DETECTION OF SOCIAL NETWORK SPAM BASED ON MACHINE LEARNING WITH NAIVE BAYES ALGORITHM

Dr. K. Anuradha[1], Dr. T. Guhan[2], Dr. N. Revathy[3], Dr. K. Jegadeeswaran[4]

Article History: Received: 05.01.2023 **Revised:** 20.03.2023 **Accepted:** 05.05.2023

Abstract

Social media is an internet platform that allows users to effortlessly build social connections with other users (Facebook, Email, LinkedIn). People today share their personal information, habits, career interests, and hobbies with their peers on social networking platforms. With online social networks, different information is spreading, both good and bad. Sharing information online is getting more commonplace every day. The naive Bayes method is utilised in this study to identify false information, such as internet rumours, as well as to solve distinguishing and predicting issues. Moreover, the Naive Bayes method is utilised for collaborative filtering, hybrid recommender systems, spam filtering, and text categorization. The premise behind many social networks is that a user's online information represents who they really are. Members of these networks who fill in their name fields with made-up names, corporate names, phone numbers, or just random characters are breaking the terms of service, tarnishing search results, and lowering the site's value for legitimate users. Identifying and banning these accounts based on their spammy names might enhance actual users' experiences on the site and stop more abusive behaviour. This project's primary goal is to identify and categorise email communications into spam and junk using various machine learning approaches. The NLP needs to be used (Natural Language Processing). This study aims to understand how different machine learning algorithms can categorise email spam with ham. We must now create machine learning-based techniques for identifying email spammers. Any emails that include unsolicited content and show up in a user's email box are referred to as spam. Spam is frequently responsible for network bottlenecks, blocking, and even harm to the electronic messaging system. We must integrate several machine learning techniques int our workflow, including support vector machines and naive bayes.

Keywords: Online social networks, Spam detection, rumours.

[1]Associate Professor, Department of Master of Computer Applications, Karpagam College of Engineering, Coimbatore, [2]Associate Professor, Department of Information Technology, Karpagam College of Engineering, Coimbatore

[3]Professor, Department of PG research in Computer Applications, Hindusthan College of Arts and Science, Coimbatore

[4]Professor, Department of Computer Engineering, Sri Ramakrishna Polytechnic College, Coimbatore.

Email: [1]k_anur@yahoo.com

DOI: 10.31838/ecb/2023.12.s3.265

1. Introduction

Online social networking services like Email, Facebook, and LinkedIn are prevalent in today's modern culture. One of the most popular websites is the social networking site Email. Email is used by many individuals to interact with one another. There is a lot of spam on the quickly expanding social network Email. Email spam is different from traditional spam, such as spam on blogs and emails, hence typical spam filtering techniques are ineffective and inappropriate for detecting it. Consequently, there is a need to identify Email spammers since many scholars have presented ways to detect spammers there. Social networking sites are being used much more often these days to communicate opinions and thoughts. Email is the social media platform used for disseminating news about accomplishments in the real world. Nowadays, nevertheless, a lot of people are utilising Email for marketing purposes and to disseminate spam messages on online social networks. Spammers send unwanted messages for a variety of reasons. The tweets promoting a product or linking to an online retailer's website make it abundantly evident that for some, the objective may be financial benefit. Right now, having access to knowledge is a must for everyone. Every person's essential desire for information as a means of communication. Information may be distributed in a variety of ways, one of which is through email (Email). Email, as its name suggests, is a tool for sending informational communications through the internet network in the form of text, letters, files, photos, etc. One of the outcomes of technical advancements that use the internet for communication is email. Email is highly common because using it is simple, quick, and affordable, among other benefits. Due to the convenience of usage, a large number of emails are utilised by some parties to send pointless communications that are known as spam email. Spam emails are frequently sent out and may contain ads, marketing messages, fraud, viruses, or malware. Spam emails are typically sent by careless individuals and often include unlawful material. They are also frequently sent carelessly with the intent to annoy and hurt the receivers. If spam emails contain viruses or malware, potential losses include drowning other crucial email communications, fraud, criminality, and even system security issues from utilised machines. Nowadays, millions of individuals use email in their daily lives. People utilise email for a variety of things, including business, school, and other things. Spam emails, which might contain ads, marketing messages, fraud, viruses, or malware, are extensively used. Spam emails are typically sent carelessly and include unlawful material. They are also frequently sent in large volumes, which may be annoying and harmful to receivers. In the event that spam emails are infected with malware or viruses, potential losses might include drowning other crucial email communications, fraud, criminality, and even system security risks from the machines utilised. Millions of individuals use email regularly these days. Individuals utilise email for a variety of reasons, including work, school, and other uses. Email spam, often known as electronic mail spam, is the practise of sending unwanted emails or commercial emails to a list of subscribers. Unsolicited emails signify that the receiver has not given consent to receive them. Throughout previous decade, utilising spam emails has grown in popularity. Spam has grown to be a significant online problem. Spam wastes space, time, and message delivery. Although automatic email filtering may be the best way to stop spam, modern spammers may quickly get through all of these apps. Prior to a few years ago, the majority of spam that came from certain email addresses could be manually stopped. For spam detection, a machine learning technique will be utilised. "Text analysis, white and blacklists of domain names, and community-based procedures" are three major strategies that have been implemented closer to junk mail filtering. Text analysis of email content is a widely used spam prevention technique. There are several solutions that may be used based on server and buyer considerations. One of the most well-known algorithms used in these processes is naive Bayes. In the event of false positives, it may be challenging to reject sends mostly based on content evaluation. Clients and organisations would typically not require any important messages to go lost. The boycott strategy was most likely the first one used for the separation of spam. The strategy is to acknowledge every send, excluding those from local or electronic mail ids. Expressly boycotted. This method no longer functions as effectively as it formerly did since more modern regions entered the category of spamming domain names. The "white list approach" is a method of accepting emails from domain names and addresses that have been publicly whitelisted while placing all other emails in a much lower priority queue. This method works best when the sender confirms their identity in response to a "junk mail filtering system" request. The use of electronic messaging systems to transmit unwanted bulk communications, including mass advertisements, harmful links, etc., is referred to as spam, according to Wikipedia. Unsolicited refers to communications from sources that you did not request. Thus, if you don't recognise the sender, the email may be spam. Most of the time, when downloading any free services, software, or when upgrading the programme, people are unaware that they have just signed up for such mailers. Emails

