# Improving Performance of Mobile SMS Classification Using TF-IDF & Multinational Naive Bayes Classifier

ISSN (e) 2520--7393
ISSN (p) 5021-5027
Received on 22nd Jan, 2018
Revised on 28th Feb, 2018
www.estirj.com

Mehr-Un-Nisa Manjotho[1], Tariq Jameel Saifullah Khanzada[2], Liaquat Ali Thebo[3], Ali Asghar Manjotho[4]

[1]*Institute of Information and Communication Technologies, Mehran University of Engineering and Technology Jamshoro*
[2,3,4]*Department of Computer Systems Engineering, Mehran University of Engineering and Technology Jamshoro*

**Abstract:** Text classification is a technique of categorizing the document in to predefined classes on the basis of its contents. Text classification is easy for human being but it's still a challenging task for machine. There are many models presented for text classification but it is still a major area of research and needs further refinement.The aim of the study was to improve the accuracy of TF-IDF weighting scheme and analyze the effectiveness of Bayesian email spam classifier in classifying Mobile SMS into different categories like Greetings, urgent, harassing, etc. The study has been conducted in two phases. The training phase, in which the classifier was trained with 5574 SMSs. Preprocessing was done in which different techniques were applied like shorthand completion technique to replace short words with its full form, Tokenization, stop words were removed, stemming technique was applied, N-grams were generated, Vector space model was created and finally the output was given to classifier. In the second phase the classifier was tested with the testing data set of 852 SMSs and the results were obtained. Multinomial Naïve Bayesian classifier was used to categorize the new SMS into the suitable category. The classifier gave more promising results with the True Positive rate of 93.77%. The overall accuracy obtained for the system was 93.74% with the precision of 94.03%.

**Keywords:** *SMS, Text Classification, Naïve Bayesian, TF-IDF, Multinomial*

## 1. Introduction

Nowadays, Mobile SMS is one of the cheapest means of communication. We receive lots of SMSs every day. Many of which are not of our interest [1]. And because of those undesired SMSs, our inbox become massive and if a person is not use to of mobile phone and he/she wants to read only urgent messages or the SMSs of their interest, then Text Classification is best suitable to overcome this problem[2].Text classification is the process of categorizing or classifying a text into predefined classes on the basis of its contents. It needs a classifier to classify a document. There are many classification techniques used for classification such as Naive Bayes, K-NN, Neural Network, Decision Tree, SVM, and many more [3]. Text classification is used in many applications such as document indexing, spam filtering, web page prediction etc. [4]. Linear classifiers works best for text classification as they are effective in computing in both training and testing and at classification phase [5].

## 2. Related Work

Many of the techniques were introduced like Naive Bayes classifier which is a simple probabilistic classifier. Decision tree is another tool for classifying text in spam/ ham or any other category. It has a tree like structure in which leaves are the class labels or categories. Support Vector Machine (SVM) is a binary classifier that uses margin to find the categories. Likewise, there are so many other tools introduced to classify the email, document or SMS into predefined categories. The author in [7] uses methodology to compare the accuracy of naïve Bayes classifier and SVM classifier to classify Nepali SMS. They proposed a hybrid model for classification. Both Naïve Bayes and SVM classification techniques are used to assess the accuracy of classification methods.

The classification accuracy of 92.74% was obtained for the Naïve Bayes classifier and 87.15% for SVM. The restriction of this research was that, it was only limited to categorize the Nepali SMS into spam and ham messages. In [8] the researchers uses SVM classifier for classification of SMS using Document Frequency Threshold. The major purpose of the research was to select many number of features in order to classify the document accurately. The limitation of this research was that, it was selecting more features when the terms were low and when the number of terms in a document were increasing, the system was abstracting less features. In order to get good accuracy, the number of terms must be increased, in the result the size of the document was also increasing.

