E-Mail Spam Detection Using Machine Learning Naive Bayes Theorem

Karra.NAGA SIVA SURYA DILEEP [1]
Krovvidi.KARTHIK SAI SRI RAMA RAJU [2]
Karella.JOHNY [3]
Kombathula.VENKAT [4]
P.SRINU VASA RAO [5]
Swarnandhra College of Engineering & Technology.

Abstract:- Spam, sometimes called spam, is unsolicited email that is typically sent to large lists of recipients. Although real individuals can send spam, botnets (computer networks infected by an attacker known as a "bully") are often responsible for sending spam. While most people view spam as a problem, they believe it is a result of email communication. In addition to being annoying, spam can also be dangerous because it can clog email inboxes if not filtered properly and deleted frequently.

Spammers or spammers often change their methods and content to trick victims into downloading malware, sharing personal information, or feeding money. Most spam is commercial in nature and financially motivated. Spammers attempt to deceive recipients by making false claims, selling questionable products, and promoting false information.

Unwanted emails, such as phishing and spam, cost businesses and individuals billions of dollars each year. Many models and techniques for automatic spam detection have been introduced and developed, but 100% accuracy has not yet been found. Among all designs, machine and deep learning algorithms are more successful. Natural language processing (NLP) improves model accuracy. This study presents the effectiveness of word embedding in spam classification.

Preliminary study Transformer model BERT (Bidirectional Encoder Represented by Transformers) is well tuned to accomplish the task of identifying spam from non-spam (HAM). BERT uses a color layer to place the content of the text into its perspective. The results were compared with the basic DNN (Deep Neural Network) model consisting of BiLSTM (Bidirectional Long Term Memory) layer and two thick layers.

- **Here are some of the most popular spam topics:**

Pharmaceuticals, financial services, working from home, porn, online courses and cryptocurrency.

Keywords:- *Machine Learning, Natural Language Processing, Spam, Ham, Email, Naive Bayes, Logistic Regression.*

I. INTRODUCTION

Spam has become a major problem in today's digital age and poses a challenge to individuals, businesses and organizations. Spam is unsolicited messages that fill inboxes, waste time and resources, and potentially expose users to malicious or fraudulent content. To solve this problem, machine learning has become a powerful tool in spam detection. The purpose of spam detection is to identify legitimate emails (spam) or spam. Due to the evolution of spam, the effectiveness of legal procedures is limited. Machine learning provides powerful and flexible processing using patterns and features extracted from large email databases. Machine learning algorithms can learn from email text files to create patterns that can identify spam patterns. This model can be used to identify new, unseen emails. By analyzing various email elements such as sending information, sentences, content, and embedded URLs, machine learning algorithms can identify spam characteristics and act accordingly.

In today's digital age, where email is one of the most common means of communication, spam has become a huge problem for people and organizations. Spam filters are essential for managing and monitoring important emails in your inbox. Machine learning (ML) and natural language processing (NLP) techniques can be used for spam classifiers to accurately identify and filter spam. In this project, we aim to create a spam classification system based on machine learning and natural language processing to accurately classify emails as spam or not spam. The performance of the system will be evaluated according to various parameters such as accuracy, precision, recall, and F1 score.

Spam is the use of email to send unsolicited email or broadcast email to multiple recipients. Spam emails indicate that the recipient has not given permission to receive such emails. Spam has become a big problem on the internet. Spam wastes storage and media. Automated email filters are still the best way to catch spam, but nowadays spammers can easily block any spam filtering application. A few years ago, most spam from specific email addresses was blocked manually. Machine learning techniques will be used for spam detection, so Naive

Bayes is one of the methods used in this process. Naive Bayes algorithm is a supervised learning algorithm used to solve classification problems and helps create fast learning models for fast prediction.

Spam and Spamming: Spam refers to the content of e-mail and the use of electronic communications to send unsolicited messages, especially advertisements; Bad links are called spam. Therefore, if you do not know the sender, the message may be spam. Many users do not realize that they are signing up for some emails only after downloading free services, software or updates. "Raw" is not spam email.

Machine learning is more complex and focuses on developing computer programs and algorithms to access data. So this template using training data is primarily a set of emails. Machine Learning Methods There are many algorithms that can be used for email filtering. In this article, Naive Bayes algorithm was used to detect spam and it gave the highest accuracy.

