IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Special Issue 2, April 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

Spam Hamin Text Data Classification Using Multinomial Naive Bayes

P Mohammed Moin Khan, Padala Bala Veera, Dr. E. Srividhya Siva Subrahmanyam

Department of CSE, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India

Department of CSE, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India

Department of CSE, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India

ABSTRACT- To categories spam messages more accurately, a classification algorithm based on the naive Bayes method is suggested. Text preprocessing based on regular expression, extraction of features based on Jiabao classification and the TF-IDF (term - document matrix document frequency) algorithm are steps involved in the construction of spam message classification models based on the nave Bayes algorithm for both the multi-classification and multi-two-classification. The nave Bayes approach using multi-two-classification is proved to be the best by further evaluating the performance of the classifier against the supported vector machine algorithm. Using the Naive Bayes classifier and algorithm, we provide a hybrid SMS classification method to identify spam or ham. Even though this method is entirely logic-based, the database's statistical nature will determine how well it performs. Naive Bayes is thought of A key technique in information retrieval, naive Bayes is regarded as one of the most effective data mining methods. However, by using the minimum support and confidence levels that the user specifies, we significantly increase the effective accuracy of the standard Naive Bayes approach while experimenting on Data Repository.

KEYWORDS- spam detection, classification, cyber-attack, malware, logistic regression, messaging.

I. INTRODUCTION

In the Naive Bayes classification, every word in a particular SMS is considered to be independent of one another. It is the most basic type of network that is conditionally independent. We have included the frequent item concept in our suggested algorithm, which significantly improves overall accuracy. We have considered not only the independence and mutual exclusivity of every word, but also the independence and mutual exclusivity of common words. This paper's key contribution is improved text classification accuracy compared to current methods. The structure of this essay is as followsSection II discusses related studies, such as how the Naive Bayes classifier categories SMS as spam or ham. Our suggested method is described in Section III. The performance evaluation of our recommended method is covered in Section IV.Currently employed spam filtering technology frequently uses black-and-white list technology. the guidelines for matching, etc. The black-and-white listing is maintained by a third party when using black-and-white list technology.This technique uses DNS to dynamically check whether a specific IP address is on the list. If the adversary uses a dynamic or concealed IP, the approach will be constrained. The core idea behind the principles of trying to match is to compare the results with the presupposed rules and decide whether the message is spam or not. These presumptive rules, however, are typically established statically without a reliable knowledge acquisition technique, which results in poor filtering efficiency and low filtering accuracy in application sectors.These presumptive rules, however, are typically established statically without a reliable knowledge acquisition technique, which results in poor filtering efficiency and low filtering accuracy in application sectors. The naive Bayes algorithm is a type of content-based filtering technique that is highly praised for its straightforward and understandable theoretical rules, quick classification speed, and high classification accuracy; as a result, it is commonly used in text filtering.

In this paper, a Naive Bayes-based message classification model is introduced. The model, which focuses on phishing message data from a mobile company, first divides the initial data set into seven categories of message data. The model then analyses and studies the performance of the classifier of the naïve Bayes classification method with number of co

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

and number of co after message data pre - processing and extraction of features depending on the TF-IDF (term intensity document frequency) algorithm. The model even farther studies different features with the var The results reveal that the classification and nave Bayes method has the best classifier accuracy when comparing the classification performance of the multi-two-classification nave Bayes methods.

II. LITERATURE REVIEW

In order to quickly detach SPAM and HAM, Amandeep Singh Rajput, Vijay Athavale, Sumit Mittal, and others [4] developed a community-oriented methods using bunch figuring and the assistance of identical machines. By accelerating handling without incurring additional costs, a group can double the calculating power many times over with already-existing tools and resources. In this study, we only employ a header-based separating method, hence maintaining perfect client security. Spam Assassin's standard test suite for HAM or SPAM is applied. In this experiment, there are two different kinds of equal conditions.

