IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

SPAM MAIL FILTERING USING MULTINOMIAL NAÏVE BAYSIAN ALGORITHM

Dr. Shaik Nagul[1], Mr.A.V.D.N.Murthy[2]

Associate professor of C.S.E[1], LENDI INSTITUTE OF ENGINEERING AND TECHNOLOGY Assistant professor of C.S.E[2], LENDI INSTITUTE OF ENGINEERING AND TECHNOLOGY.

ABSTRACT

Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spam is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious links through spam emails which can harm our system and can also seek into your system. Creating a fake profile and email account is much easier for the spammers, they pretend like a genuine person in their spam emails, these spammers target those people who are not aware about these frauds. So, it is necessary to Identify those spam mails which are fraudulent. In the existing proposed system they classify the mail using Support Vector Machine (SVM) by feature extraction. In this proposed system we identify spam by using techniques of machine learning, this system will discuss the Multinomial Naïve Bayesian Algorithm which is used for supervised learning. The Bayesian classifier works on the probability of the event which occurred previously. The Naïve Bayes always calculates the probability of each class and class having the maximum probability is chosen as output. The Naïve Bayes filter very well and work effectively.

Key Words: Spam filtering, Mail classification, Support vector machines, Multinomial Naïve Bayesian.

1. INTRODUCTION

The main theme of our work is to classify unwanted commercial bulk emails called spam by using Naive Bayesian classifier by using text classification approach to prevent the spam mails that are becoming

a huge problem on the internet. By identifying the spam mail it prevents the user from making full and good use of time.

Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, and CPU power and user time.

This system consists of a huge number of sample mails that are extracted from the kaggle dataset to train the given naive Bayesian algorithm. After training we give an input mail the input mail is compared with sample data and calculate the probability of matching the words in both the mails. If the probability is greater than threshold value the input mail belongs to spam or if the probability is less than threshold value then the input mail belongs to ham. We estimate the threshold value while we train the model.

Familiarization with data set and algorithm should be done and program should be designed for a email spam filtering using main components like multinomial naive Bayesian algorithm and some python libraries like pandas,numpy and scikit learn package to run machine learning algorithms and sample data set that should be trained and program should be designed and tested on them Components Required.

By filtering the spam mails, we avoid the problems like, users who receive spam emails that they did not request. It is also resulted to untold financial loss to many users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails pretending to be from reputable companies with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers.

2. CURRENT APPROACHES OF CLASSIFYING DATA

Machine learning (ML) researchers have developed many approaches in order to tackle email spam filtering. Within the context of machine learning, support vector machines (SVM) have made a large contribution to the development of spam email filtering. Based on SVM, different schemes have been proposed through text classification approaches (TC). A crucial problem when using SVM is the choice of kernels as they directly affect the separation of emails in the feature space.

This paper presents a thorough investigation of several distance-based kernels and specifies spam filtering behaviors using SVM. The majority of used kernels in recent studies concern continuous data and neglect the structure of the text. In contrast to classical kernels, we propose the use of various string kernels for spam filtering. We show how effectively string kernels suit spam filtering problems. On the other hand, data preprocessing is a vital part of text classification where the objective is to generate feature vectors usable by SVM kernels. We detail a feature mapping variants in TC that yield improved performance for the standard SVM in filtering tasks.

Support Vector Machines are known to give accurate discrimination in high feature space and it received a great attention in many applications such as text classification. SVMs have out-performed other learning algorithms with good generalization, global solution, number of tuning parameters, and its solid theoretical background. The core concept of SVMs is to discriminate between two or more classes with a hyperplane maximizing the margin by solving quadratic programming (qp) problems with linear equality and inequality constraints. Machine learning (ML) researchers have developed many approaches in order to tackle email spam filtering. Within the context of machine learning, support vector machines (SVM) have made a large contribution to the development of spam email filtering. Based on SVM, different schemes have been proposed through text classification approaches (TC).