that are not often wanted but are not classified spam are referred to as "ham" by Spam Bayes, who coined the phrase in 2001. Machine learning techniques are more effective because they employ training samples, which are sets of emails that have already been categorised. There are several methods available in machine learning techniques that may be utilised for email screening. "Naive Bayes, support vector machines, neural networks, K-nearest neighbour, random forests, etc." are some of these methods.

RELATED WORKS

Tuteja S K (2016) [1] For email categorization, the author has surveyed a variety of machine learning methods, including Neural Networks (NN), Support Vector Machines (SVM), SVM Decision Tree-based classifiers, and Naive Bayes. The author utilised the Spam Base dataset as his source of data. The author of this study failed to discuss the benefits and drawbacks of any algorithm.

Mujtaba G. et al.,. (2017) [2] suggested the fundamental three phases that are used in all categorization processes. Pre-processing is the initial phase, during which the provided text is transformed into tokens and stop words are removed. The second phase is the learning process, during which a feature set that is crucial for classifying emails is developed. The final stage is to classify emails as spam or junk using an effective algorithm. For classification, techniques including support vector machines, logistic regression, regression trees, and random forests are taken into account. They categorised the emails as spam or ham using the Phishing Corpus dataset and the Bag of Words feature extraction technique. The many tools for reduction strategies for email categorization were not mentioned in this study.

Ajaz S et al., (2017) [3] They gathered email dataset from publicly accessible websites and filtered emails using Nave Bayes. He suggested a hybrid strategy that filters email data using a secure hash algorithm and Naive Bayes, but he was unable to say how to prevent the waste of storage space and network traffic. The email is regarded as a message M while utilising Secure Hash Algorithm because of a produced function. The letter M is further divided into S and L categories, where L refers for legitimate email or ham email and S stands for spam email.

Abdulhamid Mushih Shafi et al., (2018) [4] performed analyses for several machine learning classification methods, including the SVM, Random Tree, Bayesian Logistic Regression, Lazy Bayesian Rule, and Radial Basic Function (RBF) Network. Based on Precision, Recall, Root Mean Squared Error, F-Measure, and Accuracy, they compared all of the provided methods. They made use of the UCI Machine Learning Repository dataset. They used the F-measure approach to calculate the accuracy and recall value. The Rotation Forest method produced the greatest measure, whereas the Naive Bayes approach produced the lowest. The best result was reached for the Rotation Forest Algorithm with 87.9 using the Kappa Statistics for the statistical results. Rotation Forest technique produced the highest accuracy (94%), whereas REP Tree algorithm produced the lowest accuracy (89%). Other algorithms, such Naive Bayes and SVM, provided accuracy of 88.5% and 92.3%, respectively.

Rusland NF et., al., (2017) [5] performed analysis on email categorization using Naive Bayes algorithm based on Accuracy, Precision, F-Measure, and Recall on two separate datasets. Three phases made up the procedure. The first phase is data pre-processing, which involves eliminating all articles, conjunctions, and unwanted words from the text. Following feature extraction, the Naive Bayes model is trained. The machine estimates if the provided text is spam or ham based on its training. The accuracy obtained by the author using the spam data dataset was 91.13%, while accuracy obtained using the other spam base dataset was 88%. The author came to the conclusion from his investigation that the Nave Bayes algorithm performs better on the Spam data dataset than Spam Base.

Yuksel SF et al., (2017) [6] For email filtering, they contrasted Support Vector Machine (SVM) with Decision Tree. A training set and a testing set were created from the provided dataset. Each model is trained independently, and after training, its accuracy is evaluated. The accuracy of the SVM algorithm, which the author used for both methods, was 92%, while the accuracy of the Decision Tree approach, which he used, was 82%. The author came to the conclusion that SVM outperformed Decision Tree based on his research.

