International Journal of Civil Engineering and Technology (IJCIET)
Volume 9, Issue 10, October 2018, pp. 1402–1411, Article ID: IJCIET_09_10_140
Available online at http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=9&IType=10
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

TOWORDS BUILDING AN ANTI-SPAM FOR ARABIC SMS

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

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 in using SMS leads to the increase of SMS spam problem. Hence, SMS developing spam filter become a goal of many organizations to deal with this problem. This work present a work on filtering SMS using "Naïve Bayesian" (NB) classifier. This is a content-based classifier that operates on the body of an Arabic input SMS samples. The work discusses issues related to availability and collection of data needed for principle classification, so no resources yet available for Arabic SMS, hence are collected manually. Performance of proposed filter is measured under certain setting of working parameters. A total of 400 SMS are considered; 70% for training and 30% for testing, and with a length of 15-featuers vector, A level of 85% of accuracy is reached. Using features selection, methods, accuracy level is raised up to 88% .Experimented result motivates further work towards analyzing a large corpus of Arabic SMS sample, yet improving classification rate. But for the time being these result exhibit a satisfactory performance of NB classification and hence is reworded for use in android platform due to simplicity and ease of implementation.

Keywords: SMS, Anti-spam, *Naïve Bayesian"* (NB) classifier

Cite this Article Hiba Adel Majeed and Dr. Maha A. Bayati Towords Building an Anti-Spam for Arabic Sms, *International Journal of Civil Engineering and Technology*, **10**(10), 2018, pp. (1402)-(1411).
http://www.iaeme.com/IJCIET/issues.asp?JTypeIJCIET&VType=9&IType=10

1. INTRODUCTION

Nowadays, mobile phone is the most popular device is used to manage our daily activities like writing notes, social networking, email, calculator,etc. The primary purpose for mobile manufacturing is that to contact people with each other's in many ways like voice call, video call, short message (a known as SMS Short Message Service).

For a long , SMS and social networking applications have become the most effective tool of communication between people in the world .It becomes a goal of many organization to deal with, when SMS is one such a publicized such threat that needs to be dealt with.

The spam messages (hence SMS) received via mobiles are junk, unwanted, and unsolicited text messages; particularly when they are sent for advertising purposes. Moreover, the number of such messages increased with the increase of mobiles since 2000s and up to date (wikipedia, 2016).

Like email spam, SMS spam can zone from social engineering hoaxes to unsolicited advertising to harmful attempts to theft subscriber's financial and personal details. (gsma, n.d.)

2. TOWARDS SMS FILTER FROM EMAIL FILTER

The problem with SMS Spam is growing more due to the increase in the use of text short messaging (SMS). To control such spams, a number of security measures need to be available. As a case in point is the filtering mechanism, which focused first on e-mail spams, a very old and popular problem that has been facing mobile phones. Second, it focused on SMS spams, which represent the watchword these days (Neelam Choudhary, 2017). It seems that the process of filtering e-mail spams is somehow similar to that of SMS spams. For instance, the content-based technologies used in filtering e-mail spams1 could be similarly invested in SMS spam filtering. This is because such technologies contain both direct content filtering1 and collaborative content filtering techniquesl (Sarah Jane Delany, 2013).

The prevalence of the SMS spam problem has pushed specialists in the area to devise advanced filters. However, the problem of concept drift in SMS spam filtering has evolved (Sarah Jane Delany, 2013). In addition, SMS spam filter has further got a number of additional problems. The first of these additional problems was related to the length of the message, which 160 characters. This means there was some space for content-based filtering. Having a short-length message, SMS users started using a local language subset with abbreviations, phonetic contractions1, emoticons,. .etc. . Such a language differs from the traditional written language used in emails. The problem of email spam filtering1 was improved using the contextual information found in the email headers. However, SMS is said to have less information in the headers; a matter that leaves less context to work with. Recently, a number of successful attempts have been conducted with regard to the use of email spam filtering techniques in solving problems related to SMS spam filtering (Sarah Jane Delany, 2013).