Corresponding author Email address: *manjotho2006@gmail.com*

In the reference [9], the researchers tried to improve the accuracy of TF-IDF weighting scheme on the basis of importance of words to reduce the classification problems. To improve the accuracy the abbreviations were replaced by its full forms. Stop words were removed, parts of speech tagging (POS) and stemming technique was applied and vector space model was created using TF-IDF technique which was used to find the importance of a word in SMS and finally the result was given to the classifier for assigning related classes. The model attends the accuracy of 91.94%.This paper was restricted to only 4 categories or class labels i.e. occasion, sales, greetings, friendship. This can be further extended to many categories and its accuracy can be further improved. In the study [10], the researchers proposes a model for SMS text classification and discussed the applications of SMS classification in different areas. It uses entropy term weighting system and Principal Component analysis (PCA) with neural network to classify the Mobile SMS to different categories like poetry, jokes, festival etc. The model was not implemented yet, the results were only bases on analysis.

## 3. Methodology

We have used UCI Machine Repository to generate Training and Test dataset. Our model comprises of 12 steps i.e. Data collection, shorthand completion, Tokenization, Removing stop words, Stemming, Generating N-Grams, Calculating TF-IDF, Applying Naïve Bayesian Model, Vectorization, Training Classifier, Testing Classifier and at last SMS Classification using Multinomial Naïve Bayes algorithm. Figure: 1 shows the methodology used in this research.

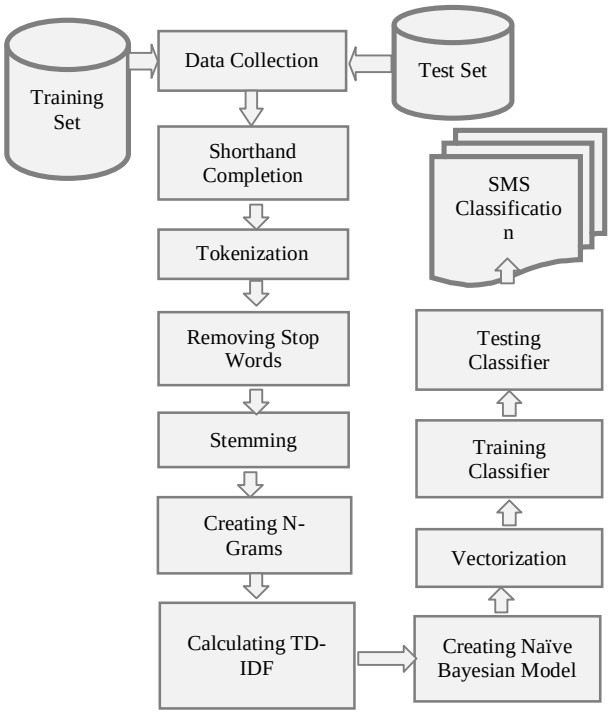


Figure.1. SMS Text Classification Model

### *3.1. Shorthand completion*

Most of the SMS text contains shorthand language words [11]. These words are replaced by corresponding full word from dictionary. We have used dictionary of 1463 shorthand word.

| I m w8ing 4u 8 bus stop. |
|---|

| I | am | waiting | for | you | at | bus | stop |
|---|---|---|---|---|---|---|---|

Figure.2. SMS with Shorthand words

### *3.2. Tokenization*

Tokenization is the process of breaking down the text corpus in to individual elements [12].

| Hello! Dear, I am on leave today. |
|---|

| hello | dear | i | am | on | leave | today |
|---|---|---|---|---|---|---|

Figure.3. SMS after Tokenization

### *3.3. Removing Stop Words*

Stop words are unnecessary word that commonly appear in the text [13]. Example, words such as so, and, or, the etc. All stop words are removed first. In the figure below the stop words are: you, are, that and the. Which are removed by using this technique.

| You are lucky, that you have won the cash prize |
|---|

| lucky | won | cash | prize |
|---|---|---|---|

Figure.4. SMS after removing Stop Words

### *3.4. Stemming*

In English language, there are different forms of verbs/words being used. Stemming is the process of converting the word in to its root form [14]. SMS shown in Figure 4 contains two words “raining” and “coming”, whose root word can be found. After applying Stemming technique, “raining” is converted into its root word “rain” and “coming” is converted into “come”, which is its root word. This technique reduces the number of characters in an SMS/Document, hence the number of words being processed is reduced which in turn increases the performance of the proposed model.

