iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36801

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Junk Filtering through Naive Bayesian Algorithm

Anushka Srivastava[1], Ayushi Saxena[2], Tanya Dhyani[3]
[1, 2, 3]*Student, Department of Information Technology, Computer Science, Electronics and Communication, Raj Kumar Goel Institute of Technology, Ghaziabad, India*

Abstract: ***As the world is seamlessly developing at a very high pace, we have been seeing enormous growth in various sectors of Technology. Networking has played a crucial part in the exchange of technological culture around the globe, and the Internet being the sole medium of Network enhancement has taken over every aspect of our society. Today, most of the professional communications are done through emailing. As far as email has proven to be an efficient, professional and easy way of communication, it also comes with the disadvantage of unwanted bulk bombarding of spam content. This has been a critical concern for email users. Consequently, it has become very difficult for spam filters to efficiently filter the unwanted emails, since nowadays emails are written in such a manner that any existing algorithm cannot give 100% accuracy in predicting spam. This paper deals with Naive Bayesian Classifier that is a Machine Learning algorithm for antispam filtering, which gives satisfactory results by automatically constructing anti-spam filters with extended conduct. The review over the researched performance of Naive Bayes algorithm is done by the investigations of Spam ham csv datasets. The performance of the algorithm is evaluated based on the accuracy, recall and precision it shows on the mentioned datasets. This technique gives 96-97% accuracy and 89% precision on the investigated dataset. The result also highlights that the content of the email and the number of instances of the dataset has an apparent effect on the performance of the algorithm.***
Keywords: Internet, E-mail, Spam, Networking, E-globe, Bayes Theorem, Naive Bayesian Classifier

I. INTRODUCTION

With the advent of a global pandemic: Covid-19, the world has ceased to depend thoroughly on technology. This had led to unbroken and exponential elevation of technology. The expeditious growth of technology has made the Internet an integral part of everyone's lifestyle. In recent years, the Internet has expanded its reach from an 80-year-old businessman to a 2-year-old infant. In the era of E-globe all the communications and transactions are partly or wholly reliant on "E-mails and messages" since it is economical as well as fast. However, due to the evident increase of networking and advertisers, bulk email spamming has also seen a dramatic shoot up in the past years. Spammed emails may contain divergent unsolicited files ranging from many copies of the same message, viruses and malware links, commercial advertisements, fraudulent schemes to pornographic content. Spam emails are sent to an indiscriminate list of recipients, whose emails have been gathered and/or bought from various sources. These are mainly for marketing and commercial purposes, often advertising some schemes from banks, courses from universities and coaching centres around the world, products and platforms, et cetera. Spams are even used for some malicious practices like spreading malware and viruses with the aim of gaining access to the user's system. The spammers even try to illegally highjack online transactions through email and text spamming. Spamming even hinders the efficient utilization of CPU, storage spaces and system applications. Hence, it is very essential to solve the issues encountered due to email spamming. These issues can be resolved using varied machine learning techniques that can not only detect but also filter the spammed emails. It is also necessary to identify which technique is best suited for giving the maximum accuracy while working on considered datasets.

II. PROBLEM BULLETIN

Spammers collect a huge amount of contact details: mobile number and email addresses of varied recipients through different websites, applications, UIs, webinars, chatbots, gaming platforms, and sometimes even by purchasing data illegally. And then this data is circulated among large dimensions of commercial and industrial range where bulk spamming gets under way. This huge volume of spammed messages flowing through the networks have destructive consequences ranging from getting victimized to financial losses, fraudulent schemes, privacy issues, less security, low accuracy and productivity to suffering from great business and data losses.

III. LITERATURE SURVEY

The domain of Machine Learning and Data Science includes multiple algorithms for spam filtering with varied levels of performances and accuracies. Currently, different researchers are working in the field of spam filtering by classifying emails through prediction mechanism on datasets. They have been working in the area of categorizing emails as advertisement, fraud, phishing, et cetera. Some of the review studies in the literature on spam filter classification are as mentioned.