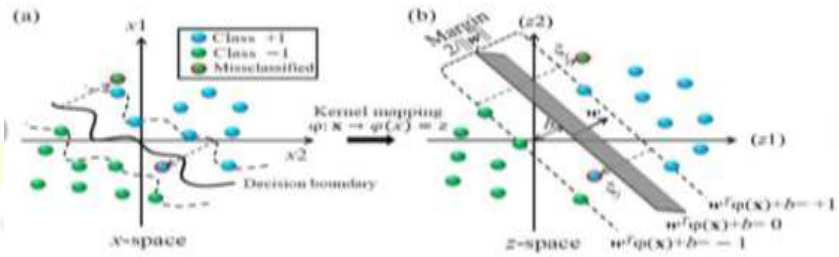


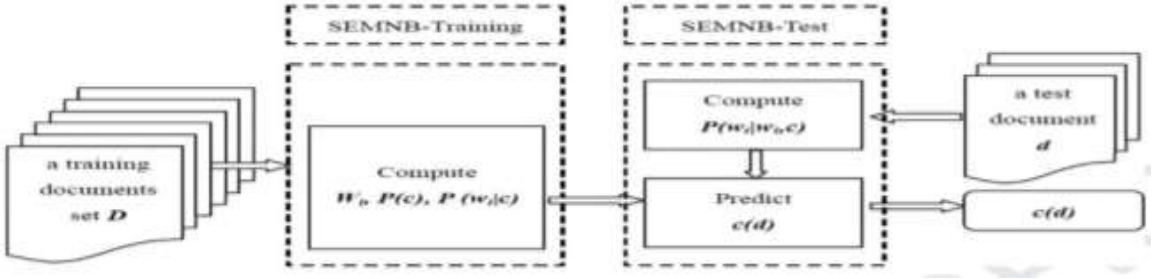
Fig: SVM classification using kernels

A crucial problem when using SVM is the choice of kernels as they directly affect the separation of emails in the feature space. This paper presents a thorough investigation of several distance-based kernels and specifies spam filtering behaviors using SVM.

The majority of used kernels in recent studies concern continuous data and neglect the structure of the text. In contrast to classical kernels, we propose the use of various string kernels for spam filtering. We show how effectively string kernels suit spam filtering problems. On the other hand, data preprocessing is a vital part of text classification where the objective is to generate feature vectors usable by SVM kernels. We detail a feature mapping variants in TC that yield improved performance for the standard SVM in filtering tasks. Support Vector Machines are known to give accurate discrimination in high feature space and it received a great attention in many applications such as text classification .

SVMs have out-performed other learning algorithms with good generalization, global solution, number of tuning parameters, and its solid theoretical background. The core concept of SVMs is to discriminate two or more classes with a hyperplane maximizing the margin by solving quadratic programming (qp) problem with linear equality and inequality constraints

3. PROPOSED METHOD FOR FILTERING EMAIL



Although there are many types of spam emails, the underlying characteristics of spam cannot be changed. We keep track of words in each email to start our filtering. For any text based big data analysis, data preprocessing is the most important step. This step helps us to remove meaningless words and text symbols. This step also includes Natural Language Processing (NLP). NLP is the ability for a computer to understand human languages and it can help us to classify text. In many approaches, Naive Bayes Classifier always does surprisingly well, so it has been widely used in several spam filtering. Naive Bayes has two popular models: Bernoulli model and Multinomial model. For Bernoulli, the significant difference from Multinomial is not only because it does not take into consideration the number of occurrences of each word, but also because it takes into account the non-occurring terms within the document. Thus, we use the Multinomial Naive Bayes model. After data pre-processing, each distinct word in the dataset is defined as an attribute and the value of the attribute is the English word found in the dataset.

The following steps are used to spam mail filtering using Naive Bayes . We use the kaggle dataset for training and testing. There are 6000 emails consisting of [500 legitimate (ham) emails and 4500 spam

Fig: A Classification approach

emails. Emails have already been labeled as spam or ham. Since we apply each distinct word in the dataset as an attribute, emails should be clean and categorized.

To overcome some of these drawbacks of existing system by replacing them with Multinomial Naive Bayes Classifier.This approach basically follows 4 steps: PreProcessing, Tokenizing, count vectorization, Classification of mail either spam or not spam which is commonly used in text classification.

