International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

SPAM DETECTION OF EMAIL

Tejas Raut[1], Aniket levharkar[1], Gaurav Dahate [1], Sahil Ande[1], Prof. T.G. Ghongade[2]

[1]Computer Science And Engineering, P.R.Pote College Of Engineering and Management , Amravati Maharashtra, India

[2]Prof. T.G. Ghongade, Computer Science And Engineering, P.R.Pote College Of Engineering and Management , Amravati Maharashtra, India

---

Abstract – Nowadays, a big part of people rely on available email or messages sent by the stranger. The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam message about our different interests .Spam fills inbox with number of ridiculous emails . Degrades our internet speed to a great extent .Steals useful information like our details on our contact list. Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. Email spam is an operation to send messages in bulk by mail .Since the expense of the spam is borne mostly by the recipient ,it is effectively postage due advertising. Spam email is a kind of commercial advertising which is economically viable because email could be a very cost effective medium for sender .With this proposed model the specified message can be stated as spam or not using Bayes' theorem and Naive Bayes' Classifier and Also IP addresses of the sender are often detected **.**

Keywords: Spam Content, Machine learning, Deep learning, Natural language processing, Social media analysis.

1.INTRODUCTION

In recent years, internet has become an integral part of life. With increased use of internet, numbers of email users are increasing day by day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam. Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. a large number of identical messages are sent to several recipients of email. Spam usually arises as a result of giving out our email address on an unauthorized or unscrupulous website .There are many of the effects of Spam .Fills our Inbox with number of ridiculous emails .Degrades our Internet speed to a great extent .Steals useful information like our details on you Contact list .Alters your search results on any computer program .Spam is a huge waste of everybody's time and can quickly become very frustrating if you receive large amounts of it .Identifying these spammers and the spam content is a laborious task . even though extensive number of studies have been done, yet so far the methods set forth still scarcely distinguish spam surveys, and none of them demonstrate the benefits of each removed element compose .In spite of increasing network communication and wasting lot of memory space ,spam messages are also used for some attack . Spam emails, also known as non-self, are unsolicited commercial or malicious emails, sent to affect either a single individual or a corporation or a bunch of people. Besides advertising, these may contain links to phishing or malware hosting websites found out to steal confidential information. to solve this problem the different spam filtering techniques are used. The spam filtering techniques are accustomed protect our mailbox for spam mails.

2.METHODOLOGY

We begin with conversion, preprocessing and splitting of datasets as per the requirement of considered algorithms. The various models are then trained and tested, followed by evaluation and comparison based on performance metrics.

A. Dataset Description

Dataset#1: SMS Spam Collection V.1

This is a collection of 5574 spam and legitimate English text messages gathered from the following free research sources: National University of Singapore SMS Corpus (3,375 Ham SMS), Grumbletext Website (425 Spam SMS), Caroline Tag's PhD Theses (450 Ham SMS), and SMS Spam Corpus

v.0.1 Big (1002 Ham SMS and 322 Spam SMS). With a totalof 4827 legitimate messages and 747 spam messages, the corpus is hosted at the UCI Machine learning repository and also available in raw format publicly [14].

Dataset#2: Spam SMS Dataset 2011-12

The SMS database contains 1,000 Spam and 1,000 Ham SMS. For collecting SMS spam data, Yadav et al. ran an incentivized crowd-sourcing scheme in their campus. Due to large influence of regional words, SMS with both Hindi & English words were collected from 43 participants. The dataset is available on request at [15].

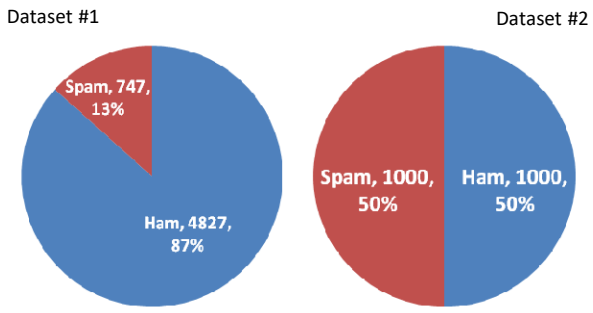

Fig. Distribution of Datasets

B. *Preparing Readable Datasets*

The datasets have been prepared as comma-separated values (CSV) files. These files contain one text message per line. Each line has two columns - *v1* is the label (ham or spam)and *v2* is the raw text.

