CONTENT-BASED SMS SPAM FILTERING USING MACHINE LEARNING TECHNIQUE

Article · May 2018

CITATIONS
5

READS
3,986

3 authors, including:

Dipak Kawade
Sangola College
15 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Kavita Oza
Shivaji University, Kolhapur
110 PUBLICATIONS **216** CITATIONS
SEE PROFILE

International Journal of Computer Engineering and Applications,
Volume XII, Issue IV, April 18, www.ijcea.com ISSN 2321-3469

CONTENT-BASED SMS SPAM FILTERING USING MACHINE LEARNING TECHNIQUE

Dr. Dipak R. Kawade[1], Dr. Kavita S. Oza [2]

[1]Department of Computer Science, Sangola College, Sangola (MS) 413307, India
E-mail: dipakkavade@gmail.com
2Department of Computer Science, Shivaji University Kolhapur (MS) 416004, India
E-mail: skavit.oza@gmail.com

ABSTRACT:

In the modern wireless communication world, Short Message Service (SMS) is one of the important communications way followed by people. It has the powerful economic impact on the user as well as service provider. Spam SMS is one of the problem faced in this world. Lots of SMS Spam filtering techniques are used to identify Spam SMS. Present Study focuses on Spam SMS identification using Machine Learning technique which is implemented using open source software Python. The experimental result shows that approximately 98 % Spam SMS's are identified.

Keywords: Short Message Service (SMS), SMS Spam filtering, Machine Learning, Python

[1] 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 are some problems faced by the people such as Spam SMS by using this SMS technique. Spammers take advantages of this wireless world and reach to potential customers. Today most of the SMS's are Spam SMS 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 calling time for users. Unfortunately, if the user accesses such Spam SMS they may

face the problem of virus or malware. When SMS arrives at mobile it will disturb mobile user privacy and concentration. It may lead to frustration of user. So Spam SMS is one of the major issue in wireless communication world and it grows day by day.

To avoid such Spam SMS; people use white and black list of numbers. But 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 SMS.

Text mining technique is useful for Spam SMS identification. It analyzes text content and find patterns which are used to identify Spam and Non-Spam SMS. Generally, Spam SMS filtering is considered as text classification technique. Size of SMS is limited and normally it is just 160 characters. This SMS includes uppercase as well as lowercase letters, special symbols, and some URL. Most of Spam SMS includes URL's, Special characters such as "$", "!' etc.[2] This paper considers all these characters to find features from Bag of Words in Spam SMS.

SMS Spam identification is Binary classification technique having two types such as Spam and Ham. In the present study, text classification is used for identification of Spam and Ham SMS.

In this paper section II focus on related research in SMS Spam filtering area. Section III discusses experimental work which consists of dataset collection, working environment etc. Section IV discusses observation of experimental work. Section V presents conclusion and future work in SMS Spam filtering area.

[2] RELATED WORK

In the paper, a author has focused on SMS Spam filtering. For this purpose, they considered top email spam filtering techniques on mobile spam. The author used different email Spam filtering techniques such as TREC etc. The author has not made any decision on SMS Spam filtering. The author suggested that SMS Spam filtering is critical procedure so need more experiments and more study of large size dataset. [3]

Author has used WEKA tool for classification of Spam SMS. He has applied different classification algorithms which are available in WEKAon the basis of accuracy and time. The output of study showed that FilteredClassifier with unsupervised discredited filter and the NaiveBayes algorithm has the highest accuracy among other algorithms. [4]

Authors discussed index based text classification method for SMS Spam filtering technique. They have defined two different models and compared the performance of these models. Based on different text features, they have implemented six different classification algorithms and then combined to form collaborative algorithm. Experimental result showed that classification algorithms performance was increased for English corpus. The result showed that performance of the word-level index is higher than document level index for the Chinese language. Time complexity was reduced after applying collaborative learning. [5]

Author has expressed their opinion that; distinguishing characteristics of SMS contents are not effective or efficient for identifying Spam SMS. In the paper, Author has analyzed different Spam SMS filtering techniques and identified good algorithm. The experiential result shows that Bayesian algorithm was best fit for SMS Spam identification. [6]

In the paper, a author has used numerous feature extraction and feature selection methods for identification of Spam SMS having Turkish and English languages. The Model framework has considered features from Bag of Words and structural feature for identification purpose. The Model framework uses different classification algorithm for classifying Spam SMS. For classification purpose; author uses different combinations of distinctive features of Bag of Words and Structural feature. Experimental result of the study shows that combination of Bag of Words and a Structural feature was higher performance than only using Bag of Words. It has also shown that for each language effectiveness is different. [2]

