|  |
| --- |
| **Spam Message Filtering possible application of NLP** |
|  |
| **Akash Chandra Das,**  Department of Computer Science and Engineering, BRAC University,  Dhaka, Bangladesh |
|  |
|  |
|  |
|  |

Abstract

|  |
| --- |
| 000  001  002  003  004  005  006  007  008  009  010  011  012  013  014  015  016  017  018  019  020  021  022  023  024  025  026  027  028  029  030  031  032  033  034  035  036  037  038  039  040  041  042  043  044  045  046  047  048  049 |

Email is an important method for all communications that is fast, cheap, accessible and easily replicated. But spam emails that is referred to as junk email or simply SPAM, is unsolicited messages sent in bulk by email can cause huge problem for individuals. To stop it we need a spam filtering method. Naive Bayes classifier is a very effective method to detect such spam emails.

Introduction

Email has become very popular after its invention. Because of too much of spam email it is causing a lot of fraudulent activities. So, many techniques have been proposed to check and filter not only the spam emails but also other texts such messages and comments in twitter and Facebook. In these purpose naïve bayes classification is a very effective algorithm to filter out the spam emails.

Literature Survey

|  |
| --- |
| 050  051  052  053  054  055  056  057  058  059  060  061  062  063  064  065  066  067  068  069  070  071  072  073  074  075  076  077  078  079  080  081  082  083  084  085  086  087  088  089  090  091  092  093  094  095  096  097  098  099 |

From statistics we have found that around 306 billion emails have been sent and receive in 2020. Total user of email is now about 4 billion and it will be 4.6 million in 2025. But it is also causing a problem.

Project Plan

For this NLP application problem we will use Multinomial Naive Bayes which can easily detect any spam. A standard dataset that acts as a model recipient mailbox with both the collections of 5695 ham and spam email messages. In this experiment we will use Kaggle dataset which contains about 4327 ham mails and 1368 spam mails in total as a

sample to process. Implementation of this work can done using tools such as Python 3.8, PyCharm professional, Windows 10 and 8 GB RAM for smooth working purpose on the big data set. We will use Count Vectorizer for the feature extraction. We can also use BERT to this for better outcome. We have to use 10-fold cross validation to measure the performances. 20 percentage of data will use for testing then our multinomial naive bayes will perform and it will show ham and spam detection accuracy a great percentage. Then we can find MNB takes the lowest time for training. And our application can developed like this.

Potential Challenges

We can face challenges with huge type of data. Multinomial naive bayes work better when we will use small percentage of data for leaning. Otherwise we may not get our expected accuracy for spam checking. Plus we have to process this data values to numeric format and have to check this dataset is there any duplicate values, if their then we have to drop duplicate data’s.

|  |
| --- |
|  |

Conclusion

The multinomial naive bayes (MNB) is an email spam classifier which can capable of classifying with an average of 99.5% accuracy. Moreover, it requires a lesser amount of data for training and to give its standard performance with a very low training time of 1.049 seconds. MNB is a fast and reliable classifier because of its nature of relating.

Bibliography

Weimiao Feng proposed a new algorithm namely Support Vector Machine – Naive Bayes (SVM-NB).Then Dr Devendra Tayal Modified K means algorithm to remove empty clusters. Nikhil Mathew used N-grams technique for creating feature set. Rohit Kumar Solanki proposed an algorithm which involves usage of local and global classifier. Simranjit Kaur Tuteja proposed an algorithm using BPNN.