The machine learning method that uses training data, which is the pre-processing of the email, is more efficient. “Naive Bayes, Support Vector Machines, Neural Networks, K-Neighbour Communities, Random Forests, etc.” that can be used for email filtering. There are many machine learning methods, including: Why choose machine learning: Machine learning allows users to feed large amounts of data into computer algorithms and the computer analyzes it based on input data and makes data-driven recommendations and decisions. What is a dataset: A dataset is collected from data or information regarding individual content. The spam email profile includes spam and non-spam emails. What is training data and test data: The main difference between training data and test data is that training data is a part of the original data used to train the learning model while test data is used to check the accuracy of the model. . The size of training data is usually larger than testing data. Training and testing datasets are two important concepts in machine learning based on training data.

ROC Curve: The ROC Curve (Receiver Operating Characteristic Curve) is a graph that shows the performance of each distribution model.

PR Curve: PR curve is a graph that has the correct value on the y-axis and remembers the value on the x-axis. In other words, the PR curve has TP/(TP+FP) on the y-axis and TP/(TP+FN) on the x-axis. It is worth noting that accuracy is also known as positive predictive value (PPV).

Confusion Matrix: Confusion matrix is a table used to evaluate the effectiveness of classification models in machine learning and statistics. It shows the results of classification by calculating positives, negatives, negatives and false positives.

II. LITERATURE SURVEY

- L.F. Cranor ve B.A. Lamacchia said that , Spam is not a marketing email that wastes our time; this is spam. It also uses network traffic and mail servers. It has also become an integral part of many attacks, including spam, phishing, cross-site scripting, cross-site request errors, and malware attacks. Statistics show an increase in spam containing malicious content, compared to spam promoting legitimate products and services. This paper studies the spam detection problem and develops an effective spam detection method based on email content analysis. We see that many features have many disadvantages. Our goal is to examine the effectiveness of these features in aiding classical spam detection techniques. To further complicate the problem, we developed a spam classification model based on randomness; Here spam is a small group, accounting for only 16.5% of all emails. Use different measurements to evaluate the design. The results show that the spam classification model improves well when learning about devices with poor characteristics.
- J. Goodman, G. V. Cormack, and D. Heckerman said that, Antispam researchers and developers are working to improve spam filtering software to solve problems caused by the complex techniques spammers use to bypass filters and enter users' mailboxes. In 1998, the volume of unsolicited emails (or spam) increased from approximately 10% to 80% of all messages sent, causing problems for email service providers (ESP). Major email services send more than a million spam messages every day. Use learning algorithms to find characteristics of spam and good emails. Spammers can also quickly learn the most obvious words to avoid and the safest words to add to the filter. Algorithms based on logistic regression and support vector machine can reduce the number of missed spam messages by half.
- TO. Blanzieri and A. Bryl said that , Spam is a big problem online today, causing financial losses for companies and affecting consumers. Filtering is an important and popular way to prevent spam. In this article, we provide an overview of the state of the art in machine learning for spam filtering, as well as methods for evaluating and comparing filtering methods. We will also briefly describe other branches of anti-spam and discuss the use of various commercial and non-commercial anti-spam software solutions.
- L. Zhang, J. Zhu, and T. Yao said that , This article evaluates five evaluation methods in the context of spam filter statistics. We use the estimated values to examine the effects of different pruning methods and sizes on each student's performance. As can be seen, the importance of feature selection varies from classifier to classifier. In particular, we found that the support vector machine AdaBoost and the maximum entropy model performed best in this test, with similar properties: insensitivity to feature selection strategies, easy size for advanced features, eight-scaling, and performance on many datasets. yield. appeared. In contrast, Naive Bayes

(a classifier commonly used in spam filtering) has been shown to be sensitive to the selection process of small elements and works poorly without any boost, resulting in a large penalty. Experiments also show that when creating filters for legitimate email applications that are more expensive than spam (e.g., λ = 999), aggressive feature pruning should be avoided to better preserve performance. An interesting finding is the impact of email headers on spam filtering, which has often been overlooked in previous research. Experiments show that a classifier using name features can perform as well as or better than a filter using text only. This means that email headers can be a reliable and unique source for spam filtering.

