ISSN No:-2456-2165

E-Mail Spam Detection by using NLP and Naïve Bayes Classification Through Machine Learning

Mr. A. Ravi Kiran[1]
[1]Assistant Professor, NRI Institute of Technology, A.P, India-521212

T. Sai Sowjanya[2]
[2]UG Scholar, Dept. Of IT, NRI Institute of Technology, A.P, India-521212

A. V. CH. Sai Pavan[3]
[3]UG Scholar, Dept. Of IT, NRI Institute of Technology, A.P, India-521212

I. Naveena[4]
[4]UG Scholar, Dept. Of IT, NRI Institute of Technology, A.P, India-521212

Abstract:- Internet has become to be an integral part of lifestyles. With multiplied use of internet, numbers of e-mail customers are growing day by day. This growing use of e-mail has created troubles induced. Through unsolicited bulk email messages normally called spam. Electronic mail has now come to be one of the satisfactory methods for commercials due to which junk mail emails are generated. Unsolicited mail emails are the emails that the receiver does not preference to gather. A massive quantity of equal messages is sent to numerous recipients of e-mail. Direct mail usually arises as a result of giving out our email address on an unauthorized or unscrupulous internet website. There are a few of the consequences of junk mail. Fills our Inbox with type of ridiculous emails, that will reduce our internet speed. Steals beneficial records like our info on you contact list. Alters your seek consequences on any laptop software. Junk mail is a huge waste of every body's time and can quickly turn out to be very frustrating if you get hold of big quantities of it. Figuring out these spammers and the junk mail content is an onerous challenge. Despite the fact that full-size variety of studies were executed, but to this point, the Techniques set forth nevertheless scarcely distinguish spam surveys, and none of them show the Benefits of each eliminated detail compose. Despite increasing network verbal exchange and Losing lot of reminiscence space, spam messages also are used for some attacks.

Keywords:- *Spam Detection, Python, Machine Learning, Stop Words, Tokenization, Potter Stemmer, Count Vectorizer, Label Encoder, Naïve Bayes Algorithm, NumPy, Pandas.*

I. INTRODUCTION

In these days' s globalized international, e-mail is a number one source of conversation. This communication can vary from personal, commercial enterprise, corporate to government. With the rapid growth in email utilization, there has additionally been growth in the unsolicited mail emails. Spam emails, additionally referred to as junk e-mail includes nearly equal messages dispatched to several recipients via electronic mail. Apart from being stressful, spam emails can additionally pose a safety chance to our systems. Its miles envisioned that junk mail fee Organizations on the order of one hundred billion dollars in 2007.

On this challenge, we use textual content mining to perform automated junk mail filtering to use emails effectively. Most spam emails divert people's attention away from genuine and important emails and direct them towards detrimental situations.

We strive to pick out Styles using statistics-mining type algorithms to allow us classify the emails As HAM or SPAM. **[Fig:1]**

To solve this trouble, numerous unsolicited mail detection techniques are used now. The Most commonplace approach for junk mail detection is the usage of Naïve Bayesian method and function sets That verify the presence of unsolicited mail key phrases in the incoming mil.

