International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 1, Jan-Feb 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Spam SMS Detection Using Naive Bayes Classifier

Vijay*, Shubham Kumar**

*(Department of Computer Science and Engineering, Neelkanth Institute of Technology, Meerut)

** (Department of Computer Science and Engineering, Bharat Institute of Technology, Meerut)

----------**********----------

Abstract:

Spam SMS (Short Message Service) has become a problem for mobile phone users, nowadays. When mobile phone gets flooded with spam SMS then important and genuine messages can be skipped from the sight of users. User can be a victim of phishing and fraud because of spam SMS. Therefore, there is a need to separate spam SMS from genuine SMS. In this paper we have used Naive Bayes algorithm for detecting spam SMS. In this paper it is discussed how Naive Bayes algorithm is implemented on a data set for detecting spam SMS and its performance on detecting spam SMS on test data is also shown. Naive Bayes algorithm is a supervised machine learning algorithm.

Keywords **— Naïve Bayes classifier, document term matrix, text classification, spam SMS, confusion matrix.**

----------**********----------

I. INTRODUCTION

Text messages are sent from one mobile device to another through SMS(Short Message Service). Spams are junk text messages or unsolicited messages [1]. Nowadays, people use their mobile phones not only for making and receiving audio calls but also for various other purposes like banking (like transferring money, checking balance etc.), sending and receiving e-mails, accessing Facebook, online shopping etc. which require their confidential information like password, PIN, bank account number, credit card or debit card number etc.. Apart from this people also keep their personal information in their mobile phones like phone number of their friends and relatives, photos, images of their IDs and other important documents. People can be victim of cyber attack through spam SMS and the information stored in their phone can be leaked. Mobile users are disturbed by spam SMS and may be frustrated [2]. Spam SMS wastes network bandwidth and cause loss of productivity [3]. National Customer Preference Registry (NCPR) was set up by Government of India, junk calls have been reduced to some extent by it but spam SMS are not filtered by it [1].

Text classification techniques are widely used for spam filtering [4]. In text classification a category (from a set of predefined categories) is assigned to a document [5]. SMS are text messages and our goal is to classify a SMS as spam or genuine. In text classification, supervised machine learning approach is used. Therefore, in text classification already labeled data set is required for constructing a classifier. In our data set genuine messages were labeled as ham. Many issues of SMS spam detection are inherited from email spam detection [6]. Since there is similarity in email spam filtration and SMS spam filtration, the techniques used for spam email filtration can be used for spam SMS filtration [7]. Naïve Bayes classifier is a popular method for spam email filtration [8]. In this paper we have used Naïve Bayes algorithm for spam SMS detection.

II. NAIVE BAYES ALGORITHM

Naïve Bayes algorithm is commonly used for text classification. It is a probabilistic machine learning

algorithm. Naïve Bayes algorithm is based on Bayes theorm which is based on conditional probabilities. Conditional probability refers to the probability of occurrence of an event when it is given that another event has already occurred. By Bayes theorem the probability of a text message being spam is evaluated by finding and using the probabilities related to every word [7]. Bayes theorem describes a simple mathematical formula by which conditional probability can be calculated. According to Bayes theorem the probability of occurrence of an event A when event B has been already occurred, represented by P(A|B) is calculated by using the formula:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

The probability P(A|B) is called *posterior probability*. P(B|A) is the probability of occurrence of event B when event A has already occurred, it is called *likelihood*. P(A) is the probability of occurrence of event A, it is called *class prior probability*. P(B) is the probability of occurrence of event , it is called *predictor prior probability.* In other words, by Bayes theorem we can calculate posterior probability from likelihood, class prior probability and predictor prior probability. Naïve Bayes assumption is that contribution of each feature is independent and equal for the outcome.

III. METHODOLOGY

We cleaned the data and then we split the dataset into training dataset and test dataset. Training data set was used to train Naïve Bayes classifier. Performance of trained classifier was tested on test dataset.

A. Dataset Description

We used SMS Spam Collection v.1 dataset [9]. We downloaded this dataset from [10]. This dataset has 5572 text messages which were classified as ham or spam. It has two columns two labeled as v1 and v2. First column v1 has only two values spam or ham describing whether the text message in second column v2 is spam or genuine. The dataset is available as CSV (comma-separated values) file. The messages in this dataset are collected from these sources: Grumbletext Web site, NUS SMS Corpus (NSC), Caroline Tag's PhD Thesis, SMS Spam Corpus v.0.1 Big. In this dataset 4825 text messages were labeled as ham and 747 text messages were labeled as spam.

