# SMS Spam Detection for Indian Messages

Sakshi Agarwal
Department of Computer Science
Thapar University
Patiala, India
09sak.agarwal@gmail.com

Sanmeet Kaur
Department of Computer Science
Thapar University
Patiala, India
sanmeet.bhatia@thapar.edu

Sunita Garhwal
Department of Computer Science
Thapar University
Patiala, India
sgarhwal@thapar.edu

***Abstract*— The growth of the mobile phone users has led to a dramatic increase in SMS spam messages. Though in most parts of the world, mobile messaging channel is currently regarded as "clean" and trusted, on the contrast recent reports clearly indicate that the volume of mobile phone spam is dramatically increasing year by year. It is an evolving setback especially in the Middle East and Asia. SMS spam filtering is a comparatively recent errand to deal such a problem. It inherits many concerns and quick fixes from Email spam filtering. However it fronts its own certain issues and problems. This paper inspires to work on the task of filtering mobile messages as Ham or Spam for the Indian Users by adding Indian messages to the worldwide available SMS dataset. The paper analyses different machine learning classifiers on large corpus of SMS messages for Indian people.**

***Keywords*— *Mobile Phone Spam; SMS Spam; Spam Filtering; Supervised machine learning; Text classification.***

## I. INTRODUCTION

The **Short Messaging Service (SMS)**, commonly referred to as "text messaging" is a service for transmitting short length messages of around 160 characters to different devices such as cellular phones, smartphones and PDAs using standardized communications protocols [1]. SMS is used as an alternate for voice calls in positions where voice communication is either not possible or not desired between the end phone users. It is one of the most flourishing phone service engendering millions of dollars in perquisite for mobile operators yearly. Today's estimates signify that billions of SMS's are sent per day. According to the Portio Research [2], in 2010 the cost of total global cellular messaging market was around USD 179.2 billion which hiked to USD 200 billion in 2011, and has passed USD 253 billion till Sep, 2014. As the huge rise has occurred in SMS market, its profit has also increased in the direct proportion. Fig. 1 shows the increase in profit in SMS comparing with the decrease in profit in Email.

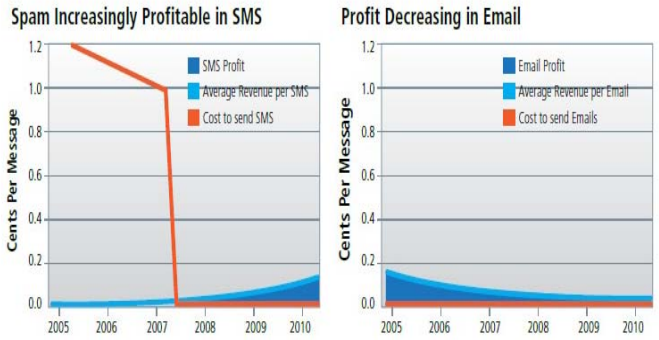


Figure 1. Profits comparison between SMS and email [2]

Spams are undesirable and unwelcomed messages which are sent electronically. These messages are sent by spammers for different ill wills of taking a hold over user's personal data or tricking them into the subscription of their premium tariff facilities. Such messages have the capability of imposing the same threats or even more dangerous aftereffects as of Email spams. In the zone of Email, though spam is a properly handled obstacle but SMS spams are increasing at a high rate of more than 500% in an year. It is an evolving setback especially in the Middle East and Asia.

According to the Cloudmark Report [3], "the amount of mobile phone spam is not same in every region. For instance, in North America, much less than 1% of SMS messages were spam in 2010, while in parts of Asia up to 30% of messages were represented by spam. Moreover 200 billion spam messages have been received by people of China in just one week of 2008."