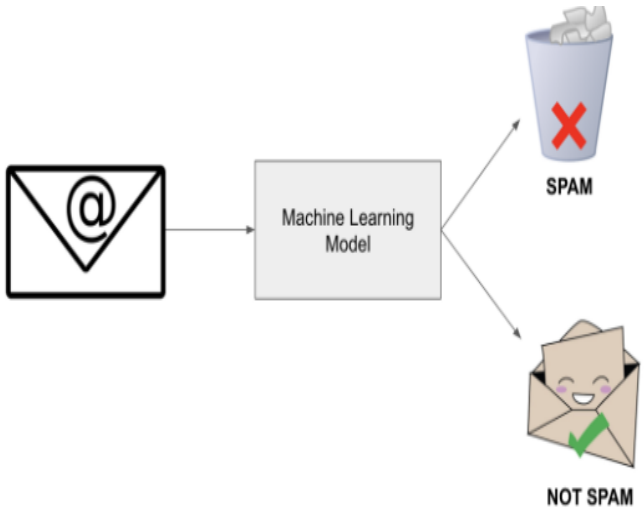

Fig 1 ML model for Spam Detection

II. TECHNOLOGIES USED

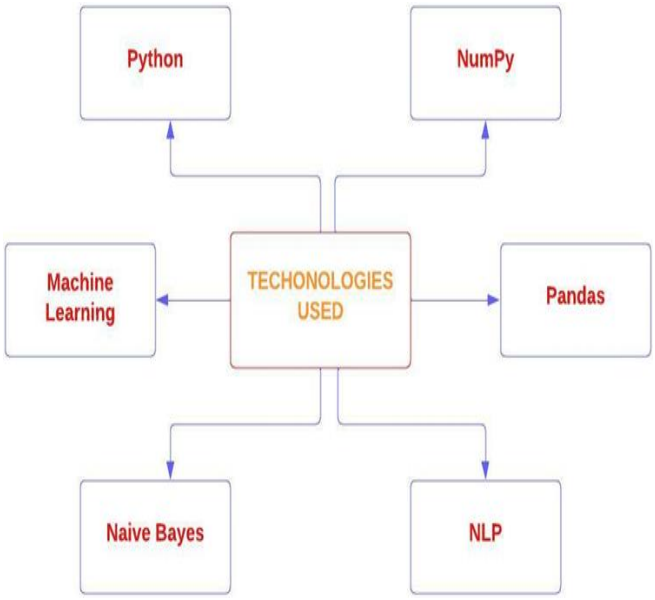

Fig 2 Technologies used

➢ *Machine Learning:*

Machine Learning is a way of training a model to be able to understand the pattern in the output with respect to the output. In one word it is an advanced algorithm that can process the data and can generate a AI program that can capable to run independently in individual systems with different inputs. Now-a-day machine learning being used in everywhere.

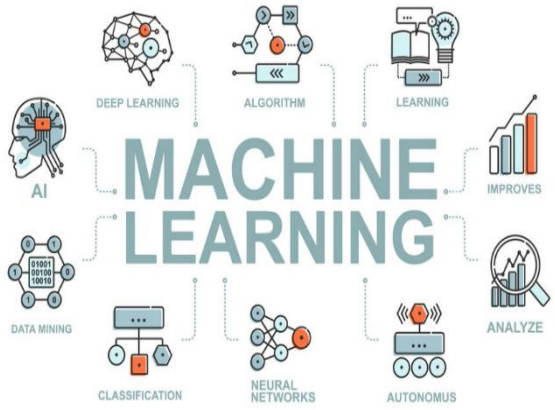

Fig 3 Machine Learning

➢ *Python:*

Python is a totally popular general-cause interpreted, interactive, object-orientated, and high-level programming language. Python is dynamically-typed and programming language. It created by means of Guido van Rossum in the course of 1985- 1990. Python includes various machine mastering libraries that could cope with massive files and can get entry to every phrase in it. [Fig: 2]

In our project we have used the python libraries like NUMPY and PANDAS that has several sub-libraries that are as follows.

- *Word Tokenize.*
- *NumPy.*
- *Pandas.*
- *Porter Stemmer.*
- *Stop words.*
- *Label Encoder.*
- *Count Vectorizer.*

➢ *Naive Bayes:*

It is a supervised learning-based algorithm, this algorithm is mostly used for classification areas like text classification. It is the simplest and effective algorithm that allows the ML models to make faster predictions.

The naïve bayes is a Probabilistic classifier i.e., it predicts the output based on the probability of an object in the text or input data. [Fig:3]

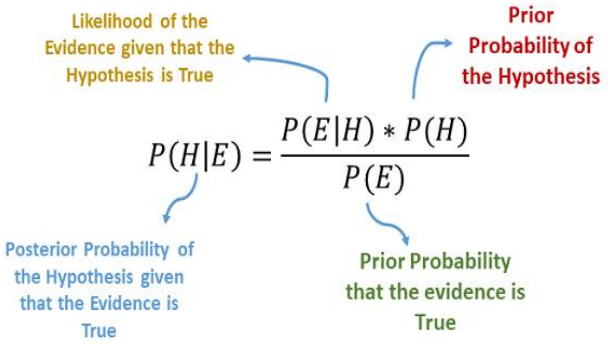

Fig 4 Naïve Bayes

➢ *Natural Language Processing:*

Natural language processing (NLP) refers to the branch of Computer science and more especially, the department of Artificial intelligence or AI concerned with giving computers the ability to understand textual content and spoken words in a good deal the same manner humans can understand. [Fig:4]

NLP combines computational linguistics rule-based modeling of human language with statistical, gadget studying, and deep gaining knowledge of models. Collectively, those technologies allow computers to method human language in the form of textual content or voice statistics and to 'apprehend' its complete which means, entire with the speaker or creator's intent and sentiment.

