WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Spam Detection Using Machine Learning Techniques

Abishek Sharma, Arjun N

St Joseph's College

ABSTRACT

Spam emails are bulk ed emails or fake emails sent to a particular person or organization . These spam mails can be detected using machine learning techniques and different methodologies. The spam mails can lead to loss of actual private data. There are new techniques to classify whether the mail is a spam or a ham. Ham are valid mails whereas spam are unwanted mails. . Present day researcher shave implemented some features of text messages are used classify them into ham or spam. This study compares several classification methods using data gathered from diverse sources, assessing each method's accuracy along the way. Our NLP system filters mail and classifies it as spam or junk. An example of a learning-based model is the Extreme Learning Machine (ELM). (is a state-of-the-art machine learning method for feedforward neural networks with a single hidden layer. It has neither slow training speeds nor overfitting concerns as traditional neural networks do. In ELM, just one iteration cycle is required. Due to its enhanced generalisation potential, durability, and controllability, this technique in particular is becoming more and more popular across a wide range of fields. Various machine learning methods for spam detection are taken into account in this paper.

KEYWORDS : Email spam detection, Spam detection, Support vector Machine Learning, Natural Language Processing, analysis.

INTRODUCTION

Technology advancement both have become the most important part of development today. The utilization of internet has been increasing every day and for the exchange of information and communication the use of email has also been increasing. Emails are important for both professional and personal purpose and these along bring unnecessary bulk mails also known as spam . These technically contain promotion of any product or scam or purchasing anything over an unsecured site, the promise of great prizes or lotteries these mails contain a large amount of our memory . At times these mails draw our attention and that leads to neglecting important mails . these lower the speed of your internet and are capable of bringing in virus to your device identification of such mails is a tough job to do and may lead to loose ones patience . Spam detection can even be done manually but it consumes a lot of time and for a world like ours time is the most important thing therefore the need for spam detection software's is the necessity of large organizations and companies. The most common technique for spam detection Is by the implementation of Naive Bayesian and finding out spam keywords.

According to estimates from social networking experts, 40% of social network accounts are exploited for spam. Spammers target particular demographics using popular social networking technologies review pages or fan pages to which to submit text with buried links sites that sell pornographic or other products something obtained through false accounts. obscene emails that are sent to groups or organizations of the same kind discuss frequent highlights. Through examination of these highlights, the detection of these emails can be improved. By We can classify emails using artificial intelligence no spam emails and spam emails. the possibility of a solution Using the subject, headers, and body of the messages to extract features body, too. Based on the nature of the data we extracted, we can classify them.

Extreme is an illustration of a learning-based paradigm. machine that learns (ELM). is a contemporary machine learning model for the only feedforward neural networks It removes the slow training speed and one hidden layer. difficulties with overfitting when compared to standard neural In ELM, only one cycle of iteration is necessary. Due to increased robustness, generalization potential, and This technique, in particular, is now utilized in In this paper, we take into consideration many machine gaining knowledge of spam detection methods.

i) This paper examines the architecture of various machine learning-based spam filters as well as their advantages and disadvantages. We also talked about the fundamental components of spam email.

(ii) By conducting an extensive study of the suggested methodologies and the makeup of spam, some attractive research gaps in the spam detection and filtering area were discovered.

(iii) Open research issues and potential future research directions are explored in order to improve email security and spam email filtration using machine learning techniques.

(iv) This paper talks about the difficulties spam filtering models now encounter and how those difficulties affect the effectiveness of the models.

(v) The role of machine learning in spam detection is explained through a thorough comparison of machine learning ideas and methodologies.

SPAM MESSAGES

the term "spam" in relation to email is misleading because everyone has their opinions of it At the moment, everyone's attention is focused on email spam. Email spam typically

consists of certain spontaneous messages sent in bulk by people. You are ignorant. The term "spam" comes from Monty Python animation [23] featuring the Hormel canned meat product has a lot of pointless emphasis. While the phrase "spam" allegedly first used in 1978 to make an unwelcome reference As we move to the mid-1990s, email grew significantly. becomes increasingly commonplace outside of academic and research groups [24]. The growth is one notable model.

7.1 Spam filtering techniques in IoT platforms and email.

Spam emails are becoming more prevalent in politics, financial market information, chain communications, and marketing and instruction [24]. At the moment, different businesses create We discuss some filtering tactics in various methodologies and algorithms for effective spam detection and filtering to comprehend the filtering procedure, read this section.