| It is raining, so I am not coming today, |
|---|

| It | is | rain | so | I | am | not | come | today |
|---|---|---|---|---|---|---|---|---|

Figure.5. SMS after Stemming

### *3.5. N-Grams*

In the n-gram model, a token can be defined as a sequence of n items. Unigram (1-gram) where each word consists of exactly one word, letter, or symbol. Unigram is most commonly used N-Gram [15]. N-grams of size 3 and 4 yield good performances [16]. It is used for supervised learning. Can be used in text categorization, spelling correction and breaking the words to find the relationship between words. We usually move in forward direction to create next N-Gram [17]. In the given SMS "I am not coming today", bi-Gram can be produced by moving 2 terms forward i.e. "Iam", "amnot", "not coming" and "coming today". In the same way, we can create N-Gram of any size.

| | | | | | |
|---|---|---|---|---|---|
| Unigram (1-gram) | I | am | not | coming | today |

| | | | | |
|---|---|---|---|---|
| Bigram (2-gram) | I am | am not | not coming | coming today |

| | | | |
|---|---|---|---|
| Trigram (3-gram) | I am not | am not coming | not coming today |

Figure.6. SMS after removing Stop Words

### *3.6. 6. Multinomial naive Bayesian*

Multinomial Naïve Bayes is the modified form of traditional Naïve Bayes classifier. It operates on binary values(i.e. 0 and 1). Either the term is present in the SMS or does not occur in it. Here we use *Term Frequency - Inverse Document Frequency (TF-IDT)* approach to characterize text documents instead of having binary values in Naïve Bayesian Classifier. It is a probabilistic classifier which computes the probabilities of every class by applying Naïve Bayes theorem. Then it will calculate the conditional probability of each SMS in the particular category.

$$\mathbf{TF-IDF} = \boldsymbol{tf_n(t,d)} \,.\, \boldsymbol{idf(t)}$$

- $\boldsymbol{tf_n(t,d)}$is normalized term-frequency

$$\boldsymbol{idf(t)} = \mathbf{log}\left(\frac{\boldsymbol{n_d}}{\boldsymbol{n_d(t)}}\right)$$

Where,

- $tf_n(t,d)$ : term frequency of term t in document d
- Idf : Inverse document frequency
- $n_d$ : Total no. of documents
- $n_d(t)$ : Total no. of documents in which t appears

The term frequencies can then be used to compute the maximum-likelihood estimate based on the training data to estimate the class-conditional probabilities in the multinomial model:

$$\hat{P}(x_i \mid w_j) = \frac{\sum tf(x_i,\ d \in w_j).idf(x_i) + \propto}{\sum N_{d \in wj} + \propto . V}$$

Where,

- $x_i$ : A word from the feature vector x of a particular sample.
- $\sum tf(x_i,\ d \in w_j).idf(x_i)$ : The sum of TF-IDF of word $x_i$ from all documents in the training sample that belong to class$w_j$.
- $\sum N_{d \in wj}$ : The sum of all term frequencies in the training dataset for class$w_j$.
- $\propto$ : An additive smoothing parameter ($\propto$ = 1 for Laplace smoothing).
- $V$ : The size of the vocabulary (number of different words in the training set).

## 4. Results and Discussion

The classifier was trained with the training dataset of 5574 SMSs. It was then tested with a test dataset of 852 different SMSs. The with True Positive rate of 93.77%. The overall accuracy obtained for the system was 93.74% with the precision of 94.03%.

Table 1 presents the class variables, class labels and number of records for each class. Table 2 presents confusion metrics parameters.

There are 7 class labels along with their class variables (w1 to w7) and the number of records associated with every class.

Confusion matrix shown in Table 2 represents the actual class and the classes predicted by classifier. Class variable w1 has a total of 105 SMSs, among which 98 were correctly identified under class w1, none was predicted as class w2, w4 and w7, two SMSs were predicted wrongly under class w3, three SMSs as Class w5 and two as class w6. Likewise, classifier predicted the categories for remaining SMSs as shown in Table 2.

