IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Spam Detection using Naive Bayes Classifier

Pooja[1*], Komal Kumar Bhatia[2]

[1]Computer Engineering, YMCAUST, Faridabad, India
[2]Computer Engineering, YMCAUST, Faridabad, India

Available online at: www.ijcseonline.org

Accepted: 03/Jul/2018, Published: 31/Jul/2018

Abstract — In digital world, there is a drastic increase of the websites that encouraged users to give their reviews on products, services, policies. This task of different data gathering and analysis of review is known as Opinion Mining. It analyses the text written in a natural language and classify them as positive or negative based on the human's sentiments, emotions, opinions expressed on any product. Nowadays user reviews and comments are very important for further evaluating and making decision for new products or policies. This gave the chance to spammers to spread malicious reviews with a target to misguide users. Spam is the unwanted similar content flooded on the internet. There is a need to detect spam efficiently. This work focused on training words and finding out whether further sentences are spam or not spam to improve accuracy. This paper discuss and implements naive bayes classifier to detect spam reviews.

Keywords— Opinion mining, naive bayes, spam

I. INTRODUCTION

With the explosive growth of social media (i.e, reviews, forum discussions, blogs and social networks) on the Web, individuals and organizations are increasingly using public opinions to make decisions and improve the quality of social media. Opinion mining is a technique which is used to identify and select useful information from large set of text. Spammers take advantages of these opinions to spread spam all over the network.

Spam messages are the unwanted, irrelevant text messages which are forcefully posted, sent to socially active user. Many copies of these messages are inundated all over the internet. These bulk messages could contain malicious links, fraudulent reviews, insults, blackmail letters, fake jobs, viruses, personal information of any authenticated user. These messages are free of cost and time. These messages are sent with the intention that few people might fall in their trap and respond to spam messages.

Initially spam messages are common visitors on email, they further expanded their roots to social networking sites. Emails received piles of messages from a fake account claiming itself as validated user. Fraud emails stated that you have win an amazing world tour, or send your money to get it double, or asking your bank details by affirming that they are bank employees, or deceiving job seekers. Now days spam messages are frequently spread in networking sites. Mostly spam messages are used for advertising their product, brand or organization. Spam messages are total waste of time for legitimate users, also it consumed network's bandwidth completely.

Social media attracted spammers because spam messages circulate easily to millions of people, if they got viral. Spam advertisements are so appealing to users that stimulate them to open it. Once you open the link you moved yourself in a pit of malicious content. These advertisements contained miscellaneous links, that provided more fruitful results to spammers as users do what they have planned.

A lot has been done to detect spam but it requires a lot more. Email filters are made advanced that stopped unwelcomed spam messages excellently. Since the data is in abundance so it impossible to stop spam completely from spreading. It is found that only 15% of the links are detected as spam. It is very difficult to perceive spam. Consequently, many people are engaged to develop new techniques to fight with spam content. Some policies are also encouraged to prevent spammers.

II. LITERATURE SURVEY

Web spam which is a major issue throughout today's web search tool; consequently it is important for web crawlers to have the capacity to detect web spam amid creeping. The approach used in [1] aimed to improve Paul Graham's Naive Bayes to separate the spam and non-spam mails. Emails are the easiest source for spammers. Big Data analyzed framework which is also outline for spam detection. Extricating the feeling from a message was also

one of the method to get the valuable data. Review Spam Detection is an important task in opinion mining. The paper [2] discussed the machine and non - learning machines advantages and disadvantages which can be used for spam detection of email messages.
Paper [3], [4] and [5] discussed the opinion mining techniques and their challenges. Reviews, sentiments, opinions, etc are considered most important for analysis in today's online world. The paper [6] represented an algorithm to measure junk mails and to separate spam and non - spam mail. This algorithm was based on the Public consultation and voting System. It has been found in their work that the error rate has been improved by 39%. The problem in product reviews is presented for the first time in [7]. It detected spam reviews on product reviews. In it duplicate and near duplicate reviews are assumed to be fake reviews. Supervised learning algorithms has been considered as one of the best algorithms to detect spam review.
In [8], the impact of single reviewer to the online store, and anomaly pattern of rating are analyzed to detect the spam reviews. In [9], the unusual review patterns which can represent suspicious behaviors are identified, and unexpected rules are formulated.

In [10], rating score of reviews is calculated. This rating score helped to detect spammer who tend to deviate user from correct product.
Reviews are ranked according to the rating score. In [11], spam is detected based on sentiments using five classifiers such as BayesNet, Naive Bayes, Random Forest, Support Vector Machine, J48. In [12], dataset of reviews are analyzed based on Natural language processing(NLP). Based on the information it tried to enhance the classification of reviews. The paper [13], used classification methods for text classification, pattern matching, feature extraction to increase accuracy. It used methods like Neural network, decision trees and increment component analysis In [16], three approaches to detect deceptive opinion spam by integrating work from psychology and computational linguistics.

