**CLASSIFICATION OF SMS MESSAGES AS USING MACHINE LEARNING TECHNIQUES**

M.Aakash1 R.Nandhini1 J.Pranitha1

MSC Data Science

Dr.j.Vijayarangam2 M.Srividya2

Assistant Professor

Department of computing

Coimbatore Institute of Technology

**ABSTRACT:**

The popularity of mobile devices is increasing day by day as they provide a large variety of services by reducing the cost of services. Short Message Service (SMS) is considered one of the widely used communication service. However, this has led to an increase in mobile devices attacks like SMS Spam. In this paper, we present a novel approach that can detect and filter the spam messages using machine learning classification algorithms. We study the characteristics of spam messages in depth and then found features, which can efficiently filter SMS spam messages from ham messages. Our proposed approach achieved 97% true positive rate and 3% false positive rate for Naïve Bayes classification algorithm.

**Keywords: SMS spam, Mobile devices, Machine learning, Feature selection**

**INTRODUCTION:**

Short Message Service (SMS) is one of the popular communication services in which a message is sent electronically. The reduction in the cost of SMS services by telecom companies has led to the increased use of SMS. This rise attracted attackers which have resulted in SMS Spam problem. A spam message is generally any unwanted message that is sent to user’s mobile phone. Spam messages include advertisements, free services, promotions, awards, etc. People are using SMS messages to communicate rather than emails because while sending SMS message there is no need of internet connection and it is simple and efficient.

The SMS Spam problem is increasing day by day with the increase in the use of text messaging. There are various security measures available to control SMS Spam problem but they are not so mature. Many android apps are also on play store to block spam messages but people are not aware of these apps due to lack of knowledge. Other than apps the filtering techniques available mainly focuses on email spam as email spam is one of the oldest problem but with the popularity of mobile devices, SMS spam is the one of the major issue these days. SMS is one of the cheapest ways to communicate and can be considered as the simplest way to perform phishing attacks as mobile devices contain sensitive and personal information like card details, username, password, etc. Attackers are finding different ways to steal this information from mobile devices and SMS is one of the easiest ways. Smishing i.e. SMS based Phishing is more popular these days in which user sends malicious link via SMS and asks user to visit that link and steals sensitive information from user’s mobile device. There are various detection approaches available for detecting mobile phishing like QR code, machine learning based, biometric based, matrix code reader based, knowledge based and authentication based. SMS spammers can purchase any mobile number with any area code to send spam messages so that it becomes difficult to identify the attacker. US tatango learning center provided the list of top 25 SMS Spam area codes used by spammers.

**RELATED WORK:**

Spam filtering system is associated to text classification problem which we need a classifier to group the similar text together in one class label. There are two stages for classification [1]. First stage is a feature selection process by generation features that needed to classify texts. Second stage is a classifications process itself. In [1], they select features as keywords. And every class here is represented by a set of keywords, and every keyword associated to weights, and these weights represent the importance of these keywords to make the classification process. Because of huge number of features that should be extracted, there are several popular feature selection tools such as text frequency threshold [2] chi square, and term strength and information gain [1].

In [1], they propose a system, which contains three modules: an extracting feature module, a learning module, and a filtering module. The first module contains a vector of keywords that assign 1 if the word presented in that document and 0 otherwise. After presenting these keywords, they select the keywords that have the largest weight. The learning module creates a learned vector and uses the naïve bayes classifier. SMS messages create values in the vector by using two stages. If the vector data have the same word on SMS then it assigns 1 on that word. After that it compares the address of SMS message and the contact list in the cell phone. If it matched then it will assigns to 1. The filtering module used to separate between spam’s and not spam SMS where 0 indicates that SMS is not spam and 1 indicates that SMS is a spam message.

**OBJECTIVE:**

The Objective of this study is to identify the SMS as either SPAM or HAM Message.

**DATASET DESCRIPTION:**

This collection of SMS spam is built from public websites where labeled messages are collected for mobile phone spam studies. It contains 5,574 examples of SMS messages including 747 spam’s and 4,827 hams. Those SMS messages are collected from different sources and manually labeled. The data set characteristics are: multivariate, text, and domain theory. The associated data mining tasks on this data are: classification and clustering. Each row starts with class label of the SMS message followed by the text message itself, whether it is spam or ham. For instance:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ham | Fine if that's the way u feel. |  |  |  |  |  |  |  |  |  |  |  |
| Spam | England v Macedonia - don’t miss it | | | | | | | | | | | |

The Dataset is divided into 75% of training and 25 % of testing. The dataset is random, first we use 4180 is used for training set-there is no need to randomize it.

**METHODOLOGY:**

Naïve Bayes - Naïve Bayes classification algorithm is based on Bayes theorem. In Naïve Bayes assumptions between predictors is independent. It is simple and easy to use so it can be used for large datasets.

LOADING DATA

TEXT MINING

DATA PREPERATION

VISUALIZATION

PREPARATION FOR NAÏVE BAYES

TRAIN MODEL ON DATA

PREDICT AND EVALUATE THE MODEL

**RESULTS:**

|  |  |  |
| --- | --- | --- |
| Prediction | Ham | spam |
| Ham | 1190 | 20 |
| Spam | 9 | 160 |

Accuracy: 0.979

95% CI: (0.9699, 0.9859)

No Information Rate: 0.8695

P-Value [Acc > NIR] : < 2e-16

Kappa: 0.9049

Mcnemar's Test P-Value: 0.06332

Sensitivity: 0.9925

Specificity: 0.8889

Pos Pred Value: 0.9835

Neg Pred Value: 0.9467

Prevalence: 0.8695

Detection Rate: 0.8629

Detection Prevalence: 0.8774

Balanced Accuracy: 0.9407

'Positive' Class: ham

Cross table:

Total Observations in Table: 1379

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted | Ham | Spam | Row total |
| Ham | 1190  0.983  0.992 | 20  0.017  0.111 | 1210  0.877 |
| Spam | 9  0.053  0.008 | 160  0.947  0.889 | 169  0.123 |
| Column Total | 1199  0.869 | 180  0.131 | 1379 |

**INFERENCE:**

Out of 29/1379 messages were classified incorrectly. This means that the algorithm classified the testing set as spam or ham with 98% accuracy. To improve the model, one might tamper with the Laplace value, collect more SMS data, or try splitting the dataset randomly into training and testing.

The accuracy would increase as the dataset gets bigger. The more data there is to train the algorithm, the more effective it would be in predicting Spam or Ham.

**CONCLUSION:**

The SMS Spam problem is increasing nowadays with the increase in the use of text messaging. SMS Spam filtering is the big challenge these days. In this paper, we propose a technique for SMS Spam filtering based on 10 feature using five machine learning algorithms namely Naïve Bayes, Logistic Regression, J48, Decision Table and Random Forest. The dataset that has been used in our work consists of 2608 messages out of which 2408 messages were collected from the SMS Spam Corpus v.0.1 publically available and 200 messages collected manually. Out of all classification algorithms, Random Forest Classification Algorithm gives best results with 96.1% true positive rate. In the future work, we will try to add more features as best spam features help in detecting spam messages more accurately. We will also try to collect more and more datasets from the real world.

**References:**

[1] Inwhee Joe\_ and Hyetaek Shim, "An SMS Spam Filtering System Using Support Vector Machine," Division of Computer Science and Engineering, Hanyang University, Seoul, 133-791 South Korea.

[2]Y. Yang, J. Pedersen. A comparative study on feature selection in text categorization. In: The Fourteenth International Conference on Machine Learning (ICML1997), pp. 412–420. Morgan Kaufmann, San Francisco (1997).