SMS spam detection and comparison of various machine learning algorithms

Paras Sethi
B.tech
Computer Science and Engineering
HMR Institute of technology and Management New Delhi, India
E-mail:
Parassethi234@gmail.com

Vaibhav Bhandari
B.tech
Computer Science and Engineering
HMR Institute of technology and Management New Delhi, India
E-mail:
Vaibhav.vb24@gmail.com

Bhavna Kohli
B.tech
Electronics and Communication
HMR Institute of technology and Management New Delhi, India
E-mail:
Bhav2411@gmail.com

ABSTRACT Past few years have seen increase in the number of spam emails and messages. Legal, economic and technical measures can be used to tackle spam sms's nowadays. A key role is being played by Bayesian filters in stopping this problem. In this paper, we analyzed and studied the relative strengths of various machine learning algorithms in order to detect spam messages which are sent on mobile devices. We have acquired the data from on open public dataset and prepared two datasets for our testing and validation purposes. Accuracy in detecting spam messages was the first priority in ranking these algorithms. Our results clearly demonstrate that different machine learning algorithms under different features tend to perform differently in classifying spam messages.

Keywords: **Natural Language Processing (NLP), spam detection, machine learning.**

1.INTRODUCTION

Since the year 2001, mobile spam messages have been a consistent problem in Far East countries. In fact, it was quite amazing to note that the amount of spam messages exceeded the number of spam emails. In some countries SMS spam problem was so disturbing, that legal action was taken in order to stop this practice. Two acts were filed by Japanese government in the year 2002 for definition and penalization of email and mobile abuse. With the effect of these laws, a huge relief has been obtained in the masses regarding spam message abuse and eradicated some of the problems in the society. Experts also feel that technical and legal measures need to be taken in order to control this widespread disturbing abuse of mobile spam messages. Since the effort and expense of sending a spam message is more than the cost of sending spam emails, this practice is still considered a minor problem in Western countries. But in Europe, where a user sends an average of 10 texts a day and nearly most of the population owning a smartphone, spam message is the fashion.

In fact, in countries like Russia, real users are being simulated and emulated by botnets and zombie PC systems when sending a message through free messaging service . So, we can clearly assume that there has been a drastic decrease in the cost of SMS spam. In other words, mobile spam have a good return on investment (ROI). There has been a constant discussion on this topic, and people have come up with concrete technical measures in order to tackle this problem. Most of these measures and practices can be used in order to deal with SMS spam messages. Bayesian filters have been most prominent ones and most widely accepted. These methods discriminate and classify legitimate messages from the pile of normal and SMS spam messages with a clever use of machine learning algorithms. This paper is based upon testing the results obtained with Bayesian filtering and compare those results with other machine learning algorithms in order to find the best alternative that would be useful in classifying SMS spam messages. We, also proposed some feature sets that could be utilized by different machine learning algorithms in order to classify those text messages and we present our findings, with accuracy of the algorithm being the deciding factor in choosing the best method.

2. TECHNICAL MEASSURES AGAINST SPAM EMAIL

Activities like phishing, selling drugs and advertising pornographic sites use on an average of millions of spam email messages [1]. There has been a drastic disturbing increase in the growth of this problem in the recent years which has sparked a keen motivation in scientists all over the world to take a keen interest in this problem and come up with good solutions to tackle this problem.

2.1 Spam Email Filtering

Following ones are the best techniques nowadays to reduce the amount of spam messages.
White and black listing. The messages coming from people that are considered spammers are blocked. This is made

possible by blacklisting those people. However, this does not disrupts the normal flow of messages as the other persons are still whitelisted, and their messages received are acted upon normally by a mobile device.

Address management. This phenomenon rests upon a system, Which contains automatic addresses that are produced by machines that are generated at some particular time intervals. [3].

Collaborative filtering. This type of algorithm is use in recommendation systems where when one user marks one thing as “something”, the other user gets some information based upon the tagging of first user. This can be applied here also : if one person marks message as “spam”, then other person can be aware while opening that message. [4].

Digital signatures. If there is no digital signature present in the message, it is safe to assume that it is a spam message. Service provider can be used to obtain the digital signatures of the sender [4].

Content-based filtering [6]. Here, a careful approach is taken. A well annotated dataset is maintained of spam messages and analysis is done by observing some patterns in those messages. Then machine learning models are built on top of those annotated classified messages.