Shradhanjali and Verma (2017) [7] proposed a technique for screening emails using the SVM algorithm and feature extraction. This process involves a number of processes, including Email Collecting, where data is taken from the dataset. It is then routed via preprocessing, wherein extraneous material is eliminated and only desired content is passed on to the next step. Then comes feature extraction followed by SVM model training. The dataset from the Apache Public Corpus was used by the author. Special symbols, HTML elements, URLs, and extraneous alphabets were deleted from the suggested solution. All of the dictionary terms were mapped by the author using the vocabulary file. A 98% accuracy was achieved using the SVM method on a pre-processed dataset.

V. K. Singh Bhardwaj, S. (2018) [8] worked on a method of integrating classification techniques to improve spam filtering results. The author used data mining to collect all the data on spam

filtering's past successes, present issues, and problems in the past. The approach was based on binary categorization, with 0 denoting legitimate emails and 1 designating spam. For the purpose of email filtering, they merged the two methods of machine learning and knowledge engineering. For the combined KNN and SVM algorithm, the performance of the suggested solution was quite subpar.

Preriti Sharma A. Uma Bhardwaj (2017) [9] compared the SVM decision tree method to the Naive Bayes technique for email categorization. They made use of a 1000 element dataset. They conducted three tests, and the algorithms were evaluated based on the findings by assessing several performance characteristics including accuracy, recall, precision, true negative rate, and F-measure. In the first trial, a Naive Bayes classifier was used, and accuracy was reached at 83.5%, along with precision and recall values of 85.26% and 85.26 respectively. The accuracy of the second trial utilising the SVM decision tree classifier was 91.5%, with precision values of 93.68% and 89% recall. The hybrid bagged technique was used for the third trial. The accuracy attained in the previous trial was 87.5%, with a precision value of 89.47% and a recall value of 85%. The boosting strategy may be used to replace the weak classifier's learning features with those of the strong classifier in the future.

Manmohan Singh et al., (2018) [10] used Support Vector Machines to classify emails and compared with Gaussian Kernel with the Linear Kernel. A class can be divided using a linear decision boundary in a linear separable problem. There may be multiple different decision boundaries for a given problem, but a good and effective decision boundary is the one that accurately matches the provided data and is also capable of classifying any additional data. When the presented dataset cannot be properly suited by a linear decision boundary, a gaussian boundary might be utilised. They utilised the 4k spamTrain.mat dataset, which includes both ham and spam emails, for training purposes. 1000 entries in the SpamTest.mat file were utilised for testing. A portion of the Spam Assassin Public Corpus includes both the training and test files. They concentrated on the training time and testing time for both methodologies in addition to the testing accuracy. With a linear kernel, the accuracy was 98.5%, and the training time was 134 seconds. Using a Gaussian kernel, the accuracy was 97.1%, and the training time was 190 seconds (in sec). As a consequence, the author came to the conclusion that linear kernel training takes much less time than gaussian kernel training and that linear kernel accuracy is higher than gaussian kernel accuracy. The dataset utilised has a big number of features, therefore even though the Gaussian kernel is more sophisticated and better fitting than the linear kernel, a dataset with a large number of features fits more effectively using the linear kernel than the Gaussian Kernel.

Cindy Huang et al [11] suggested a fix to improve Naive Bayes' accuracy and lower the false positive rate. Naive Bayes is a Bayes theorem-based supervised machine learning technique that may be used as a probabilistic model for email categorization. Leetspeak and diacritics are used by spammers to get beyond filters, despite Naive Bayes classifier's greater accuracy. Leetspeak is a coded spelling system and language used in highly casual online communication. It includes creative misspellings, jargon, and slang and uses letters coupled with numbers or special symbols in place of letters that they may resemble. A diacritical mark, such as an accent or cedilla, is a symbol that, when written above or below a letter, indicates that the pronunciation of that letter differs from that of the same letter whether it is not marked or is marked differently. They modified Naive Bayes to turn the symbols in the text into potential letters, applied a spell check to confirm that the corrected symbol is a word, and then put the data through the classification method. They increased accuracy by doing this from 23.9% to 62%.

Prachi Gupta et al.,(2019) [12] examined how well the Naive Bayes and the Support Vector Machine algorithms classified emails. They have utilised a dataset of 5574 rows and 2 columns. Two columns are used: one for labelling and the other for storing emails (Ham or Spam). For the categorization of emails, they employed a total of 4 steps: data collection, data preprocessing, data transformation, and classification system. The data was cleaned and made free of all ambiguities, mistakes, and redundancies using data pre-processing. Pre-processed data is translated into lowercase and into the format required by the classification algorithm during data transformation. Finally, the relevant qualities are found, and an algorithm uses feature extraction to categorise the material as either Ham or Spam. Naive Bayes' accuracy was 99.49%, and Support Vector Machine's accuracy was 86.35%. The author came to the conclusion that the Naive Bayes method outperformed SVM in terms of email classification.