Table.1. Class Labels

| Class Variable | Class label | No. of records in corpus |
|---|---|---|
| W1 | Greetings | 105 |
| W2 | Urgent | 75 |
| W3 | Invitation | 109 |
| W4 | Harassing | 106 |
| W5 | Request | 92 |
| W6 | Advertisement | 205 |
| W7 | Unknown | 160 |

Table.2. Confusion metrics parameters

| | | Predicted | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | W1 | W2 | W3 | W4 | W5 | W6 | W7 | TP | FN |
| Actual | W1 | 98 | 0 | 2 | 0 | 3 | 2 | 0 | 98 | 7 |
| | W2 | 1 | 72 | 0 | 0 | 1 | 0 | 1 | 72 | 3 |
| | W3 | 0 | 0 | 107 | 0 | 0 | 1 | 1 | 107 | 2 |
| | W4 | 2 | 0 | 0 | 102 | 0 | 0 | 2 | 102 | 4 |
| | W5 | 0 | 0 | 1 | 0 | 85 | 3 | 3 | 85 | 7 |
| | W6 | 2 | 2 | 1 | 0 | 2 | 185 | 13 | 185 | 20 |
| | W7 | 2 | 3 | 0 | 1 | 1 | 3 | 150 | 150 | 10 |
| FP | | 7 | 5 | 4 | 1 | 7 | 9 | 20 | | |
| TN | | 740 | 772 | 739 | 745 | 753 | 648 | 672 | | |

| | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | W1 | W2 | W3 | W4 | W5 | W6 | W7 |
| Actual | W1 | 98 | 0 | 2 | 0 | 3 | 2 | 0 |
| | W2 | 1 | 72 | 0 | 0 | 1 | 0 | 1 |
| | W3 | 0 | 0 | 107 | 0 | 0 | 1 | 1 |
| | W4 | 2 | 0 | 0 | 102 | 0 | 0 | 2 |
| | W5 | 0 | 0 | 1 | 0 | 85 | 3 | 3 |
| | W6 | 2 | 2 | 1 | 0 | 2 | 185 | 13 |
| | W7 | 2 | 3 | 0 | 1 | 1 | 3 | 150 |

Table.3. True Positives (TP), False Positive (FP), True Negatives (TN) and False Negatives (FN)

Table 3 shows 4 parameters. True positive, True negative, False positive and False negative parameters of all the 7 classes were calculated using confusion matrix.

Below figures shows the graphs of detection parameters (i.e. True positive, True negative, False positive and False negative) of all the 7 classes W1 through W7 individually.

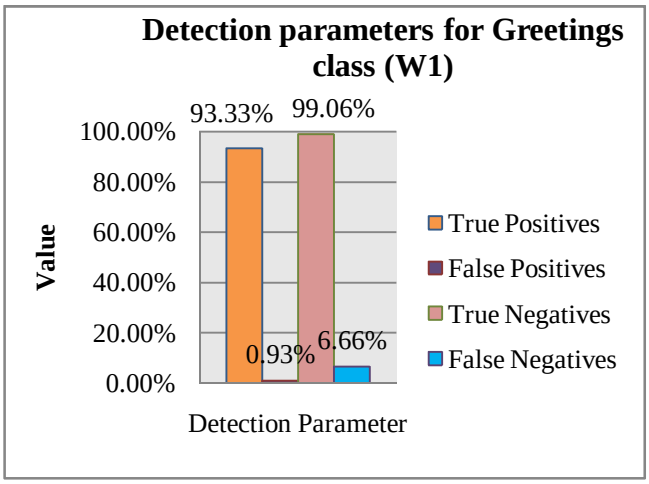


Figure.7. Detection Parameters for class W1

Figure 7 shows the detection parameters for Greetings class. It shows the true positive value of 93.33%, False positive rate of 0.93%, True negative 99.06% and False negative rate 6.66%. The overall accuracy of class W1 comes out to be 0.99%.