3. SENDER OF SPAM SMS

Most of the sources of spam SMS are from web1 site or companies or organizations for the purpose of publicity and commercial advertising and may also be from persons who tend to be doing threats or making a jokes. It is also possible that the telecommunications companies themselves send the advertising messages for the purpose of material1 gain from advertisers. The probability of received spam SMS:

- If the message is from an unknown number (unknown) and there is no SMS preceded by this number (Iosif Androulidakis, 2013).
- When the first digits of an SMC match neither the recipient nor the sender, this means that the message is sent from1an SMC that belongs to a country other than the sender’s, and so it represents a spam (Iosif Androulidakis, 2013).
- As for the ID of the1 sender, it can either be a purely numerical containing as a result other characters, or a non-numerical, and so chances of being spams will be great; (Iosif Androulidakis, 2013).
- When the first digits of an SMSC match between recipient and sender,1but the time of the SMSC is more than one minute ahead, this will be a clear indicator that the received or sent SMS is a spam (Iosif Androulidakis, 2013).
- SMS that do not contain a reply (Iosif Androulidakis, 2013).

- SMS that contain spam words or web site URL that may be as spam (Iosif Androulidakis, 2013).

4. DESIGN OF THE PROPOSED ARABIC SMS FILTER

The proposed system is a step forward towards building Arabic Antispam SMS Filter. The System can classify SMS into two categories spams and legitimate SMS. System's workflows the following basic as shown in figure (1):

- Reading SMS (Arabic Datasets)
- Preprocessing phase
- Extraction Feature phase
- Normalization phase
- Features Selection phase
- Classification phase

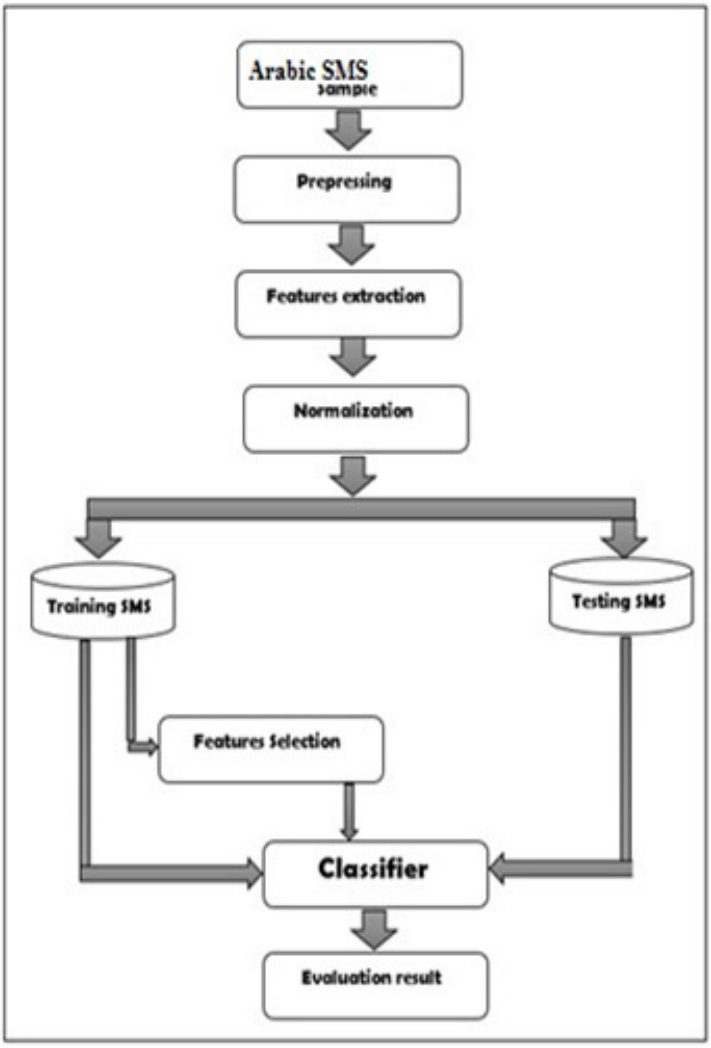


Figure 1 Block diagram of Arabic SMS filter

4.1. Reading SMS

No resource is found to provide for Arabic SMS dataset. Even the National company communication and media commission of Iraq (NCMC) failed to support getting a real data from one of the local mobile communication company. Hence, Arabic dataset was collected manually in a way or another. The 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.

