International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 1 Volume No.7 January - February – 2023
DOI:10.46647/ijetms.2023.v07i01.030 ISSN: 2581-4621

USING MACHINE LEARNING AND NLP TECHNIQUES FOR EFFICIENT SPAM EMAIL DETECTION

Dr. K. Suresh Babu, Dr.G.Murali, Paul John Maddala
Professor, Computer Science and Engineering, VVIT, Nambur, Andhra Pradesh, India.
Professor, Computer Science and Engineering, JNTUACEP,Pulivendula,Andhra Pradesh,India.
CSE Graduate, Computer Science and Engineering, VVIT, Nambur, Andhra Pradesh, India.

***ABSTRACT*- Email spam has become a prevalent issue in recent times, with the growing number of internet users, spam emails are also on the rise. Many individuals use them for illegal and unethical activities such as phishing and fraud. Spammers send dangerous links through spam emails, which can harm our systems and gain access to personal information. It has become easier for criminals to create fake profiles and email accounts. They often impersonate real individuals in their spam emails, making them difficult to identify. This project aims to identify and detect fraudulent spam messages. The paper will explore the use of machine learning techniques, algorithms, and apply them to data sets. The goal is to select the best methods for maximum precision and accuracy in email spam detection.**

KEYWORDS: Machine Learning, Naïve Bayes, Support Vector machine, k-nearest neighbour, Random Forest, Bagging, Boosting, Neural networks, Jupiter Notebook.

I. INTRODUCTION

Spam email refers to sending unsolicited commercial messages or unwanted emails to a group of recipients through electronic mail. Such emails are considered unsolicited as the recipient has not given permission to receive them. Spam has become the biggest misfortune on the Internet. Spam wastes space, time and transmission speed of messages. Automatic spam filtering can be the most effective way to detect spam. Spammers may now easily bypass all of these spam blocking programmes easily. Several years ago, most spam coming from specific sources could be manually prevented. Electronic mail addresses. For spam detection, a machine learning approach will be utilised. Major "Text analysis, white and blacklisting" are some of the most recent ways to junk mail filtering. Domain name blacklists, as well as community-based techniques. "Spam and Ham: According to Wikipedia, the term "spam" refers to the use of electronic messaging systems to send unwanted bulk messages, such as mass advertisements and harmful links. These messages are unsolicited, meaning they are not requested by the recipient. If you are not familiar with the sender, the email may be considered spam. Many people unknowingly sign up for these emails when downloading free services or software, or when updating existing software. . "Ham" refers to emails that are not typically unwanted and are therefore not considered spam, as defined by Spam Bayes in 2001". Machine learning techniques are more effective as they utilize a pre-classified set of email samples as training data. A variety of algorithms can be employed for email filtering, including Naive Bayes, SVM, Neural Networks, KNN, Random Forests, etc.

II. LITERATURE REVIEW

There have been some previous studies that have employed machine learning techniques in the detection of email spam, such as the research done by A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab. They present the findings of a comprehensive literature review on the use of Artificial Intelligence and Machine Learning techniques for identifying email spam [1]. T. Kumar and K. Agarwal. Harisinghaney et al. and Mohammad & Selamat used the "picture and textual dataset for email spam detection" with a variety of algorithms. Harisinghaney et al. experimented with the KNN algorithm, Nave Bayes, and Reverse DBSCAN algorithm with their dataset. They also utilized an OCR library for text recognition, but found that it did not perform well. On the other hand, Mohammad Selamat employed a hybrid technique combining TF-IDF (Term Frequency Inverse

Document Frequency) and basic mathematical feature selection. This model uses email data sets from multiple internet sources including Kaggle and sklearn, as well as some data sets created by the author. Additionally, the model utilizes Kaggle's spam email data service.

Data Set:

This model utilized email data sets from various online platforms such as Kaggle and sklearn, as well as some data sets created by the author. To train the model, Kaggle's spam email data service was used, and other email data sets were utilized to obtain results. The "spam.csv" data set includes 5573 lines and 2 columns, and other data sets consist of 574, 1001, and 956 lines of email data in text format.

III. METHODOLOGY

A. Data Preprocessing

When examining data, it is common to encounter large datasets with a significant number of rows and columns. However, data can come in various forms, such as images, audio, and video files. Machines are not able to process images, videos, or text data as they are; they can only understand binary language (1's and 0's).

1. Data Cleaning

In this step the work like filling of “missing values”,”smoothing of noisy data”, “identifying or removing outliers”, and “resolving of inconsistencies is done.”

