International Research Journal of Computer Science (IRJCS)
Issue 12, Volume 07 (December 2020)
ISSN: 2393-9842
https://www.irjcs.com/archives

MOBILE SMS SPAM DETECTION USING MACHINE LEARNING TECHNIQUES

Samadhan Nagre,
Department of CS and IT,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad- 431004, India
samadhannagre340@gmail.com
Sachin N. Deshmukh,
Department of CS and IT,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad- 431004, India

Publication History
Manuscript Reference No: IRJCS/RS/Vol.07/Issue12/SPCS10089
Received: 08, September 2020
Accepted: 18, November 2020
Published: 05, January 2021
DOI: https://doi.org/10.26562/irjcs.2020.v0712.004
Citation: Samadhan, Sachin N(2020). Mobile SMS Spam Detection Using Machine Learning Techniques. IRJCS:: International Research Journal of Computer Science, Volume VII, 331-334.
https://doi.org/10.26562/irjcs.2020.v0712.004
Peer-review: Double-blind Peer-reviewed
Editor: Dr.A.Arul Lawrence Selvakumar, Chief Editor, IRJCS, AM Publications, India

Abstract: This paper analyses the method of intelligent spam filtering techniques during the SMS (Short message Service) text paradigm, in the context of mobile text messages spam. The unique characteristics of the SMS contents be indicative of the fact that all approaches cannot be equally effective or efficient. This paper compares some of the trendy mobile SMS spam filtering techniques on a publically available SMS spam corpus, to categorize the methods that work best in the SMS text context. This can give hints on optimized SMS spam detection for mobile text messages.
Keywords: machine; detection; mobile; sms ; spam;

I.INTRODUCTION

In the wireless communication age, Short Message Service (SMS) is one of the easiest and affordable communication way. SMS is popular worldwide due to high response rate, "secure" personal service and lowest prize [1]. But there some problems faced by the people such as spam SMS through using this SMS technique. Spammers get advantages of this wireless world and make to potential customers. Today most of the SMS's are Spam which consists of Credit Card offer, discount offers, traffic plans, promotions etc. Due to Spam SMS, Mobile service providers suffer from some sort of financial problems as well as it reduces the calling time for users. Unfortunately, if the client accesses such Spam SMS they May face the problem of virus before malware. When arrives as mobile it resolve disturb mobile client privacy and concentration. It may lead to frustration of user. So Spam SMS is one of the most important problems in wireless communication world and it grows day by day. To avoid such as Spam SMS. And People use of white and black list of numbers. Other than this technique is not adequate to completely avoid Spam SMS. To tackle this problem it is needful to use a smarter technique which correctly identifies Spam

II. RELATED WORK

In this paper author describe as assessment of case reasoning approach for long text messages to short text messages. In this evaluation it determines appropriate feature types and feature representation of short text messages and then compares performance of classifier algorithm. In this paper authors uses Naïve Bayes classifier and support vector machine algorithms [2]. In this paper authors uses Bayesian classifier for filtering spam SMS. Author analyses what extent Bayesian classifier to block email spam also can be applied for detecting and stopping mobile spam. It is uses Machine learning algorithm and Bayesian filtering technique can be effectively worked [3]. Authors consider that for email filtering it requires some adaptation to reach good level of performance on SMS spam.

To prove these assumption authors performer experiments on SMS filtering using proper feature representation [4]. In this paper author describe compare performance of machine learning methods on new non encoded SMS spam collection. On these algorithms, support vector machine outperformance then other method [5]. In this paper author contrast performance of several machine learning algorithms. The procedures which describe in this paper build collection of non-Unwanted messages [6].

Introduction to Spam

A. Spam.

The definition of a spam do not very much within the case of emails or SMS Spam in simple terms, it can be described as "Unwanted Bulk Messages" These are usually unsolicited information being pushed to the users, as advertisement, or by tricksters and level for fraud [8]. The spammers may be businessmen and they send spam because it works, in the form of responses that they accepted to their messages [7].

Table I. The distribution of spam and non- spam in the dataset

Label	Percentage in dataset
Spam	0.13
Hams	0.87

B. Spam Filtrations

It is very easy for anyone to identify a mobile SMS spam message just by reading through it. Our challenge in mobile SMS spam filtration is to solve this problem using fairly simple algorithms [9]. The most common classical approaches that uses white-lists and black lists does not work as it is only capable of blocking an entire server (source) from transfer messages, which can include too many legitimate messages (false positives) as well. Hence the problem of spam filtration is essentially a case of text classification [9]

C. SMS Spam

Spam in the SMS context is very similar to email spams, typically, unsolicited bulk messaging with some business interest [1]. However, the limitation of SMS imposes restrictions to the message, in that, there can only be limited number of characters, which includes alphabets, numbers and a few symbols. This restricts the amount and format of information that a spammer can send. A look through the messages reveals a very clear pattern which can attribute to this restriction.