The Spam SMS Dataset 2011-12 was procured as a zipped file with numerous text files containing a message each. The name of the file indicated whether the message it contained was legitimate or spam. A script was prepared and executed to accurately label the messages and unify them into a single CSVfile.

For any classifier to be able to use this data, we need to do some preprocessing.

C. *Data Preprocessing*

Different preprocessing approaches have been applied to different classifiers based on their requirement of input data. Following is a brief description of these approaches.

1) Using Term Frequency—Inverse Document Frequency (tf-idf): In a given document, the count of the number of times a word appears is called Term Frequency. In a given corpus of documents, the number of times a word appears is called Inverse Document Frequency. Words are weighed according to the importance in tf-idf. Frequently used words have a lower weight, while words used infrequently have higher weight.

We started with removal of stop words, capital letters, non alpha-numeric characters and any unnecessary punctuation. Wethen collected similar words (for example, desks will be transformed to desk). We then converted the cleaned text to tf- idf features (5000 features for an entry) using sklearn's TfidfVectorizer to create a bag of words i.e., a count vector, followed by tf-idf matrix.

2) Using Tokenizer: When working with text, it is always good to start with splitting the text into words. Words are known as tokens. Tokenization is the process of splitting text into words or tokens. Keras' Tokenizer is a class for vectorizing texts. It is used to turn texts into sequences. A sequence is a list of word indexes where the word of rank *i* in the dataset (starting at 1) has index *i*.

Text can be split into a list of words using the text_to_word_sequence() function provided by Keras. This function by default splits words by space, converts text to lowercase and filters out punctuation. Tokenization is restricted to the top most common words in the corpus. We defined 5000 as the maximum number of words to work with. After applying either of the approaches on our datasets for different classifiers, the representation of first column was changed. Ham and spam were transformed to values 0 and 1 using sklearn's LabelEncoder.

D. *Training and Test Datasets*

The datasets were split into two parts - the data that will be used to train our classifiers and the data that will be used to testthem. We split our datasets such that 80% of the data was usedfor training while 20% of the data was used for testing.

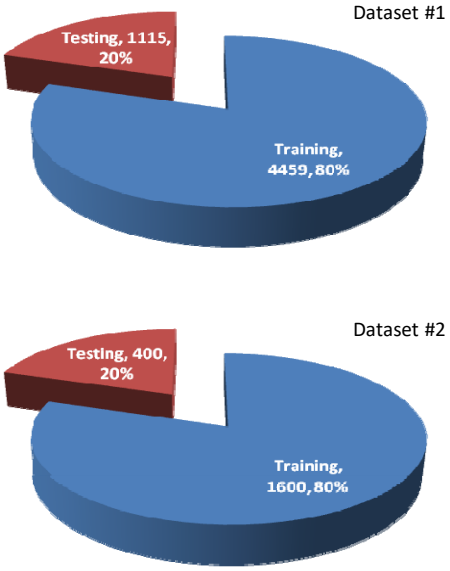

Fig. Splitting of Datasets

E. *Different Classifiers*

Different preprocessing approaches have been applied to different classifiers based on their requirement of input data. Following is a brief description of these approaches.

1) Support Vector Machine (SVM): SVM is a discriminative classifier which is widely used for classification task. The algorithm plots each data item as a point in *n*- dimensional space assuming the value of each feature as the value of a particular coordinate. It then forms a line that splits the whole data into two differently classified groups of data. The closest points in the two groups will be the farthest away from this line.

2) Naive Bayes (NB): This classification technique is based on Bayes' theorem, assuming independence between predictors. The Bayesian classifier assumes that a particular feature in a class is not related to the presence of any other feature. Even if they do depend upon each other or upon the existence of the feature, the Bayesian classifier will consider all

of the desired properties to independently contribute to the probability. The classifier holds well when the desired input's dimensionality is high. It is regarded as simple and sturdy.

An advanced version of NB is Multinomial Naive Bayes (MNB). The main improvement is the independence between document length and class. It involves multinomial distribution which works well for countable type of data such as the words in a document or text. In simple terms, NB classifier involves conditional independence of each of the features in the model, whereas MNB classifier is a special case of a NB classifier which uses a multinomial distribution for each feature.