4.2. Preprocessing phase

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

4.2.12. Tokenization

The word "tokenization" "means breaking a stream of text into it's constituent meaningful units (called tokens)". Typically, tokenization processes the SMS text body. It breaks the body down into words, hence cleaning up all white space.

4.2.2. Stop words removal

This is the second step in prepressing, the SMS body, through which occurrence of any of predefined stop words removed. Note that stop words are some sort common Arabic words, those provide no useful information to help deciding the class of some SMS.

4.2.3. Stemming

Unlike emails, SMS cannot be suitably undergone stemming for two reasons :

- Short messages are almost written in local languages where abbreviations are frequently used.
- No BOW is used here, only a limited number of spam words like (فائز ، فاز) hence it is not worthier do stemming.

Figure (2) exhibit sample of the resulting Arabic data files, respectively. Upon completion of prepressing phase.

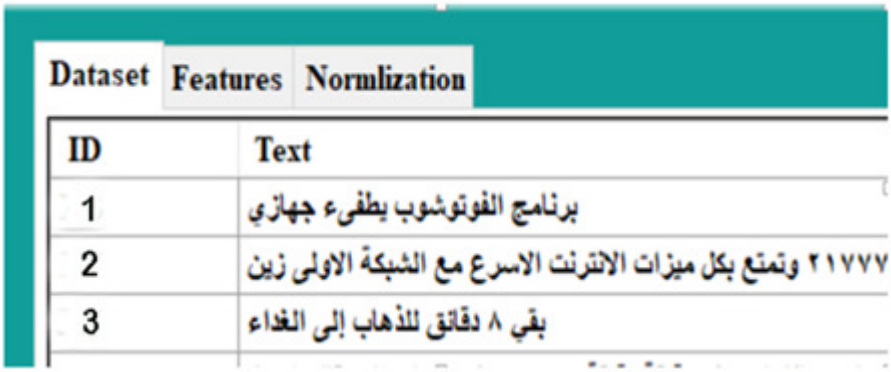


Figure 2 Sample prepressing Arabic SMS

Basic steps of prepressing phase in presented next algorithm (1):

```
Input: SMS            // SMS in the dataset (body of SMS)
       Stop words     //List of Arabic Stop Words

Output: SMS           //SMS with prepressing phase
Begin:
    For each SMS body do
        Remove white space
        Remove all stop words
    End for
End
```

Algorithm 1 Preprocessing Phase

4.3. Extraction Feature phase

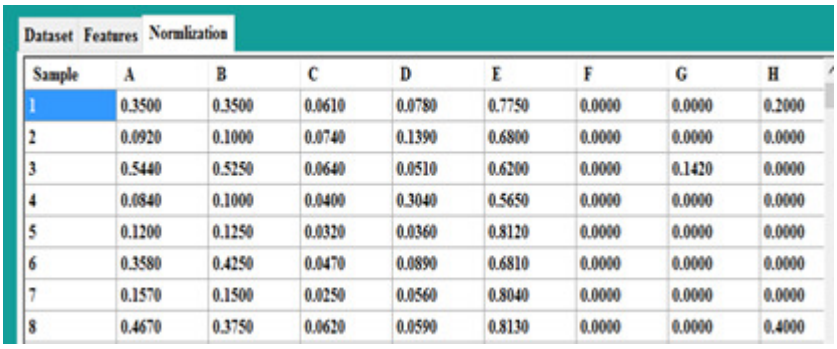
In this phase extract fifteen features are extracted from the body of each SMS. Due to lack of recourses concerning Arabic messages. Table (1) illustrate the fifteen features that is used.