The first is the area where various anti-spam techniques are used equally to test corpora and phoney positive and phoney negative precision is recorded. A single anti-spam technique is used to all computers in the second equal environment, when standard test corpora is divided into smaller groups and handled in an equal machine environment. The effectiveness of this environment is then measured using independent machines. Utilizing Weka Data Mining Software, anti-spam measures are implemented.

III. PROPOSED SYSTEM

I constructed an SMS spam classifier when I was just getting started, but I performed extensive study to know where to begin. To make it simpler for you to create your first spam classifiers using the Naive Bayes model, I'll guide you throughout my project in 10 steps.
Load the dataset and make it simpler:
If you read our dataset of SMS text messages in pandas, it comprises 5 columns: v1 (which contains the text messages' class labels of "ham" and "spam"), v2 (which contains the texts itself), and 3 Unnamed columns. See how "5572 rows x 2 columns" indicates that our database has 5572 text messages?
View the dataset in a bar chart.
Before you begin working with your data, it's a good idea to perform an exploratory data analysis (EDA) to visualize, extract information from, or identify any problems with it. We'll examine the quantity of spam/ham texts we have and make a bar graph for it.
4825 ham communications and 747 spam communications are included in our sample. This dataset is skewed since there are much more junk mails than spam ones. This might bring good fortune to our model. We can resample the data to obtain an equal amount of spam/ham communications in order to correct this.

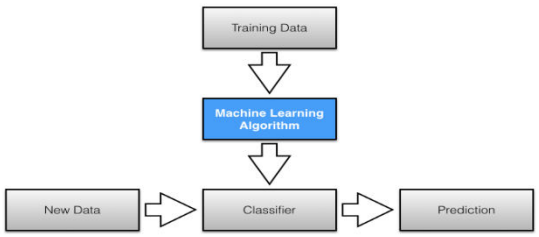

Fig.1 SPAM HAM Content of Data

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

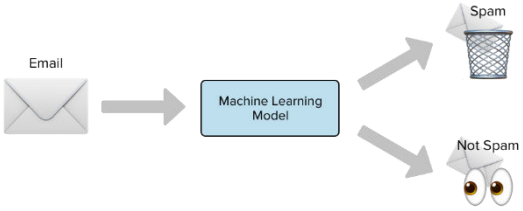

Fig.2 Process of SPAM HAM

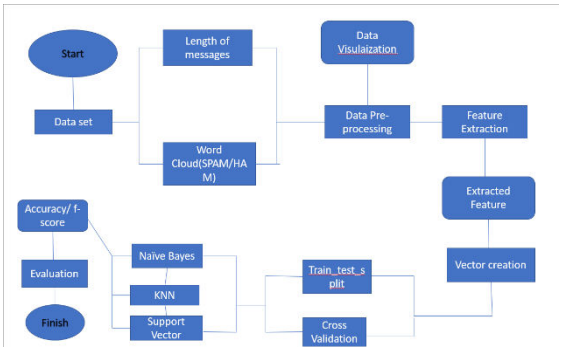

Fig.3 System Architecture of SPAM HAM

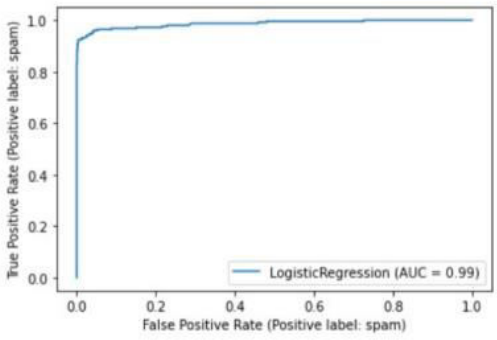

Fig.4 Logistic Regression

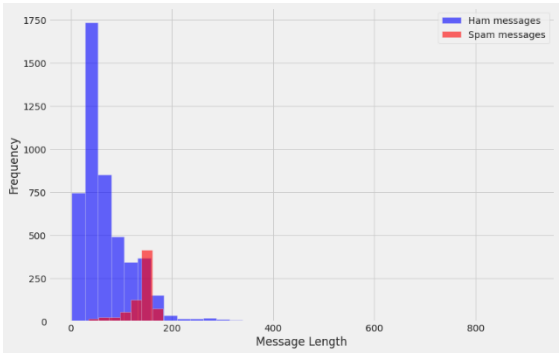

Fig 5. Visualize for Compare the length between SPAM and HAM messages.