U.K Sah] et al., (2017) [13] proposed a strategy for identifying emails as Hamor Spam using feature selection and tried to increase the spam filtering model's training time and accuracy. The Naive Bayes method and the Support Vector Machine were also compared. based on the algorithm's precision and calculation time for the specified dataset. Four steps made up the entire procedure. The first step was data preparation, during which the provided dataset was split into a training set with 702 emails and a testing set with 260 emails.

The second phase included creating a word dictionary, and the third involved selecting features by creating a feature vector matrix. The model was trained in the last phase, and based on that training, it now predicts whether an email is spam or junk. The author came to the conclusion that Naive Bayes provides superior accuracy than Support Vector Machine based on the findings gathered.

D.Ruano-Ordas et al., (2018) [14] used regular expression to discover a word or group of words that displayed some sort of pattern. They modified an existing algorithm and created DiscoverRegex, an effective algorithm. This technique, which was dynamic in nature, could generate regular expressions for a given dataset automatically.

The spam or junk emails were gathered by the author directly from various sources [15]. They painstakingly chose 23 variables from the dataset that they meticulously studied to determine if an email was spam or not. Each of the criteria was given a value, and a threshold value was set based on the analysis. Each email was assigned a total value that was computed, verified to see if it exceeded or fell short of the threshold value, and the categorization was determined. That wasn't particularly successful because the study only used a small dataset of 750 emails.

2. Methodology

Preprocessing of the data: When data are taken into consideration, especially big data sets with numerous rows and columns are always noticed. Nevertheless, this is not always the case because the data might be be in the form of image, audio, or video files. Detailed tables, etc.

Machines simply comprehend 1s and 0s; they are incapable of understanding photos, video, or text data.

Data preprocessing steps: cleansing data: The tasks of "filling in missing values," "smoothing noisy data," "identifying or deleting outliers," and "resolving discrepancies" are completed in this stage. Integration of data: Many databases, information files, or information sets are added in this stage. Transformation of data : Scaling up to a certain value is accomplished by aggregation and normalisation.

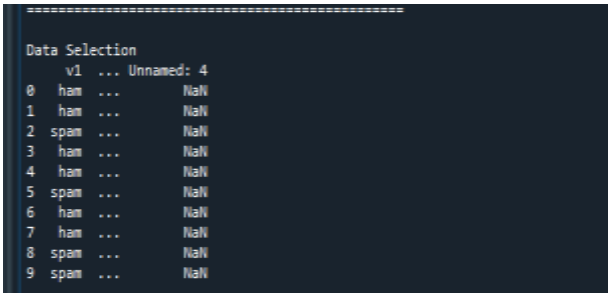

Figure 1 Data Cleaning

Data reduction: While the dataset in this part is relatively little, it has so far produced the same analytical conclusion. Stop words are English words that don't contribute much sense to a statement, according to one definition. These can be safely disregarded without affecting the sentence's meaning. For instance, if a search for "how to make a veg cheese sandwich" is attempted, the search engine will attempt to look for online sites that contain the terms "how," "to","make," "a," "veg," "cheese," "sandwich," etc. The search engine looks for websites that contain the phrases "how","to," and "a" rather than pages that feature veg cheese sandwich recipes since these terms are so often used in English. The outcome would be interesting if these three terms were dropped or halted in favour of fetching sites that had the keywords "veg", "cheese", and "sandwich" as shown in Figure 1.

Tokenization: "Tokenization is the process of dividing a stream of content into tokens," which might be words, symbols, phrases, or other expressive features. The list of tokens is also used to contribute to subsequent processing, such content mining and parsing. Tokenization is useful for lexical analysis in software engineering and construction as well as semantics (where it serves as content separation). It might be challenging to explain what is meant by the term "word" at times. As word-level tokenization takes place. A token frequently relies on simple heuristics, such as: Whitespace characters, such as "line break" or "space," or "punctuation characters," are used to separate tokens. Each adjacent group of alphanumeric letters, just like each group of digits, makes up a single token. The lists of tokens that are produced may or may not include white spaces and punctuation. The below fig 2 show the Data Pre-processing stage.

Fig 2: Preprocessing

Old school classifier: Data analysis that uses classification extracts the models characterising significant data classes. For the purpose of predicting class labels, such as "A loan application as dangerous or safe," a classifier or model is built. Data classification is a two-step process that involves learning (building a classification model) and categorization.

1. NAIVE BAYES: In 1998, the Nave Bayes classifier was utilised to identify spam.

$$P(y|x) = \frac{P(x|y)\ P(y)}{P(x)}$$

For supervised learning, there is an algorithm called the Naive Bayes classifier. The Bayesian classifier uses dependent events and calculates the likelihood that an event that has already happened may predict an event that will happen in the future. On the basis of the Bayes theorem, which presumes that characteristics are independent of one another, naive Bayes was developed. The Naive Bayes classifier method may be used to categorise spam emails since word likelihood is the key factor at play. Any term that frequently appears in spam but not in ham indicates that an email is spam. The Naive Bayes classifier algorithm is currently the most effective method for email filtering. For this, the model is extremely well trained using the Nave Bayes filter. Each class' probability is always calculated by the Naive Bayes algorithm, and the class with the highest probability is then selected as the output. Naive Bayes results are always reliable.