Table 1 Features Extractions

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-alphanumeric character	Number of non-alphanumeric characters
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

4.4. Normalization phase

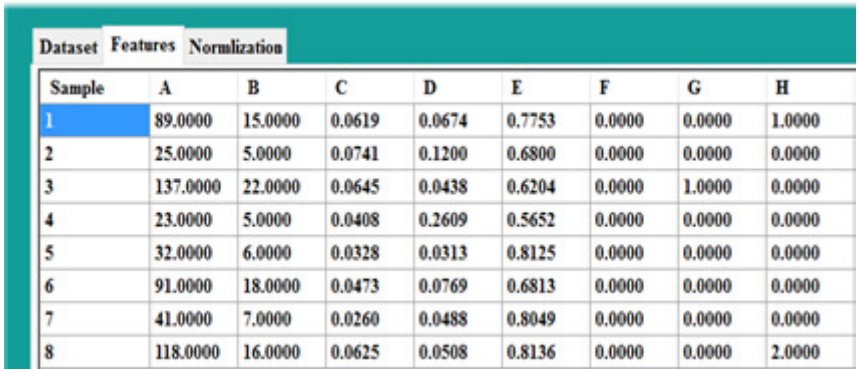
Just after extracting fifteen features from each SMS body, time to apply normalization to reduce the variance of values between those. Figure (3) presents the result of normalizing features value, for the sample messages presented in figure (4)



Dataset | Features | Normlization

Sample	A	B	C	D	E	F	G	H
1	0.3500	0.3500	0.0610	0.0780	0.7750	0.0000	0.0000	0.2000
2	0.0920	0.1000	0.0740	0.1390	0.6800	0.0000	0.0000	0.0000
3	0.5440	0.5250	0.0640	0.0510	0.6200	0.0000	0.1420	0.0000
4	0.0840	0.1000	0.0400	0.3040	0.5650	0.0000	0.0000	0.0000
5	0.1200	0.1250	0.0320	0.0360	0.8120	0.0000	0.0000	0.0000
6	0.3580	0.4250	0.0470	0.0890	0.6810	0.0000	0.0000	0.0000
7	0.1570	0.1500	0.0250	0.0560	0.8040	0.0000	0.0000	0.0000
8	0.4670	0.3750	0.0620	0.0590	0.8130	0.0000	0.0000	0.4000

Figure 3 sample of features extraction



Dataset | Features | Normlization

Sample	A	B	C	D	E	F	G	H
1	89.0000	15.0000	0.0619	0.0674	0.7753	0.0000	0.0000	1.0000
2	25.0000	5.0000	0.0741	0.1200	0.6800	0.0000	0.0000	0.0000
3	137.0000	22.0000	0.0645	0.0438	0.6204	0.0000	1.0000	0.0000
4	23.0000	5.0000	0.0408	0.2609	0.5652	0.0000	0.0000	0.0000
5	32.0000	6.0000	0.0328	0.0313	0.8125	0.0000	0.0000	0.0000
6	91.0000	18.0000	0.0473	0.0769	0.6813	0.0000	0.0000	0.0000
7	41.0000	7.0000	0.0260	0.0488	0.8049	0.0000	0.0000	0.0000
8	118.0000	16.0000	0.0625	0.0508	0.8136	0.0000	0.0000	2.0000

Figure 4 sample of normalization

4.5. Features Selection phase

Since relevant, features are often unknown a prior certain classification requires that features selection be carried out to reduce the number of irrelevant, as well as redundant features, those features whose removal would drastically improve system performance, following are features selection methods applied by the system:-

4.5.1. 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 (Forman, 2003).

$$TF=\sum_{1}^{C}\frac{H}{N}*F \ +\frac{S}{N}*F \qquad 1$$

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.