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

IV. PROPOSED WORK

Data pre-processing:
Since your data may include a lot of noisy and undesirable characters, text cleaning is a crucial stage in machine learning. Our standard practices included tokenization, lemmatization, deleting stop words, transforming all characters to lower- or uppercase, and removing punctuation. Part of our cleansed data is depicted in the figure below. Additionally, we took advantage of the built-in preprocessing functions of Count Vectorizer and Tfidf Vectorizer.

Feature Extraction:
In this procedure, we carefully examine the data (in this case, emails) to identify the features (i.e., words) that will be most helpful in the categorization. The classifier would then be trained using these features. We employed the Count Vectorizer and TF-IDF techniques for this. These can be described as numeric statistics used to show how important a term is to a piece of text found in a corpus. The values correspond exactly to how frequently a term is used in a document. A portion of our vectorizer matrix is shown below.

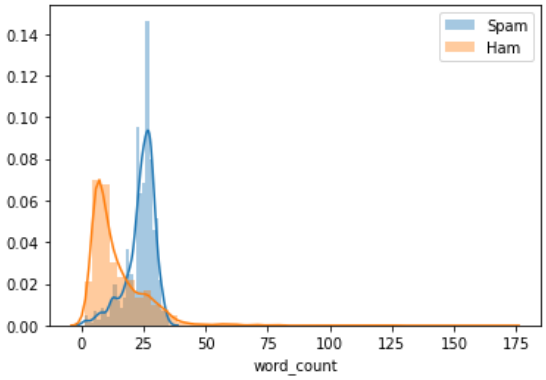

Fig.6 Visualize the distribution of word counts.

V. METHODOLOGY

A Technology Requirements
The machine learning libraries used in the project are mentioned below.

1.NLTK Library
Common Language Learning is the ability of any computer or product to control or obtain text or speech. People connecting, understanding one another's perspectives, and responding appropriately are similarities. In NLP, a computer rather than a person makes the association and understands the response.
Natural Language Toolkit is the name for it. This toolkit contains resources to help build machines understand the human language and respond appropriately to it, making it one of the most notable NLP libraries. A section of these bundles, including tokenization, stemming, lemmatization, grammar, character tally, and word tally, will be looked at in this educational exercise. The most well-known computations, such as tokenizing, grammatical form labelling, stem, concluding inquiry, point subdivision, and named substance acknowledgement, are all included in NLTK. The PC is encouraged to examine, pre-measure, and interpret the written text with the help of NLTK.
The primary building block for Python projects that work with language processing data is NLTK. It is directed through the fundamentals of writing Python programs, dealing with corpora, sorting texts, dissecting etymology construction, and that's only the beginning. It was created by the creators of NLTK.
We use Human Language Toolkit to continue with Languages Datasets in Python 3. (NLTK). The process entails:

2.Scikit-learn library
Scikit-learn is a free artificial intelligence (AI) library for the programming language Python (formerly known as scikits.learn and also known as sklearn). It features several grouping, relapsing, and arrangement computations like support vector machines, irregularity backcountry, inclined boost, k-means, and DBSCAN, and is designed to work with Python's NumPy and SciPy logical and mathematical libraries.

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

The majority of Scikit-code learns is written in Python and it frequently makes use of NumPy for advanced direct user defined math and cluster jobs. Additionally, to speed up performance, some center computations are written in Cayton. Strategy relapsed and straight support vector machines are carried out by a Cayton cover around LIBSVM, while support vector machines are carried out by a similar covering around LIBLINEAR. It might not be possible to use Python to expand these strategies in such circumstances.