Existing System

Information is one of the fundamental demands of the present, according to current study. Email is one of the many services used to convey informational messages. Email's multiple conveniences have both beneficial and harmful effects. One of the bad effects is that an abused email turns into spam email, which may be harmful and harmful to the receiver. This serves as the foundation for research comparing the Multinomial Naive Bayes Classifier algorithm (MNBC), Support Vector Machine algorithm (SVM), and Recurrent Neural Network algorithm (RNN) to find the best accurate assessment of spam in emails [7]. The evaluation's findings are provided using the Classification Report to show how each method performed in terms of accuracy, precision, memory use, and f1 score. The Support Vector Machine method, whose accuracy value is 96%, precision value is 0.92, recall value is 0.96, and fl-score is 0.94, is the algorithm that delivers the highest accuracy value in this study's study of email spam categorization.

DISADVANTAGES: Results fall short of expectations; it is inefficient for handling big volumes of data; it takes a lot of time; theoretical limits.

PROPOSED SYSTEM

The email dataset was used as input for this system. The source of the input data was a dataset repository. After that, we must carry out the data pre-processing stage. At this stage, we must deal with the missing values to prevent incorrect prediction and encode the input data's label. The next step is to put natural language processing into practise. We must eliminate punctuation, stop words, and stemming in this phase. The dataset must then be divided into test and train groups. Ratio-based data splitting is used. The majority of the data will be present in train. A reduced percentage of the data will be present during the test. The model is assessed during the training phase, and predictions are made during the testing phase. The vectorization must then be put into practise. That indicates that in order to construct feature vectors, text must be encoded as integers or numeric values. The categorization method must then be put into practise using machine learning. the SVM and Naive Bayes machine learning methods. The experimental findings demonstrate that performance indicators like recall, precision, and accuracy are important.

MODEL ARCHITECTURE

This section outlines the process used by the Spam Mail Detection (SMD) System to categorise emails as either ham or spam. The notion of the hybrid bagged method was introduced with the SMD system's powerful categorization capabilities. Correlation-based feature selection and a cutting-edge hybrid bagging methodology are used in the feature selection method. The SVM method, which is based on decision trees, and the Naive Bayes Multinomial classifier are used in the bagging technique's hybrid approach for classification. Figure 4 displays the SMD system's flowchart for

email categorization. Emails are divided into spam and ham emails by the SMD system. Initial pre-processing is done on the text-based email dataset under consideration to ensure effective feature extraction. the method of classification a hybrid bagged strategy is taken into account.

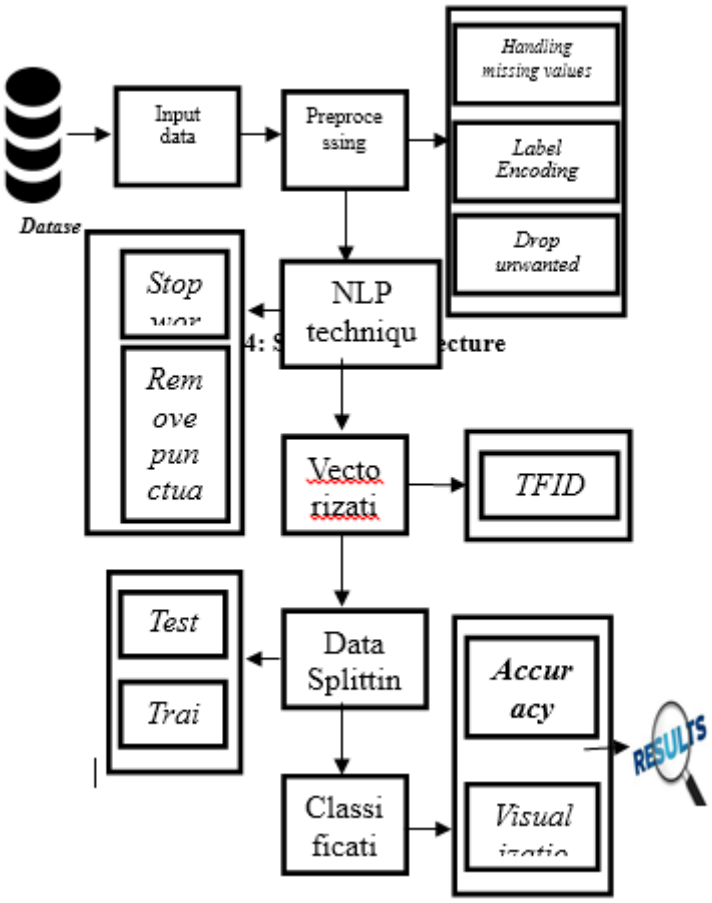

Figure 4: System Architecture

Email Dataset: For the Spam mail detection system, an email dataset is created. Randomly selected emails are gathered from the Ling spam collection. For categorization purposes, the dataset consists of a total of 1000 emails, including both ham and spam emails. The dataset is partitioned into sets for each classifier method since the strategy being explored is the bagging approach. There are two sets, each with 500 emails. 300 emails each are utilised for training the Naive Bayes and SVM algorithms, and 200 emails each are used for testing.