Naive Bayesian Classifier works by correlating the tokens with spam & non spam emails & then it uses Bayes theorem to calculate the probability that a mail is spam/not spam.

Algorithm

Algorithm 1: Multinomial Naïve Bayes

```
Initialise Input Variables;
N ← No. of Documents;
X ← Datapoints;
y ← Target Inputs;
for i = 0; i < TrX; i++ do
    if (i,y) = Spam then
        Learn i = Spam;
    else
        Learn i = Ham;
for t in testSize // Test sizes = 20,
    30 and 40
do
    for K in CV do
        X_test and y_test = testing size;
        X_train and y_train = training size;
        for i = 0; i < TeX; i++ do
            Calculate P̂(t_k|p);
            Calculate the Accuracy;
    return t_k;
```

4.RESULTS

The experimented results of the proposed approach are given below :

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

Fig: Sample dataset

Use the Jupiter notebook and import pandas to load the sample input mail dataset file.

Steps to input data and categorize the data are:

1.Import pandas as pd

2. df= pd.read_csv("spam.csv")

3. df.head()

4.df.groupby('Category').describe()

Extract the features using count vectorization

V= CountVectorizer()

X_train_count = v.fit_transform(X_train.values)

X_train_count.toarray()[:3]

After countvectorization, the model classifies the email into either ham or spam using multinomial naïve basian

algorithm. Based on the probability threshold value, the mail is classified. If the probability is > the threshold value(i.e here we took probability as 98%) then mail is categorizaed as spam. Else , ham.

```
In [11]: clf.score(X_test,y_test)
Out[11]: 0.9899497487437185
```

Predictions: Input: mail Messages dataset Output: 0 or 1

5. CONCLUSION

Naïve Bayes is a useful algorithm in many respects, especially for solving low-data text classification problems. The way in which it can make accurate predictions with limited training data by assuming naïve independence of words is its main advantage and In pre-processing phase, the number of vocabulary may be so large. It's because spam emails may have many spaces,

and one word may be split by spaces to two words. Fortunately, Naive Bayes Classifier can handle this problem. By using Naive Bayesian Algorithm, we achieve our goals and get a satisfactory conclusion with a good precision rate.

6.REFERENCES

[1] S. K. Tuteja, “Classification Algorithms for Email Spam Filtering", 2016.

[2] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues", 2017.

[3] S. Ajaz, M. T. Nafis, and V. Sharma, "Spam Mail Detection Using Hybrid Secure Hash Based Naive Classifier, 2017.

[4] Shafi’i Muhammad Abdul Hamid, M. S., Osho, O., Ismaila, I., & Alhassan, J. K. “Comparative Analysis of Classification Algorithms for Email Spam Detection”, 2018.

[5] Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H.. “AnalysisofNaïveBayes Algorithmfor Email Spam Filtering across Multiple Datasets”, 2017.

[6] Yüksel, A. S., Cankaya, S. F., & Üncü, İ. S. “Design of a Machine Learning Based Predictive Analytics System for Spam Problem”, 2017Verma, T,”E-Mail Spam Detection and Classification Using SVM and Feature Extraction”, 2017.

[7] Singh, V. K., & Bhardwaj, S,”Spam Mail Detection Using Classification Techniques and Global Training Set”, 2018.

[8] Priti Sharma, Uma Bhardwaj,”Machine learning based Spam emaildetection”,2017

[9] Manmohan Singh, Rajendra Pamula, Shudhanshu Kumar shekhar”.Email Spam Classification by Support Vector Machine”,2018.

[10] Linda Huang, Julia Jia,Emma Ingram, Wuxu Peng “Enhancing the Naive Bayes Spam Filtering through Intelligent Text Modification Detection”,2018

[11] Prachi Gupta, Ratnesh Kumar Dubey,Dr. Sadhna Mishra,”Detecting Spam Emails/Sms Using Naive Bayes And Support Vector Machine”,2019.

[12] Sah, U. K., & Parmar, N,”An approach for Malicious Spam Detection in Email with comparison of different classifiers”, 2017.