III EVALUATION AND DISCUSSION

A. Spam Detection Techniques

The list of spam filtration algorithms considered for this observation is listed in table II. The methods were selected from a list of methods implemented in the Nave bayes is a collection of machine learning algorithms for data mining task. R is open source software issued under the General Public License (GNU) [10]. We make use of this collection and apply the various algorithms on the SMS spam corpus to compare the effectiveness of each in an attempt to identify the ones that perform better in the SMS Paradigm.

Table II. Performance and analysis for different level Classification

Machine Learning Algorithms Effectiveness					
No. of test	Dataset Spam & Ham SMS	Naïve Bayes	SVM	Decision Trees	Logistic Regression
1	400	86.36%	94.00 %	84.00%	60.00%
2	800	97.03%	95.00 %	88.05%	74.00%
3	1200	96.72 %	93.33 %	86.00%	78.00%
4	1600	97.35 %	96.00 %	90.75%	85.05%
5	1672	98.03 %	94.26 %	89.47%	84.93%
6	3200	98.53 %	95.75 %	89.38%	94.75%
7	4800	98.05 %	96.42 %	92.00%	96.08%
8	5574	97.94 %	95.98 %	91.54%	95.77%

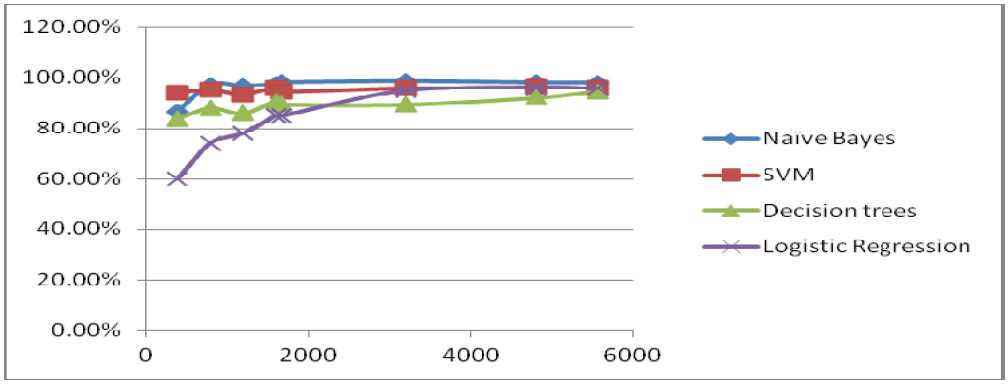

Figure1. Comparison of the preferred successful algorithms

B. The Preferred Algorithm

As we estimated, the Naïve bayes methods best within most cases. The most excellent result was thrown up by the method "Naïve Bayes" which provide a 98.53% correct classification and be also observed to be fast. However, our preferred algorithm be the Naïve Bayes method Support Vector Machine because this returned a effectiveness of 96.42% with false positive and Decision Trees 92.00%and just false positive, as a ham marked as spam be way to insulting than allowing a spam near come through when a ham[7]. Based on these, the preferred algorithms be listed in table II. As it can be seen, it is not necessarily the top performing algorithms that we have chosen, giving preference to algorithms contribution a combination of good effectiveness and low "false positive" and execution times.

1. Naïve Bayes

Bayesian is a probabilistic move toward that starts among a previous faith, observes some data and then updates that faith The probabilistic life form spam and not spam of a word can be intended through the incidence of that word in ham and spam messages with the Bayesian algorithm [11].

2. Support vector Machine

Support vector machines be supervised learning by way of linked algorithms that analyses data used future for the categorization as well as regression analysis. If a pot of teaching example containing spam and rightful SMS is known after that SVM teaching algorithm build a model that can assigns new example keen on spam and rightful group An SVM model is a demonstration of the example because a point in space, mapped so that example of the divide category are separated by a clear gap so as to is wide as achievable. [12]

3. Decision Trees

A decision hierarchy is a decision support instrument that uses a hierarchy similar to or model of decisions and their likely penalty, counting possibility of event outcomes. A decision tree can be used to make choice to whether a fresh message is spam or ham [13]

4. Logistic Regression

A logistic regression is a prognostic analysis. Logistic regression be used to explain data and to explain the association flanked by single reliant binary changeable and single or additional supposed ordinal, interval or percentage-level independent variables. From time to time logistic regression be hard to understand, the intellects statistics instrument without difficulty allows you to conduct the analysis, after that in simple English interprets the production.