Pre-processing of the dataset: The email dataset under consideration is unprocessed. So, it has to be prepared before being further considered. Three phases make up the pre-processing stage. To begin with, the text data is tokenized. The statement is broken

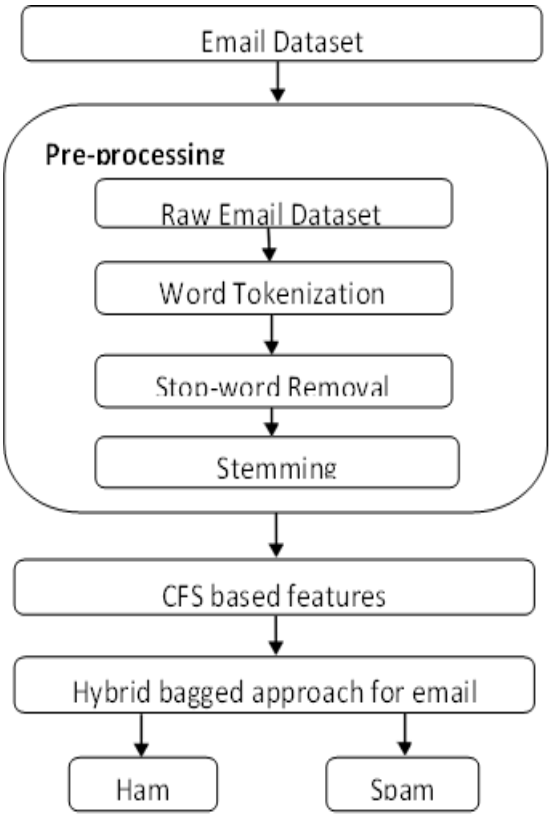

Figure 5: System Flow

up into token words. Stop words are eliminated from the tokenized words. Stop words are undesirable words without any linguistic significance. Around 670 stop words are manually added to a text file, and during pre-processing, words are taken out of the text. The stemming process is the third stage in the pre-processing module. The term is reduced to its basic word by the stemming process. Stemming and stop word removal are crucial pre-processing processes because they help to narrow the search field for effective feature extraction and selection.

Feature Selection: Every categorization system relies heavily on features. The SMD method operates with the presumption that spam mail has different content than ham mail. Alphanumeric words, language, grammatical or spelling mistakes, improper terms (words associated with the promotion of goods or services, words associated with dating, adult words, etc.), frequency count, document length, and other features are included in the feature set. Correlation feature selection (CFS) is a mechanism used in SMD systems. CFS only selects the top features from a pool of characteristics that are beneficial to enhancing system performance. According to the premise that "Good feature subsets comprise characteristics highly correlated with the classification, yet uncorrelated to each other", correlation-based feature selection methodology is based on features that are chosen.

3. RESULTS

By assessing the performance metrics, the effectiveness of the suggested spam mail detection system is discovered. In order to assess the effectiveness of the Spam mail Detection system, metrics like precision, recall, accuracy, F-measure, true negative rate, false negative rate, and false positive rate are calculated. Based on the metrics listed in table 4, the SMD system's performance is assessed. The system's overall accuracy, which is the average of the two classification algorithms' accuracy levels, is 87.5%. The Naive Bayes classifier successfully achieves an accuracy of 83.5%, with precision and recall values of 85.26% and 81%, respectively. The SVM algorithm [1][3], on the other hand, achieves an accuracy of 91.5%, with precision and recall values of 93.68% and 89%, respectively.

Evaluation Measures	**Naïve Bayes**	**SVM**
TP	81	89
FP	14	6
TN	86	94
FN	19	11
Precision (%)	85.26	93.68
Recall (%)	81	89
Accuracy (%)	83.5	91.5
F-Measure (%)	89.27	84.8
TNR (%)	86	94
FPR (%)	19	11
FNR (%)	14	6

Table 1: Comparison between Naive Bayes and Support Vector Machine

The assessed outcomes of the three tests, which used the Naive Bayes, SVM algorithm, and hybrid bagged technique, are shown in above table.

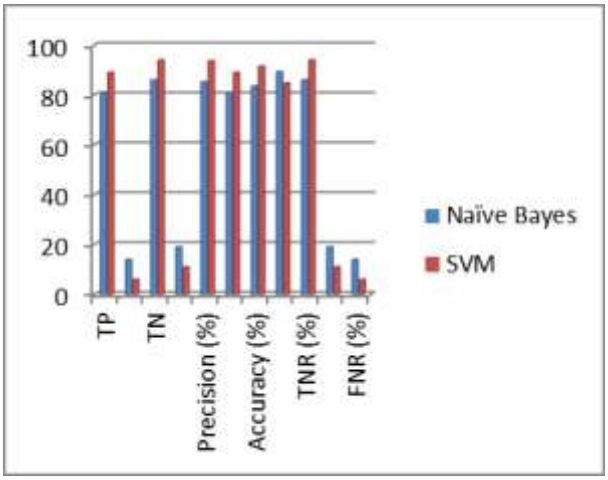

Chart 1: Chart showing the comparison between Naive Bayes and Support Vector Machine

The SVM decision tree method outperforms the Naive Bayes and the hybrid bagged approaches in terms of precision, recall, and accuracy, according to a comparative study of the findings that is shown in table. The F-measure percentage value for Naive Bayes, however, is greater (89.27%) than that for SVM (84.8%) and the hybrid bagged technique (87.03%). In above Figure, the comparison between the outcomes obtained by the SMD system and the related classification algorithms is shown graphically.

