ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS FOR SMS SPAM DETECTION

[1] Nithish D S, [2]Naveen K S, [3] Rashmi R, [4] Sharadadevi S Kaganurmath
[1]Student, [2]Student, [3] Assistant Professor, [4] Assistant Professor
[1]Information Science
[1]R V College of Engineering, Bengaluru, India

Abstract : There has been a dramatic increase in number of mobile phone users in recent years. It has brought anything and everything at the tip of fingers. Five billion people globally send and receive SMS messages. That's about 65% of world's population. At the same time, irrelevant and unnecessary information have also dominated the content that flows through the network. There is a serious need to reduce this SMS spam which disturbs the mobile users a lot.

This paper analyses some of the algorithms which are applied in detecting SMS spam detection. Different algorithms were applied on the same dataset to determine their accuracy in spam detection. The results clearly indicate that different ML algorithms tend to perform differently in classifying spam messages. Certain conclusions are drawn based on their results and future enhancement is discussed.

***IndexTerms* -** Mobile Phone Spam, SMS, Feature extraction, Classification, Bag of words

I. INTRODUCTION

Text messaging has enormously increased in market in the recent past. Although SMS spam is less prevalent than email spam, it is 1% of texts sent in US and 30% of SMSs in Asian countries. In the United States, SMS spam messages have been illegal under the Telephone Consumer Protection Act since 2004. Citizens who receive unsolicited SMS messages can now bring the solicitors to small claims court. Short messaging services (SMS) are one of the important things that bridge the communication among millions of people around the world.

According to statistics brain research institute, the number of monthly texts sent has increased by more than 7,700% over the last decade [1]. "Over 83% of millennial consumers informed that they usually text more than talking on their smartphones", as per GFK study [2]. Flow route nationwide survey has found that 58% of consumers indicated they prefer a business which offered SMS capabilities [3].

Spam has also increased parallelly to the good use of SMS. It is the use of messaging options to send unwarranted and unsolicited bulk messages, especially marketing. According to a survey by Tatango blog, 68% of survey respondents said they received text message spam and also that people are likely to be the recipients of text message spam irrespective of gender [4]. Figure 1 shows this statistic in graph also with respect to age.

The spam messages that come from various senders like our own service providers ,the merchants where we usually buy something, unknown people and bots cause a lot of annoyance with various kinds of messages like "Amazing friends that are close to you. Call 5630035", "Spicy talks with lovely friends. Call 5630066", "To Get interesting LOVE tips for your loved ones, Dial 504042 at Rs.6/Min" etc,. To solve this problem, there is no better way than Machine learning. A key problem that has significant amount of data and a proper way to apply the machine learning approach gives a reasonable solution the problem.

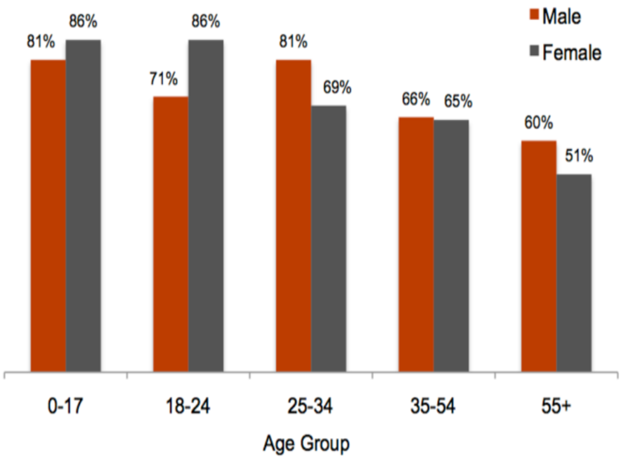

Figure 1 Recipients of Text Message Spam

This paper is structured as follows. Section II gives more background information on mobile phone spam and machine learning approach. Section III discusses the related work in the domain of machine learning and its application in case of SMS spam detection. Section IV gives brief details on the dataset used and about the methodology employed to conduct this research. Section V presents results and analysis of the effectiveness of different algorithms. Section VI gives a conclusion and briefly explains the future scope of the work.