4.5.2. 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 (Daniel I. MORARIU).However the entropy of Y is

$$H(Y) = -\sum_{y \in Y} p(Y) \log_2 p(Y)$$ 2

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

$$H(Y|X) = -\sum_{x \in X} p(X) \sum_{y \in Y} p(Y|X) \log_2 p(Y|X)$$ 3

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:

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y)$$ 4

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 (Novakovic, 2009).

4.5.3. 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 (R. Praveena Priyadarsini1, 2011) **.** Present the formula used to find GR of Certain features.

$$GR(feature) = \frac{\boldsymbol{Gain}\ (\boldsymbol{featuer})}{\textbf{intrinsic info.}(\textbf{featuer})}$$ 5

4.6. 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 Arabic SMS as by Spam or Ham. 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 (NB, n.d.) (Jabbar, 2015).

4.6.1. Training phase :

This phase attempts to calculate the following probabilities:

- ✓ Calculate the probability of Spam SMS class and Ham SMS class to the total number of SMS sample:

$$P(C_i) = \frac{|C_i|}{N} \quad 6$$

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

- ✓ 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 , as shown below:

$$P(X|C_i) = \prod_{j=1}^{n} P(x_j|C_i) \quad 7$$

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.

4.6.2. 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"* as bellow:

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

Equ.8

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

5. ARABIC SMS AND FEATURES

The effect of 15 features F0-F14 that were extracted from Arabic SMS samples illustrates in bar chat below for each of Spam and Non-Spam class.

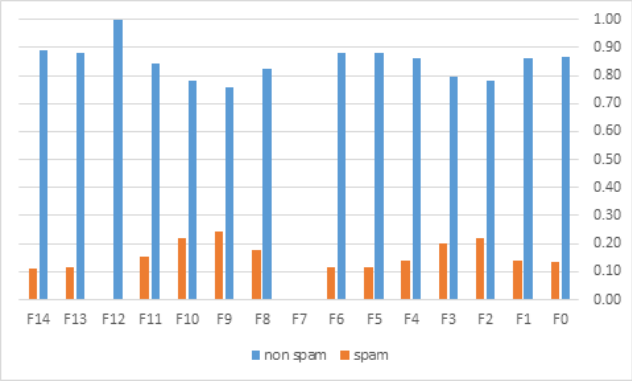


Figure 5 Probability of Arabic (Spam Non-Spam) Features

The feature F7 that will not be shown in each spam and non-spam samples, F12 have a high effective on non-spam samples when it is not shown in spam sample. F9 has a highest effective on spam samples.

6. RESULTS AND EXPERIMENTS:

The proposed filter was exterminated to measure it's efficiency under different settings of working permeates. Arabic SMS dataset, a total of 400 SMS were considered; 70% for training and 30% for testing. For a total of 15-featuers, extracted from each SMS, an accuracy of 83% was reached. Using features selection, accuracy level was raised up to 88%. Experiments for testing the proposed NB classifier are presents in next section. Note that corresponding test results are listed together with that SVM library for comparison purposes (Basic evaluation measures from the confusion matrix, n.d.).

6.1. Accuracy Results

The ratio1 of the total number1 of SMS that were classified correctly (Accuracy) also the highest percentage in Lisvm with fifteen features, while The NB classifier recorded the lowest percentage within that range , The TF-NB with twelve features recorded the highest among NB classifier with the other features selection methods. Table (2) and figure (6) shown the overall results of Accuracy.