2. Data Integration

In this step addition of several databases, information files or information set is performed.

3. Data Transformation

Aggregation and normalization is performs to scale to a specific value.

4. Data Reduction

This section obtains summary of the dataset which is very small in size but so fat produces the same analytics results.

B. NLP

1. Stop Words

"Stopwords are common English words that do not contribute significant meaning to a sentence and can be disregarded without altering the overall meaning of the sentence."

2. Tokenization

"Tokenization involves breaking down a piece of text into smaller units, known as tokens, which can include phrases, symbols, words, or other meaningful elements. The resulting list of tokens may or may not include white spaces and punctuation."

3. Bag of Words

"Bag of Words is an extraction approach. "Text documents have characteristics. These features can also be used to train machines. Algorithms for learning Bag of Words compiles a list of all the unique words found throughout the document. The dataset for training" CLASSIFIERS OF OLD: The term "classification" refers to a type of data analysis. A classifier or model is created to predict class labels, such as determining whether a loan application is risky or safe, by identifying key data classes.

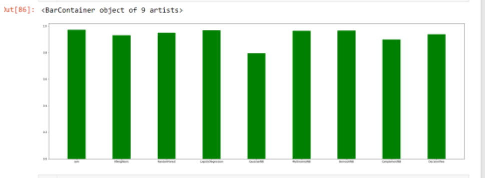
Data Classification is a two step

1. learning step
2. classification step

C. *Naïve Bayes*

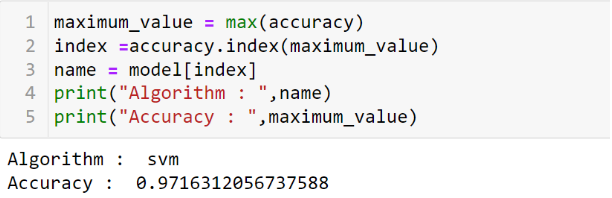
The Naive Bayes classifier algorithm was first utilized for identifying spam emails in 1998. It is a supervised learning method that is based on the probability of past events and their relationship to future events. The algorithm assumes that the features of the data are independent of one another. The technique of Naive Bayes classification is particularly useful in email filtering because it takes into account the probability of certain words appearing in spam emails. If there are words that are typically found in spam emails but not in legitimate emails, the email is classified as spam. The Naive Bayes classification algorithm has proven to be the most effective method for filtering spam emails.

IV. RESULTS AND ANALYSIS



```
Out[86]: <BarContainer object of 9 artists>
```

Fig 1. Comparison of various models



```
maximum_value = max(accuracy)
index =accuracy.index(maximum_value)
name = model[index]
print("Algorithm : ",name)
print("Accuracy : ",maximum_value)
```

```
Algorithm :  svm
Accuracy :  0.9716312056737588
```

Fig 2. Accuracy measure of model

The spam mail can be detected easily using the implementation of naïve bayes algorithms . The text processing involves in clearing the unwanted chunks of text data and helping work with only the useful content . The usage of nltk is encouraged to filter the text . Then various machine learning algorithms are used to perform spam detection . Among many svm seemed to work great efficiency compared to other algorithms.

V. CONCLUSION

The visualization helps in better understanding of the various machine learning models for the given text dataset. The svm algorithm has a good accuracy score of 97 percent.

VI. REFERENCES

[1] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab, “A Comprehensive Surevey for Intelligent Spam Email Detection,” in IEEE Access, vol. 7, pp. 168261-168295, 2019.

[2] T. Kumar and K. Agarwal. The "picture and textual dataset for the email spam detection with the usage of several algorithms"

[3] M. Sasaki and H. Shinnou, “Spam detection using text clustering,” in *Proceedings of the 2005 International Conference on Cyberworlds (CW’05)*, IEEE, Singapore, November 2005.View at: Google Scholar

[4] K. Narisawa, H. Bannai, K. Hatano, and M. Takeda, “Unsupervised spam detection based on string alienness measures,” in *Proceedings of the International Conference on Discovery Science*, Springer, Sendai, Japan, October 2007.View at: Google Scholar

[5] P. Lison, *An Introduction to Machine Learning*, Language Technology Group, Edinburgh, UK, 2015.

[6] S. Smadi, N. Aslam, and L. Zhang, “Detection of online phishing email using dynamic evolving neural network based on reinforcement learning,” *Decision Support Systems*, vol. 107, pp. 88–102, 2018