3. EXPERIMENTS AND OBSERVATIONS

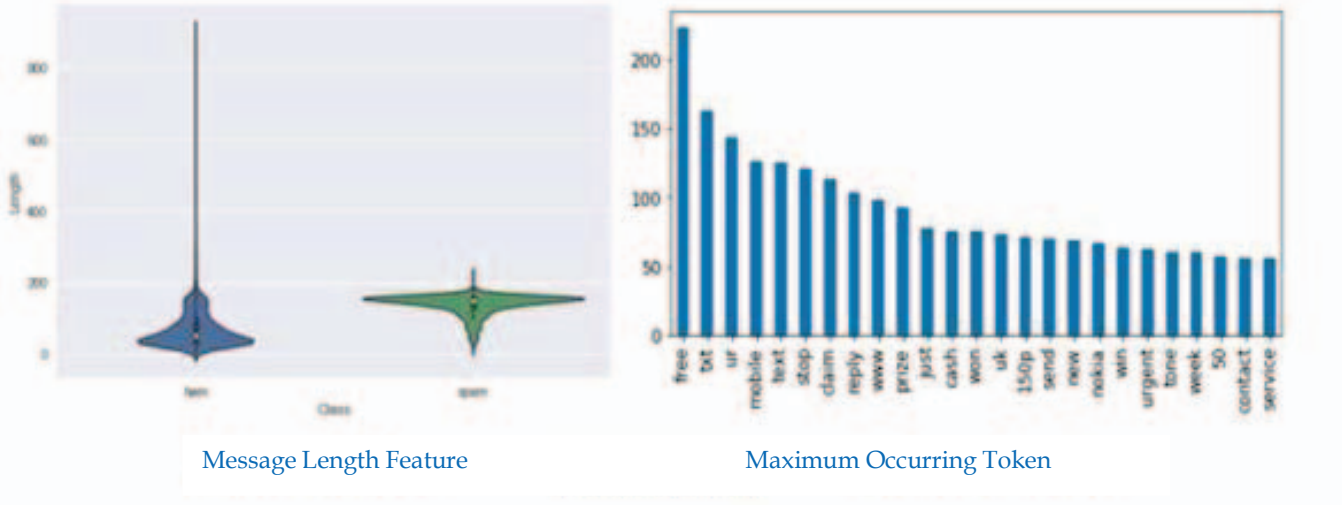
We performed some tests and analysis with various machine learning algorithms with different feature sets obtained and finally, we have represented our findings by ranking different machine learning algorithms in order of their accuracy so as to determine the best possible machine learning algorithm which will be able to classify SMS spam messages with highest accuracy.

3.1 Corpus

The database which was used for this research is currently publicly available on Kaggle under the name "SMS spam collection database". Kaggle is a website which is an transparent repository for public datasets and this dataset was provided to the website by the UCL machine learning repository and other public open datasets.

3.2 Database Description

Our database is a freely available dataset of 5,574 classified short messages (SMS’s) that are real and non-encoded. The language in these messages is English and they are labelled as ham or spam. Firstly, we tokenized the text and obtained the IG matrix. We also removed all the stop words from the text which might impact the modelling procedure, as the stop words don’t have any impact on the model.



Message Length Feature

Maximum Occurring Token

3.3 Feature Selection

In a data mining inspired approach, we decided to feed the algorithm basically two most important features: the message length and the count vectorizer matrix. We use length of the message as attribute quality metric. During our exploratory analysis we figured out that messages with spams have different message lengths as compared to other non-spam messages. This is evident from the plot shown(*fig.1*) , where we can easily see that the spam messages tend to have a mean length of about 176 characters while the ham messages tend to have a message length of about just 55 characters .We use Information Gain (IG) [6, 7](*fig.2*) as attribute quality metric. The experience in learning-based text classification is that IG can reduce substantially the number of attributes, without no loss (or even some improvement) of accuracy.

3.4 Machine Learning Algorithms

For our experiments we have used the following algorithms:

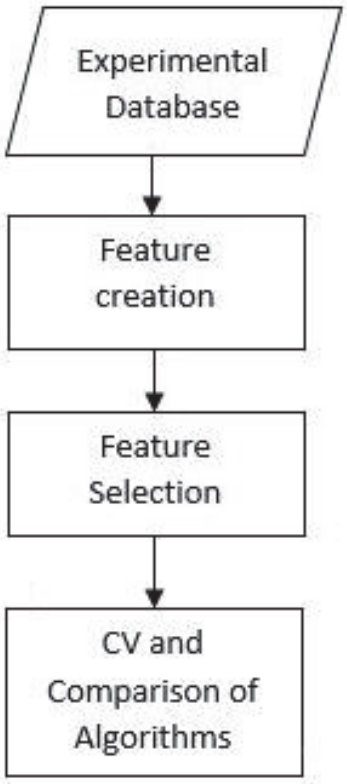
• *Naive Bayes (NB):* This is the most immediate and simple Bayesian learning method. It is based on the Bayes theorem and they learn parameters by observing each feature independently, regardless of all the other features in the data set and derive statistics from each class for each class.They also provide generalization performance that is slightly more worse than that of linear classifiers.

• *Random Forests:* A random forest is just and ensemble of various decision trees grouped together,which is used in order to remove the provlem of overfitting in decision trees.The main idea behind this algorithm is that each tree is capable of producing predictions that are sligghlty different from other trees.By producing different results, they tend to perform differently from each other and in the end , we generalize their results by averaging them.

• *Logistic regression:* Logistic regression is one of the most basic algorithms used for binary classification and is a linear algorithm.The outcome here is decided with the help of a dichotomous variable(which may possiblty has 2 outcomes). Instead of selecting parameters that minimize the sum of squared errors (like in ordinary regression), prediction in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

1) Experimental Database: This step mostly comprised of selecting our primary dataset and involved applying data preprocessing techniques that can be used to make the database more understandable and develop some extra information about the dataset.