Table 2 Accuracy Results

No. of features	TF-NB	IG-NB	GR-NB	NB	SVM
15	-	-	-	83%	91.8 %
14	85%	86.8%	86%	-	-
13	85%	86.8%	85.7%	-	-
12	88%	86.7%	86.7%	-	-
11	85%	85.7%	86%	-	-
10	85%	85.7%	85.3%	-	-

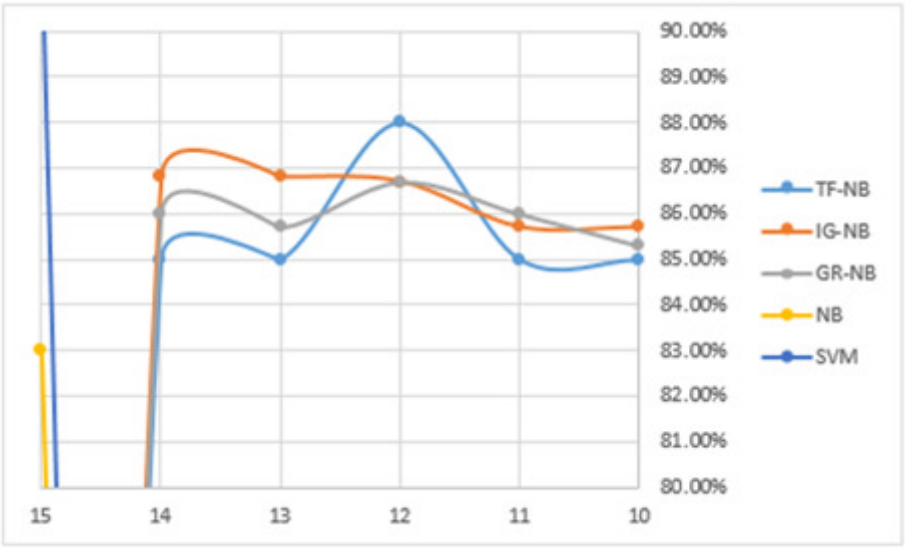


Figure 6 Accuracy Result

7. CONCLUSIONS

- Implementing Naive Bayesian classifier gives very good but not excellent accuracy result of SMS classification
- To improve the accuracy result of Naive Bayesian there is a need to implement a feature selection approach, this proposal suggest TF IG and GR algorithms as a feature selection. Using these selection methods raise the accuracy result of classification
- The accuracy of IG-based NB classifier is better than those obtained by using NB classifier.
- Stemming in this work not effective because deal with local language and using spam words exists in more than shape.

ACKNOWLEDGEMENT

It is our pleasure to express our appreciation and thanks for Computer Science Department/ College of Science / Mustansiriyah University / Baghdad / Iraq for the valuable assistance and encouragement to accomplish this research. Seeking more scientific researches at Mustansiriyah University in the future.

REFERENCE

[1] Basic evaluation measures from the confusion matrix. (n.d.). Retrieved from https://classeval.wordpress.com/introduction/basic-evaluation-measures/

[2] Daniel I. MORARIU, R. G. (n.d.). Feature Selection in Document Classification. “Lucian Blaga” University of Sibiu.

[3] Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification . Journal of Machine Learning Research 3.

[4] gsma. (n.d.). Retrieved from http://www.gsma.com/managedservices/spam-management-prevention/mobile-spam/what-is-mobile-spam/

[5] Iosif Androulidakis, V. V. (2013). FIMESS: Filtering Mobile External SMS Spam.

[6] Jabbar, S. F. (2015). A Spam Email Classifier Based on Naive Bayesian Approach. Baghdad.

[7] NB. (n.d.). Retrieved from http://www.saedsayad.com/naive_bayesian.htm

[8] Neelam Choudhary, A. K. (2017). Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique.

[9] Novakovic, J. (2009). Using Information Gain Attribute Evaluation to Classify Sonar Targets. Serbia, Belgrade.

[10] R. Praveena Priyadarsini1, M. a. (2011). GAIN RATIO BASED FEATURE SELECTION METHOD FOR PRIVACY PRESERVATION. 1, 3Department of Computer Science and Engineering, Avinashilingam Deemed University for Women, Tamil Nadu, India.

[11] Sarah Jane Delany, M. B. (2013). SMS spam filtering: Methods and Data. researchgate.

[12] wikipedia. (2016). Retrieved from https://en.wikipedia.org/wiki/Mobile_phone_spam