7.2 Common Spam Filtering Technique

Standard spam filtering is a type of filtering that employs a set of rules and uses them as a classifier. demonstrates a typical method for spam detection. The second step entails putting in place content filters that, using artificial intelligence techniques, detect spam . a header filter that extracts the header information from emails Implementing email is the second stage. The backlist follows that. In order to stop spam emails, emails are put through filters to pick out spam in the backlist file.

After this stage, the sender is identified using rule-based filtering. utilise the topic line and the user-defined parameters. In the end, a job and allowance filter implementation is employed.

7.3 Spam filtering on the client side.

A filtering system that uses protocols as well as a set of rules to accomplish An individual with access to an email network or the Internet who can send or receive email is received. Spam detection at the client point offers various rules and methods for assuring safe communications transmission between individuals and organisations. For data transfer, a client must install multiple functional frameworks on their machine. Such systems filter the client's inbox by establishing connections with client mail agents and writing, receiving, and managing the incoming emails.

7.4 Commercial-Grade Spam Filtering.

Enterprise-level email spam detection involves installing various filtering frameworks on the server, interacting with the mail transfer agent, and categorising the gathered emails into one spam or ham is system. The email is rated using a criterion that is currently used by spam detection techniques.

This concept allows for the scoring of each post and the construction of a ranking system. Each piece of spam or unsolicited mail is given a specific grade or score. Since spammers employ a variety of strategies, all jobs are regularly adjusted by the introduction of a list-based way to automatically block the messages.

Table 1 : Spam Categories

Categories	**Descriptions**
Health	the proliferation of bogus drugs
Products Promotion	The proliferation of counterfeit clothing, bags, and watches
Adult content	The proliferation of pornographic and prostitution-related adult content
Marketing and accounts	The overabundance of loan packages, tax strategies, and stock kiting.
Fraud	Fraud mails to access you wealth.

LITERATURE SURVEY

1. Email:

Electronic mail also known as email is a messaging method by which we can send electronical messages across computer network. Anyone from anywhere can access email by using and internet connection it is mostly used for business and formal purposes and even used for personal use.

2. SPAM:

The unnecessary bulk mails can be classified as spam mails. These spam mails can lead to corrupt a device by occupying all space in the inbox and decreasing the speed of the internet.

3. SPAM DETECTION:

A method known as spam filter is used to check unwanted and virus-infected emails and prevents them from reaching a your inbox. There are a number methods for detecting spam, including blacklists and whitelists, naive bayes, machine learning methods, support vector machines, and classification using neural networks.

4. MACHINE LEARNING ALGORITH:

Machine learning algorithm is a branch of the broader field of artificial intelligence that makes use of statical models to develop predictions. The expected outputs are going to be the most accurate for the respective system.

5. Naïve Bayes Classifier:

Naïve Bayes Classifier Algorithm, Which is based on Bayes theorem and used for solving classification problems.

6. SUPPORT VECTOR MACHINE:

One of the most well-liked supervised learning algorithms, support vector machine is utilised for classification issues in machine learning. In order to simply place additional data points in the appropriate category in the future, SVM aims to establish the optimal line or decision boundary that can divide n-dimensional space into classes.