In the study; the author has used four different classification algorithms namely Neural Network (NN), Naïve Bayesian, Support Vector Machine (SVM) and Relevance Vector Machine (RVM) for classification of Spam SMS. Experimental performance is counted on the basis of the different size of training dataset and feature extracted size. Author has shown that NN algorithm is not suitable for Spam SMS identification. Performance of SVM and RVM algorithm was good but among these two algorithms, RVM is faster than SVM. The result of the study shows that RVM algorithm is best suited for filtering Spam SMS. [7]

[3] EXPERIMENTAL STUDY

[3.1] Working Environment

This experiment is implemented in Windows 7 operating system with machine configuration as Intel Core i3, 3.3 GHz, and 2 GB RAM. Python 3.4 is used for this experiment.

[3.2] B.Dataset

Dataset used in the study is freely available on the internet. It is created by Tiago A. Almeida and José María Gómez Hidalgo[8]. Dataset consists of a collection of 5568 SMSand 2 attributes. The first attribute is class attribute whereas the second attribute is text attribute i.e. SMS. Class attribute has two possible values namely Span and Ham. Among 5568 SMS, 746 SMS are of type Spam and 4822 SMS are of Ham type.

[3.3] Python

Python is a programming language developed by Van Rossum and his team. Initially, it was started in 1989. For present study python 3.4 is used because it is open source and better community development. Python is downloaded from internet freely. [9] Also it has extensive library supports, easy learning, and user-friendly etc. and many more advanced feature.

Present study uses NLTK, pandas and re package. Present study uses text mining technique on SMS text and then cluster into two clusters namely spam and ham.

[3.4] Data Pre-processing

For gaining more accuracy; data must be cleaned. Present dataset consists of SMS text which is not useful for text processing. Also, SMS consists of stop words i.e. some words are frequently used such as conjunctions, numbers, prepositions, names, base verbs, etc. Such

type of words doesn't play any role in text mining. This unwanted text generates processing overhead so there is need to clean such data. To achieve this; data pre-processing is important. Data pre-processing consist of following steps

1. Filtration:-Filtration can remove links, URLs, special words etc.
2. Tokenization:- This step separates text into different tokens.
3. Remove Stopwords:-some words having no any analytical values are removed in this step, which reduces processing overheads.
4. SteamDocumnet:- This step removes some common ending words such as "ing","es" etc.
5. Remove White Space:- Generally, text contains lots of white spaces, which are removed in this step.
6. Conversion to lower case:- this step converts all text into lower case letters.

Present study uses regular expression technique for removing white space, numbers etc. from SMS text. Stopwords are collected and removed by using NLTK package.

[3.5] What is Spam?

Simply Spam SMS is defined as "Unsolicited Bulk Messages". [6][10]. By using Spam SMS, unwanted information is posted to user. This information contains some sort of advertisements, tricks and cheating information.

[3.6] SMS Spam Detection Process

It is very easy to identify Spam SMS just by reading SMS. But our task is to automatically identify Spam SMS by using some algorithm. This can be achieved by machine learning technique. Present Study uses machine learning technique to identify Spam SMS.

Present study identifies Spam SMS from SMS data. For detection of Spam SMS; present study list out words related to Spam in SMS. There are total 225 words listed which are mostly available in Spam SMS. These words are collected from different web sites as well as form different SMS.

In initial stage; pre-processing of text has been taken place and generated clear data. This data then used for Spam detection purpose. Text mining techniques are applied to SMS text to find bag of words [11] and also counts occurrences of spam word. If count is greater than 0 then such SMS's are considered as Spam otherwise it is considered as Ham.

[3.7] Experiment Work

Present study focuses on detection of Spam SMS by using text mining technique. SMS dataset is pre-processed first to obtain clear data. After pre-processing of SMS data; it counts Spam word available in each SMS. Depends upon count value obtained by experiment identifies Spam SMS. If count value greater than 0 then such SMS are considered as Spam.

Following table shows count of Spam SMS from original dataset and present algorithm.

Type	Original Dataset count	Present Algorithm count
Spam SMS	746	732
Ham SMS	4822	4836

Figure 1:- Shows Spam SMS count obtained by present study algorithm and original Dataset

Present algorithm correctly identifies 732 spam SMS. Percentage of correctly matching Spam SMS is 98.12% and mismatching Spam SMS is 1.88%. For finding error rate of algorithm Mean Absolute Error (MAE) technique is used. A formula for MAE is as follows :

$$\text{MAE} = \frac{\text{ABS}(\text{Predicted} - \text{Actual})}{\text{Total Count}}$$

MAE value for present study is 0.091.