II. BACKGROUND

In the previous section, SMS and SPAM SMS have been introduced briefly, this section will provide more insight on the increase of SPAM SMS and also discuss about machine learning approach. This will also briefly explain on Bag of Words Model which is an important algorithm which is used in feature extraction. There are three different kinds of implementation for Bag of Words model which is discussed.

2.1 Geographical Origin

In 2011 the origins of spam were analysed by Cisco Systems. They provided a report that shows spam volume originating from countries worldwide. Table 2.1 gives details on the geographical origin of spam messages.

Table 2.1: Geographical origin of SMS spam

Rank	Country	Spam Volume (%)
1	India	13.7
2	Russia	9.0
3	Vietnam	7.9
4	South Korea	6.0
5	Indonesia	6.0
6	Chine	4.7
7	Brazil	4.5
8	USA	3.3

2.2 Reason Behind increase of mobile spam in India

According to an article in Medianama, The problem in India is that the network companies' trade has scaled marginally. This has increased the competition, brought new entrants into the market, eventually bringing down the prices. There are lot of people who have took undue advantage of this fall down of prices to turn the bulk SMS service from useful to spam. The below figure indicates that cost of sending Bulk SMS has become incredibly low in India [5].

Table 2.2: SMS packs rates in last two years

SMS Credits	Prince in INR	Per SMS Rate	Validity
5	Free	Free	30 days
1 Lakh	3500/-	3.5P	Unlimited
3 Lakh	9000/-	03P	
10 Lakh	12500/-	2.5P	
10 Lakh	20000/-	02P	Unlimited
25 Lakh	45000/-	1.8P	Unlimited
50 Lakh	80000/-	1.6P	Unlimited
1Cr and Above		1.5P	Unlimited

2.3 Machine Learning

2.3.1 Support Vector Machines

A Support Vector Machine (SVM) is a classifier algorithm which is defined by a separating n-dimensional spaces. SVM provides an optimal n -dimensional spaces as output which categorized new examples for given training data set. SVM constructs n-dimensional space which can be used for regression or classification. SVM Algorithm is used for solving both linear and non-linear classification problems.

2.3.2 Random Forest

Random forest algorithm also called as random decision forests are a group of learning methods, during training time it constructs a multitude of decision tree for operation. This algorithm outputs mean prediction of the tree. In regression or classification problem Random forest can be used to rank the importance of the variable.

2.3.3 Adaboost

Adaboost is a type of algorithm used to improve the performance, which can used with other algorithms. For preparing haar-like list of capabilities. Gentle AdaBoost algorithm is used to improve hub classifier capacity. Thus, the face discovery execution of the face detector is improved.

2.3.4 Multinomial Naive Bayes (MNB)

Multinomial Naive Bayes (MNB) is a probabilistic classifier based on applying Bayes theorem. For constructing classifier naive Bayes is the easiest technique that assign class labels to problem instances. MNB is a conditional probability algorithm it assigns given problem instance with n features it assigns instance probabilities.

2.3.5 K-Nearest Neighbors

K-Nearest Neighbors is a algorithm which is normally used for regression and classification which uses non-parametric method. K closest training examples are fed as input for algorithm. For better output assign weight to the contribution of neighbors, so the neighbor which is far has less contribution and the neighbor which is near has more contributions.

2.3.6 Logistic Regression

Logistic regression algorithm is statistical based algorithm used in various kinds of applications and generic scenarios which may cover social sciences, genetic engineering, medical application, machine learning and artificial intelligence. Logistic regression is basically a supervised classification algorithm.

III. RELATED WORK

Gomez Hidalgo and others have done tremendous research and experiment on SMS spam filtering. They emphasize on attribute selection to help classification of spam messages. Their work concludes that support vector machines is by far the suitable algorithm based on running time [5].

H S Mehar applied different machine learning algorithms and found that different features are required for different classifiers. It also states that combination of features would lead to good results and Tf-Idf vectorizer among others yields better accuracy [6].

Duan and Huang have explained the dual filtering approach which makes use of KNN classifier algorithm and rough set in order to classify messages as ham or spam. This resulted in less time for classification while accuracy was still retained [7].