1) *Decision Tree (DT):* DT is a supervised learning algorithm which is normally preferred for classification tasks. The algorithm works well for both types of variables i.e., categorical and continuous. It starts with splitting the population into multiple homogeneous sets which is done on the basis of most significant attributes or independent variables. DT is non-parametric and hence the need for checking outlier existence or data linearity separation is not required.
2) *Logistic Regression (LR):* It is considered as the go-to method for classification involving binary results. It is mainly used in estimating discrete values which are based on set of variables which are independent. In more relative terms, LR outputs the probability of an event by fitting it into a logistic function which helps in prediction. The logistic function whichis mostly used is sigmoid.
3) *Random Forest (RF):* It is a term used for an ensemble of decision trees. The Random Forest classifier is a ensemble learning method which involves collection of decision trees. Voting is done to classify a new object which is performed by each tree i.e., the trees mark their votes for that class. The class having most number of votes decides the classification label.
4) *AdaBoost:* AdaBoost or Adaptive Boosting is a meta-machine learning algorithm, used to increase the performance of a classifier by simply using the weak classifiers to combine them into a strong one. The final output of the boosted classifier depends upon the weighted sum of the output of all the weak classifiers. A drawback of this technique is that although it predicts more accurately, it takes more time for building the boosted model.
5) *Artificial Neural Network (ANN):* ANNs are nonlinear statistical data modelling techniques defined over complex relationships between inputs and outputs. They have various advantages, but among them, learning by observing datasets is the most recognised one. It is considered as tool for random function approximation, which helps in estimating the most effective methods to come to solutions. One such network is CNN.
6) *Convolutional Neural Network (CNN):* A convolutional neural network (CNN) is a particular type of artificial neural network which uses perceptrons for supervised learning. The supervised learning is used to analyze data. There are a wide of range of applications involving CNNs. Traditionally used for image processing, CNNs are nowadays used natural language processing as well. A CNN in relative terms is known as a ConvNet. Similar to other ANNs, a CNN also has an input layer, some hidden layers and an output layer, but it is not fully connected. Some layers are convolutional, that use a mathematical model to pass on the results to layers ahead in the network

3.TRANING AND TESTING

Principally the informational index is separated into two sections i.e., test informational collection and train informational index. Preparing information is utilized to fabricate the AI model and afterward we test it with test informational index to check its exactness and accuracy and numerous different components.

Here Multinomial Naive Bayes used in training the model. Multinomial NB calculation is probabilistic learning technique which is for the most part utilized in Natural Language Processing. The calculation depends on the Bayes hypothesis and predicts the tag of a book like a piece of email or paper article. It figures the likelihood of each tag for a given example and afterward gives the tag with the most noteworthy likelihood as yield. Naive Bayes classifier is an assortment of numerous calculations where every one of the calculations share one basic guideline, and that is each component being arranged isn't identified with some other element. The presence or nonattendance of an element doesn't influence the presence or nonappearance of the other element.

In process of evaluation of result generate confusion matrix as the output which method is imported from sklearn. It is used to calculate accuracy, sensitivity, precision, recall etc.

4.RESULTS AND DISCUSSION

In this final step, on our prepared dataset, we will test our classification model and also measure the efficiency of SMS spam detection on our dataset. To assess the efficiency of Our defined category and make it comparable to existing approaches .SMS Spam detectors are benefital and used to future enhancement as this will detect the spam messages and network resources many upcoming detectors are upcoming in future enhancement.