IMPLEMENTATION

Data Selection:

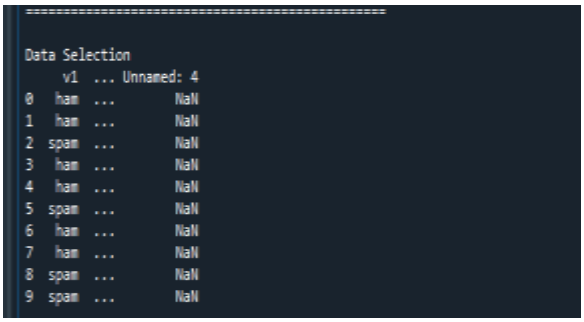

Figure 6: Data Selection

The process of choosing the proper data source, data type, and tools to gather the data is known as data selection. The real activity of collecting data comes before data selection. This definition distinguishes between active/interactive data selection (using gathered data for monitoring activities/events or undertaking secondary data analysis) and selective data reporting (excluding data that is not supportive of a study premise). The method used to choose appropriate data for a research endeavour may have an effect on data integrity. Finding the proper data kind, source, and instrument that enables researchers to appropriately address research issues is the main goal of data selection. This decision is frequently discipline-specific and is largely influenced by the nature of the inquiry, the body of prior research, and the ease of access to relevant data sources.

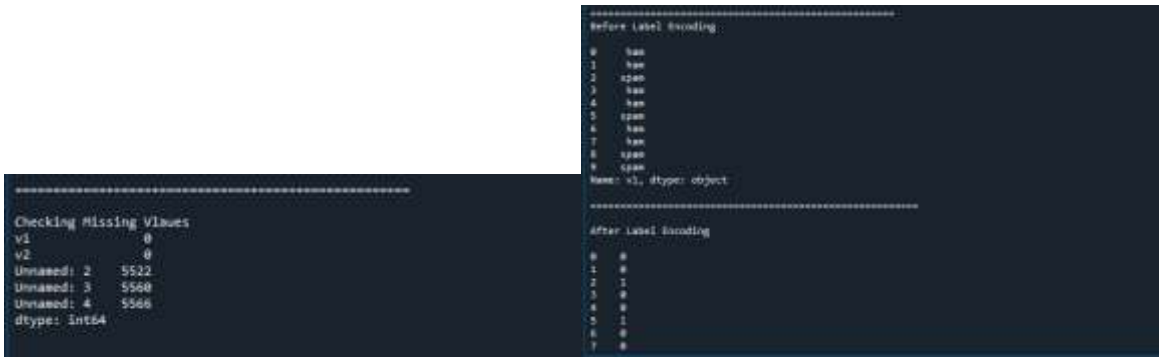

Figure 7: Data Preprocessing

Preparing the raw data to be acceptable for a machine learning model is known as data preprocessing. In order to build a machine learning model, it is the first and most important stage [16]. It is not always the case that we come across the clean and prepared data when developing a machine learning project. Also, every time you work with data, you must clean it up and format it. Thus, we employ a data pre-treatment activity for this.

Steps in Data Preprocessing:
To obtain the dataset
bringing in libraries
Bringing in datasets
Encoding Lost Data and Discovering It Data Categorical
dividing the dataset into a test and training set
Aspect sizing

Figure 8: Natural Language Processing

Recent advances in the domains of machine learning and natural language processing have made them significant subfields of artificial intelligence. Turning an artificial agent into an artificial "intelligent" agent depends heavily on machine learning and natural language processing. Because of improvements in natural language processing, an artificially intelligent system can take in more information from its surroundings and respond to it in a way that is user-friendly. Similar to this, by using machine learning techniques, an artificially intelligent system may analyse the information it receives and make better predictions about how it will respond.

The system can learn from examples and prior experiences thanks to machine learning. Generic algorithms cannot tackle unknown situations since they only carry out a predetermined set of operations as specified by their programming. Also, the majority of situations in the actual world have a lot of unknown variables, which makes the standard algorithms incredibly inefficient. Here, machine learning takes centre stage. A machine learning method is far more suited to address such unsolved issues with the aid of prior instances.

One of the well-known examples given is the detection of spam mail. There are numerous unknowns involved in determining whether a message is real or spam. There are several techniques to get around spam filters.

Figure 9: Prediction

The above viewed is the result screens of the project where we given an input which is analysed using the above techniques and returns it as a spam message.... In the below screen it is plotted using matplotlib.