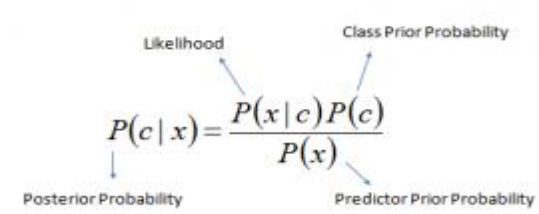
In Machine learning has made innovations from the prepared datasets furthermore anticipated the choice making framework hence they are broadly utilized as a part of feeling order with the exceptionally precision of framework.

III. NAIVE BAYES ALGORITHM

Naive Bayes algorithm[20] is the simplest and easiest technique commonly used for text classification. It is one of the effective and accurate supervised machine learning algorithm. It is a simple probabilistic classifier based on Bayes theorem. A more precise term for this probabilistic model is termed as "Independent feature model". The features known as words are not dependent on the other data. It computed the calculation of probability of trained data.

It used a file as a bag of features for text classification. It does not classify on the one or two words of data but take account of each and every relevant word. Bayes theorem is used for calculation of probability of features.

Bayes Equation[21] :-



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

.....**(1)**

where P(c|x) = Probability of class when the attributes are given,
P(x|c) = Probability of a attribute given that attribute belongs to that class
P(c) = Probability of class
P(x) = Probability of attributes

The conditional probability used in Naive Bayes is :-

$$\mathbf{P(Sentiment/Sentence)} = \frac{\mathbf{P(Sentiment)\, P(Sentence/Semtiment)}}{\mathbf{Sentence}}$$

.....(2)

IV. IMPLEMENTATION

This work used the method called Naive Bayes classifier which is used to perform spam classification on reviews. Naive Bayes is the most used supervised machine algorithm for spam detection. The aim to use this algorithm in the proposed work is it is one of the well organized, well planned, efficient and reliable algorithm. It is one of classification algorithm which gives accurate result. It not only classify one or two word rather it takes each word for consideration.

It is a probability model for classification that assigned probability values to words, represented these values as feature. It showed that features are independent, that is no feature is connected with the other feature. It is based on

bayes theorem with the independence assumptions between predictors.

According to this theorem, if there are two events say, e1 and e2 then the condition probability of occurrence of event e1 when e2 has already occurred is given by the following mathematical formula:

$$\mathbf{P}(e1/e2) = \frac{\mathbf{P(e2|e1)P(e1)}}{\mathbf{P(e2)}} \quad \textbf{.....(3)}$$

This algorithm is implemented to calculate the probability of a data to be spam or not spam.

$$\mathbf{P(Sentiment/Sentence)} = \frac{\mathbf{P(Sentiment)\ P(Sentence/Semtiment)}}{\mathbf{Sentence}}$$

.....(4)

Sentence in the equation (4) showed the spam or non spam messages in the dataset and sentiment represent the each word in the sentence.

A) Algorithm

S1: Initialize P(positive) num as (not spam)/ num_total
S2: Initialize P(negative) num as (spam) / num_total
S3: Convert sentences into words for each class of {not spam, spam}:
S4 :for each word in {phrase}
P(word/class) =

$$\frac{\textbf{No of word occurence in class + 1}}{\textbf{Number of words belonging to a class + Total no of words}}$$

S5 : Classify a new sentence according to:
S5 (i) : Final value for spam, Vs = P(spam)*P(word1/spam)*P(word2/spam)* ..... *P(word i/spam)
S5 (ii): Final value for non-spam, Vns = P(non-spam)*P(word1/non-spam) * P(word2/non-spam)*..... *P(word i/non-spam)
S6 : Returns max {Vs, Vns }

This paper described the method which is used to perform spam classification. The first step is to select the file from the dataset and apply the feature extraction technique for extracted feature.

The Feature Extraction :- The features are extracted and words are counted. This word count gave the accurate result. In this work the probability of each word coming in spam and non spam sentences was calculated. And then it counted the total number of unique word out of the total words for both spam and non-spam and found the frequency of that word in a particular document. The main thing about this algorithm is to makes a dictionary. In that dictionary probability of occurring a word in the sentence is stored.

In this work words are counted for extracting features from the dataset. Here we using data set which contains around 100 messages which are trained as spam and non-spam messages.

Here, spam filtering was done such as spam review detection and non-spam review detection. It aimed to detect review by calculating their probability values. Probability of each word is calculated for spam and non-spam review. On the sentences ,we train naive bayes algorithm.