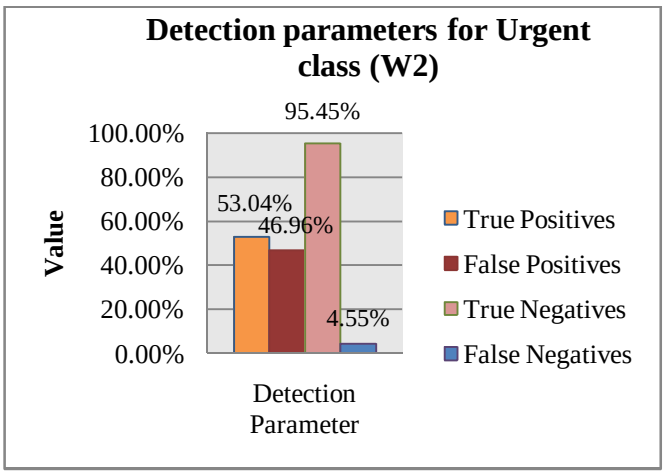


Figure.8. Detection Parameters for class W2

In the above figure 8, detection parameters for Urgent class are shown with True positive rate of 53.04%, False positive rate of 46.96%, True negative rate 95.45% and False negative rate of 4.55%. The accuracy calculated is 0.8755% which means almost all the urgent SMSs were correctly classified under urgent messages class.

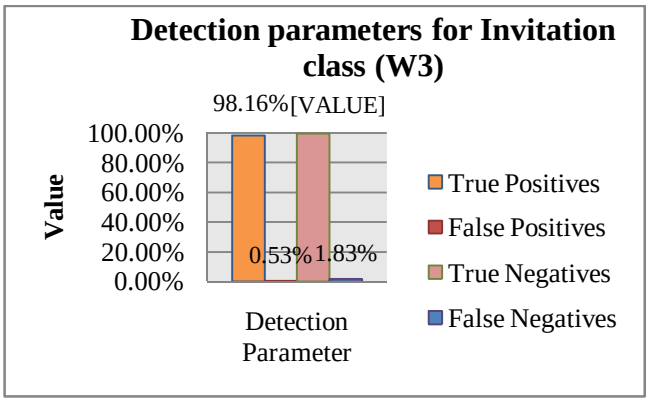


Figure.9. Detection Parameters for class W3

In Figure 9, detection parameters for Invitation class were calculated. The True positive rate comes out to be 98.16%, False positive rate of 0.53%, True negative rate 99.46% and False negative rate of 1.83%. The overall accuracy of the class W3 is 0.9941%.

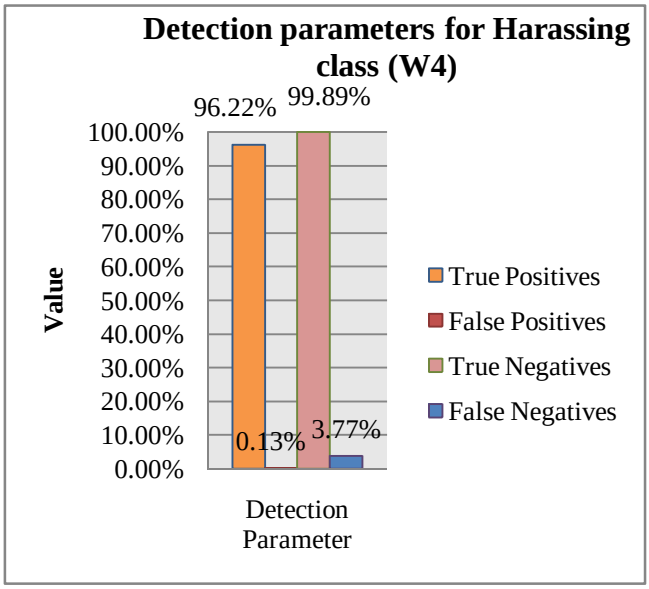


Figure.10. Detection Parameters for class W4

Above figure 10 shows the detection parameters for Harassing class with True positive rate of 96.22 %, False positive rate of 0.13%, True negative rate 99.89% and False negative rate of 3.77%. The calculated accuracy is 0.9835%.