3.4 Proposed Model



Detailed description of Phases

2) Feature Creation: This is the portion where we create some extra meaningful information about the database which can be used to easily help us classify the text as either spam/non-spam. We were able to craft two features that were able to very easily differentiate between spam and other non-spam messages. These were message length and information gain matrix.

3) Feature selection: Here, we fed all the above features as well as the raw data to all the algorithms and we asked the algorithms to select features that were best for its use. The algorithm was fed all the features and the features selected depended from one algorithm to another.

4) CV and Comparison of Algorithms: Here, a 5-fold cross validation technique was used to prevent the algorithms from overfitting and we wanted to get a very good precision of the accuracy of the algorithm working underneath. Also, in the end we compared different algorithms according to their working

3.5 Experimental Results

We fed following information to all the 3 algorithms: raw data, information gain matrix and the message length of the message. After the parameter optimization using a cross validation technique with 5-folds we obtained following results and accuracies with our models (*fig.3*). We gave full choice to the algorithm to pick its own features and it could either discard both, and just pick the raw text, or go with either of them or go with both. All the algorithms didn't pick all the features as the naïve Bayes algorithm performed best with a marginal difference than its variant with features and picked just the information gain matrix, while the other two algorithms went for all the features.

4. Conclusion

From the following set of experiment, we could say that naïve Bayes outperforms random forest algorithm and logistic regression algorithm when it comes to classifying SMS spam. The naïve Bayes algorithm achieved a high accuracy of 98.445 % just with the information gain matrix and easily classified the text as either spam/non-spam.

Although we haven't provided detailed analysis on the running times of these algorithms but we checked that naïve Bayes also has running time which is the least among these algorithms. We can also say that Random forest algorithm using both the features also performed excellently and could very well be a good alternative. We have experimentally verified that Naïve Bayes algorithm outperforms Random forest and Logistic regression algorithm.

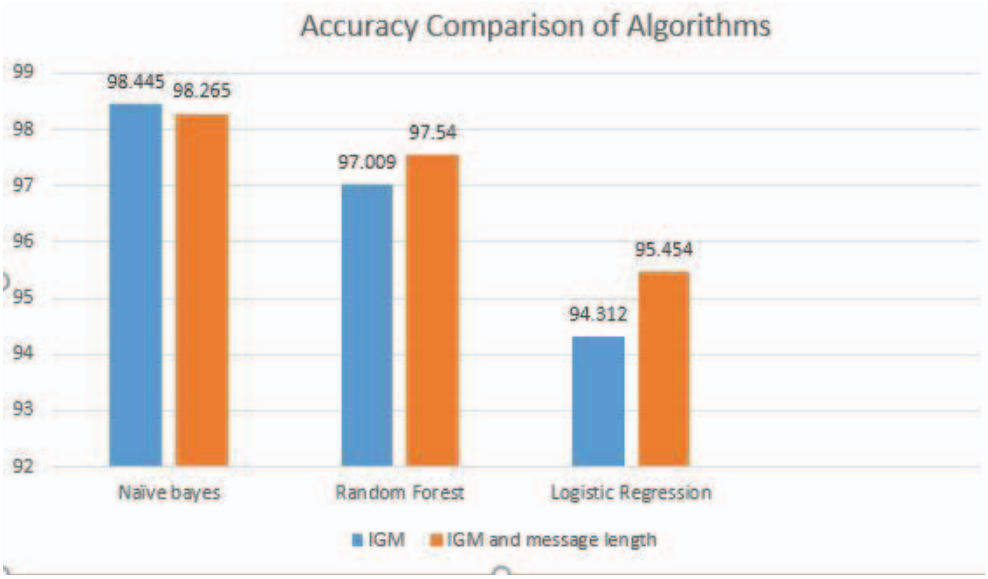


Fig.3 Comparative study of different machine learning algorithms

REFERENCE

[1] Christine E. Drakeand Jonathan J. Oliver, Eugene J. Koontz. Anatomy of a Phishing Email. Proceedings of the First Conference on Email and Anti-spam (CEAS), 2004.

[2] Dwork, C., Goldberg A., Naor M.. On memory-bound functions for fighting spam. In Proceedings of the 23rd Annual International Cryptology Conference (CRYPTO 2003), August 2003.

[3] R.J. Hall. How to avoid unwanted email. Communications of the ACM, March 1998.

[4] Golbeck, J., Hendler, J. Reputation network analysis for email filtering. In Proceedings of the First Conference on Email and Anti-Spam (CEAS), 2004.

[5] Tompkins T., Handley D. Giving e-mail back to the users: Using digital signatures to solve the spam problem. First Monday, 8(9), September 2003.

[6] Yang, Y. 1999. An evaluation of statistical approaches to text categorization. Information Retrieval, 1(1/2):69-90.

[7] Yang, Y., J.O. Pedersen. 1997. A comparative study on feature selection in text categorization. En Proceedings of the 14th International Conference on Machine Learning.