METHODOLOGY USED

- **DATASET**
- **DATA CLEANING**
- **EDA**
- **PREDICTION**

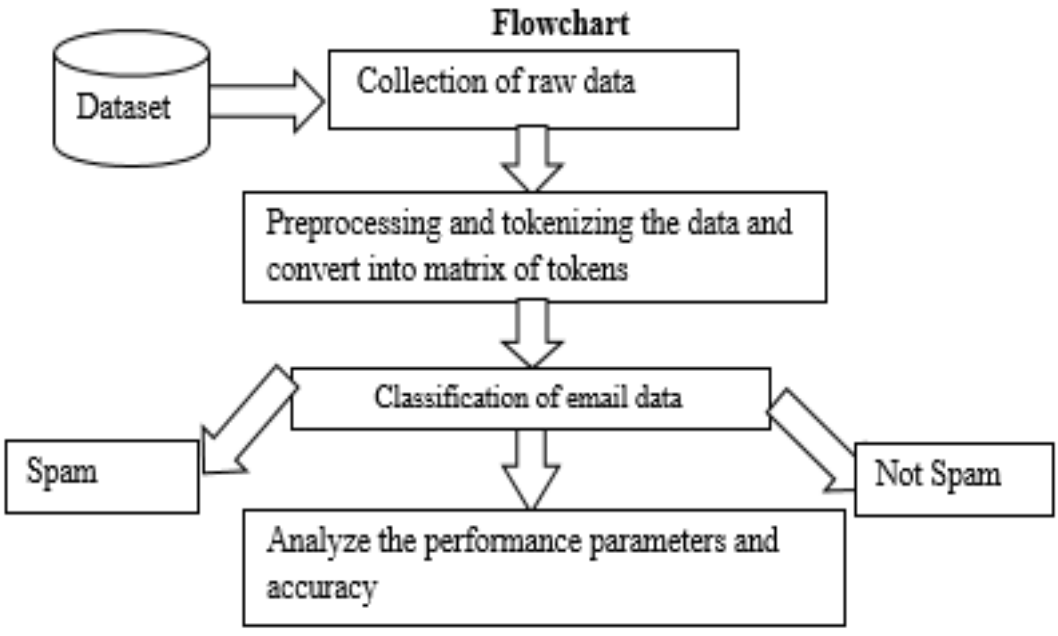

Figure 1: Flowchart of Email Spam detection

DATASET

The dataset used in this research paper is available on Kaggle, which is also a machine

Learning repository. The dataset 'Spam' contains 5572 instances and 4 attributes. In the given data 4900 mails are ham and 672 mails are spam. The dataset can be seen in table 2.

DATA CLEANING

The cleaning of data is the most important part of machine learning, if the data is not cleaned properly the accuracy of the model will not be accurate.

Benefits of data cleaning are ,

- Better decision making
- Saves time
- Increases the productivity
- Helps in streamline business practices
- It gives boost the revenue

Steps to clean your data

STEP 1: Drop the unwanted data from the data set

	D1	D2
1938	ham	Bravo! Are you ready to scream in...
529	ham	Raj says that you're a genius
2137	ham	why would come to hostel.
3296	spam	Hurray!! U won 100,000...
3383	spam	FREE home for 200rs...

Table 2: Cleaned Data Table

STEP 2: Re-naming data accordingly, we rename the

	Sender	Text
1425	ham	I'll be there in few seconds
4456	ham	Aight should I come up later toni...
4482	ham	True lov n care wil nevr go unrecognized. thou...
1879	spam	U have a secret admirer who is looking 2 make ...
3145	ham	SHIT BABE.. THASA BIT MESSED UP.YEH

STEP 3: labeling the given data for making the model less complicated. We convert spam as 1 and ham as 0.

	Sender	Text
0	0	Can we meet tomorrow babe...
1	0	Ok lets meet at cafeteria
2	1	Register to win trip to London!!!...
3	0	Mate my day was a bang one I had a blast...
4	0	Lts go for a night out...

Table 3: Labeled Data Table

STEP 4: In this process we fixing the missing data.

Sender 0

Text 0

Unnamed: 4 5566

dtype: int64

STEP 5: In this process we remove the duplicate data from the given dataset.

```
In [36]: df.duplicated().sum()
Out[36]: 403

In [37]: df = df.drop_duplicates(keep='first')

In [38]: df.duplicated().sum()
Out[38]: 0
```

EDA (Exploratory Data Analysis)

Exploratory Data Analysis is a process of examining or understanding the data and extracting insights or main characteristics of the data. EDA is classified as, i.e. graphical analysis and non-graphical analysis.

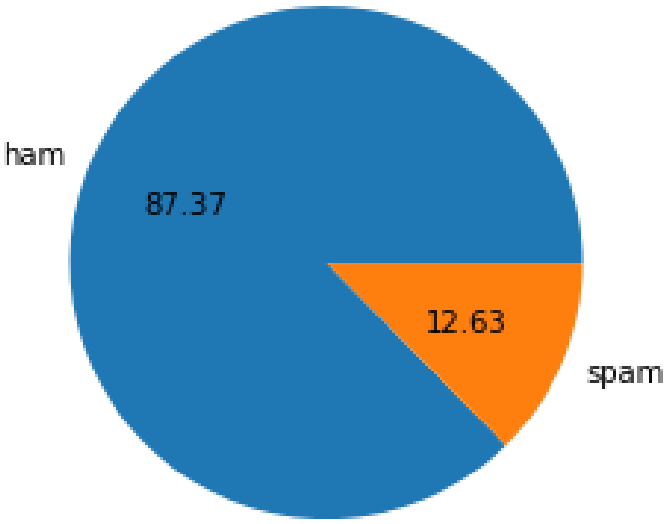

Figure 2: Pie plot of Ham and Spam Percentage

1. The given pie plot shows the percentage of spam and ham in our data set.
2. We add various attributes to our data set for better modeling for example we find the total number of characters, number of words and number of sentences present in a particular mail.