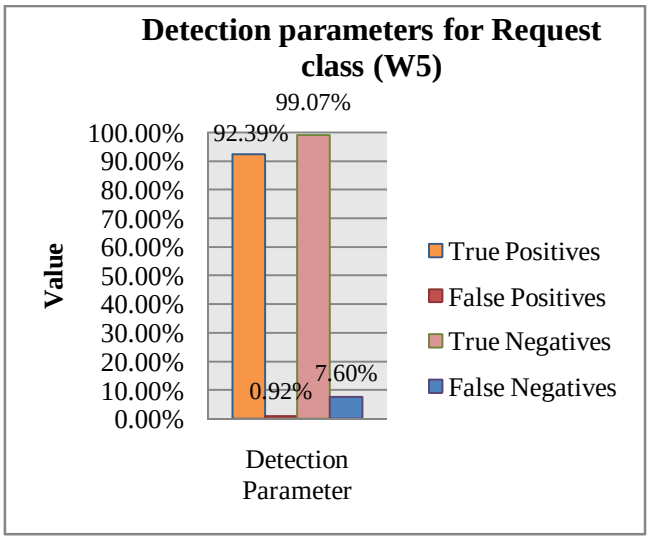


Figure.11. Detection Parameters for class W5

Detection parameters for Request class are shown in Figure 11. True positive rate is 92.39, False positive rate is 0.92%, True negative rate 99.07% and False negative rate is 7.60%. The accuracy of class W5 is 0.9663%.

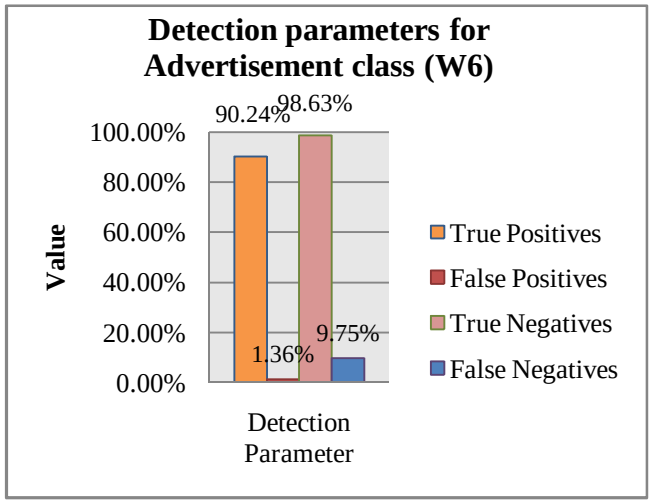


Figure.12. Detection Parameters for class W6

Figure 12 shows the detection parameters for Advertisement class. It shows the true positive value of 90.24%, False positive rate of 1.36%, True negative 98.63% and False negative rate 9.75%. The overall accuracy of class W6 comes out to be 0.9647%.

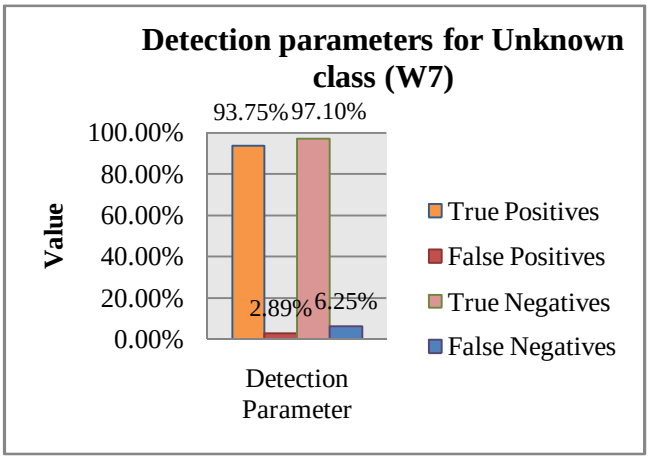


Figure.13. Detection Parameters for class W7

Detection parameters for Unknown class are shown in Figure 13. True positive rate is 93.75, False positive rate is 2.89%, True negative rate 97.10% and False negative rate is 6.25%. The accuracy of class W7 is 0.9263%.