Haiyi Zhang, Di Li elaborated Email spam detection by classifying the text/message using Naive Bayes algorithms. The paper is focused on training, testing and later classification of datasets. We have selected this paper because the categorization is explained in detail and various strategies are utilized for probability computation and prediction. [1]

Tianda Yang described the process that Naive Bayes Classifier utilizes to detect and categorize emails as spam via three main steps of pre-processing, testing and training on the taken dataset. [2]

S.M. ELseifi and W.A. Awaad stated the different ML techniques to filter spam emails at an efficient level. Descriptions of the used methodology and contrast of the analogy of their execution done on the SpamAssassin spam corpus is given. With respect to accuracy of algorithms, we have the Naive Bayes algorithm with a satisfactory result. [3]

Ms. Ashwini Athawale, Mrs. Deepali M. Gohil explained the procedure of distinguishing spam on twitter using content. In the given paper, detection is carried out by Naive Bayes rule on mistreatment trained data set. The accuracy of the results are shown by comparing the performance of Naive Bayes Algorithm with SVM. The conclusions state that Naive Bayes is more efficient than SVM as the error rate of the SVM is very high. [4]

Blanzieri and Bryl presented a well organized summary of existing machine based techniques of spam filtering. In addition, a survey on datasets, text and image-based features, performance measures, and spam filtering algorithms were mentioned as well. [5]

Rohit Kumar Solanki described the classification of spam emails by two local-global classifier techniques that use training and learning of messages. The training comprised two parts namely, pre-processing and tokenization.The technique is supposed to be achieving an accuracy of 93% for correctly classified datasets. [6]

IV. BAYES THEOREM

Thomas Bayes was the developer of Bayes Theorem which is still one of the preliminary probabilistic inference algorithms. Bayes theorem is not a single algorithm that waves path to every problem, instead it is a family of algorithms where each of them shares a common working principle, i.e. every classified feature must be independent of the others. [12] [13]

Bayes theorem provides a way that we can calculate the posterior probability of class c(target) and given data x(attributes) P(c|x) from P(c), P(x) and P(x|c) where c is the best hypothesis and x is given data. The equation for P(c|x) is written as:

$$P(\text{class}|\text{data}) = [P(\text{data}|\text{class})P(\text{class})]/P(\text{data})$$

OR

$$P(c|x) = [P(x|c)P(c)]/ P(x)$$

where, P(x|c)= Likelihood, P(c)= Class prior probability and P(x)= Prior probability. In order for the above equation to exist and the proof to continue one essential condition that must be satisfied is that whenever the Bayes Theorem is applied neither P(c) nor P(x) should be equal to zero under any circumstances. Other than this, the alternative form of Bayes Theorem is generally encountered when looking at two competing statements or hypotheses:

$$P(c|x)= [P(x|c)P(c)]/P(x|c)P(c) + P(x|c')P(c)$$

where, P(c') is the corresponding probability of the initial degree of belief against c,

where, $P(c') = 1 - P(c)$.

For some partition {c} of the sample space, the extended form of Bayes Theorem is:

$$P(c|x)= [P(x|c)P(c)]/ \Sigma i\ P(x|ci)P(ci)$$

V. NAIVE BAYES MODEL FOR FILTERING

The basic concept of Naive Bayes classifier (NBC) is applying Bayes theorem where the objects or attributes have independence.

The Naive Bayes model is one of the two most used, classified and categorized models. It is easy to set up and principally useful for very large volume data sets. It solely works on the Bayes Theorem that surmount even extremely sophisticated classification methods. The process is classified as:

A. Consider C and C' as two possible classes to categorize each email into the application as spam and Ham.

B. Consider ***X*** as the vector that uses bag of words feature to identify spam emails. Here, the search is simply for individual words or phrases.

C. And let's take ***x'*** as a vector denoted for every email. The filter will then search and scan for the particular words that are used for the spam emails in the whole sentence or paragraphs and form the vector of each email using the presence or absence of the attributes if detected.