IV. RESULT ANALYSIS

R programming language is used for the implementation of the proposed framework. Support Vector Machine light [9] is used for as classification tool for Support Vector Machine and Naïve Bayes is implemented in R. Naïve Bayes and Support Vector Machine algorithms have been implemented for the Spam filtering task. The study has gone through the empirical analysis of the performance of both the Spam filters (Support Vector Machine and Naïve Bayes) for India messages. It is observed from the experiment that the Spam Filter based on Naïve Bayes outperforms the Spam Filter based on Support Vector Machine. Extensive tests have been performed with varyring numbers of data set size. The success rates reach their maximum using all the messages and all the words in training corpus.

V. CONCLUSION

SMS identification is one of the dangerous issues in today`s world. To identify spam messages is very important task which will decrease user`s time and amount. For this purpose present study uses Naïve Bayes, Support Vector Machine Decision Trees Logistic Regression classification algorithm. In this study we apply algorithm on Mobile SMS dataset. Our present study identifies best algorithm for classification or identification of SMS. The results of our evaluations presented here have shown that for different algorithms, accuracy and time are different. Present study shows that Filtered Classifier with unsupervised discredited filter and Naïve Bayes algorithm has great accuracy.

ACKNOWLEDGMENT

I am indebted to the Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad and University Grant Commission, Delhi for providing all facilities to me related to my research work.

REFERENCES

1. Delany, S. J., Buckley, M. & Greene, D. (2012) SMS Spam Filtering: Methods and Data, Expert Systems with Applications, vol. 39 (10),p9899-9908 https://doi:10.1016/j.eswa.2012.02.053
2. Delany, Sarah Jane, and Pádraig Cunningham. "An analysis of casebase editing in a spam filtering system." Advances in Case-Based Reasoning. Springer Berlin Heidelberg, 2004. 128-141.
3. Gómez Hidalgo, J.M., CajigasBringas, G., PuertasSanz, E., CarreroGarcía, F. ,"Content Based SMS Spam Filtering". Proceedings of the 2006 ACM Symposium on Document Engineering (ACM DOCENG'06), Amsterdam, The Netherlands, 10-13, 2006

4. Cormack, G. V., Gómez Hidalgo, J. M., and PuertasSánz, E., "Feature engineering for mobile (SMS) spam filtering". Proceedings of the 30th Annual international ACM Conference on Research and Development in information Retrieval (ACM SIGIR'07), New York, NY, 871-872, 2007.
5. Gómez Hidalgo, J.M., Almeida, T.A., Yamakami, A. ,"On the Validity of a New SMS Spam Collection". Proceedings of the 11th IEEE International Conference on Machine Learning and Applications (ICMLA'12), Boca Raton, FL, USA, 2012.)
6. Almeida, T.A., Gómez Hidalgo, J.M., Silva, T.P. "Towards SMS Spam Filtering: Results under a New Dataset". International Journal of Information Security Science (IJISS), 2(1), 1-18.
7. Paul Graham, (August 2002), A plan for spam,viewed:2September2011,http://paulgraham.com/spam.html
8. Cai,J,Tang, Y., & Hu,R. (200S). "Spam filter for short messages using winnow". 200S International Conference on Advanced Language Processing and Web Information Technology, 454-459.
9. Cormack, G. v., Hidalgo, 1. M. G., & Sanz, E. P. (2007). "Feature engineering for mobile (SMS) spam filtering 00. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07,S71.
10. Weka The University of Waikato, Weka 3: Data Mining Software in Java, viewed on 2011 September14 http://www.cs.waikato.ac.nzlml/wekal
11. http://fastml.com/bayesian-machine-learning/ [Last Accessed: 05-11-2016
12. https://en.wikipedia.org/wiki/Support_vector machine [last Accessed: 05-11-2016]
13. http://en.wikipedia.org/wiki/K
14. V. T. Joachims, "Making Large-Scale SVM Learning Practical," In: B. Schölkopf, C. Burges and A. Smola, Eds., Advances in Kernel Methods Support Vector Learning, MIT-Press, Cambridge, 1999.