Figure 14 shows the overall detection parameters using Multinomial Naïve Bayes classifier. The evaluation metrics shows that the model attended the accuracy of 93.74, which means that approximately 94 out of 100 SMSs were categorized under actual class. The 93 out of 100 is quite good accuracy with the precision of 94.03%.

The false negative ratio is 6.2% which shows that approximately 7 out of 100 SMSs will be classified under wrong category.

It is evidence from results that the proposed model works successfully to classify a mobile SMS under particular category. Also an Email spam classifier effectively classified a Mobile SMS

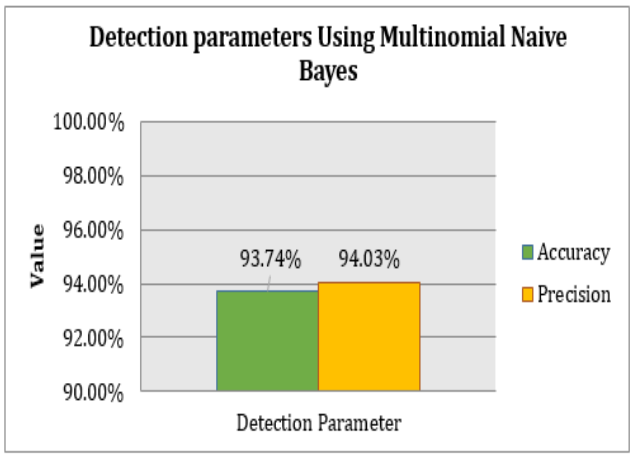


Figure.14. Detection Parameters

## 5. Conclusion

This research presents a new model to classify an SMS into different categories like Greetings, Harassing, Urgent, Invitation etc. A person who is not addicted of cellphone, can go to specific category and read the SMS without going through a complete bulky inbox.The proposed model uses TF-IDF technique with Multinomial Naïve Bayes classifier. It attended an overall accuracy of 93.74% that determines that approximately 94 out of 100 SMSs can be correctly classified to their correct classes. The 94 out of 100 is quite good accuracy. The false negative ratio is 6.2% which means that approximately 7 out of 100 SMSs will be classified under wrong category. The overall precision of model is 94.03%. And ensures that an email spam classifier could be used to classify a Mobile SMS into different categories.

## References

[1] Ahmed, Ishtiaq, Guan, Dhongi, Chung, Choong, Tae, "*SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset,"* International Journal of Machine Learning and Computing, vol. 4, no. 2, April 2014, pp.183-187.

[2] Ghayda A. Al-Talib, Hind S. Hassan., "A Study on Analysis of SMS Classification Using TF-IDF Weighting," International Journal of Computer Networks and Communications Security, vol. 1, no. 5, October 2013, ,pp.189-194.

[3] Parimala, R., "A Study on Analysis of SMS Classification Using Document Frequency Threshold," Information Engineering and Electronic Business, April, 2012, pp. 44-50.

[4] Patel, Deepshikha, Bhatnagar, Monika., *"Mobile SMS Classification: An Application of Text Classification."*International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, vol. 1, Issue 1, March 2011, pp. 47-49.

[5] Shahi, B. Tej, Yadav, Abhimanu.,*"Mobile SMS Spam for Nepali Text Using Naïve Bayesian and Support Vector Machine."* International Journal of Intelligence Science, April, 2014, pp. 24-28.

[6] Raschka, Sebastian. *"Naive bayes and text classification introduction and theory.",* 2014, pp.72-81.

[7] UCI Machine Learning Repository, SMS Spam Collection Data Set,https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection, [Accessed: 10 Dec, 2016]

[8] Ruchika B., Najuka S., Rakhi S., Sunita G., *"Mobile SMS Classification,"* International Journal of Computer Science and Information Technology Research, 4(2), Jun 2016, pp.182-185.

[9] Padhiyar, Hiral.; Rekh, Purvi., *"Improving Accuracy of Text Classification for SMS Data."*International Journal for Scientific Research & Development| vol. 1, Issue 10, 2013, pp. 181-189.

