to the average probability of two closest features values of x.

c. For each sample X, find the posterior probabilities using *"Bayes theorem"* (Equ.1) as bellow:

ii. $P\,(C|X) = \dfrac{P(X\,|c\,)P(X)}{P(X)}$

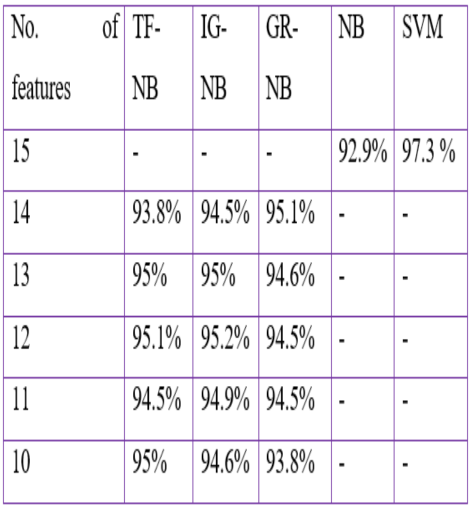
d. Decided the class of X based on result from c. The class would be the one with largest probability for X.

V. Results and Conclusion

A The proposed filter was exterminated to measure it's efficiency under different settings of working permeates. For English SMS dataset, a total of 5574 SMS were considered; 70% for training and 30% for testing. For a total of 15-featuers, extracted from each SMS, an accuracy of 93% was achieved. For Arabic SMS, a total of 400 SMS were considered and under the same specifications for the English SMS, an accuracy 85% was reached. Using features selection, accuracy level was raised up to 95% for English SMS and 88% for Arabic SMS.

For this relatively low accuracy of Arabic SMS, an LibSVM was set to operate properly with Arabic as well as English SMS. Results showed that SVM classifier behaved similarly with both language. It gave 91% accuracy level for Arabic SMS and 97% for English ones. The bellow tables shown the accuracy results1 for both English and Arabic set

(a) English set



No. of features	TF-NB	IG-NB	GR-NB	NB	SVM
15	-	-	-	92.9%	97.3 %
14	93.8%	94.5%	95.1%	-	-
13	95%	95%	94.6%	-	-
12	95.1%	95.2%	94.5%	-	-
11	94.5%	94.9%	94.5%	-	-
10	95%	94.6%	93.8%	-	-

References

[1] A. K. J. Neelam Choudhary, "Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique," 2017.

[2] M. B. ,. D. G. Sarah Jane Delany, "SMS spam filtering: Methods and Data," researchgate, 2013.

[3] José María Gómez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sánz, "Content Based SMS Spam Filtering".

[4] A. M. M. Tarek M Mahmoud1, "SMS Spam Filtering Technique Based on Artificial Immune System," 2010.

[5] D. G. C. C. Ishtiaq Ahmed, ""SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Item set"," 2014.

[6] V. V. ,. A. P. Iosif Androulidakis, "FIMESS: Filtering Mobile External SMS Spam," 2013.

[7] P. G. Baris Coskun, " Mitigating SMS Spam by Online Detection of Repetitive Near-Duplicate Messages".

[8] A. K. U. S. G. S. E. E. S. Gunal, The Impact of Feature Extraction and Selection, 2013.

[9] M. S. A. L. H. C. O. O. G. A.-S. A. I. A. a. T. H. Shafi'i Muhammad Abdulhamid, "A Review on Mobile SMS Spam Filtering Techniques," IEEE, 2016.

[10] A. Y. Tej Bahadur Shahi, "Mobile SMS Spam Filtering for Nepali Text Using Naïve Bayesian and Support Vector Machine," International Journal of Intelligence Science, 2014.

[11] G. C. B. P. S. José María Gómez Hidalgo, "Content Based SMS Spam Filtering," jmgomez, 2013.

[12] "Collaborative filtering," recommender-system, 2012.

[13] "Content-based SMS Spam Filtering based on the Scaled Conjugate Gradient Backpropagation Algorithm," International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2015.

[14] "https://www.meaningcloud.com/developer/text-classification/doc/1.1/what-is-text-classification," [Online].

[15] S. Y. Liu Xing, "An Adaptive Spam Filter Based on Bayesian," School of Computer Science and Technology.

[16] S. Y. Liu Xing, "An Adaptive Spam Filter Based on Bayesian," School of Computer Science and Technology.

[17] S. KAUSHIK, "Introduction to Feature Selection methods with an example (or how to select the right variables?)," Analytics Vidhya, 2016.

[18] J. Brownlee, "An Introduction to Feature Selection," machine learning mastery, 2014.