The probability of the classification word with prior knowledge is calculated as P(word/spam) for spam detection and P(word/non-spam) for non-spam detection. Probability of these values is known prior to us.

In this step, dataset is trained. For training the data we calculated the probability of spam and non-spam words in the document. This trained data contained probability of words which are further used to calculated values according to the formula given below.
Conditional probability of a word is given as:

$$\mathbf{P(Word/Sentiment)} = \frac{\textbf{No of word occurence in class} + \mathbf{1}}{\textbf{Number of words belonging to a class} + \textbf{Total no of words}}$$

....(5)

The next step was to calculate the probability of words with the help of Naïve Bayesian Classifier for calculating the probability of new spam and non-spam words and make a decision which value is higher. If spam words are greater than non-spam words then the message is spam otherwise non-spam message.

This was calculated using equation (5). In Equation (5) total number of words occurred are added to 1. 1 is added to the total number of words because if suppose there is no word in the dictionary then the result would be zero. The values of these words is used further in naive bayes algorithm to calculate whether the word is spam or not.

Hence, some random words are trained. These words are extracted from messages which are the reviews which are given by users. These words are trained to calculate the polarity of new words. These words are provided with trained probabilitc values which are further used by the formulas given above.

The advantages of naive bayes are :-

- It is the best classifier for training small dataset.
- It gives accurate result as it works on probabilistic values.
- It computes efficiently especially for text classification.
- It is easy to interpret for Binary and Multiclass classification.
- It is consider over other algorithms like logistic regression because of its independent, scalable and flexible nature.

The disadvantages of naive bayes:-

It can learn independent features but it cannot establish a relation between them.

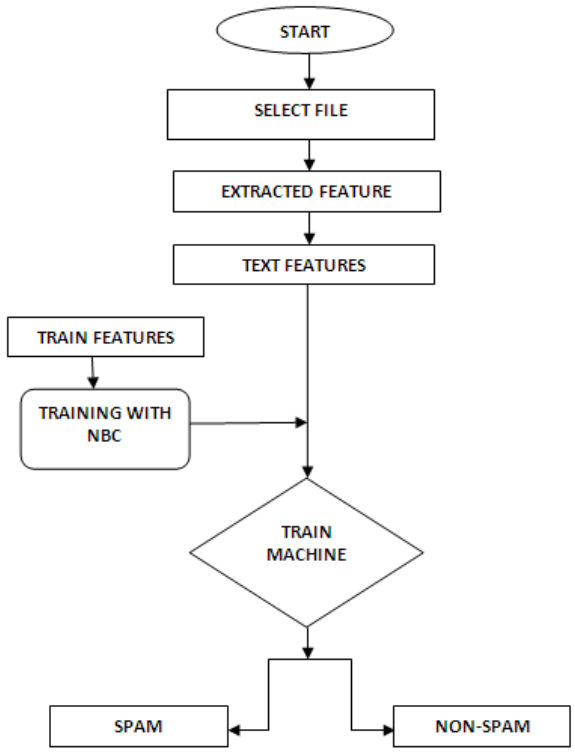


Fig 4.1 Flow chart for classification of naive bayes

V. RESULTS

5.1 For single value

Initially a single word is checked whether it is spam or not spam. This word belonged to the trained dataset. A new sentence is taken as a input from which spam and non - spam words are found.

```
Enter the new sentence
review us with your unique id and password

Input the word from given sentence which you want to check
password

Finding out whether the values are spam or not
spam value is :0.0035576925
Non Spam value is :0.0037472527
Maximum is nonspam =0.0037472527
Accuracy is =0.9999743
```

Fig 5.1 Checking Spam Or Non - Spam For A Single Value

5.2 For multiple values

A new sentence was given input to find out the spam and non - spam values. In this set of words are considered from the sentences to find out whether it is spam or non - spam.

```
Enter the new sentence
give us your mail password

Input the word from given sentence which you want to check
mail password

Finding out whether the values are spam or not
spam value is :2.7366865E-5
Non Spam value is :2.3677696E-5
Maximum value is spam =2.7366865E-5
Accuracy is =0.9999997
```

Fig 5.2 Checking Spam or Non - Spam For A Multiple Value

5.3 For word not present in the sentence

The following figure shows that the feature is not present, which means this is not trained feature. Hence, it return the value not found when the feature is not trained earlier in Naive Bayes.

```
Enter the new sentence
review us

Input the word from given sentence which you want to check
and
The string is not present

Finding out whether the values are spam or not
spam value is :0.37
Non Spam value is :0.62
Maximum is nonspam =0.62
```

Fig 5.3 Result

VI. CONCLUSION

Spam classification has major issue in today's electronic world. Spam is most crucial matter in a social network. There are many problem created through spam. The spam is unwanted message or email which the end user clients are receiving in our daily life. Spam messages are nothing but it is the advertisement of any company, any kind of virus etc. Because of these spam the performance of the system can be degraded and also affected the accuracy of the system.