[10] Gaurav, Sethi.; Vijender, Bhootna.; *"SMS Spam Filtering Application Using Android."* International Journal of Computer Science and Information Technologies, vol. 5(3), ISSN: 0975-9646 , 2014, pp. 4624-4626.

[11] S. M. Kamruzzaman., F. Haider ., A. Ryadh Hasan.,*"Text Classification Using Data Mining,"* International Journal of Machine Learning and Computing, vol. 4, no. 2, April 2014, pp. 79-85.

[12] A. Faraz., *"An elaboration of text categorization and automatic text classification through mathematical and graphical modelling,"* An International Journal (CSEIJ), Vol.5, No.2, June 2015 , pp. 239-248.

[13] C.Ramasubramanian., R.Ramya., V. Tamilnadu.,*"Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm,"* International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013, pp. 4536-4538.

[14]V.Srividhya.,R.Anitha,*"Evaluating PreprocessingTechniques in Text Categorization."* International Journal of Computer Science and Application Issue 2010, pp. 49-51.

[15] G. V. Cormack.; J.M.Gómez Hidalgo.; E. Puertas Sánz., *"Feature Engineering for Mobile (SMS) Spam Filtering"* Amsterdam, The Netherlands. ACM 07, July 2007, pp. 271-276.

[16] F. Johannes., *"A Study Using n-gram Features for Text Categorization."* International Journal of Computer Science & Communication Networks,Vol 5(1), ISSN:2249-5789, pp. 7-16.

[17] G.Forman., *"An extensive empirical study of feature selection metrics for text classification."* The Journal of machine learning research, Vol 3: ISSN: 1289–1305, pp. 1289-1295.

[18] C. Silva.; B. Ribeiro., *" The importance of stop word removal on recall values in text categorization,"* International Joint Conference, Vol 3, ISSN: 1098-7576, pp 1661–1666

[19] A. M. Kibriya.; E. Frank.; B. Pfahringer.; G.Holmes., *"Multinomial Naïve Bayes for Text Categorization"* International Conference on machine learning, pp 616-623.

[20] Hiral D. Padhiyar; Dilipsinh N. Padhiar., "Improving Accuracy of Text Classification for SMS " International Journal of Computer Applications, Vol 169, ISSN: 0975 – 8887, July 2017, pp. 19-21.

[21] J. Anvik, "Automating bug report assignment," Proc. .Intl. Conf.Software Engineering, ACM, 2017, pp. 937-940

[22] Sheetal A; Sable; P.N. Kalavadekar., "SMS Classification Based on Naïve Bayes Classifier andSemi-supervised

Learning" International Journal of Modern Trends in Engineering, ISSN: 2349-9745, 2016, pp. 561-564.

## About Authors

**Mehr-Un-NisaManjotho** received her B.E degree in Computer Systems Engineering Department from Mehran University of Engineering & Technology Jamshoro and also received M.E degree in Computer and Information Engineering from Mehran University of Engineering & Technology. Her research area is Text Classification.

**Prof. Dr. Tariq Jameel Saifullah Khanzada** is currently working as a professor at Computer Systems Engineering Department, MUET Jamshoro. He received his B.E degree in Computer Systems Engineering Department from Mehran University of Engineering & Technology Jamshoro and also received his Ph.D. from Germany. His research area is Wireless Communication System.

**Dr. Liaquat Ali Thebo** is currently working as anAssistant professor at Computer Systems Engineering Department, MUET Jamshoro. He received his B.E degree in Computer Systems Engineering Department from Mehran University of Engineering & Technology Jamshoro and also received his Ph.D. from Mehran University of Engineering and Technology, Jamshoro. His research area is computer networking.

**Engr. Ali Asghar Manjotho** is currently working as an Assistant professor at Computer Systems Engineering Department, MUET Jamshoro. He received his B.E degree in Computer Systems Engineering Department from Mehran University of Engineering & Technology Jamshoro and also received his Master's degree in Information Technology from Mehran University of Engineering and Technology, Jamshoro. His research area is computer security and machine learning.