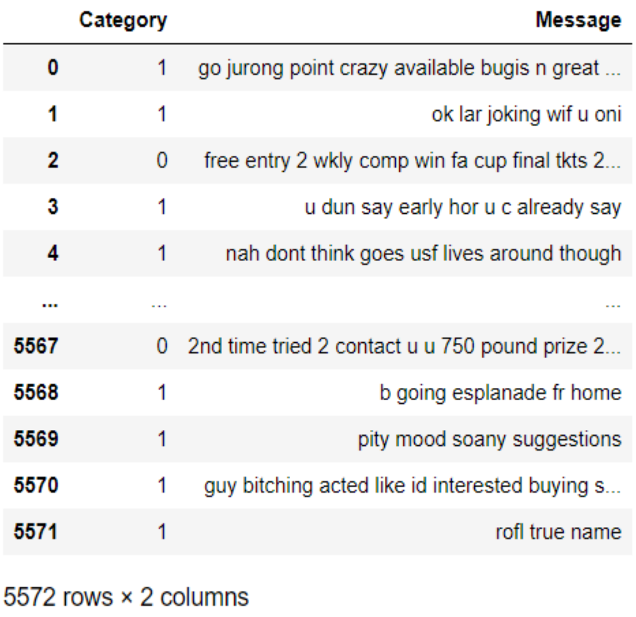
- J.-J. Sheu said that , The contact header of an email is usually the email name, sender's name, email address, delivery date, etc. It includes the following key features. These features help classify emails. This article uses decision tree data mining techniques to identify important features of email headers to identify spam organization strategies and propose spam filtering methods to accurately identify spam and emails. Based on spam filter testing using many emails from China, we got the following positive numbers: 96.5% accuracy, 96.67% accuracy and 96.3% bounce rate. Therefore, the methods mentioned in this article only need to analyze the session header to effectively identify spam emails and thus reduce the computational cost.

III. METHODOLOGY

Naive Bayes is an effective method of data analysis based on Bayes theorem used for email spam filtering. If you have an email account, you will notice that emails are divided into different groups and classified as important, spam, advertisements, etc. We are sure you have seen it classified as. Wouldn't it be nice to see a smart machine working for you?

In general, the tags added by the system are correct. So does this mean that our email software reads every communication and now understands what you are doing as a user? This is true! In the age of data analysis and machine learning, automatic filtering of emails is done by algorithms such as Naive Bayes classifiers, which use the simple Bayes theorem for the data.

Current spam filtering software is constantly trying to classify emails correctly. Avoiding spam and promotional communications is the hardest part of all. As the battle between spam filtering software and anonymous spam and email marketing continues, spam communication algorithms must also continue to evolve. In data analysis, the Naive Bayes algorithm forms the basis for filtering messages on Gmail, Yahoo Mail, Hotmail and all other platforms. As online consumption of goods and services increases, consumers are facing a major problem with too much spam in their inboxes. Whether it's advertising or fraud. But this is why very important messages/emails are compressed in spam emails . In this article, we will create an email spam detection model that will help you detect whether an email is spam or not using Naive Bayes and Natural Language Processing (NLP).



	Category	Message
0	1	go jurong point crazy available bugis n great ...
1	1	ok lar joking wif u oni
2	0	free entry 2 wkly comp win fa cup final tkts 2...
3	1	u dun say early hor u c already say
4	1	nah dont think goes usf lives around though
...	...	...
5567	0	2nd time tried 2 contact u u 750 pound prize 2...
5568	1	b going esplanade fr home
5569	1	pity mood soany suggestions
5570	1	guy bitching acted like id interested buying s...
5571	1	rofl true name

5572 rows × 2 columns

Fig 1 : Email Spam Dataset

Naive Bayes classifier is a popular algorithm for email filtering. They often use the message bag feature to identify spam, a method used to classify letters.

The Naive Bayes classifier works by correlating usage tokens (usually words, sometimes other) with spam and non-spam emails, and then using Bayes' theorem to calculate whether the email is associated with spam.

Naive Bayesian spam filtering is a basic technology for processing spam that can be adjusted according to the user's email needs and provides the negative result of spam detection generally available to users. It is one of the oldest spam filtering systems, launched in the 1990s. Naive Bayes is one of the simplest yet powerful classifier algorithms. Given hypothesis A and evidence B, Bayes' theorem calculates the relationship between the probability of the previous hypothesis obtaining with evidence P(A) and the probability of obtaining the final hypothesis if we accept evidence P(A/B) is :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Naïve Bayes Equation.

Where:

- A, B = events
- P(AB) = probability of A if B is true
- P(BA) = probability of B if A is true
- P (A), P(B) = independence of A and B

IV. RESULT

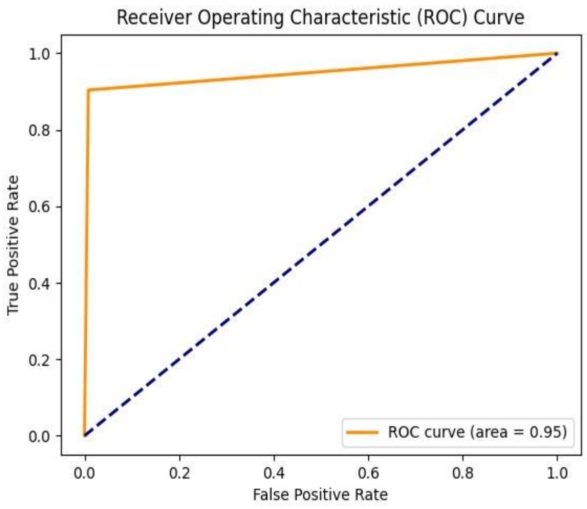
Naïve Bayes Method:

```
data['Spam']= data[ 'category'].apply(lambda x:1 if
x=='spam' else 0)
X_train,X_test,y_train,y_test=train_test_split(data.Message
, .Spam , test_size=0,20)
model=Pipeline([ ('vectorizer',CountVectorizer()),
('nb',MultinomialNB()) ])
model.fit(X_train,y_train)
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(accuracy)
```