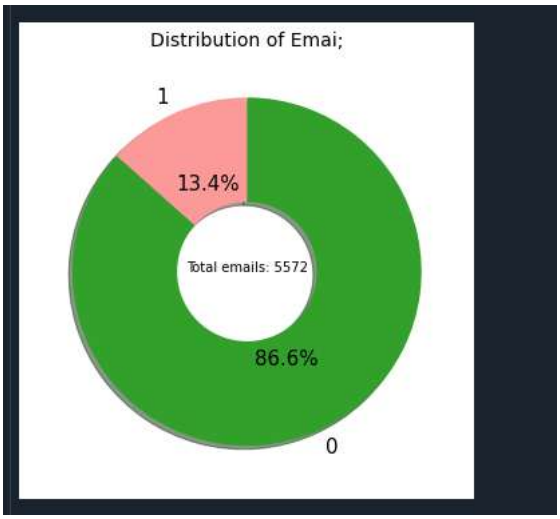

Figure 10: Output visualisation image using Matplotlib

4. CONCLUSION

One of the most demanding and problematic internet concerns in today's communication and technological environment is spam email. Spammers abuse this communication tool by sending out spam emails, which has an impact on businesses and numerous email users. In this research, a hybrid bagged approach-based spam mail detection system is described, including its implementation. Nave Bayes and SVM are the classification algorithms employed in this method. Naive Bayes and the SVM algorithm both reach accuracy of 83.5% and 91.5%, respectively. The hybrid bagged method based SMD system's overall accuracy of 87.5% demonstrates that the experimental outcomes are better when the SVM algorithm is used exclusively.

FUTURE WORK: To improve the system's efficiency and output, Future work could take the boosting strategy into consideration. By substituting the learning characteristics of the strong classifier for those of the weak classifier, the

boosting strategy improves the performance of the entire system.

5. References

Tuteja S K (2016), A Survey on Classification Algorithms for Email Spam Filtering, International Journal of Engineering Science and Computing, 6(5), 5937 – 5940.

Mujtaba G, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi (2017), Email Classification Research Trends: Review and Open Issues, IEEE Access, IEEE Transactions and Content Mining, 5, 9044 – 9064.

Ajaz Sana, Nafis, Md, Sharma, Vishal. (2017). Spam Mail Detection Using Hybrid Secure Hash Based Naive Classifier, International Journal of Advanced Research in Computer Science, 8(5), 1195 – 1199.

Shuaib Bobi Maryam, Osho Oluwafemi, Idris, Ismaila, Alhassan, John, Abdulhamid, Shafi'i. (2018). Comparative Analysis of Classification Algorithms for Email Spam Detection. International Journal of Computer Network and Information Security(IJCNIS). 1. 60-67. 10.5815/ijcnis.2018.01.07.

Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim, Hanayanti Hafit, (2017), Analysis of Naive Bayes for Email Spam Filtering across Several Datasets, International Research and Innovation Summit (IRIS2017), IOP Conf. Series: Materials Science and Engineering, 226, doi:10.1088/1757-899X/226/1/012091

Yüksel, A. S., Cankaya, S. F., and Üncü, Design of a Machine Learning Based Predictive Analytics for spam problem, Special issue of the 3rd International Conference on Computational and Experimental Science and Engineering (ICCESEN 2016), 132(3), ACTA PHYSICA POLONICA A. 500 – 504.

Shradhanjali, Toran Verma (2017), E-Mail Spam Detection and Classification Using SVM and Feature Extraction", International Journal of Advance Research, Ideas and Innovations in Technology, 3(5), 1491 – 1495.

Singh, V. K., Swetha B., (2018), Spam Mail Detection Using Classification Approaches and Global Training Set, Intelligent Computing and Information and Communication (pp.623-632).

Priti Sharma and Uma Bhardwaj (2017). Machine learning based spam email detection, Proc. of Data Science and Security.

Manmohan Singh, Rajendra Pamula, and Shudhanshu Kumar Shekhar (2018). Email Spam Classification by Support Vector Machine, International Conference on Computing, Power and Communication Technologies (GUCON), 878 – 882.

Julia Jia, Wuxu Peng, Linda Huang, Emma Ingram (2018). Enhancing the Naive Bayes Spam Filtering using Intelligent Text Change Detection, 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications.

Prachi Gupta, Ratnesh Kumar Dubey, and Dr. Sadhna Mishra (2019). Detecting Spam Emails/Sms Using Naive Bayes And Support Vector Machine, International Journal of Scientific & Technology Research, 8(11), 1 – 4.

Sah, U. K., Parmar, N (2017). An approach for Malicious Spam Detection in Email with comparison of different classifiers, International Research Journal of Engineering and Technology (IRJET), 4(8), 2238 – 2242.

Ruano-Ordas D, F. Fdez-Riverola, and J. R. Mendez (2018). Using evolutionary computing for detecting spam patterns from e-mail samples, Information processing and Management, 54(2), 303 – 317.

A.S. Aski and N.K. Sourati (2016). Proposed Efficient method to filter spam using machine learning approaches, Pacific Science Review A: Nature Science and Engineering, 18, 145 – 149.

Anuradha K, Senthil Kumar P, Naveen Prasath E, Vignesh M, Sneha S, "Fake News Detection using Decision Tree and AdaBoost", European Chemical Bulletin, 12(S3), 570 – 582, 2023.