D. Attributes here are represented into binary form i.e. ***x*** is ON or "1" if the word or phrase exists in the email and ***x*** is OFF or "0" if the word or phrase doesn't exist in the email.

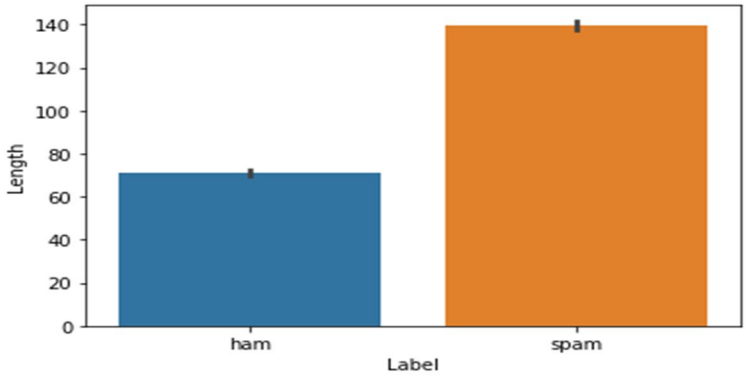
So to calculate the probability that the email received is spam or ham by the help of Bayes Theorem is written as:

$P(C|\boldsymbol{X}=\boldsymbol{x}) = [P(C).P(\boldsymbol{X}=\boldsymbol{x}|C)] / [P(C).P(\boldsymbol{X}=\boldsymbol{x}|C) + P(C').P(\boldsymbol{X}=\boldsymbol{x}|C')$

where, $(\boldsymbol{X}=\boldsymbol{x})$ is the Word or phrase that is to be detected and is written in this form because ***x*** represents the value of attribute of ***X.***

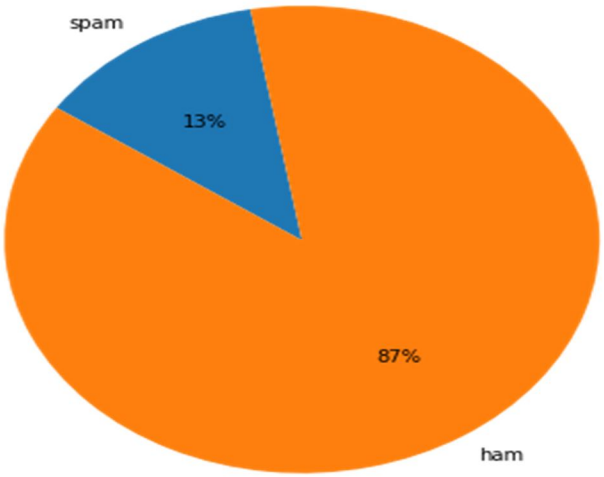
1) CASE 1 -: (Training Set Design)

```
sns.barplot(x="Label",y="Length",data=sms_spam)
```



2) CASE 2-: (Test Set)

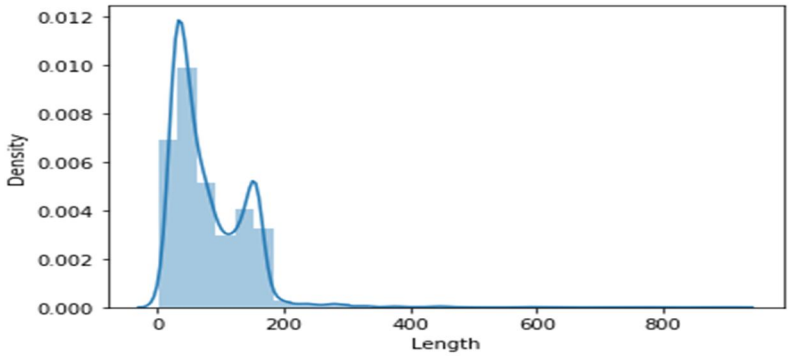
```
amount_of_spam=test_set['Label'].value_counts()[1]
amount_of_ham=test_set['Label'].value_counts()[0]
category_name=['ham','spam']
sizes=[amount_of_ham,amount_of_spam]
plt.figure(dpi=120)
plt.pie(sizes,labels=category_name,textprops={"fontsize":6}, startangle=100,autopct='%1.0f%%')
plt.show()
```