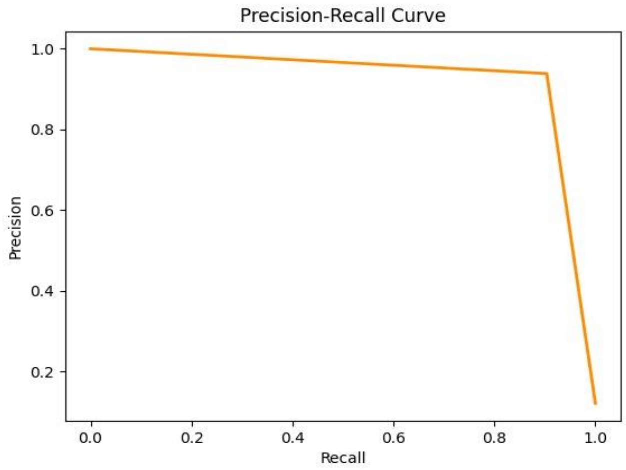
V. OUTPUT

Array(1,0), dtype : int64
The probability of receiving an e-mail is calculated correctly according to the data set.
So accuracy = 0.9856502242152466 (OR) 98%.

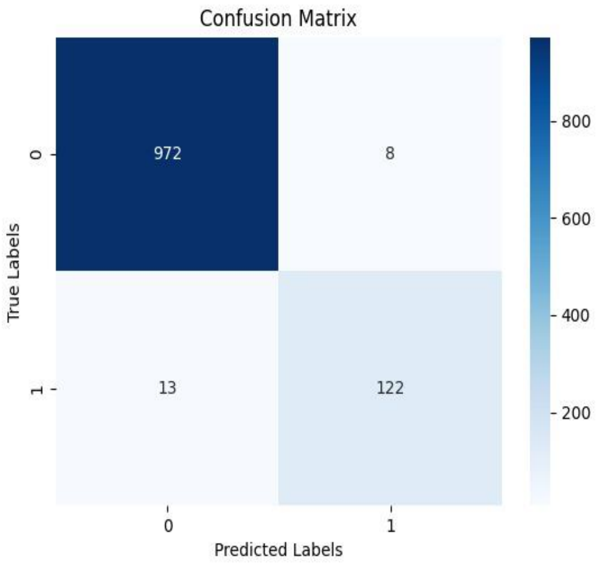
Classification Report :
Precision recall f1 support score\n\n 0 0.99 0.99 0.99 965\n 1 0.97 0.92 0.94 150\n\n Accuracy 0.98 1115\n Macro average 0.98 0.96 10199810.n 0 91099.1099 1091099.100 99 9110999.10998 19910999.10998 19910999.10998 199.109989 10999.10998 199'



Output 1 : True False Positive Rate



Output 2 : Recall and Precision Graph



Output 3 : Confusion Matrix

VI. CONCLUSION

We conclude that the Naive Bayes algorithm is the best for classification in spam detection and it is worth examining the Polynomial Naive Bayes algorithm because it has many applications in many industries and the algorithm Predictions are very fast. Media classification is one of the most popular users of the Naive Bayes algorithm. News political, regional, international etc. It is widely used to divide into different sections such as political , regional and so on.

REFERENCES

[1]. L.F. Cranor ve B.A. Lamacchia: "Spam!" Communications of the ACM, vol. 41. No. 8, p. 74-83, August 1998.

[2]. J. Goodman, G. V. Cormack, and D. Heckerman, “Spam and the ongoing war for the inbox,” Communications ACM, vol. 50, no. 2.00 pm 24-33, February 2007.

[3]. TO. Blanzieri and A. Bryl, “A study on learning-based email spam filtering strategies,” Artificial Intelligence, vol. 29. No. 1 p.m. 63-92, Three. Year 2008.

[4]. L. Zhang, J. Zhu, and T. Yao, “Evaluation of spam filtering techniques,” ACM Transactions on Asian Studies, vol. 3, p. 243-269, 2004

[5]. J.-J. Sheu, “An effective two-stage spam filtering method based on email classification,” International Journal of Cyber Security, vol. 9. No. 1, s. 34-43, 2009.