QUARTZ INDIA’s survey reports that “Around 96% of Indians receive unwanted text messages (SMS) every day, according to a survey of over 12,000 people by online community platform Local Circles. Nearly half of the respondents get between four and seven such messages in a day” [11].

Coskun and Giura present a network-based dynamic and online SMS spam detection technique by calculating the number of messages sent in single network over a period of time and which have same kind of data [10]. Their solution had bloom filters to have the tentative count of message occurrences.

Freund,Y, Schapire R. E.,& Hill.M have showed how a learning algorithm can be boosted to yield greater results. They demonstrate the same with adaboost and conclude that boosting along with complex algorithm can improve performance especially when data is huge [8].

T. A. Almeida, J. M. G. Hidalgo and A. Yamakami have made strong points about the problems that hinder the development in this research field and thus present a large mobile spam collection, making it publicly available. Their work indicates that SVM performs better than any other classifier [9].

IV. METHODOLOGY

4.1 Dataset Used

After a detailed research on data, finally a dataset having 5574 classified SMSs was used. This dataset is publicly available in kaggle website, which was originally collected from grumble text website, NUS SMS corpus etc., and collaborated in UCI machine learning repository. Different machine learning algorithms/classifiers are used. The language in these messages is English and there are two columns namely message and class. The class can either be ham or spam. Around 30% i.e. 1600 messages were used for validation purpose and the other messages for training the model.

4.2 Experimental Flow

From different sources data was collected and studied which yielded in selecting the dataset already mentioned before. Once the dataset is ready, it is fed to the model. For better output and more accuracy, a lot of preprocessing is done for the data. Then pre-processed data is used to extract and create features using either of vectorizers in scikit learn library.

In the next step, a classifier is applied to the dataset and then trained thoroughly. In this particular dataset, text and class are the attributes. Validation of the model is done using the test data already mentioned before. Once the classification is done, results from the classifier are analyzed, and those results are compared with the previously applied algorithms.

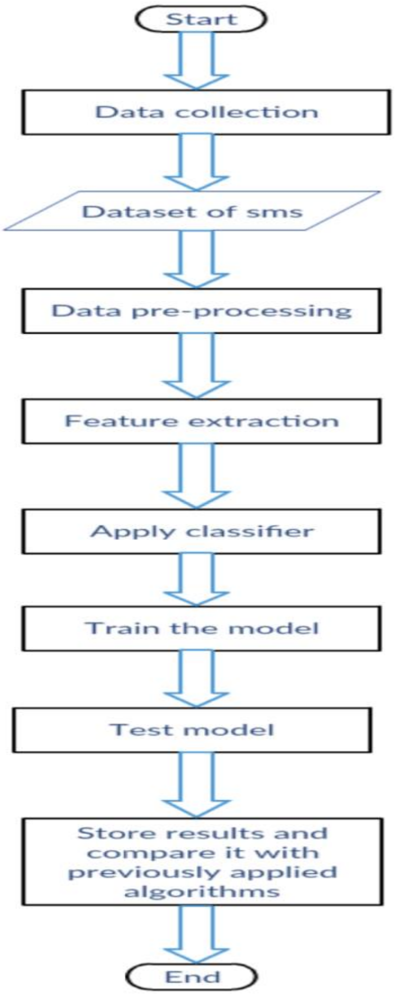

Figure 4. Flowchart of methodology of experiment

V. RESULTS AND ANALYSIS

Different classification algorithms are applied on the data set and have different accuracies. From the results we can say that Support vector machine algorithm and Multinomial Naive Bayes algorithm are best classifiers to detect spam SMS. K nearest neighbor's algorithm has the lowest accuracy among all the classifiers used. Logistic Regression classifier does better than k nearest Neighbors with accuracy of 95.99% but performs worse than the other four classifiers applied to dataset. The result of the classifier algorithms used are tabulated in Table 3.

Table 3. List of algorithms and respective accuracies

Algorithm	Score
Naive Bayes	0.989833
Support Vector Machine	0.985048
K nearest Neighbors	0.944976
Random Forest	0.980263
Logistic Regression	0.959928
Adaboost	0.980861

The accuracy of the different classifiers is plotted in Figure 2. In the graph X-axis shows different classifiers applied and Y-axis shows the accuracies of classifiers.