Once you have done all of the above, you can start running the API by either double click temp.py or executing the command from the Terminal and open Application so the output will be in following:

```
(base) PS C:\Users\Tejas Raut> conda activate spam
(spam) PS C:\Users\Tejas Raut> cd E:\Download\Spam-Ham
(spam) PS E:\Download\Spam-Ham> python app.py
 * Serving Flask app 'app' (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 141-473-469
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Fig:- :- img command exe

OUTPUT:-

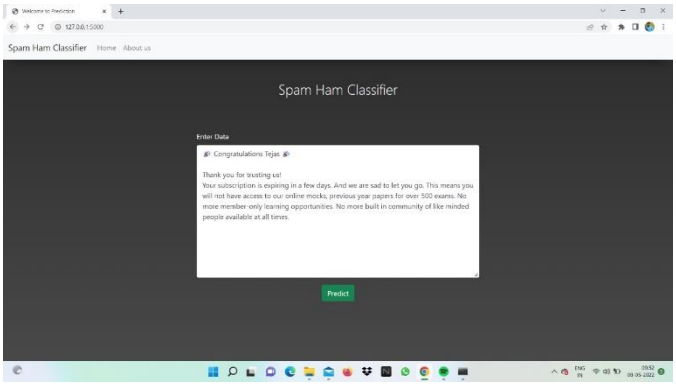

Fig:- Predict the Message Spam Or Not Spam

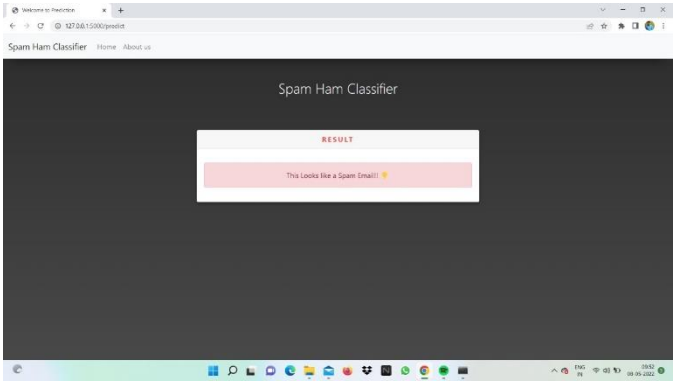

Fig:- OutPut

5.CONCLUSION

In this Project, Naïve Bayes Algorithm analyses based on the factors like precision, recall, f1- score, support. Naive Bayes order calculation is viably valuable for managing clear cut information characterization. The basic hypothesis it utilizes is the Bayes contingent probabilistic model for tracking down a back likelihood given certain conditions. It is classified "Credulous" on the grounds that under the presumption that all highlights (assortments of words) in the dataset are similarly significant and free. Utilizing the Naïve Bayes grouping calculation, the venture got over 98% exactness in foreseeing a spam message dependent on the words it contains which is determined from the confusion matrix which we got as yield.

	precision	recall	f1-score	support
ham	0.98	1.00	0.99	4825
spam	1.00	0.84	0.92	747
accuracy			0.98	5572
macro avg	0.99	0.92	0.95	5572
weighted avg	0.98	0.98	0.98	5572

Fig.. Result of NB classifier

REFERENCES

1.ShukorBin Abd Razak, Ahmad Fahrulrazie Bin Mohamad Design and Applications (ISDA), 2013.

2.Mohammed RezaParsei, Mohammed Salehi "E-Mail Spam Detection Based on Part of Speech Tagging"2 nd International Conference on Knowledge Based Engineering and Innovation (KBEI), 2015.

3.Sunil B.Rathod, Tareek M. Pattewar "Content Based Spam IEEE ICCSP 2015 conference.

4.Aakash AtulAlurkar, Sourabh Bharat Ranade, Shreeya Vijay Joshi, Siddhesh Sanjay Ranade, Piyush A. Sonewa, Parikshit N. Approach for Email Spam Classification using Machine Learning

5.Kriti Agarwal, Tarun Kumar "Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Conference on Intelligent Computing and Control Systems (ICICCS), 2018.

6.CihanVarol, HezhaM.TareqAbdulhadi "Comparison of String Matching Algorithms on Spam Email Detection", International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism Dec, 2018.

7.Duan, Lixin, Dong Xu, and Ivor Wai-Hung Tsang. "Domain adaptation from multiple sources: A domaindependent regularization approach." IEEE Transactions on Neural Networks and Learning Systems 23.3 (2012).

8.Mujtaba, Ghulam, et al. "Email classification research trends: Review and open issues." IEEE Access 5 (2017). 9.Trivedi, Shrawan Kumar. "A study of machine learning classifiers for spam detection." Computational and Business Intelligence (ISCBI), 2016 4th International Symposium on. IEEE, 2016. '

10. You, Wanqing, et al. "Web Service-Enabled Spam Filtering with Naïve Bayes Classification." 2015 IEEE First International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, 2015