Root Mean Square Error (RMSE) is one another important error measure is used for checking error rate. A formula for RMSE is as follows :

$$\text{RMSE} = \sqrt{(\text{ABS}(\text{Predicted} - \text{Actual}))/(\text{Total Count})}$$

RMSE value for present algorithm is 0.3.

Following graph shows SMS Spam count of original dataset and count obtained by an algorithm.

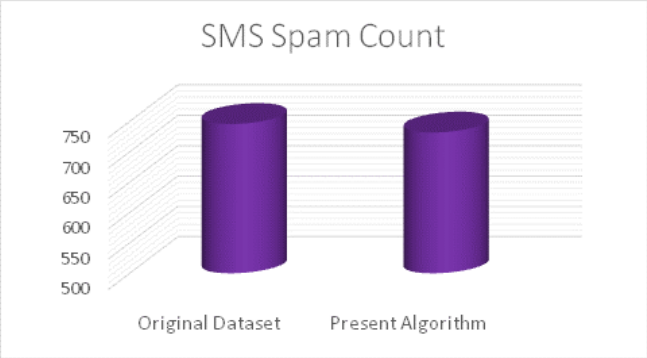

Figure 2:- Shows SMS spam count obtained by present study algorithm and original dataset

[4] OBSERVATIONS

Present work detects Spam SMS by using text mining technique, which is implemented in python freeware software. From table 1 and figure 1 it is clearly shown that present work correctly identifies Spam SMS. The percentage of correct detection of Spam SMS is 98.12 whereas misidentification percentage is 1.88.

For finding errors of algorithm MAE and RMSE techniques are commonly used and values are 0.091 and 0.3respectively, which are very less. We observed that present work correctly identifies Spam SMS.

[5] CONCLUSION

SMS Spam identification is one of the important task in present world, which is wasting user's valuable time as well as money. Present algorithm tackles this issue.

Present Work is useful to identify Spam SMS from SMS dataset. Experimental work shows that 98.12% SMS are identified correctly as Spam SMS's from the dataset.

It also checks algorithm errors by most important error checking technique MAE and RMSE. MAE of current algorithm is 0.091 and RMSE is 0.3 which is very less. Therefore present study correctly identifies Spam SMS's as compared to other algorithms. There is more scope to increase accuracy in identifying Spam SMS. The merit of our approach which lies in the various machine recognizable statistics derived from the skeleton of the document (HTML tags).

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. doi:10.1016/j.eswa.2012.02.053

[2] A. K. Uysal1, S. Gunal1, S. Ergin2, E. SoraGunal , The Impact of Feature Extraction and Selection on SMS Spam Filtering, , ELEKTRONIKA IR ELEKTROTECHNIKA, ISSN 1392-1215, VOL. 19, NO. 5, 2013, http://dx.doi.org/10.5755/j01.eee.19.5.1829, pp-67-72

[3] Gordon V. Cormack, José María Gómez Hidalgo, Enrique PuertasSánz, Feature Engineering for Mobile (SMS) Spam Filtering, SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands. ACM 978-1-59593-597-7/07/0007

[4] Dipak R. Kawade, Dr. Kavita S. Oza , SMS Spam Classification using WEKA, , International Journal of Electronics Communication and Computer Technology (IJECCT), Volume 5 Issue ICICC(May 2015), www.ijecct.org, ISSN:2249-7838, PP-43-47

[5] WuyingLiu,Ting Wang , Index-based Online Text Classification for SMS Spam Filtering, JOURNAL OF COMPUTERS, VOL. 5, NO. 6, JUNE 2010, doi:10.4304/jcp.5.6.844-851, PP-844-851

[6] K. Mathew and B. Issac, "Intelligent spam classification for mobile text message," Proceedings of 2011 International Conference on Computer Science and Network Technology, Harbin, 2011, pp. 101-105. doi: 10.1109/ICCSNT.2011.6181918

[7] A comparative study for content-based dynamic spam classification using four machine learning algorithms,B. Yu, Z. Xu,Knowl. Based Syst. (2008), doi:10.1016/j.knosys.2008.01.001

[8] http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/ Accessed:-03-Aug-2017

[9] https://www.python.org/ Accessed:-03-Mar-2017

[10] Tarek M Mahmoud and Ahmed M Mahfouz, SMS Spam Filtering Technique Based on Artificial Immune System, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012, ISSN (Online): 1694-0814, www.IJCSI.org pp-589-597

[11] Gordon V. Cormack, José María Gómez, Enrique PuertasSánz, Spam Filtering for Short Messages, Proceeding CIKM '07 Proceedings of the sixteenth ACM conference on Conference on information and knowledge management Pages 313-320 , doi:10.1145/1321440.1321486