To solve this problem spam classification system was created which identified the spam and non-spam messages. Here the Naïve Bayesian Classifier was used and extracted the word using word-count algorithm. After calculating the new probabilitic result we found that naïve Bayesian classifier has more accurate results. The error rate is very low when we are using the Naïve Bayesian Classifier. So we can say that Naïve Bayesian Classifier produce better result than Support Vector Machine.

7. REFERENCES

[1] Sharma K, Jatana N, "Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach". IEEE 2014 pp. 939-942.

[2] Sandeep Negi, Rekha, "A Review on Different Spam Detection Approaches" International Journal of Engineering Trends and Technology (IJETT) – Volume 11 Number 6 - May 2014.

[3] P.Kalarani, Dr.S. Selva Brunda, "An Overview on Research Challenges in Opinion Mining and Sentiment Analysis" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, October 2015.

[4] Nidhi R. Sharma , Prof. Vidya D. Chitre, "Opinion Mining, Analysis and its Challenges" International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 3, Issue 1 April 2014.

[5] Ayesha Rashid, Naveed Anwer, Dr. Muddaser Iqbal, Dr. Muhammad Sher, "A Survey Paper:Areas, Techniques and Challenges of Opinion Mining" International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013.

[6] Ali M. et al, "Multiple Classifications for Detecting Spam email by Novel Consultation Algorithm" CCECE 2014, IEEE 2014, pp. 1-5.

[7] Jindal Nitin, Liu Bing, "Opinion spam and analysis" Proceedings of the 2008 International Conference on Web Search and Data Mining. New York: ACM Press , 2008:219-230.

[8] Xie Sihong, WANG Guan, LIN Shuyang, et al, "Review spam detection via temporal pattern discovery" Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. New York: ACM Press , 2012:823-831.

[9] Jindal Nitin, Liu Bing, Lim Ee-peng, et al, "Finding unusual review patterns using unexpected rules" Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM Press , 2010:1549-1552.

[10] Lim Ee-Peng, Nguyen Viet-An, Jindal Nitin, et al, "Detecting product review spammers using rating behaviors" Proceedings of the 19th ACM international conference on Information and knowledge management. New York: ACM Press , 2010:939-948.

[11] Nasira Perveen, Malik M. Saad Missen, Qaisar Rasool, Nadeem Akhtar, "Sentiment Based Twitter Spam Detection" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016.

[12] Swati N. Manke, Nitin Shivale, "A Review on: Opinion Mining and Sentiment Analysis based on Natural Language Processing" International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 4, January 2015.

[13] Anchal, Abhilash Sharma, "SMS Spam Detection Using Neural Network Classifier" International Journal of Advanced Research in Computer Science and Software Engineering Research Paper, Volume 4, Issue 6, June 2014.

[14] Behrouz Minaei-Bidgoli, Saeedeh Sadat Sadidpour, Hossein Shirazi, Nurfadhlina Mohd Sharef, Mohammad Ebrahim Sanjaghi, "Context-Sensitive Opinion Mining using Polarity Patterns" International Journal of Advanced Computer Science and Applications, Vol. 7, No. 9, 2016.

[15] Nidhi Mishra and C K Jha, "Classification of Opinion Mining Techniques" International Journal of Computer Applications 56 (13):1-6, October 2012, Published by Foundation of Computer Science, New York, USA.

[16] Oded Z. Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook" Springer, 2005.

[17] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Sentiment classification using machine learning techniques." In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

[18] Myle Ott, Yejin Choi, Claire Cardie, et al. Hancock, "Finding deceptive opinion spam by any stretch of the imagination" Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011: 309-319.

[19] Haseena Rahmath P, "Opinion Mining and Sentiment Analysis - Challenges and Applications" International Journal of Application or Innovation in Engineering & Management (IJAIEM). Volume 3, Issue 5, May 2014.

[20]https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[21] Nikhila Zalpuri, Meena Arora, "An Efficient Model for S.M.S Security and SPAM Detection: A Review", International Journal of Computer Sciences and Engineering, volume - 3, Issue - 12,Dec2015.

[22] S. Nagaparameshwara Chary, B.Rama, "Analysis of Classification Technique Algorithms in Data Mining" International Journal of Computer Sciences and Engineering, volume-4, Issue - 6, june 2016.