A Python module called Scikit-learn provides a variety of unsupervised and guided learning computations. It's a never-ending source of innovations like Python language, pandas, and Matplotlib that you may already be familiar with. Run the following order in the terminals for pip establishment:

Import scikit-gain from sklearn import datasets. • import sklearn.

Print the information's current status to confirm that it is stacked. iris=datasets.load iris() (iris.data.shape).

Probably the most useful Python AI library is scikit-learn. Characterization, relapse, grouping, and dimensionality reduction are just some of the powerful tools for AI and factual presentation that are available in the Sklearn toolkit.

VI. RESULT & CONCLUSION

Simple to construct and especially helpful for very big data sets is the naive Bayes model. Along with being straightforward, Naive Bayes is well-known for performing better than even the most complex classification methods. From P(c), P(x), and P(x|c), the posterior probability P(c|x) can be calculated using the Bayes theorem. Look at the following equation:

• P(c|x) represents the conditional probabilities of the given class (c, target) (x, attributes).

The probability of the class is P(c).

• The likelihood, or P(x|c), measures the likelihood of a predictor for a given class.

• P(x) represents the predictor's previous probability.

Despite its name, **logistic regression** is a linear method for categorization as opposed to regression. In the literature, logistic regression Regression, maximum-entropy classification (Maxent), and the log-linear classifier are also used to refer to logistic regression. In this model, a logistic function is used to simulate the probability representing the possible results of a single experiment

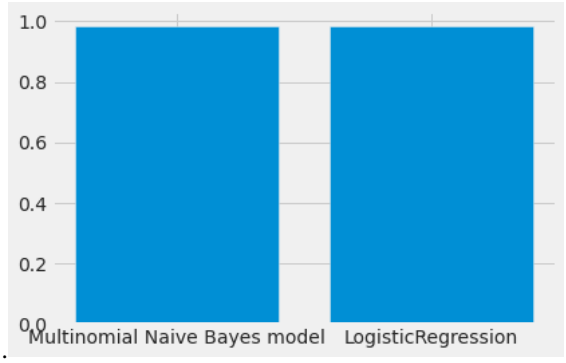

Fig.7 Final Output of my SPAM and HAM

Although Logistic Regression has a better AUC than Multinomial Naive Bayes, the numbers are nearly identical.

REFERENCES

1. Alom, Carminati & Ferrari (2020) Alom Z, Carminati B, Ferrari E. A deep learning model for Twitter spam detection. *Online Social Networks and Media.* 2020;18(8):100079. doi: 10.1016/j.osnem.2020.100079. [CrossRef] [Google Scholar]
2. Abiramasundari (2021) Abiramasundari S. Spam filtering using semantic and rule-based model via supervised learning. *Annals of the Romanian Society for Cell Biology.* 2021;25(2):18. [Google Scholar]
3. Basyar, Adiwijaya&Murdiansyah (2020) Basyar I, Adiwijaya, Murdiansyah DT. Email spam classification using gated recurrent unit and long short-term memory. *Journal of Computer Science.* 2020;16(4):559–567. doi: 10.3844/jcssp.2020.559.567. [CrossRef] [Google Scholar]

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

4. Bathla& Kumar (2021) Bathla G, Kumar A. Opinion spam detection using Deep Learning. 8th International Conference on Signal Processing and Integrated Networks (SPIN); 2021. pp. 1160–1164. [Google Scholar]
5. Dedeturk&Akay (2020) Dedeturk BK, Akay B. Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Applied Soft Computing*. 2020;91(16):106229. doi: 10.1016/j.asoc.2020.106229.[CrossRef] [Google Scholar]
6. Hossain, Uddin & Halder (2021) Hossain F, Uddin MN, Halder RK. Analysis of optimized machine learning and deep learning techniques for spam detection. 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS); 2021. pp. 1–7.[Google Scholar]

INNO SPACE
SJIF Scientific Journal Impact Factor

SJIF: 8.423

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details