	Sender	Text	Unnamed: 4	num_characters	num_words	num_sentences
0	0	Go until jurong point, crazy.. Available only ...	NaN	111	24	2
1	0	Ok lar... Joking wif u oni...	NaN	29	8	2
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	155	37	2
3	0	U dun say so early hor... U c already then say...	NaN	49	13	1
4	0	Nah I don't think he goes to usf, he lives aro...	NaN	61	15	1

3. Describing the given data based on Ham

	num_characters	num_words	num_sentences
count	4300.000000	4321.000000	4310.000000
mean	69.459699	15.123339	2.815545
std	50.358868	11.491315	2.364098
min	1.000000	1.000000	1.000000
max	908.000000	205.000000	34.000000

4. Describing the given data based on Spam

	num_characters	num_words	num_sentences
count	556.000000	558.000000	420.000000
mean	124.891271	25.667688	3.969372
std	26137753	4.008418	4.488910
max	224.000000	42.000000	8.000000

5. Histoplot based on number of words used.

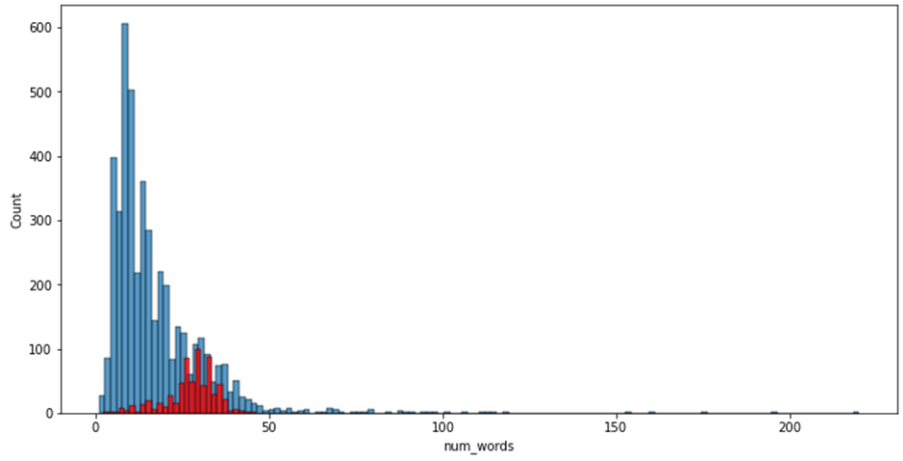

Figure 3: Histoplot based on number of words used in ham and spam

6. Pairplot using filter Spam or Ham

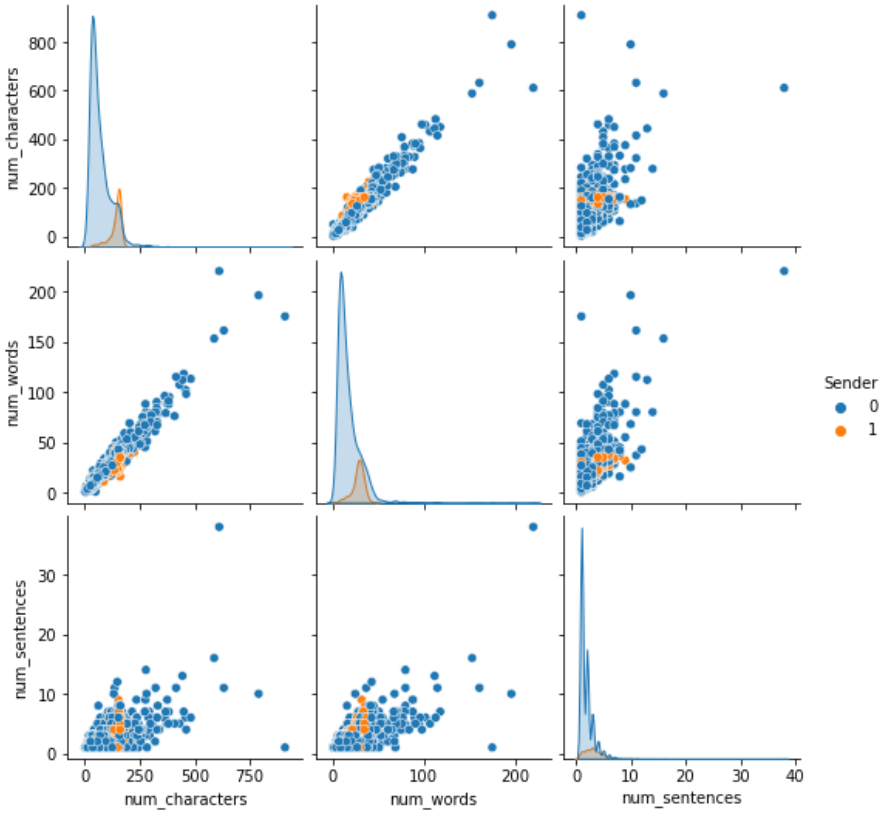

Figure 4: Pairplot using filter Spam or Ham

1. Heatmap to check the correlation

The given heatmap gives the correlation of 0.38 with respect to sender as the number of words characters the tendency of it being a spam also increases.

And 0.26 and 0.27 respectively.

The highest correlation lies with num_characters and num_words

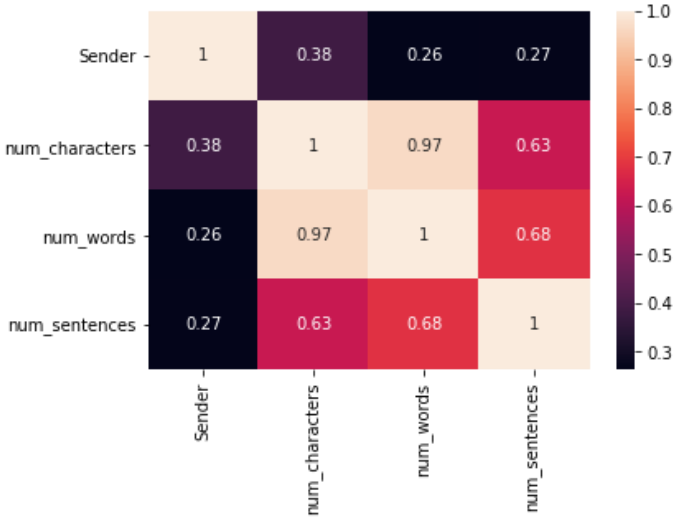

Figure 5: Heatmap to Check Correlation

2. We use wordcloud to find the most common words used in Spam and Ham respectively.

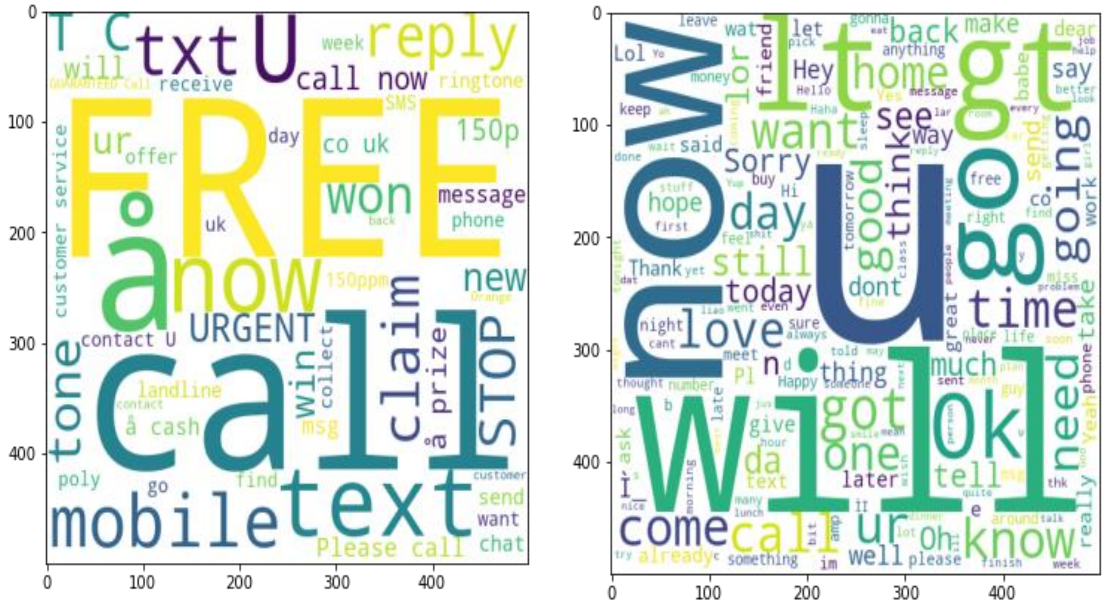

Figure 6: Wordcloud To Find Most Used Words in spam and Ham.