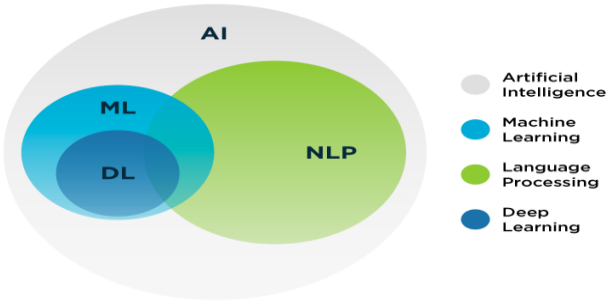

Fig 5 NLP

ISSN No:-2456-2165

➢ *NumPy and Pandas:*

NumPy and Pandas are the libraries of Python, these are used for the purpose of data analysis operations in Python. NumPy is used while working with the numerical data as it makes easy to handle mathematical functions like matrices and multi-dimensional arrays. Pandas is an Open-source Python package, it is mostly used for the data analysis and data science and ML operations.

Another main advantage of NumPy and Pandas are, these two Libraries have high Processing Speed.

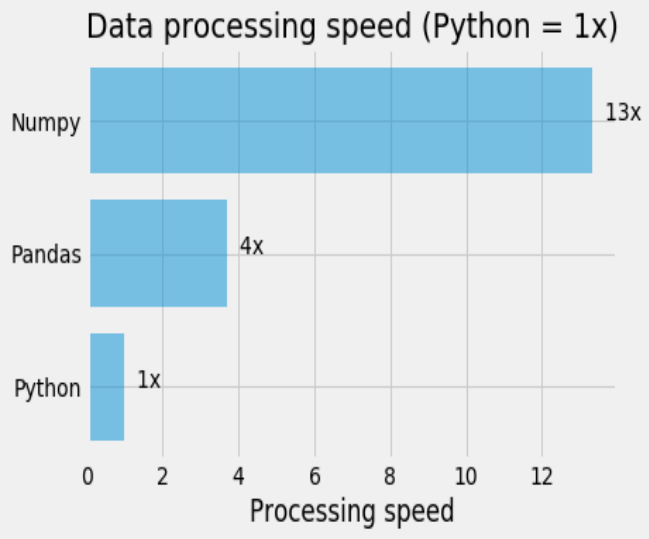

Fig 6 NumPy and Pandas

III. SOFTWARE REQUIREMENTS SPECIFICATION

➢ *Python 3.0 or later or Jupiter Notebook.*

- *Windows XP, 7, 8, or 10 operating system.*
- *NumPy, pandas, sklearn, and matplot; and an*
- *Internet browser (google chrome or Firefox)*
- *8GB ram and minimum ROM with Intel i3 or higher.*

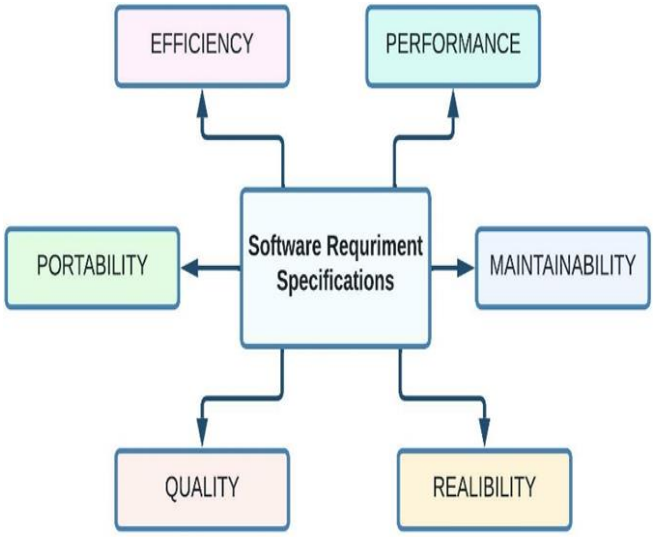

Fig 7 Software Requirement Specifications

IV. EXISTING SYSTEM

The methods are currently used by the most anti-spam software are static, mean that it is fairly easy to avoid by them to tweaking the message a little. To effectively battle the spam, an adaptive new technique is needed. Any small modification in the data can easily manipulate the spam filter. This method must be familiar with spammer's tactics as they change over time.