B. Data Preprocessing

We renamed the column v1 as *class* and v2 as *text*. After renaming the columns we shuffled the dataset to reduce overfitting. After shuffling, dataset was cleaned. To clean the dataset all text messages were converted to lowercase, and punctuations, numbers, stopwords and URLs were removed.

C. Naive Bayes Classifier

After data preprocessing, dataset was split into training dataset and test dataset. There are 5572 text messages in dataset in which 747 text messages are labeled as spam and 4825 text messages are labeled as ham. The data were split into two datasets. Training dataset had 4000 text messages in which 3461 were labeled as ham and 539 were labeled as spam. Test dataset had remaining 1572 text messages in which 1364 were labeled as ham and 208 were labeled as spam. For classification, a model or classifier is constructed then this model or classifier is further used for predicting the class labels [11]. First we converted text messages of training dataset into document term matrix and the terms having frequency less than five were removed. The entries 0 of document term matrix were replaced by "No" and other non-zero entries were replaced by "Yes". So, this document term matrix had only two values: "Yes" and "No". The Naïve Bayes classifier was trained by using this document term matrix and class labels of text messages of training dataset. In same manner, document term matrix for text messages of test dataset was also created and used for predicting the class labels of text messages by Naïve Bayes classifier.

IV. CONCLUSIONS

Test dataset had 1572 text messages. Among 1364 ham text messages of test dataset 1336 were correctly classified as ham by Naïve Bayes classifier and remaining 28 were wrongly classified as spam.

```
Confusion Matrix and Statistics

          Reference
Prediction  ham spam
      ham  1336   15
      spam   28  193

               Accuracy : 0.9726
                 95% CI : (0.9633, 0.9801)
    No Information Rate : 0.8677
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8839

 Mcnemar's Test P-Value : 0.06725

            Sensitivity : 0.9795
            Specificity : 0.9279
         Pos Pred Value : 0.9889
         Neg Pred Value : 0.8733
             Prevalence : 0.8677
         Detection Rate : 0.8499
   Detection Prevalence : 0.8594
      Balanced Accuracy : 0.9537

       'Positive' Class : ham
```

Fig. 1 Confusion matrix and statistics for test dataset

Among 208 spam messages of test data 193 were correctly classified as spam and remaining 15 text messages were wrongly classified as ham. Fig. 1 shows confusion matrix and statistics for test dataset.

REFERENCES

[1] M. Gupta, A. Bakliwal, S. Agarwal and P. Mehndiratta, "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, 2018, pp. 1-7, doi: 10.1109/IC3.2018.8530469.

[2] Kawade, Kavita Oza, and Kavita S. Oza. "Content-based SMS spam filtering using machine learning technique." *International Journal of Computer Engineering and Applications* 7 (2018): 4

[3] A. Lambert, ``Analysis of SPAM," M.S. thesis, Dept. Comput. Sci., Univ. Dublin, Trinity College, Republic of Ireland, 2003, pp. 1_100.

[4] Chen Ye-wang, Yu Jin-shan. An Improved Text Classification Method Based on Bayes. *Journal of Huaqiao University (Natural Science)*. 2011; 32(4): 401-404.

[5] Ikonomakis, M., Sotiris Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." *WSEAS transactions on computers* 4.8 (2005): 966-974.

[6] Ji, Hua, and Huaxiang Zhang. "Analysis on the content features and their correlation of web pages for spam detection." *China Communications* 12.3 (2015): 84-94.

[7] Sethi, Gaurav, and Vijender Bhootna. "SMS spam filtering application using Android." *Int. J. Comput. Sci. Inf. Technol* 5.3 (2014): 4624-4626.

[8] Awad, W. A., and S. M. ELseuofi. "Machine learning methods for spam e-mail classification." *International Journal of Computer Science & Information Technology (IJCSIT)* 3.1 (2011): 173-184.

[9] http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/

[10] https://www.kaggle.com/uciml/sms-spam-collection-dataset

[11] Nikhil Kumar, Sanket Sonowal, Nishant. "Email Spam Detection Using Machine Learning Algorithms." 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA).doi:10.1109/icirca48905.2020.9183098