We find out that the most used words in spam are free, call, now, Mobile text and the words used in ham are will, u, go, respectively.

3. Number of common words used in a spam mail.

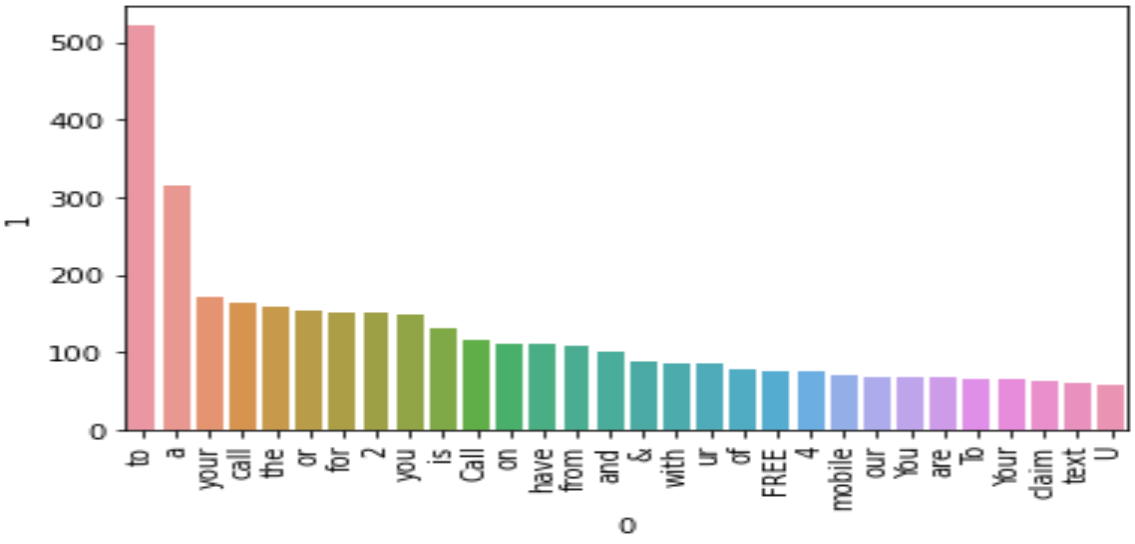

Figure 7: Number of common words used in a spam mail.

1. Number of common words used in a ham mail.

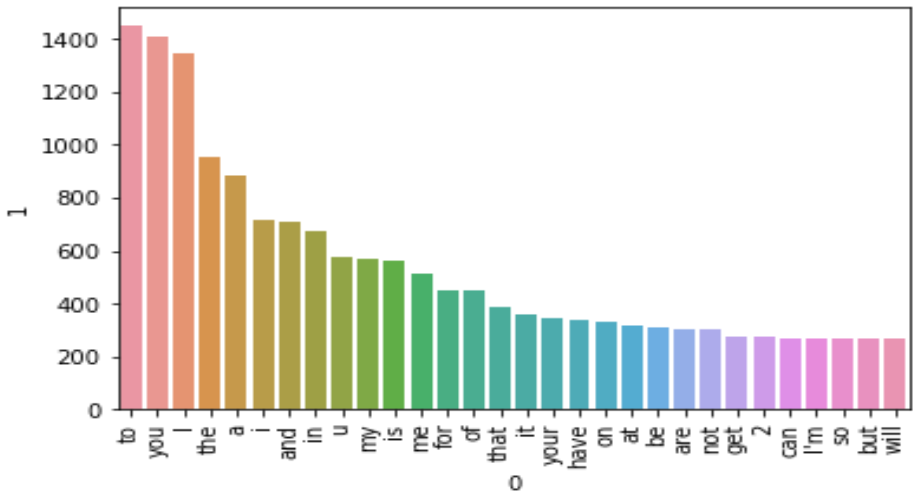

Figure 8: Number of common words used in a Ham mail.