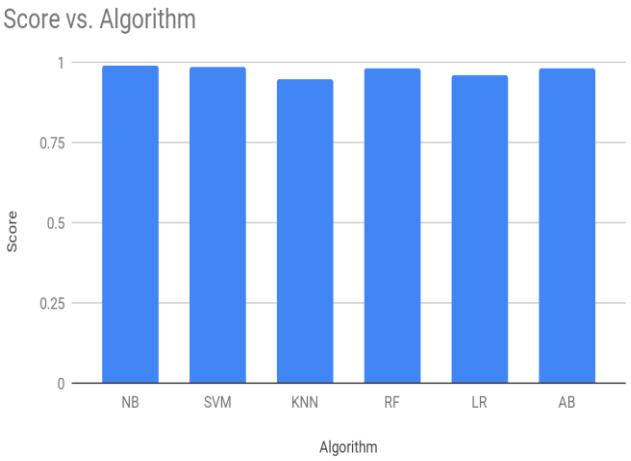

Figure 2. Accuracy graph for different classifiers

It was also noticed that the length of a message can be a distinguishable feature as the spam messages are lengthier than ham messages.

VI. CONCLUSION

The current generation is very much used to Digital Media where everything is available at fingertips. Although there are a lot of advantages of having the option to send SMS to reach out to masses, a serious concern arises when some organizations or people take undue advantage of this feature. SMS SPAM continues to annoy mobile phone users by distracting them from seeing genuine messages received.

Classification algorithms are applied to the dataset for SMS Spam detection keeping the complexity and effectiveness of each algorithm in mind. Accuracy with which the classifiers classify the SMS as spam or ham was the main evaluation metric. Multinomial Naive Bayes classifier clearly does classification better than any other classifier. Even though different algorithms prove their efficiencies in detecting spam, the integration of these mechanisms at the root of mobile phones is not completely successful.

6.1 Future enhancement

The current works done on SMS spam detection don't recognize false positive cases much. Although many mobile manufacturers have implemented some design to separate spam SMS as 'Notifications', usually which are sent by senders not in our contacts, the problem is that the real worthy message which the user needs are also going in 'Notifications' bunch. This causes a serious problem to the user especially when he immediately wants to check a part of that worthy message but its in the group of spam messages. It will be extremely difficult to do that when searching option is also limited.

Example of false positive: "KSRTC Bus : PNR :J76216971,Journey Date :23-Apr-2019 16:05,Trip Code :1605DVGBNG,Bus Number :F1829,Depot :DAVANAGERE, Crew Mobile No:9113997433 . Happy Journey."

When there is a need to open this message and check the PNR but this message is in spam collection, it annoys to a great extent. Hence, its very much necessary to work on reducing or eliminating the false positive cases.

REFERENCES

[1] Statistic Brain. (2017, September). Text Message Statistics. Retrieved from: https://www.statisticbrain.com/text-message-statistics/

[2] Press release | GFK (2016, February). Retrieved from: https://www.gfk.com/en-us/insights/press-release/smartphone-users-spend-as-much-time-on-entertainment-as-texting-gfk-mri-study/

[3] Text Message Spam Statistics | Tatango. (2011, August). Retrieved from: https://www.tatango.com/blog/text-message-spam-statistics/

[4] Nikhil Pahwa. Why SMS Spam Has Increased In India: Business Model Changes – MediaNama. Retrieved from: https://www.medianama.com/2010/09/223-sms-spam-increase-india/

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

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

[7] Duan, L., Li, N., & Huang, L,"A new spam short message classification" 2009 First International Workshop on Education Technology and Computer Science,16S-171.

[8] Hill M and Freund Y. Schapire R. E."Experiments with a new boosting algorithm". Thirteenth International Conference on Machine Learning,San Francisco,14S- 156.

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

[10] BP Giaura and Coskun“Mitigating SMS spam by online detection of repetitive near-duplicate messages,” in IEEE International Conference on Communications, 2012, pp. 999–1004.

[11] Kuwar Singh.. Telecom, realty firms, banks send most SMS spam in India — Quartz India. Retrieved from https://qz.com/india/1573148/telecom-realty-firms-banks-send-most-sms-spam-in-india/