Sometimes the important mails are also getting into spam folder, this is because of, these spam filters more likely to consider the repeated sender as the spam, which may contain the important data**.** [9].

➢ *Disadvantages***:**

- *Time consuming.*
- *Memory wastage.*
- *No standard Classifier.*
- *Filtration done based on sender.*
- *Less accuracy.*
- *Small modifications in the incoming mail can easily manipulate the filter.*
- *Sometimes Ham mails are also sent to spam folder.*

V. PROPOSED SYSTEM

In this we are proposing a system which detects the spam messages using NLP (Natural Language Processing) Our system mainly deals with the format of the spam mail data and will try to analyze the information by using ML (machine learning) algorithms, we will make use of Natural language processing technique train a model and perform spam detection. here we follow the various ways to process the data by using Tokenization, Stemming, StopWords. etc.

In our project we will train the ML model the understand the format of the spam mails and the ham mails (not a spam mail) the mostly used words in spam and ham are analyzed, the accuracy could be increased by given new spam mails to ML algorithm. It is utmost necessary to stop all the unwanted messages as they contain viruses which harms the computer. It is reliable means it can be accessed in multiple systems.

This helps the user to avoid the spam messages by predicting the spam mail before it entering into the Inbox. The ML and NLP model can produce accurate results when compared to previous models i.e., LR model (Logistic Regression). We are going to develop this model by using all the above concepts.

➢ *Advantages of Proposed System:*

- *High Accuracy.*
- *Block mail from known spam sources.*
- *It is effective and easy to implement.*
- *The presence of single Token should not cause the e-mail to be classified as spam.*
- *Low rate of false positives.*