MODEL BUILDING

1. Transforming text data to array data for model.

```
In [100]: from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
          cv = CountVectorizer()
          tfidf = TfidfVectorizer(max_features=3000)

In [102]: X = tfidf.fit_transform(df['Text']).toarray()

In [103]: X.shape
Out[103]: (5169, 3000)

In [105]: y = df['Sender'].values
```

2. We use train test split with respect to different algorithms like

mnb = MultinomialNB()

bnb = BernoulliNB()

to calculate the accuracy, confusion matrix and precision of the model and check for the best model for spam classification.

BernoulliNB()

```
In [111]: bnb.fit(X_train,y_train)
          y_pred3 = bnb.predict(X_test)
          print(accuracy_score(y_test,y_pred3))
          print(confusion_matrix(y_test,y_pred3))
          print(precision_score(y_test,y_pred3))

0.9845261121856866
[[893   3]
 [ 13 125]]
0.9765625
```

MultinomialNB()

```
In [112]: mnb.fit(X_train,y_train)
          y_pred2 = mnb.predict(X_test)
          print(accuracy_score(y_test,y_pred2))
          print(confusion_matrix(y_test,y_pred2))
          print(precision_score(y_test,y_pred2))

          0.9738878143133463
          [[896   0]
           [ 27 111]]
          1.0
```

We see that multinomial model does not give false positive value and the the accuracy of the model is moderate , therefore the given model is best for spam detection as it doesn't classify any ham as spam.

We still go through other algorithms for checking on a better model for our problem.

- knc = KNeighborsClassifier
- mnb = MultinomialNB
- dtc = DecisionTreeClassifier
- lrc = LogisticRegression
- rfc = RandomForestClassifier
- bc = BaggingClassifier
- etc = ExtraTreesClassifier
- xgb = XGBClassifier

	Algorithm	Accuracy	Precision
0	KN	0.917157	1.000000
1	NB	0.983888	1.000000
2	RF	0.921954	1.000000
3	ETC	0.933559	1.000000
4	xgb	0.955822	0.974790
5	LR	0.940348	0.944954
6	BgC	0.933250	0.896825
7	DT	0.989072	0.894737

RESULT ANALYSIS

The NB Algorithm proposed the highest accuracy and precision together, because of this the number of spam mails detected is high and the accuracy of the algorithm increases and does not give us false positive value thus the best algorithm for spam detection is NB algorithm.

CONCLUSION

Today, email is the most significant form of communication because it allows for the delivery of any message anywhere in the globe thanks to internet connectivity. Every day, more than 270 billion emails are sent and received, of which 57% are spam. Spam emails, also referred as as "non-self," are unwanted commercial or harmful emails that damage or compromise personal information, such as bank account information, financial information, or anything else that harms a single person, a business, or a group of people. In addition to advertisements, they could have connections to websites hosting phishing or malware created to steal personal data. Spam is a severe problem that end consumers find bothersome but is also financially harmful and a security risk.

This project's spam detection is capable of identifying emails that include specific information. The use of reputable and verified domain names can be used to identify scam emails. The spam email categorization is very important for classifying emails and identifying whether they are spam or not. Naive Bayes has low false positive spam detection rates that are typically acceptable to consumers, making it a baseline technique for regulating spam to the email needs of individual users. The Naive Bayes approach's parameters are further optimized, which improves the accuracy of the entire classification process. The Nave Bayes Classifier can improve the accuracy of spam detection.

12.REFFERNCES

1. Elchouemi, P. W. C. Prasad, A. Alsadoon, and M. K. Chae. Gain and the graph mining method are used in spam filtering and email categorization (sfecm). The 7th IEEE Annual Computing and Communication Workshop and Conference will be held in 2017.
2. "Logistic Regression for Machine Learning," by Jason Brownlee April 1, 2016, The Machine Learning Mastery. Logistic regression using machine learning is available at https://machinelearningmastery.com.
3. ianying Zhou, Wee-Yung Chin, Rodrigo Roman, and Javier Lopez, (2007) "An Effective MultiLayered Defense Framework against Spam", Information Security Technical Report 01/2007.
4. For email spam classification in a distributed context, K.R. Dhanaraj, V. Palaniswami, Firefly, and Bayes classifier, Aust.J. BasicAppl
5. A review of machine learning techniques to spam filtering by Guzella, T. S., and Caminhas, W. M. Appl. Expert Syst.
6. An experimental comparison of naïve Bayesian versus keyword-based anti-spam filtering using personal email communications by Androutsopoulos, J. Koutsias, K. Chandrinos, and C.