The rest of this paper is structured as follows. Section 2 contains some important related works in the field of creation of SMS spam dataset as well as in the SMS spam detection. Section 3 discusses about the types of messaging attacks which are occurring throughout the world and which are needed to get eliminated. Section 4 offers details about the experimentation, describing about the SMS spam dataset that we have incorporated and describing the whole methodology of the process being carried. In Section 5, we display a performance assessments and calculations for comparing several existing machine learning algorithms. Finally, Section 6 presents the conclusion and outlines for future work.

## II. BACKGROUND

In today's scenario subscribers use phones for everyday communications, mobile applications and financial transactions are increasingly relying on their mobile phones due to which they have become tempting and magnetic target for message spam attackers. Such attackers try to make messaging attacks which have the capability to influence the entire mobile ecosystem henceforth creating a problematic situation for both operators and subscribers. For operators censorious traffic turns out to be really costly, losing worth network assets and hiking customer support. Subscribers are directly wedged by receipt

of malevolent messages and anticipate that their operators shall look into the matter to give them a safe mobile network.

Machine learning is basically the analysis of algorithms that can be trained by the data. It is a scientific branch of learning that delves into the construction of algorithms. Such algorithm works by constructing a model relying on the inputs and consuming that information for making predictions or decisions. Content based method consists of machine learning and its categorization is done below in Fig. 2.

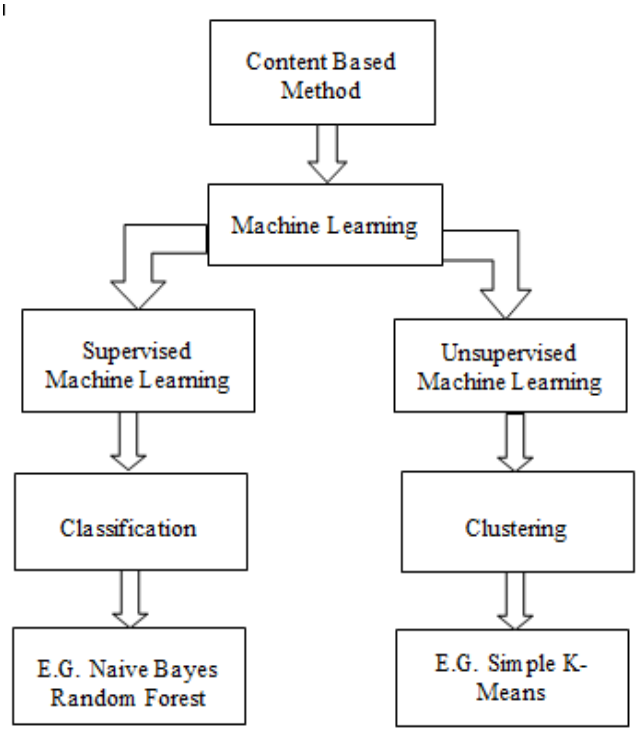


Figure 2. Categorization of Machine Learning

Classification is a primary subject while dealing with the subjects of machine learning and data mining. While performing the process of classification, the main aim of a learning algorithm is to create a classifier provided with a collection of training examples with class labels. In machine learning there are different classifiers which can be chosen on the basis of following characteristics:

- Computational cost
- Expected data types of features/labels
- Suitability for certain sizes and dimensions of data sets

Classifiers used in our experiments are listed below:

### *A. Multinomial Naive Bayes (MNB)*

Naive Bayes is one of the utmost proficient and helpful inductive learning approach for machine learning and data mining [4]. While doing classification, its reasonable presentation is amazing. Reason behind it is the conditional independence hypothesis on which it is based upon seldom exists in physical domain applications.

### *B. Support Vector Machine (SVM)*

SVMs attain significant enhancements over the presently finest performing approaches and perform stoutly over a diversity of various learning errands [5]. Additionally they are entirely involuntary, eradicating the requirement for labor-intensive factor.

### *C. Random Forest (RF)*

Random forests are an amalgamation of tree prognosticators in which every tree rest on the assessments of a haphazard vector sampled autonomously and with that same allotment for every single one tree of the forest [6].