ISSN No:-2456-2165

VI. SYSTEM ARCHITECTURE

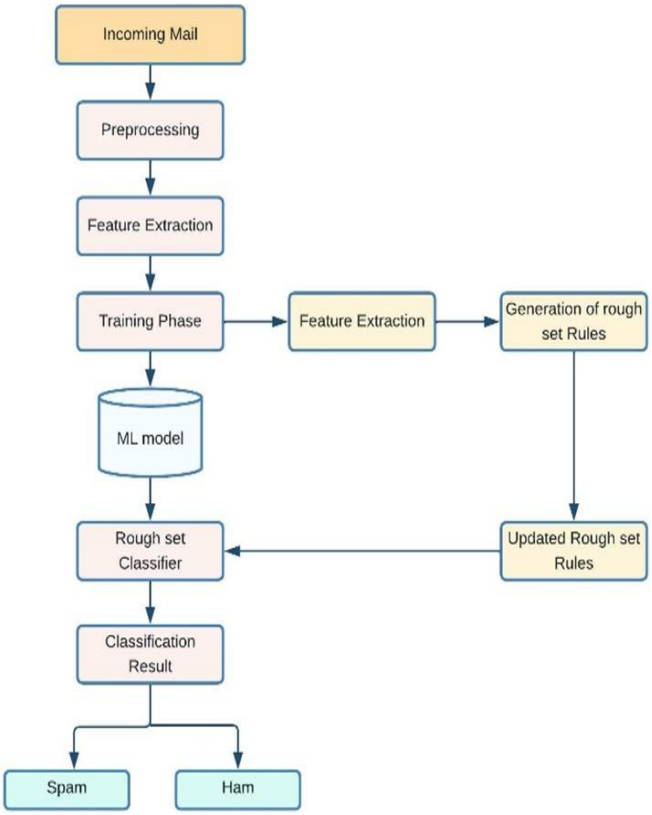

Fig 7 System Architecture

➢ *Class Diagram:*

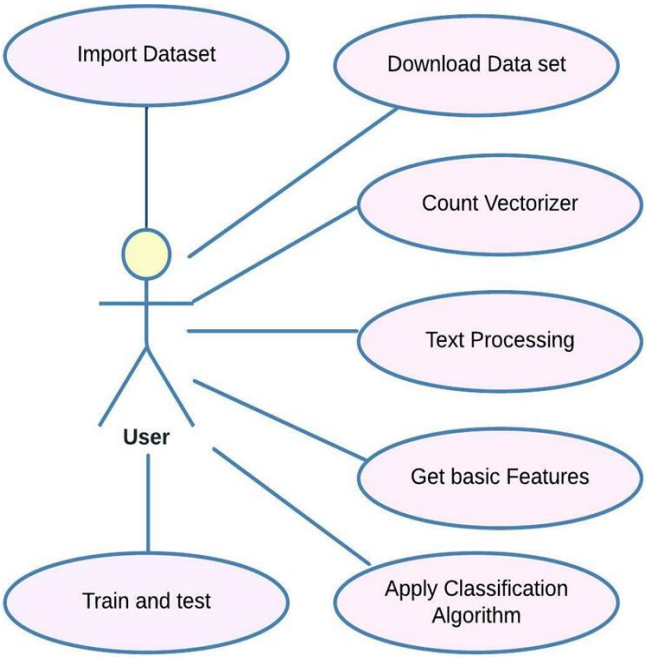

Fig 8 Class Diagram

➢ *Future Scope:*

In future, we would like to update our project with the advanced algorithms by using Neural networks, that are having the capability to update themselves for the new modified spam mails. So that the space complexity can be reduced to minimum.

VII. CONCLUSION

Spam mails are a serious concern a major displeasure for many Internet users. The mode proposed as a solution in this paper is highly beneficial because it introduces a threshold counter that helps overcome congestion on the web server and also maintain the spam filter capability, but at the same time, it also requires overhead storage space for the databases.

Since NLP is a relatively underdeveloped area for research, further enhancements can be made in the field of spam detection for online security using Natural Language Processing in the future.

REFERENCES

[1]. Qaroush, A., Khater, 1. M., &Washaha, M. (2012). Identifying spam e-mail based-on statistical header features and sender behavior. In Proceedings of the CUBE international information Technology Conference, pp. 771-778.J. Wong, “Almost all the traffic to fake news sites is from facebook, new data show,” 2016. View at: Google Scholar

[2]. M. Basavaraj, R. Prabhakar, "A Novel Method of Spam Mail Detection using Text Based Clustering Approach", International Journal of Computer Applications (0975–8887), vol. 5, no. 4, August 2010..

[3]. Comprative study of classification Algorithms for spam Email dection Aakankasha sharaff, Naresh Kumar Nagwani & Abhishek Dhadse. View at: Google Scholar

[4]. C. Baziotis, N. Pelekis, and C. Doulkeridis, “Datastories at semeval-2017 task Deep lstm with attention for message-level and topic-based sentiment analysis”, in Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), 2017, pp. 747–754.

[5]. G. Egozi and R. Verma, “Phishing email detection using robust nlp techniques”, in 2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2018, pp. 7–12.. View at: Google Scholar

[6]. Karthick veerakumar, Spam filter, 2017. [Online]. Available: https://www.kaggle.com/karthickveerakumar/spam-filter. View at:Google Scholar

[7]. Spam detection Detection Using Machine Learning Ensemble Methods (hindawi.com)

[8]. Spam Detection Using Machine Learning Algorithms (researchgate.net)

[9]. E. Blanzieri and A. Bryl, “A survey of learning-based techniques of email spam filtering,” *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63–92, 2008.View at: Publisher Site | Google Scholar

[10]. N. Udayakumar, S. Anandaselvi, and T. Subbulakshmi, "Dynamic malware analysis using machine learning algorithm," in *Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, Palladam, India, December 2017.View at: Google Scholar
[11]. S. O. Olatunji, "Extreme Learning machines and Support Vector Machines models for email spam detection," in *Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, IEEE, Windsor, Canada, April 2017.View at: Google Scholar
[12]. M. H. Arif, J. Li, M. Iqbal, and K. Liu, "Sentiment analysis and spam detection in short informal text using learning classifier systems," *Soft Computing*, vol. 22, no. 21, pp. 7281–7291, 2018.

BIOGRAPHIES

Mr. Aala Ravikiran is presently working as Assistant Professor in the department of Information Technology at NRI Institute of Technology, Vijayawada. He received his M. Tech degree from Jawaharlal Nehru Technological University, Kakinada (JNTUK). He has published over 2 research paper in international journals. He has more than 6 years of experience in Teaching.

T. Sai Sowjanya is currently studying B.Tech with specification of Information Technology in NRI Institute of Technology. She done summer internship on Spam Detection.

A. V. CH. Sai pavan is currently studying B.Tech with specification of Information Technology in NRI Institute of Technology. He done summer internship on Spam Detection.

I. Naveena is currently studying B.Tech with specification of Information Technology in NRI Institute of Technology. She done summer internship on Spam Detection.