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

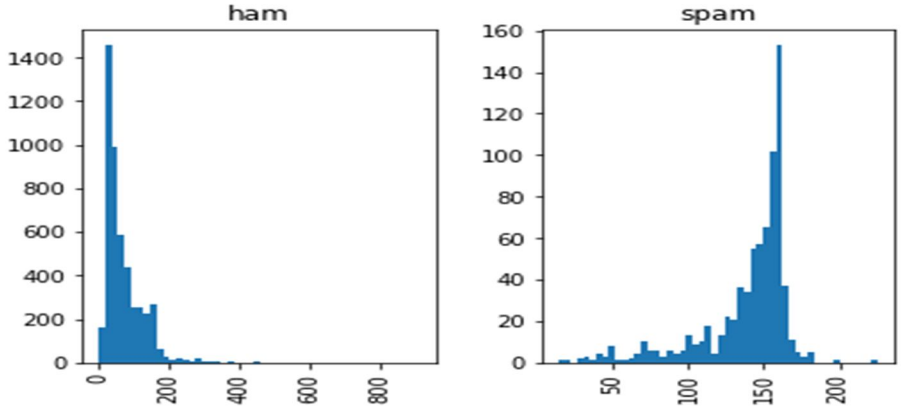
3) CASE 3-:

sns.distplot(sms_spam['Length'],bins=30)

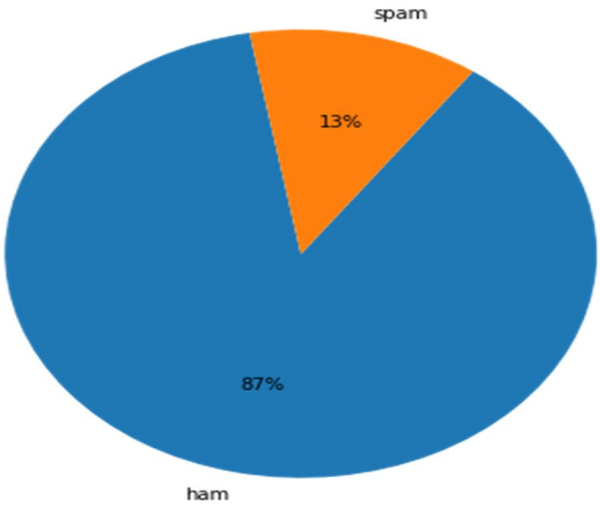


4) CASE 4-:

sms_spam.hist(column="Length",by="Label",bins=50)