### *D. Adaboost*

Adaboost approach works good for the higher noise levels and it seldom over fits for low noise regime [7]. Central to the understanding of this fact is the margin distribution. Adaboost can be considered as a restraint gradient depreciation in an error function with respect to the same margin.

## III. RELATED WORK

Gomez Hidalgo *et al.* had done a milestone work to detect mobile phone spam and assessed several Bayesian based classifiers [8]. In this work, the first two well-known SMS spam datasets, namely, the Spanish and English test databases were proposed by the authors. A number of message portrayal methods and machine learning approaches were tested by the authors on those two datasets. They came up with the conclusion that that Bayesian filter can be adequately utilized for the classification of SMS spam.

Cormack *et al*. evaluated that even content-based spam filtering cab be used for short text messages which occurs in three diverse perspectives: SMS, blog comments, and email summary information [9]. End of the line of their paper was that SMS is restricted to have less words for the sufficient support of words or word bigram based spam classifiers and thus by doing the expansion in the set of features to comprise orthogonal sparse word bigrams and character bigrams and trigrams, efficiency was increased.

Nuruzzaman *et al.* looked over the efficiency of sieving message spam on independent cellular phones using Text Classification approaches [10]. On an independent mobile various processing were done related to training, filtering, and updating. Their established outcomes display that the projected model was successful in distilling messages hams and spam with moderate efficiency, consuming less memory, and appropriate time was consumed while functioning without taking the help from a machine.

Coskun and Giura gave a network-based online detection technique for the identification of SMS spams campaign by taking the calculation of number of messages which were sent in single network over a small period of time and carry similar sort of data [11]. The approach given by them involved Bloom filters to keep a tentative count of message content occurrences.

Sarah Jane Delany *et al.* have worked on a clustering experiment on a SMS corpus [12]. In order to access the

behavior of SMS spam, they compiled 1353 spam messages and tried to use it as the dataset which comprehended of no duplicity. They applied k-way spectral clustering with orthogonal initialization. By applying spectral clustering on their own compiled dataset few clusters were produced which were ten in count with their linked top 8 terms and a presumed annotation.

Tiago A. Almeida *et al.* showcased the particulars of a new authentic, open and non-encoded SMS spam compilation which constitutes of maximum number of messages [13]. It is composed of 4,827 mobile ham messages and 747 mobile spams. Furthermore, the authors performed several established machine learning algorithms on their dataset and they came to the conclusion that according to them SVM is a better approach for advance evaluation.

Houshmand Shirani-Mehr applied different machine learning algorithms to SMS spam classification problem, compare their performance to gain insight and further explore the problem, and design an application based on one of these algorithms that can filter SMS spams with high accuracy [14]. They used a database of 5574 text messages.

## IV. IMPLEMENTATION

The cyber-crime has risen over the years. There are many techniques to deal with this Spam problem. All these techniques use different kinds of Spam filters. Basically all these filters classify the messages in to the category of Spam and non-Spam.

### *A. Dataset Used*

Since cellular messages repeatedly have a number of acronyms, it affects the efficiency of the filters. So a big and good message dataset is used in this process. The dataset provided by T. A. Almeida *et al.* [13] is already composed of 4,827 ham messages and 747 mobile spams, here we tried to convert the same dataset for Indian market by adding Indian spams and hams to the previous available messages. We added 439 legitimate messages and 748 spams from Indian perspective. The distribution of the messages is tabulated below in TABLE I. Legitimate messages are collected from volunteers, students attending the Thapar University, who were made aware that their contributions were going to be made publicly available. The spam messages which we have used are provided us by an Indian Network Operator company.