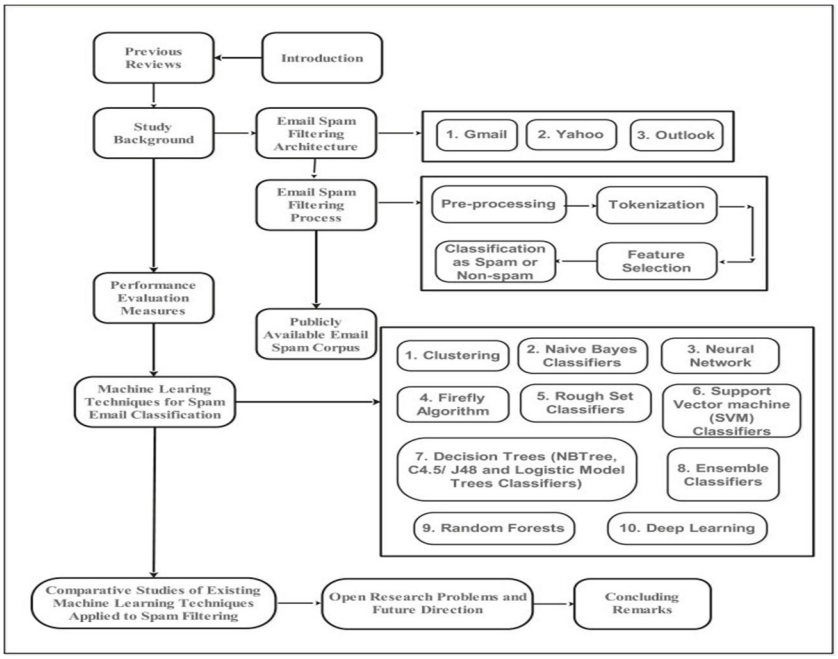
5) CASE 5 -: Final Model Prediction



Correct: 1107
Incorrect: 7
Accuracy: 0.993716337522441

VI. ARCHITECTURE OF THE SYSTEM

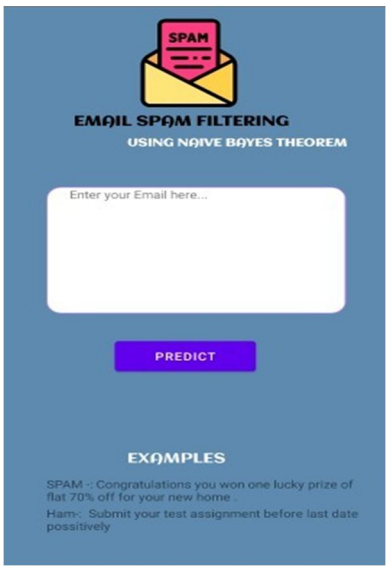
The system architecture highlights the procedure that can be followed for distinguishing between emails as spam and not spam(ham). This architecture shows various possible approaches of machine learning that can be chosen for spam filtering. The Naive Bayes Classifier for instance, when chosen, gives satisfactory results. [11]

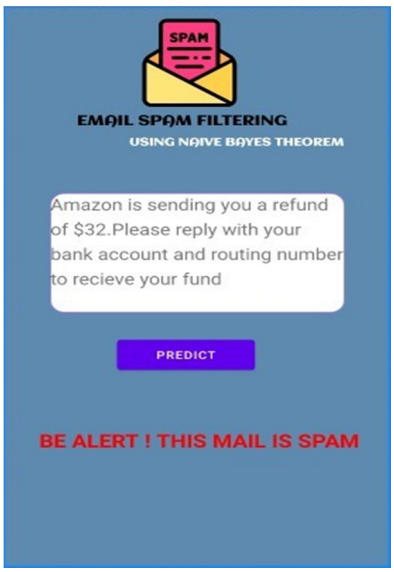


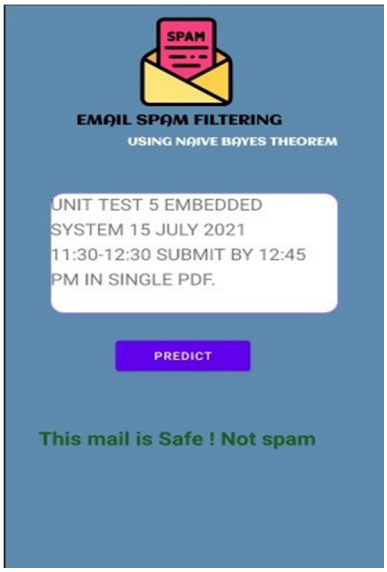
VII. RESULTS

In this paper, we have reviewed the techniques that can be applied for spam filtering. For this purpose, we have chosen the Naive Bayes Classifier for prediction. The model tested on the standard inputs taken from "Spam ham csv dataset from newly added UIC repository"[10].This classifier initially computes the likelihood probabilities of all the features that the training set possess, and after that it uses the previous instances to determine whether to label the document as spam or ham. During the training of the model, the calculated probabilities of the most representative and least representative words for spam are specified by the classifier.

We have built a user interface that can be easily utilized by people to identify the emails and predict if they are spam or not. Few instances of the same are shown below:







VIII. TECHNOLOGY UTILIZED

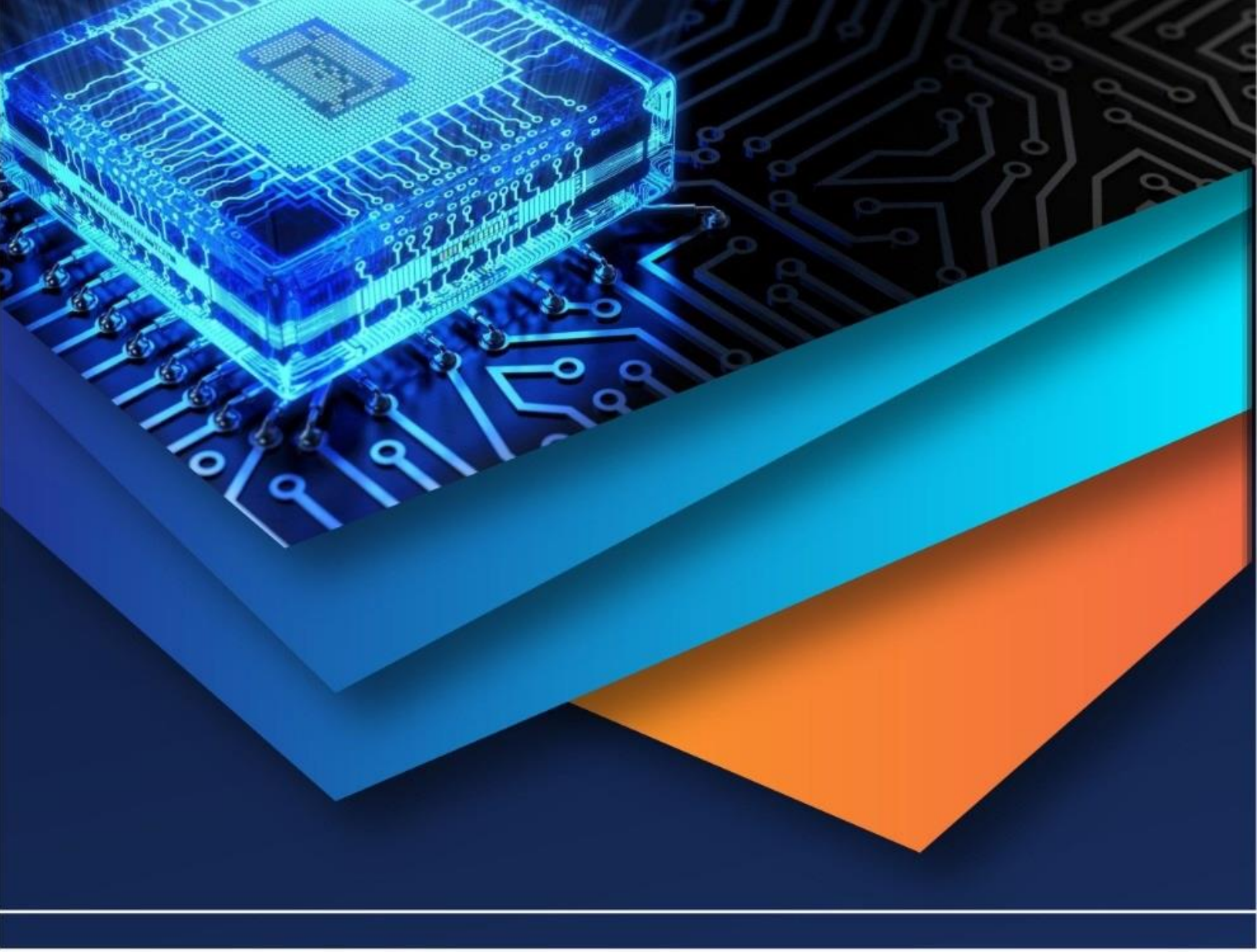
A. Interface: Jupyter notebook for inserting python libraries in a notebook format, it is typically a python code where we can easily estimate our data sets model in one single notebook.
B. Naive Bayes classifier for probability calculation.
C. Pandoc to download the data in html pdf format.
D. Spam ham csv dataset from newly added UIC repository.
E. Operating System Environment: Windows 10
F. Hardware Environment: Hardisk - 721RPM, RAM - 128 GB, GPU - 12GB VRAM, Intel core processor temperature under control
G. Software: Matplotlib, Sklearn, Scikit, Pandas, Numpy, Anaconda, Graphlab, Lucid charts, Android Studio, Firebase, Xml