TABLE I. DATA DIVISION

| Messages | Amount |
|---|---|
| Hams | 4,827 |
| Indian hams | 439 |
| Spams | 747 |
| Indian Spams | 748 |
| **Total** | **6761** |

### *B. Methodology*

This section describes the general design of workflow of the experiment. In this experiment machine learning tool is used for the analysis and classification of the dataset. At the first level data is gathered from different sources to create a good dataset of ham and spam in text format and give that data as the input in model. At the second level of the experiment we converted the data set which is earlier in the text format to CSV (Comma Separated Value). Then preprocessing is done for a better quality input by implementing various feature extractions techniques. The labeled data is opened and the attributes are listed. The attributes that are used for the analysis purpose are text and class in this dataset. After that a classifier is applied to the data set we have used. Thus the data is trained using the dataset. Testing is done on the data to get results. Finally at the last step of the experiment Confusion Matrix is obtained from the data set, and the results of the applied classifier are analyses and discussed.

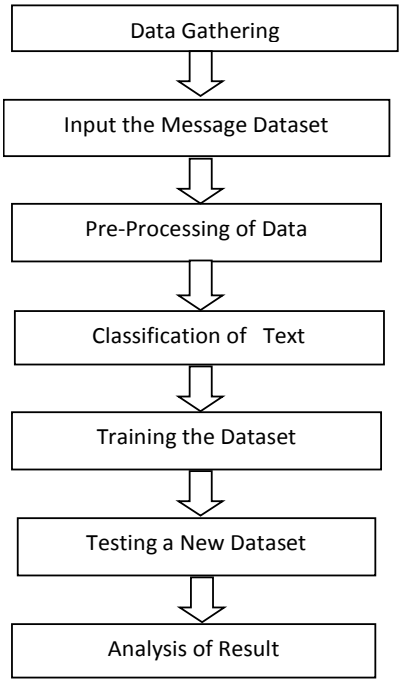


Figure 3. Flowchart of Processes of Spam Filtering

### *C. Evaluation Metrics*

The metrics measure the percentage of Spam detected by the system and how many misclassifications it makes. Few of the evaluation metrics are:

*1) True Positive (TP):* When positive instances are correctly classified it is reported by a number called True positive.

*2) False Positive (FP):* When positive instances are incorrectly classified it is reported by this number called False positive.

*3) False Negative (FN):* When negative instances are incorrectly classified it is reported by this number

*4) True Negative (TN):* When negative instances are correctly classified it is reported by this number

The above mentioned basic metrics were used for further calculations of various metrics. For evaluating the performance of spam detection system, we measured accuracy (ACC), Spam Caught (SC), Blocked Hams (BH), Matthews Correlation Coefficient (MCC), F-measure, precision, recall, Area Under the ROC Curve (AUC), True Positive Rate (TPR) and time consumed by the classifier.

*1) Accuracy (ACC):* Accuracy can be defined as the proportion of correctly classified classes namely True Positive and True Negative over the total number of classifications as depicted by formula below:

$$\text{Accuracy} = \frac{(TN+FP)}{(TN+TP+FN+FP)} \times 100$$

*2) Precision (P):* Precision is the fraction of the messages retrieved that are relevant for the user.

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

*3) Recall (R):* Recall is the fraction of the successfully retrieved messages that are relevant to the query.

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

*4) F-Measure*: It serves as the harmonic mean of precision and recall.

$$\text{F-measure} = 2 \times \frac{P \times R}{P+R}$$

*5) Matthews Correlation Coefficient (MCC):* MCC is used to determine the quality of classification for two classes even when size of the classes varies. It ranges between -1 and +1 where +1 representing the finest performance

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$$

*6) Spam Caught (SC):* It is the ratio of number of caught spams to the total number of spams exists in the dataset.

$$SC = \frac{\text{False negative data}}{\text{Total number of spams}}$$

*7) Blocked hams (BH):* It is the ratio of number of hams that are detected as spams to the total number of hams exists in the dataset.

$$BH = \frac{\text{False positive data}}{\text{Total number of hams}}$$

*8) Area Under Curve (AUC):* A Receiver Operating Characteristic (ROC) curve is a graphical plot which is used to represent the performance of a binary classifier system. It is plotted using the values of True Positive Rate on the vertical axis and False Positive Rate on the horizontal axis of the curve. The area covered under this curve is known as Area Under Curve (AUC).

## V. RESULTS AND DISCUSSION

Various classification techniques were applied on the dataset: Our Altered SMS Spam Collection Data Set including Indian content. By applying different classifiers the best and the worst classifier can be judged. After applying various combinations of feature selections and extractions, their performances were judged. Preprocessors like stemming, removal of small length words; stopwords etc. are the few main preprocessors which have been used. Count Vectorizer and TF-IDF Vectorizer are used for feature extraction in our research. Count Vectorizer converts a collection of text documents to a matrix of token counts and TF-IDF Vectorizer Convert a collection of raw documents to a matrix of TF-IDF features which is equivalent to Count Vectorizer followed by TF-IDF Transformer.

Best results of each of the four classifiers on Our Altered SMS Spam Collection Data Set including Indian content are tabulated in TABLE II. The accuracy of different classifiers is plotted in Fig. 4 for various features. In the graph, X-axis shows combinations of different features and Y-axis shows the accuracy percent of different classifiers for various features. The input for all the classifiers is our altered spam collection data set and the output is various performance metrics, out of which accuracy is plotted in graph.

TABLE II. BESTRESULTS OF VARIOUS CLASSIFIERS ON THE DATA SET INCLUDING INDIAN CONTENT

| | MNB | RF | SVM | Adaboost |
|---|---|---|---|---|
| ACC% | 97.87 | 96.04 | 98.23 | 96.21 |
| BH% | 0.82 | 0.14 | 0.55 | 1.61 |
| SC% | 92.09 | 78.76 | 92.89 | 85.88 |
| F-Measure | 0.978 | 0.959 | 0.982 | 0.962 |
| MCC | 0.929 | 0.863 | 0.941 | 0.865 |
| Precision | 0.978 | 0.962 | 0.982 | 0.961 |
| Recall | 0.979 | 0.960 | 0.982 | 0.962 |
| Time(secs) | 2.031 | 26.845 | 443.977 | 515.213 |
| AUC | 0.956 | 0.893 | 0.962 | 0.921 |
| TPR | 0.979 | 0.960 | 0.982 | 0.962 |

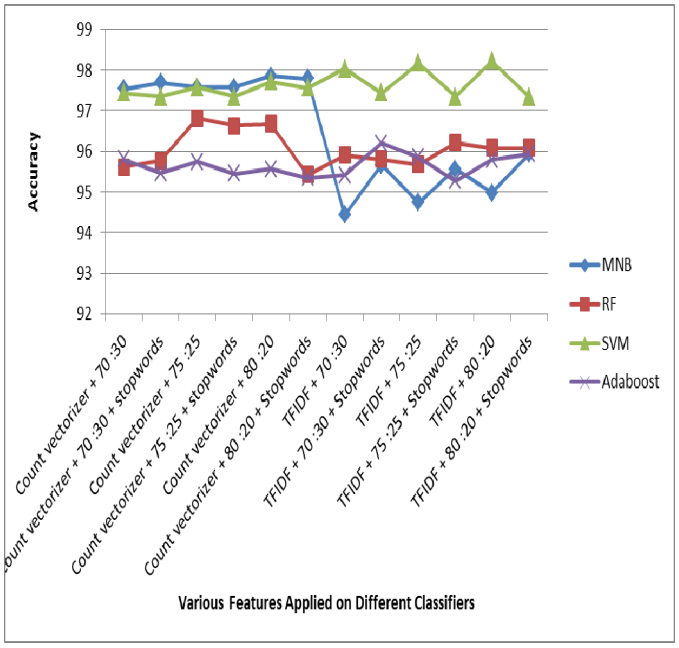


Figure 4. Comparison of Various Features with Different Classifiers

## VI. CONCLUSION AND FUTURE WORK

We are living in the age of information and technology. Now people are going away from the traditional ways of communication. These ways of information and communication are facing a serious and irritating problem known as spam. Various classification techniques were applied on the dataset: Our Altered SMS Spam Collection Data Set including Indian content. The results showed that Support Vector Machine and Multinomial Naive Bayes are among the best classifiers for the SMS spam detection. The classifier SVM with Linear kernel had the best accuracy but the time required for the process is high. On the other hand MNB with Laplace smoothing also had its accuracy very close to SVM but the time taken by MNB was far lesser than SVM. The best results of Altered SMS Spam Collection Data Set including Indian content came out to be 98.23% of ACC%, 92.88% of SC% and 0.54 % of BH% with SVM**.**

Future work must practice several approaches to escalate the aspect of the feature plot. Adding more meaningful features like certain thresholds for the length and analyzing the learning curves can contribute to the improvement in results. From the perspective of practical implementation, particularly for Indian mobile users, comparative analysis of spam detection and prevention from spam messages will bring a bright future for messaging industry. An application for the smartphones using this technique, especially for the India, can be developed for protecting the cell phones from the harmful spam messages as India is among the top countries of receiving the spam SMS messages.

## REFERENCES

[1] SMS (Nov 2010). [Online]. Available: http://searchmobilecomputing.techtarget.com [Accessed: Oct 2014].

[2] J. M. G. Hidalgo, T. A. Almeida and A.Yamakami. "On the validity of a new SMS spam collection." In *Machine Learning and Applications (ICMLA), 11th IEEE International Conference*, vol. 2, 2012.

[3] T. A. Almeida , J. M. G. Hidalgo and A. Yamakami, "Contributions to the study of SMS spam filtering: new collection and results." In *Proceedings of the 11th ACM symposium on Document engineering*, 2011.

[4] H. Zhang, "The optimality of naive Bayes." *AA* 1, vol. no. 2, 2004.

[5] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", *Springer Berlin Heidelberg*, pp. 137-142, 1998.

[6] L. Breiman, "Random forests", *Springer Berlin Heidelberg*, vol. no. 43, pp. 5-32, 2001.

[7] G. Ratsch, T. Onoda and K. R. Muller, "Soft margins for AdaBoost", *Springer Berlin Heidelberg*, vol. 42, issue 3, pp. 287-320, 2001.

[8] J. M. G.Hidalgo, G. C. Bringas, E. P. Sanz and F. C. García, "Content Based SMS Spam Filtering," in *Proceedings of the 2006 ACM Symposium on Document Engineering*, Amsterdam, The Netherlands, 2006, pp. 107–114.

[9] G. V. Cormack, J. M. G. Hidalgo and E. P Sanz,"Spam Filtering for Short Messages," in Proceedings of the 16th ACM Conference on Conference on information and Knowledge Management, Lisbon, Portugal, 2007, pp. 313–320.

[10] M. Taufiq Nuruzzaman, C. Lee, M. F. A. bin Abdullah and D. Choi, "Simple SMS spam filtering on independent mobile phone," *Security and Communication Networks*, vol. 5, no.10, pp. 1209–1220, 2012.

[11] B. Coskun and P. Giura, "Mitigating SMS spam by online detection of repetitive near-duplicate messages," in *IEEE International Conference on Communications*, 2012, pp. 999–1004.

[12] S. J. Delany and M. Buckley, "Expert Systems with Applications," *Expert Systems with Applications*, vol. 39, pp. 9899-9908, 2012.

[13] T. A. Almeida, J. M. G. Hidalgo and T. P Silva, "Towards SMS Spam Filtering: Results under a New Dataset," *International Journal of Information Security Science,* vol. 2, no. 1, 2013.

[14] H. Shirani-Mehar, "SMS Spam Detection using Machine Learning Approach.", *International Journal of Information Security Science,* vol. 2, no. 2, 2014.