IX. CONTRAST AND CONCLUSION

During the downright studies, we analysed Naive Bayes Classifier for prediction based spam filtering, so that we can utilize the theorem set by Bayes for distinguishing between spam and not spam(ham) emails. An overview on the various machine learning approaches such as SVM (Support Vector Machine), Random Forest, Decision Tree, CNN (Convolutional Neural Network), KNN(K Nearest Neighbour), MLP(Multi-Layer Perceptron), Adaboost (Adaptive Boosting) ,Naïve Bayes algorithm have been taken to contemporarily state that Naive Bayes gives satisfactory results with respect to some of the other algorithms which have been used for prediction, detection and spam filtering. Our final model of prediction on the considered data set from the newly added UIC repository, through Naive Bayes Classifier showed a sharp accuracy of 98% in accordance with the statement. In addition, we observe that MLP, Decision Tree and Adaboost give better accuracy than the rest. Further, the user interface for future scope of spam filtering using Naive Bayes Classifier was designed. The paper surveyed multiple open-source data sets for analysing purposes, which helped in reaching precise conclusions regarding the performance of the algorithm.

REFERENCES

[1] Haiyi Zhang, Di Li “Naïve Bayes Text Classifier” IEEE International Conference on Granular Computing,2007
[2] Tianda Yang, Kamal AI Nasr and Ying Qian “Spam Filtering using Association Rules and Naive Bayes Classifier” IEEE International Conference on Progress in Informatics and Computing,2015.
[3] W.A. Awad and S.M. ELseifi “machine learning methods for spam e-mail classification” International Journal of Computer Science & Information Technology Vol 3, No 1, Feb 2011.
[4] Ms. Ashwini Athawale , Mrs. Deepali M. Gohil “Spam Detection on Collection of Twitter Data Using Naive Bayes Algorithm” International Journal of Innovative Research in Science, Engineering and Technology, Vol. 7, Issue 6, June 2018
[5] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering", Artif. Intell. Rev., vol. 29, pp. 63-92, Sep. 2008
[6] Rohit Kumar Solanki, Karun Verma, Ravinder Kumar “Spam filtering using hybrid local-global Naive Bayes classifier” IEEE International Conference on Advances in Computing, Communications and Informatics,2015.
[7] E. Horvitz - A Bayesian approach to Filtering Junk E-mail.
[8] Robinson, G. Gary Robinson’s Rants. Available: http://www.garyrobinson.net
[9] Godbeck, “Reputation Network Analysis for Email Filtering”.
[10] https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv
[11] https://www.sciencedirect.com/science/article/pii/S2405844018353404
[12] https://becominghuman.ai/naive-bayes-theorem-d8854a41ea08
[13] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[14] https://www.analyticssteps.com/blogs/what-naive-bayes-algorithm-machine-learning
[15] Henderson, H. (2009). Encyclopaedia of computer science and technology /. Reference Reviews, 67 (5), 1556-65
[16] Trivedi, Shrawan Kumar. "A study of machine classifiers for spam detection." Computational and Business Intelligence (ISCBI), 2016 4th International Symposium on. IEEE, 2016.
[17] Liu B. et al (2013) “Scalable Sentiment Classification for Big Data Analysis Using Na¨ıve Bayes Classifier” IEEE 2013 pp.99-104.
[18] Cihan Varol, Hezha M.Tareq Abdulhadi “Comparison of String Matching Algorithms on Spam Email Detection”, International Congress on Big Data, Deep Learning andFighting Cyber Terrorism Dec, 2018
[19] Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization (No. CMU-CS-96-118). Carnegie-mellon univ pittsburgh pa dept of computer science.
[20] Mujtaba, Ghulam, et al. "Email classification research trends: Review and open issues." IEEE Access 5 (2017).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY