Email Spam Classification Using Machine Learning algorithm.

Heramb Devarajan

Student of Computer Science,
Birla Institute of Technology and Science, Pilani, Dubai Campus, United Arab Emirates

ABSTRACT

Email has revolutionised communication, but it has also led to the rise of spam, or unsolicited emails. As the volume of emails increases, so does the proportion of spam, necessitating sophisticated mechanisms to filter out unwanted messages. This further leads to problems like social engineering attacks, phishing attacks and so much more. A spam email classification is required to tackle this issue and to make the cyberspace secure for the users and to protect companies from attacks. This research aims to tackle email spam classification using various machine learning techniques. Traditional rule-based approaches have proven insufficient in keeping up with evolving spamming tactics. Machine learning, with its ability to discern patterns and adapt to dynamic environments, is a powerful solution to distinguishing between legitimate and spam emails. It can be used as a powerful tool to filter out the spam emails hence keeping the inbox clear of any spam or phishing emails. We can use Machine learning algorithms to do this by creating a model and training it with proper dataset and then using the model to classify spam emails. We can get open source dataset from the internet and train our model. We can use various machine learning algorithms and techniques like integrating a multifaceted feature set, including textual content, sender information, and metadata, to capture the nuanced characteristics that differentiate spam from legitimate communication. The system employs machine learning algorithms such as Support Vector Machines, Random Forest, and Naive Bayes to accurately classify emails as spam or not. The model undergoes rigorous training on a diverse dataset of spam emails and non-spam emails to increase accuracy and other evaluation metrics for better classification of the spam emails.

Keywords: Machine Learning, Email Spam, Phishing, Classification, Cybersecurity, Feature Engineering

INTRODUCTION

Email communication is an integral part of modern-day interactions, but the persistent challenge of spam emails continues to compromise the efficiency and security of this medium. This research focuses on the development of an advanced email spam classification system leveraging machine learning techniques for addressing these issues. The objective is to create a robust model capable of accurately distinguishing between spam and legitimate emails, thereby enhancing the overall email experience for users. The proposed system adopts a multifaceted approach, incorporating a diverse set of features to comprehensively analyze email data. These features encompass content-based, structural, and behavioral attributes, providing a holistic view of the email's characteristics. By leveraging a large and well-curated dataset containing both spam and non-spam emails, the machine learning model can learn intricate patterns and correlations inherent in spam content. Feature engineering plays a crucial role in extracting relevant information from the dataset. This involves transforming raw data into a format that is conducive to the learning process of machine learning algorithms. Various characteristics like the frequency of particular phrases, the existence of particular keywords, and the structure of the email, are carefully selected to enhance the model's ability to discern spam from legitimate messages.

The research evaluates a spectrum of classification algorithms, incorporating Neural Networks, Random Forest, and Support Vector Machines, to determine their effectiveness in accurately classifying emails. Each algorithm is refined and evaluated using criteria including F1 score, recall, and precision. This rigorous evaluation process ensures that the selected algorithm not only performs well on the training data but also generalises effectively to new, untrained data. In addition to traditional feature sets, the study explores the utilization of natural language processing (NLP) methods to enhance the model's understanding of textual content. By analysing the linguistic nuances of emails, the system aims to capture subtle cues that may indicate spam. NLP techniques, including sentiment analysis and semantic analysis, contribute to a more nuanced and context-aware spam classification. To validate the efficiency of the suggested approach, real-world email datasets are used in experiments. The results demonstrate a significant improvement in spam detection accuracy compared to traditional rule-based methods. The system's adaptability to evolving spam patterns is a key strength, ensuring that it remains effective even as spammers employ new tactics and techniques.

The implications of this research extend beyond the technical realm, offering tangible benefits for both individual users and organizations. By deploying a sophisticated spam classification system, email providers can enhance the security and reliability of their services, leading to improved user trust and satisfaction. Individual users, in turn, experience a reduction in the influx of unwanted and potentially harmful emails, streamlining their communication channels. In conclusion, this research makes a contribution to ongoing efforts to combat email spam by proposing a comprehensive and intelligent classification system. By leveraging machine learning techniques, including advanced feature engineering and NLP, the system demonstrates a high degree of accuracy and adaptability. The research results paves the way for the development of more robust and scalable solutions, addressing the ever-evolving landscape of email spam and fortifying digital communication channels.

LITERATURE SURVEY

So far, there has been many works on Email Spam Classifiers. Various works employ different techniques. Here is a summary of a few of the chosen works.

The paper "Spam E-Mail Classification by Utilizing N-Gram Features of Hyperlink Texts" authored by A. Selman Bozkir, Esra Sahin, Murat Aydos, Ebru Akcapinar Sezer, and Fatih Orhan focuses on innovative methods for combating email spamming, particularly the use of hyperlink texts[1]. Email spamming is a pervasive issue in digital communication, with 70% of emails sent for spamming purposes. Knowledge engineering methods, such as collaborative spam filtering, heuristic filters, whitelisting, blacklisting, greylisting, honey pots, and signature schemes, have shown success in spam mail filtering but have limitations. Machine learning methods, on the other hand, provide more adaptive and dynamic countermeasures by learning hidden patterns from real-world training data. The dynamic nature of spam filtering has driven the adoption of machine learning approaches in recent years, including Naive Bayes, Support Vector Machines, decision trees, and conventional neural networks. Bag-of-words (BOW) and n-gram feature representations have gained popularity in spam mail detection due to emails being primarily composed of textual data. In existing literature, n-gram indexes have been explored for spam/ham mail classification, with studies demonstrating the superiority of SVM over methods like Naive Bayes and k nearest neighborhood classification. Active machine learning has also been used to classify malicious spam emails, incorporating text-based features from email bodies. The URL has also been utilized to identify spam and fraudulent emails.. Despite the richness of existing literature, the authors identify a gap in the usage of hyperlinks in emails. The paper introduces a novel approach, proposing the use of hyperlink texts located in the body of emails as the primary feature source for spam/ham email classification.

The paper "Antlion Optimization and Boosting Classifier for Spam Email Detection" by Amany A. Naem, Neveen I. Ghali, and Afaf A. Saleh presents a two-stage methodology called ALO-Boosting, which combines antlion optimization (ALO) and boosting to improve the precision of machine learning algorithms for spam email classification. ALO is a meta-heuristic optimization algorithm that efficiently searches for the optimal feature subset within a given search space, ensuring that the selected features significantly contribute to the classification task[2]. In the second stage, boosting is applied to the features selected by ALO, producing a highly accurate and robust classifier. This step transforms soft learners into powerful ones, enhancing the system's ability to discriminate between spam and non-spam emails. Comparative experiments with established classification methods, such as Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Bootstrap Aggregating (Bagging), demonstrated the superior performance of ALO-Boosting in detecting optimal features while minimizing the number of selected features, leading to high precision in spam email classification. The authors also investigated the impact of key parameters on the performance of ALO-Boosting, identifying optimal values for these parameters, contributing to fine-tuning the method for enhanced results. In conclusion, the paper presents a comprehensive and efficient solution to the challenge of spam email detection, demonstrating its competitive edge in achieving high precision while minimizing feature dimensionality.

The paper "Machine Learning for Email Spam Filtering: Review, Approaches, and Open Research Problems" by Emmanuel Gbenga Dada and his colleagues provides a comprehensive review of machine learning approaches for email spam filtering. The authors highlight the pervasive nature of spam, accounting for over 77% of global email traffic, and its adverse effects on users, including disruptions in time management, storage capacity constraints, and financial losses due to internet scams[3]. They highlight the initiatives taken by major internet service providers like Gmail, Yahoo, and Outlook, which leverage machine learning techniques for efficient spam detection. The study emphasizes the adaptability of machine learning models, particularly those employed by Google, which exhibit a 99.9% accuracy in filtering out spam and phishing emails by continuously learning from extensive datasets. The review categorizes and analyzes various machine learning approaches used in email spam filtering, including case-based filtering, content-based filtering, previous likeness-based filtering, heuristic or rule-based filtering, and adaptive spam filtering. It also examines popular machine learning algorithms applied in spam filtering, including Neural Networks, Naive Bayes, Support Vector Machines (SVM), and Decision Trees, and discusses stochastic optimization techniques like Genetic Algorithms and Particle Swarm Optimization. The paper concludes with a synthesis of the paper's contributions, including a survey of crucial features of email spam, an exploration of spam filter architectures, exposure to lesser-explored machine learning algorithms, identification of open research problems, and proactive recommendations for advancing machine learning techniques to combat emerging spam variants.

The paper "Detecting Spam Email With Machine Learning Optimised With Bio-Inspired Metaheuristic Algorithms" by Simran Gibson, Biju Issac, Li Zhang, and Seibu Mary Jacob discusses the use of machine learning (ML) algorithms optimised with bio-inspired metaheuristic algorithms for spam email detection[4]. The authors conduct a literature review, highlighting various techniques and methodologies in the field, such as Naive Bayes, Multi-Layer Perceptron, Support Vector Machine, Decision Tree, and Random Forest, along with bio-inspired optimization methods like Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). They emphasise the challenges posed by spam emails and the importance of automated detection mechanisms. The proposed research aims to enhance spam detection by experimenting with ML models on different datasets, employing feature extraction and preprocessing. The bio-inspired algorithms, PSO and GA, are utilised to optimise the performance of classifiers, with Multinomial Naïve Bayes and GA identified as the most effective combination. The paper also discusses the use of these algorithms in specific datasets, focusing on Python platforms like Spyder, Jupyter Notebook, Google Collaboratory, and Kaggle. The authors recommend further exploration of bio-inspired algorithms like Firefly, Bee Colony, and Ant Colony Optimization and propose investigating deep learning Neural Networks with PSO and GA. The overall conclusion emphasises the successful implementation of models combined with bio-inspired algorithms, achieving promising results in spam detection accuracy.

The paper "A Semantic-based Classification Approach for Enhanced Spam Detection" by Nadjate Saidani, Kamel Adi, and Mohand Saïd Allili presents a novel approach to spam detection that incorporates semantic information and domain categorization[5]. The authors identify six major domains targeted by spammers, including Health, Finance, Adult, Computer, Education, and Others. They use a two-stage process for spam detection: Categorization of email by domain and semantic feature extraction based on domain. In the first stage, emails are categorised into predefined domains based on their content, using preprocessing steps like keyword recognition, word with separate letters recognition, stemming, segmentation, stop-word removal, and spell checking. Feature selection is performed using information gain (IG) to identify the most discriminative terms for each category. The second stage focuses on domain-specific semantic feature extraction, using manual and automatic methods. The paper emphasizes the importance of semantic features for effective spam detection, as they capture both morphological and semantic relationships between words. The authors evaluate their approach using a comprehensive dataset from public sources and compare the performance of six classifiers. The results show that their approach outperforms state-of-the-art methods such as eTVSM, Doc2Vec, and BoW-SF. The paper also explores the adaptability of their approach to new data by retraining classifiers with an augmented dataset.

The paper "Detection of Spam and Threads Identification in E-mail Spam Corpus Using Content-Based Text Analytics Method" by U. Murugavel and R. Santhi presents the Multi-Split Spam Corpus Algorithm (MSSCA) as a robust mechanism for spam detection and categorization. The research focuses on enhancing spam filter techniques to efficiently reduce the influx of unnecessary spam messages, providing email users with a more secure and streamlined communication experience. The MSSCA algorithm classifies different types of spam threads by checking the spam corpus database using content text analytics methods[6]. The algorithm aims to improve the performance of spam detection and provide better solutions for handling ethical hacking issues related to spammers. The proposed methodology involves tokenizing the text and removing stop words from the email dataset, then filtering and classifying the dataset into spam and ham, identifying various spam threads using content text analytics. The MSSCA algorithm aims to extract frequent spam threads and enhance the overall performance of the spam corpus. The research presents experimental results on the extraction of various spam threads using content text analytics, showcasing its effectiveness in handling spam. The authors emphasize the avoidance of hacking problems and the identification of different types of spam threads based on text analytics.

Areej Alhogail and Afrah Alsabih's paper "Applying machine learning and natural language processing to detect phishing email" presents a novel phishing email classifier model that uses deep learning algorithms, specifically Graph Convolutional Network (GCN), and natural language processing (NLP), to enhance detection accuracy[7]. The model aims to address the urgent nature of phishing attacks, which manipulate human emotions for quick actions, often resulting in financial and data losses. The study is grounded in the context of growing online services and the rise in cyber-attacks, with phishing being a prevalent and effective method. The proposed classifier is designed to address the dynamic and evolving nature of phishing emails, necessitating advanced detection mechanisms beyond traditional rule-based and signature-based approaches. The experimental evaluation of the classifier shows a high accuracy rate of 98.2% and a low false-positive rate of 0.015. The paper contributes a sophisticated phishing detection model, combining GCN and NLP, demonstrating its effectiveness in addressing the challenges posed by phishing emails.

The paper explores the development and evaluation of Enhanced Grasshopper Optimization Algorithms (EGOAs) for global numerical optimization problems and their application to spam email detection. The study compares the proposed algorithms against the standard Grasshopper Optimization Algorithm (GOA) and other traditional optimization algorithms, as well as the performance of EGOAs in training Multilayer Perceptrons (MLPs) for spam detection[8]. The authors reference existing literature on the application of nature-inspired algorithms in optimization problems, highlighting the evolutionary nature of algorithmic development. The paper also highlights the importance of integrating optimization algorithms with machine learning, particularly the training of Multilayer Perceptrons (MLPs), for practical applications. The paper positions spam email detection as a classic problem in cybersecurity and information retrieval, and the authors introduce metaheuristic algorithms as a promising approach to improving the efficiency and adaptability of spam detection systems.The paper also discusses ensemble methods and hybrid models, emphasizing the need for models that balance exploration and exploitation to overcome challenges in local optima and stagnation. The review concludes by identifying a research gap in the existing body of work, requiring a comprehensive evaluation of the proposed EGOAs compared to both the standard GOA and other traditional optimization algorithms.

The research article "Machine Learning-Based Detection of Spam Emails" by Zeeshan Bin Siddique et al. (2021) explores the issue of spam email detection in the Urdu language[9]. The authors highlight the growing prominence of Urdu in social media, websites, and emails, and the challenges posed by spam emails, including phishing URLs, advertisements, commercial segments, and indiscriminate distribution. Despite advancements in spam filtering applications, distinguishing between legitimate and malicious emails remains a complex task. The study proposes using Naive Bayes, Convolutional Neural Network (CNN), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) which are examples of machine learning and deep learning models, to detect and categorize Urdu-scripted spam emails. The authors highlight the lack of a dedicated dataset for Urdu spam emails and the uniqueness of their study using a dataset based on the Urdu script. The proposed methodology includes tokenization, stop word removal, stemming, and feature extraction for data preprocessing. The study positions the proposed study as a novel contribution, addressing the specific linguistic and technological nuances of spam detection in the Urdu language, leveraging both traditional ML and advanced DL models for enhanced accuracy and effectiveness.

The paper "Email Spam Filtering Using Machine Learning Based XGBoost Classifier Method" by Anitha, P U Rao, C V Guru Babu, D Suresh provides an in-depth analysis of the evolution and current landscape of email spam filtering[10]. It highlights the need for adaptability and robustness in distinguishing spam from legitimate communications, as well as the challenges of false positives and false negatives inherent in conventional approaches. The review highlights the transformative journey of machine learning applications in spam filtering, highlighting the strengths and limitations of different approaches. Ensemble methods, feature engineering, and labeled datasets are crucial components in the efficacy of machine learning models. The narrative emphasizes the need for models that can adapt to the dynamic and sophisticated strategies employed by spammers, highlighting the limitations of static filtering mechanisms. The XGBoost classifier is highlighted as a technological advancement and a strategic choice in overcoming the shortcomings of previous models. The review positions the XGBoost classifier as a promising solution, showcasing its adaptability to the dynamic nature of spam content. The algorithm's ability to discern patterns, adapt to changing spam tactics, and deliver superior performance metrics is discussed. In summary, the paper provides a comprehensive overview of email spam filtering, tracing its roots from early rule-based systems to more recent heuristic approaches. It also positions the XGBoost classifier as a significant contribution to the ongoing discourse on enhancing email security through advanced machine learning techniques.

The 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) presented a research paper on phishing in emails, a growing concern for cybercriminals[11]. Emails have become integral to personal and professional interactions, making them vulnerable to phishing attacks. The paper uses machine learning techniques to detect phishing emails as a classification problem, using popular algorithms like Logistic Model Trees (LMT), Multilayer Perceptron (MLP), and Decision Trees (J48, C4.5). The Weka environment, an open-source platform for data analysis, facilitates the application of these algorithms. The methodology involves a systematic process, starting with a dataset from the Phish Tank archive, which includes 11,000 instances and 2,500 identified phishing URLs. The Weka environment is used for algorithm selection, with Principal Component Analysis (PCA) for data preprocessing and analysis. The chosen algorithms are subjected to rigorous testing and evaluation for accuracy, sensitivity, specificity, and other metrics. The results show that LMT achieves a maximum accuracy of 96.9245%, positioning it as a promising method for email classification. The paper provides valuable information about the application of machine learning models for detecting phishing emails, providing a structured methodology and empirical results.

The paper "Feature Selection by Multiobjective Optimization: Application to Spam Detection System by Neural Networks and Grasshopper Optimization Algorithm" by Sanaa A. A. Ghaleb et al., published in IEEE, provides a comprehensive literature review on the challenges of spam detection in electronic communication. The authors highlight the ongoing battle between spammers and detection tools, emphasizing the need for robust protective measures and efficient feature selection[12]. The review highlights the importance of feature selection in spam classifiers, as it is crucial for accurate discrimination between legitimate and spam emails. The authors categorise feature selection algorithms into filters and wrappers, with wrappers being preferred due to their superior performance. The authors introduce the concept of multi-objective optimization algorithms and propose a wrapper technique based on the feature extraction multi-objective grasshopper optimization algorithm (MOGOA). They also discuss the limitations of traditional approaches and the shift towards natural-inspired methodologies like genetic algorithms and optimization using particle swarms. The writers also discuss the pivotal role of statistical analysis, machine learning, and artificial intelligence in knowledge detection strategies. They categorise these strategies into unsupervised, supervised machine, and semi-supervised learning techniques, with supervised learning generally outperforming others. Machine learning algorithms such as Naive Bayes, Support Vector Machines, Artificial Neural Networks, and K-Nearest Neighbour are identified as prevalent tools in spam detection. In summary, the literature review provides a solid foundation for the authors' proposed methodology by exploring the nuances of spam detection systems, highlighting the limitations of existing approaches, and underscoring the potential of innovative multi-objective optimization algorithms.

The research paper "Efficient Spam Filtering Through Intelligent Text Modification Detection Using Machine Learning" by N. Mageshkumar, A. Vijayaraj, N. Arunpriya, and A. Sangeetha explores the issue of spam emails and proposes a novel approach to enhance the accuracy of spam filters, specifically targeting text modifications[13]. The authors emphasize the growing significance of email communication despite the advent of social networks and the urgency to improve spam filters. They introduce a unique method aimed at improving Naive Bayes' accuracy by detecting and categorizing text alterations, contributing to more effective spam filtering. The proposed Python approach incorporates semantic, keyword, and machine learning algorithms, demonstrating improved accuracy compared to the established Spamassassin. The paper also explores the correlation between email length and spam score, suggesting the occurrence of Bayesian Poisoning, a controversial concept in the realm of spam emails. The literature review comprehensively surveys existing spam detection methods, evaluating the limitations of rule-based spam filtration systems and emphasizing the need for flexible strategies capable of responding to evolving spamming methods. The proposed framework's architecture includes Naive Bayes classifiers, Multinomial Naive Bayes optimization, and graphical analyses to depict the accuracy of the spam filtering system. The paper advocates for the broader implementation of the proposed method across different spam filter devices, suggesting its potential to evolve into a sophisticated spam detector for various textual alterations.

The paper "Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach" presents a unique perspective on spam email classification, focusing on the shortcomings of existing models in adapting to evolving spam techniques[14]. Traditional spam detection models have primarily focused on binary classification, but these models often overlook recent spammer tricks, such as image-based spam and hidden text or salting. The paper highlights the importance of adapting spam detection models to the evolving landscape of spam techniques, as current models often train on datasets from the early 2000s. The paper introduces a topic-based approach for multi-classification of spam emails, presenting two new datasets, SPEMC-15K-E and SPEMC-15K-S, for English and Spanish spam emails, respectively. These datasets are created using agglomerative hierarchical clustering and propose a classification pipeline that employs various text representation techniques and classifiers. The paper's unique contribution addresses gaps in existing research, such as the need to adapt to evolving spam techniques, address image-based spam and hidden text, and employ a topic-based approach for multi-classification.

The authors address the prevalent problem of cybersecurity threats brought on by spam and phishing emails in electronic communication in the research described above [15]. By concurrently classifying spam and phishing messages, the study aims to improve the efficacy of identifying harmful emails. The email body or the content were frequently the focus of earlier works when choosing features, but this research introduces a revolutionary dual-layer architecture. This architecture provides a more thorough method by including features from the email body as well as the content during model training. Several deep learning techniques include convolutional neural networks (CNN), artificial neural networks (ANN), and recurrent neural networks (RNN) which are used by the dual-layer design. The first layer categorizes phishing emails, and the second layer categorises spam emails.In order to effectively handle minority classes, the research addresses the prevalent issue of data imbalance in the classification of email phishing and spam. The suggested approach performs exceptionally well, as evidenced by experimental evaluations, which reveal excellent values for accuracy of 99.51%, recall of 99.68%, precision of 99.5%, and F1-score of 99.52%. These figures highlight the dual-layer architecture's great effectiveness in accurately identifying and categorising malicious emails. The paper's contributions include the development of a novel architecture, the successful resolving of data imbalance problems, and the proof that taking into account features from both the email body and content when training a model improves performance metrics. In conclusion, the research article proposes a complex dual-layer architecture that uses deep learning algorithms to classify spam and phishing emails. The findings point to a possible direction for enhancing email communication system security against cyber-attacks.

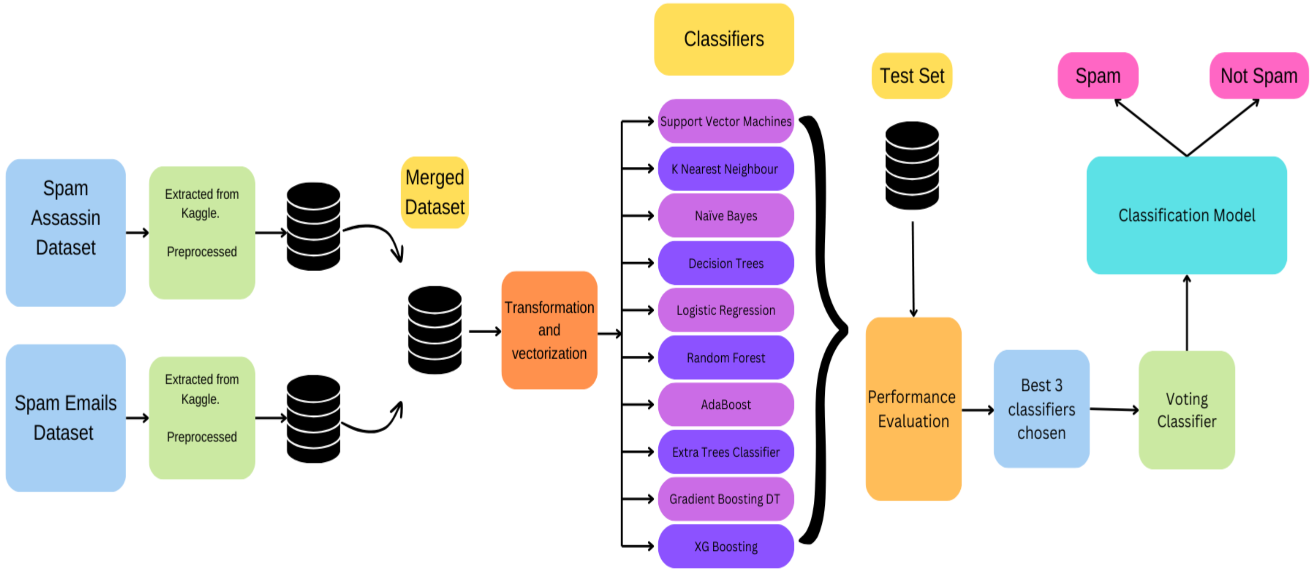
IMPLEMENTATION

This email spam classification application is implemented by extracting the voting classifier model consisting of support vector machines, extra trees classifiers and multinomial naive bayes through the pickle module which is then used by the application for classifying the email as spam or legitimate. The datasets were preprocessed then combined for the exploratory data analysis and model training. The voting classifier and the machine learning models are trained in jupyter notebook. After hyperparameter tuning of the classifiers, the best models were chosen for the voting classifiers. The final model is hosted on streamlit with a very user-friendly interface for ease of use.

The python libraries used in this project are:

1. **Pandas:** A library for data cleaning and analysis and is majorly used for transforming the data and visualising the dataset.
2. **NLTK:** This is a library which is used for analysing natural language for gaining insights from the data as a text.
3. **NumPy:** It is a library for array and matrix calculations along with computing high level mathematical functions.
4. **Matplotlib:** A library for visualising the dataset using graphs, charts, etc. It visualises the data for better understanding in similar ways to matlab.
5. **Scikit-learn:** It is a special library for machine learning which has a lot of tools for machine learning algorithms and can be used to make machine learning tasks like classification, clustering and regression simple.

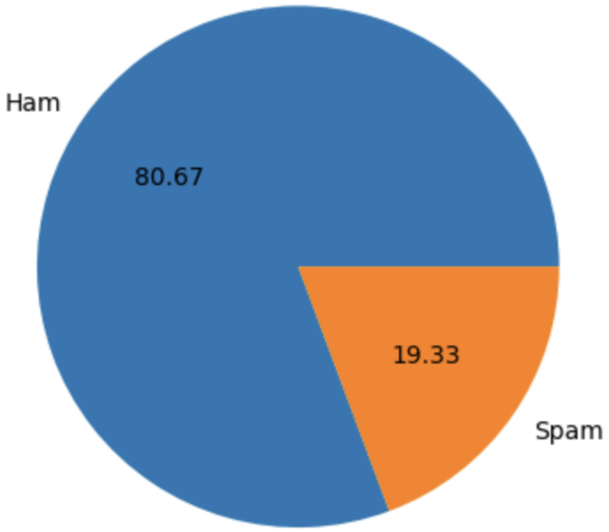
ARCHITECTURE DIAGRAM



ALGORITHM

Dataset Collection: The datasets were exported from kaggle. Two different datasets SpamAssasin containing 6072 spam and ham emails and Spam Emails from UCL containing 5572 spam and ham emails were chosen as they had a good variety of emails and very widely regarded as the better datasets through the literature reviews.

Data Preprocessing: The datasets were then cleaned and preprocessed by dropping null values and dropping some unnecessary columns to make the dataset more relevant. The datasets were then merged into a singular complete dataset which would be used for further stages in our project. After preprocessing of the new dataset, the complete dataset resulted in 10451 emails out of which 8431 were spam and 2020 ham emails.



The emails are then transformed for the model to accept as input by stemming and removing punctuations and stopwords for efficient execution. Then the email text's frequency in the text was computed using Term Frequency-Inverse Document Frequency(TFIDF). Using this the email transformed text was converted into an array for model prediction.

Exploratory Data Analysis: The dataset dataframe was then analysed on and many data trends were found. The distribution of words and sentences, the most common words in spam and ham emails were then found. The correlation between these columns were then found. These were done using the help of NLTK, a library specialised in NLP related tasks.

Model Selection and Training: The data set was split into training and test data in 80:20 ratio which is a good ratio.Ten different Supervised machine learning algorithms were used to classify the emails as spam or ham.

The models include:

1. **Support Vector Machines(SVMs):** This algorithm uses a hyperplane to classify different classes by separating it in a multi dimensional graph.
2. **K nearest neighbours:** This algorithm classifies the data as different classes using the majority class of its n nearest neighbours.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

3. **Naive Bayes Classification:** This algorithm uses a probabilistic approach for classifying the class and is usually most suitable for textual dataset.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

4. **Decision Trees:** This algorithm splits the data into different classes based on best splitting attributes and then uses this to classify test data.

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2$$

5. **Logistic Regression:** This algorithm uses a probabilistic approach along with linearly separable data for classifying the data.

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

6. **Random Forests:** This algorithm uses ensemble approach to create multiple decision trees and combines their prediction for better outcome.
7. **AdaBoost Classification:** This algorithm uses a boosting technique to combine weak learner models to make an accurate classifier model.

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

8. **Extra Trees Classification:** This algorithm also uses ensemble approach similar to random forests but the selection is highly random.
9. **Gradient Boosting:** This algorithm is used to build trees iteratively outperforming the previous tree. This uses Gradient Descent to improve its model.

$$y_i^p = y_i^p + \alpha * \delta \sum (y_i - y_i^p)^2 / \delta y_i^p$$

which becomes, $y_i^p = y_i^p - \alpha * 2 * \sum (y_i - y_i^p)$

where, α is learning rate and $\sum (y_i - y_i^p)$ is sum of residuals

10. **XGBoost Classification:** This algorithm uses an extreme version of Gradient Boosting to improve the results of the model.

Real value (label) known from the training data-set

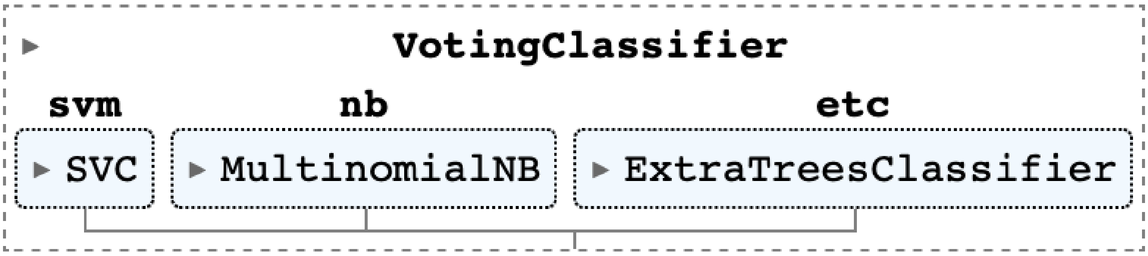
$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(\overset{\Uparrow}{y_i}, \underset{\Downarrow}{\hat{y}_i{}^{(t-1)} + f_t(\mathbf{x}_i)}) + \Omega(f_t)$$

Can be seen as f(x + Δx) where x = $\hat{y}_i{}^{(t-1)}$

All these classifiers were used to compare accuracy and precision score to find the best model. As Multinomial Naive Bayes had the highest precision and Support Vector Machines had the highest accuracy and Extra Trees Classifier performed best overall, Ensemble learning approach was used to create a voting classifier to outperform these algorithms.

Hyperparameter Tuning: The hyperparameters in these models were adjusted to give the best results. The random state of the models were set to 2. For most of the models, the hyper parameters were only modified a little bit. These hyperparameter tuning was then evaluated using a test set.

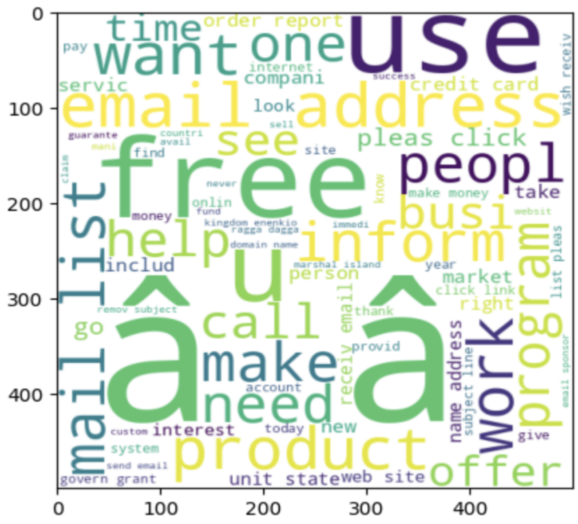
Voting Classifier Model: A voting classifier model was created using extra trees classification, support vector machine and multinomial naive bayes which is an ensemble learning approach. This Voting classifier approach of these 3 models gave a very impressive result. This model is used for further stages in our project.



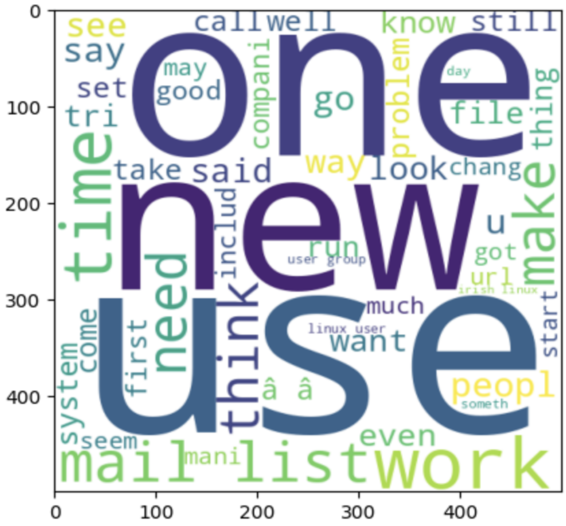
Model Testing and Evaluation: The voting classifier model is evaluated using metrics like accuracy and precision using the test data set. The main evaluation metric we have used is precision to avoid false positives in our dataset to not classify a legitimate email as spam. The voting classifier model gave an accuracy of 97.08% and a precision score of 1.0 which is a very good result. This model is then saved and used for application implementation.

Graphic User Interface development: A web based email spam classification approach is used in this project where the model is dumped into a pickle file for application usage. A simple user-friendly graphic interface was created with a text area input for the email body, and a predict button to run the model. This test data will be preprocessed and tested on the model and the results are displayed to the user along with the probability of the email being spam.

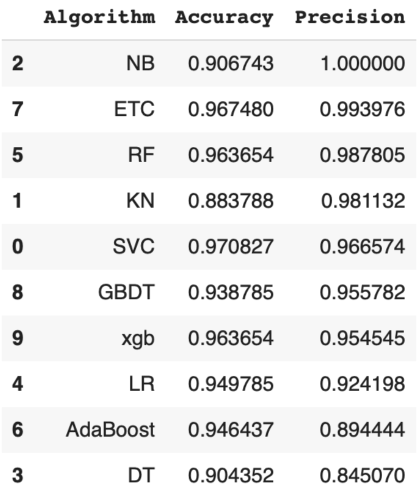
RESULTS AND DISCUSSION



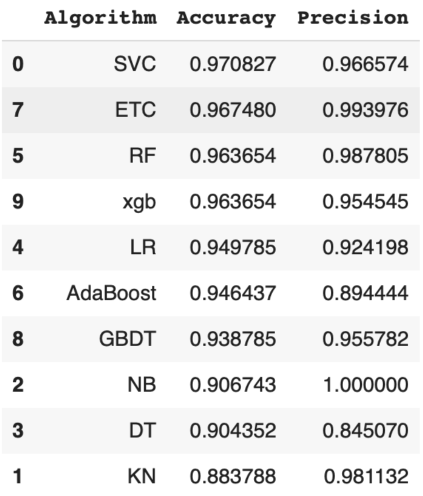
[Figure 1: Word Cloud of the spam emails]



[Figure 2: Word Cloud of the ham emails]

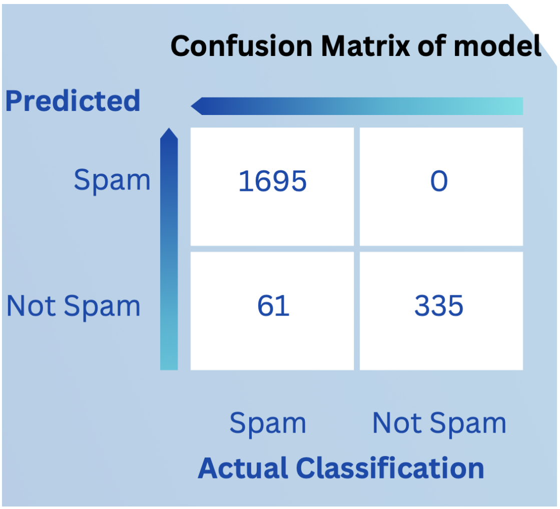


	Algorithm	Accuracy	Precision
2	NB	0.906743	1.000000
7	ETC	0.967480	0.993976
5	RF	0.963654	0.987805
1	KN	0.883788	0.981132
0	SVC	0.970827	0.966574
8	GBDT	0.938785	0.955782
9	xgb	0.963654	0.954545
4	LR	0.949785	0.924198
6	AdaBoost	0.946437	0.894444
3	DT	0.904352	0.845070



	Algorithm	Accuracy	Precision
0	SVC	0.970827	0.966574
7	ETC	0.967480	0.993976
5	RF	0.963654	0.987805
9	xgb	0.963654	0.954545
4	LR	0.949785	0.924198
6	AdaBoost	0.946437	0.894444
8	GBDT	0.938785	0.955782
2	NB	0.906743	1.000000
3	DT	0.904352	0.845070
1	KN	0.883788	0.981132

[Figure 3 and 4 shows the accuracy and precision of the algorithms sorted by max precision and max accuracy respectively]



Confusion Matrix of model

Predicted	Spam	Not Spam
Spam	1695	0
Not Spam	61	335

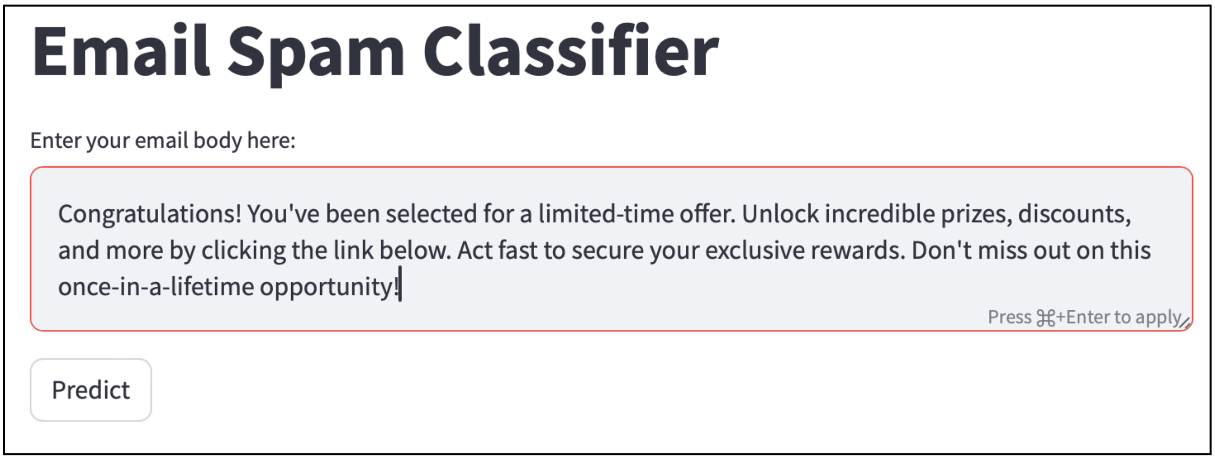
Accuracy: 97.08%

Precision: 1.00

[Figure 5 shows the confusion matrix of the voting classifier]

We can see the confusion matrix and can find out the accuracy and precision. The results of this model is very good with no no false positive detection.

SCREENSHOTS OF APPLICATION GUI

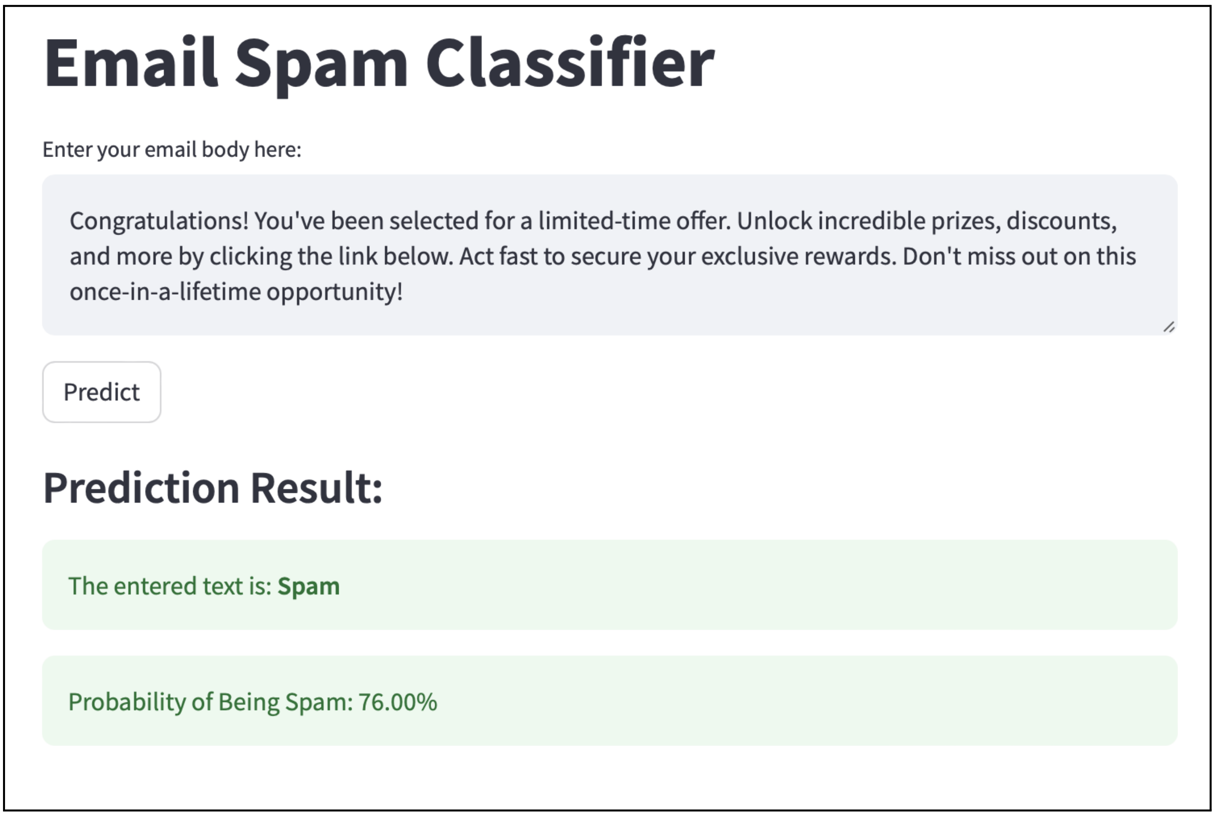


Email Spam Classifier

Enter your email body here:

Congratulations! You've been selected for a limited-time offer. Unlock incredible prizes, discounts, and more by clicking the link below. Act fast to secure your exclusive rewards. Don't miss out on this once-in-a-lifetime opportunity!

This web application contains a text area for input body by user. The above image shows the input fields and the basic GUI design of the Email spam Classification application.



Email Spam Classifier

Enter your email body here:

Congratulations! You've been selected for a limited-time offer. Unlock incredible prizes, discounts, and more by clicking the link below. Act fast to secure your exclusive rewards. Don't miss out on this once-in-a-lifetime opportunity!

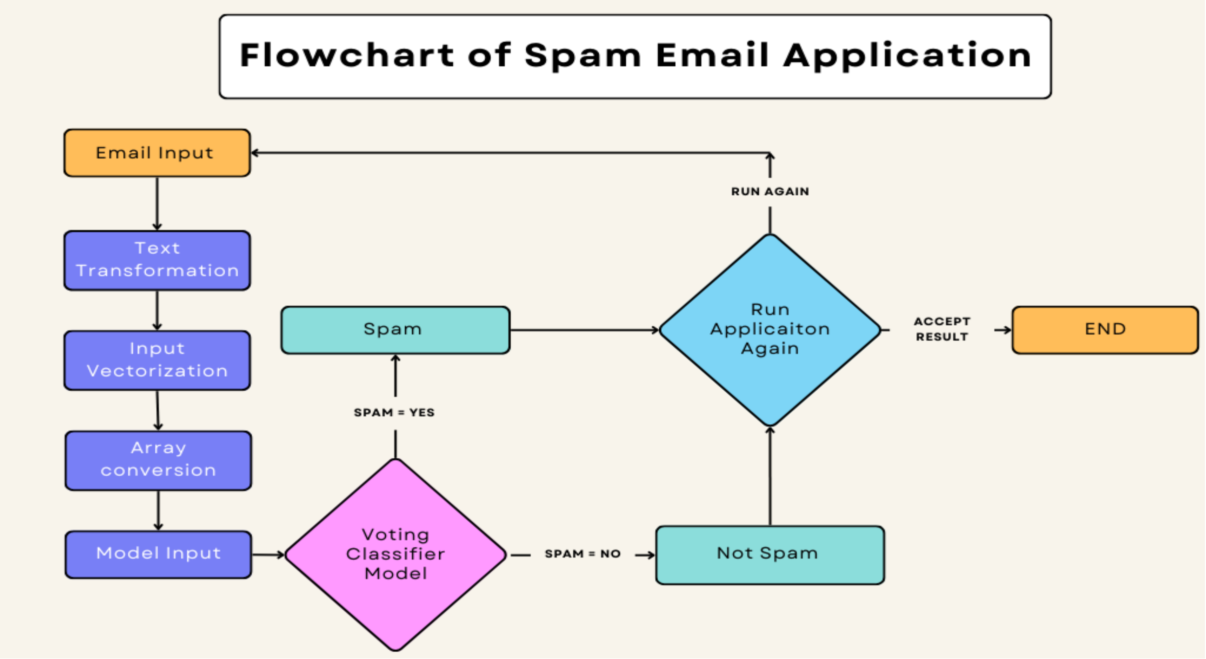
Prediction Result:

The entered text is: **Spam**

Probability of Being Spam: 76.00%

The user after entering the email body and clicking on the predict button. The web app runs the test data on the model and returns back the classification prediction. It also returns the probability of the mail being spam or legitimate. The above image shows the output of a test case.

FLOWCHART OF THE APPLICATION



CONCLUSION

Nowadays, emails are a widely used communication medium across the world. Spam mails are unwanted marketing or malicious emails that can attack an individual or a corporation by stealing confidential information. Spam poses a serious threat to security and financial risk. Hence, this paper focuses on a system that is designed to categorise unwanted emails into spam. Various machine learning classifiers are experimented on, with SVC, MultinomialNB and ExtraTreesClassifiers providing the best results. These models are then combined using a voting classifier to provide an accuracy of 97.08% and a precision score 1.0. The email spam application is hosted on https://email-spam-classification-ml-project.streamlit.app/ for implementation by other internet users. This classification can accurately classify the email as spam or not and has a very high precision. The results of this project can be improved by taking a larger body of the email. Also, categorising the spams may be done using verified domain names. The prevalent work can also be extended to other domains like e-commerce, job profile based websites where a majority of fake information exists.

REFERENCES

1. A. S. Bozkir, E. Sahin, M. Aydos, E. A. Sezer and F. Orhan, "Spam E-Mail Classification by Utilizing N-Gram Features of Hyperlink Texts," 2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT), Moscow, Russia, 2017, pp. 1-5, doi: 10.1109/ICAICT.2017.8687020.
2. Naem, A. A., Ghali, N. I., & Saleh, A. A. (2018). Antlion optimization and boosting classifier for spam email detection. Future Computing and

Informatics Journal, 3(2), 436-442. https://doi.org/10.1016/j.fcij.2018.11.006 (https://www.sciencedirect.com/science/article/pii/S2314728818300746?via%3Dihub)

3. Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802. (https://www.sciencedirect.com/science/article/pii/S2405844018353404)
4. S. Gibson, B. Issac, L. Zhang and S. M. Jacob, "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms," in *IEEE Access*, vol. 8, pp. 187914-187932, 2020, doi: 10.1109/ACCESS.2020.3030751.
5. N. Saidani, K. Adi, M.S. Allili, "A semantic-based classification approach for an enhanced spam detection," Computers & Security, Volume 94, July 2020, 101716. DOI: 10.1016/j.cose.2020.101716. (https://www.sciencedirect.com/science/article/abs/pii/S0167404820300043)
6. U. Murugavel, R. Santhi, "Detection of spam and threads identification in E-mail spam corpus using content-based text analytics method," Materials Today: Proceedings, Volume 33, Part 7, 2020, Pages 3319-3323. DOI: 10.1016/j.matpr.2020.04.742. (https://www.sciencedirect.com/science/article/abs/pii/S2214785320333903)
7. Alhogail, A. and Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. Computers & Security, Volume 110, 102414. https://doi.org/10.1016/j.cose.2021.102414 (https://www.sciencedirect.com/science/article/abs/pii/S0167404821002388?via%3Dihub)
8. S. A. A. Ghaleb, M. Mohamad, S. A. Fadzli and W. A. H. M. Ghanem, "Training Neural Networks by Enhance Grasshopper Optimization Algorithm for Spam Detection System," in *IEEE Access*, vol. 9, pp. 116768-116813, 2021, doi: 10.1109/ACCESS.2021.3105914.
9. Zeeshan Bin Siddique et al. "Machine Learning-Based Detection of Spam Emails." AI-enabled Decision Support System: Methodologies, Applications, and Advancements 2021. Research Article, Open Access, Volume 2021, Article ID 6508784. https://doi.org/10.1155/2021/6508784. (https://scholar.google.com/scholar?q=Zeeshan%20Bin%20Siddique%2C%20et%20al.%2C%20Machine%20Learning-Based%20Detection%20of%20Spam%20Emails%2C%20Scientific%20Programming%202021%20(2021).)
10. Anitha, P. U., Rao, C. V. Guru, & Babu, D. Suresh. (2021). "Email Spam Filtering Using Machine Learning Based Xgboost Classifier Method." Turkish Journal of Computer and Mathematics Education, 12(11), 2182-2190. Trabzon. doi: (https://scholar.google.com/scholar_lookup?title=Email%20spam%20filtering%20using%20machine%20learning%20based%20xgboost%20classifier%20method&publication_year=2021&author=P.U.%20Anitha&author=C.V.%20Guru%20Rao&author=D.%20Suresh%20Babu)

11. R. Abdulraheem, A. Odeh, M. Al Fayoumi and I. Keshta, "Efficient Email phishing detection using Machine learning," *2022 IEEE 12th Annual*

Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2022, pp. 0354-0358, doi: 10.1109/CCWC54503.2022.9720818.

12. S. A. A. Ghaleb *et al*., "Feature Selection by Multiobjective Optimization: Application to Spam Detection System by Neural Networks and Grasshopper Optimization Algorithm," in *IEEE Access*, vol. 10, pp. 98475-98489, 2022, doi: 10.1109/ACCESS.2022.3204593
13. Mageshkumar, N., Vijayaraj, A., Arunpriya, N., Sangeetha, A. (2022). Efficient spam filtering through intelligent text modification detection using machine learning. Materials Today: Proceedings, Volume 64, Part 1, Pages 848-858. https://doi.org/10.1016/j.matpr.2022.05.364 (https://www.sciencedirect.com/science/article/abs/pii/S2214785322036550?via%3Dihub)
14. Jáñez-Martino, F. et al. (2023). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. Applied Soft Computing, Volume 139, 110226. https://doi.org/10.1016/j.asoc.2023.110226 (https://www.sciencedirect.com/science/article/pii/S1568494623002442?via%3Dihub)
15. Doshi, J., Parmar, K., Sanghavi, R., Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. Computers & Security, Volume 118, 103378. https://doi.org/10.1016/j.cose (https://www.sciencedirect.com/science/article/abs/pii/S0167404823002882?via%3Dihub)

Email Spam Classification Using ML Technique

NO	Reference	Objective	Problem Statement	Methodology	Dataset	Algorithm	Advantage	Disadvantage	Performance measure value
1	A. S. Bozkir, E. Sahin, M. Aydos, E. A. Sezer and F. Orhan, "Spam E-Mail Classification by Utilizing N-Gram Features of Hyperlink Texts," 2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT), Moscow, Russia, 2017, pp. 1-5, doi: 10.1109/ICAICT.2017.8687020.	1. Gather dataset from COMODO. 2. Extract n-gram features and train the model using SVM Pegasos. 3. Classify the link as spam or not. 4. Refine the model for improved accuracy.	1.Advanced spam filtering is crucial for cybersecurity resilience, as it helps counteract evolving email-based threats. 2.The ever-evolving nature of spam tactics needs the implementation of adaptive email filtering systems to maintain personal security. 3.Email spam classification systems should optimize resource utilization to prevent bottlenecks and ensure filtering processes don't negatively impact overall email system performance.	The study used a large-scale dataset from COMODO Inc. to classify spam emails using hyperlink texts. N-gram indexing and machine learning methods like Naïve Bayes, SVM, and SVM Pegasos were used to achieve up to 98.75% accuracy in trigram-based bag-of-words representation.	1.COMODO Inc. provided a dataset containing 150,000 spam and 141,414 ham hyperlinks from emails collected between August and November 2016, focusing on hyperlink texts for classification.	The study utilized Naive Bayes, Support Vector Machine (SVM), and SVM Pegasos to classify spam/ham emails based on hyperlink texts using n-gram features.	1.The SVM Pegasos algorithm, particularly in trigram configuration, achieved a high accuracy of 98.75% in distinguishing spam and ham emails based on hyperlink text features. 2.The paper introduces a new method for spam/ham email classification, utilizing hyperlink texts exclusively, demonstrating creativity in feature selection compared to traditional methods.	1.The paper's effectiveness in identifying email spam beyond hyperlink texts is limited by its narrow scope, making its generalizability uncertain. 2. The paper uses COMODO Inc.'s dataset, but lacks transparency about its representativeness, diversity, and potential biases, raising concerns about its applicability to real-world scenarios.	Accuracy: SVM Pegasos Algorithm: 98.75%
2	Naem, A. A., Ghali, N. I., & Saleh, A. A. (2018). Antlion optimization and boosting classifier for spam email detection. Future Computing and Informatics Journal, 3(2), 436-442.	1.Create a ALO algorithm model for spam classification.. 2. Incorporate Booster Classifier to improve accuracy. 3. Classify the	1.Large-scale data breaches from email-based exploits highlight the need for comprehensive spam detection strategies. 2.The evolving threat landscape of spam emails necessitates adaptive filters that can effectively tackle new and intricate attack strategies.	The ALO-Boosting methodology uses antlion optimization for optimal feature selection, mimicking predatory behaviour of antlions, and then applies boosting to selected features, transforming soft learners into a powerful ensemble classifier for spam email detection.	1.The CSDMC2010 and SpamAssassin datasets are used in data mining competitions, containing 4,327 and 6,047 emails respectively, with 2,949	1.The ALO-Boosting algorithm uses Antlion Optimization (ALO) for optimal feature selection and Boosting for spam email classification based on the selected features.	1.The ALO-Boosting method achieves high accuracy rates of 99.80% on the CSDMC2010 and 98.91% on the SpamAssassin datasets, demonstrating its effectiveness in identifying spam and non-spam emails. 2. The paper introduces	1.The paper assesses the ALO-Boosting method on specific datasets, highlighting potential concerns about its generalizability across diverse datasets, email characteristics, and real-world scenarios. 2.Antlion Optimization and Boosting's	Accuracy: SpamAssassin Dataset: 98.91% CSDMC2010 Dataset: 99.8%

	https://doi.org/10.1016/j.fcij.2018.11.006	email spam as spam or not. 4. Refine the model for improved accuracy.	3.The increasing frequency of identity theft incidents necessitates the development of advanced email spam filters to safeguard personal data and prevent financial fraud.		solicited and 1,378 unsolicited emails respectively.		an efficient feature selection mechanism using Antlion Optimization (ALO), enhancing classification precision by imitating antlions' predatory behaviour.	complexity may enhance precision but also pose challenges in computational resources, implementation, and user understanding for practitioners without deep understanding.	
3	Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802.	1. Gather Dataset from any of the source mentioned 2. Train a model using any of the algorithms given. 3. Classify email as spam or not 4. Refine the model for better accuracy	1.For email filters to combat changing spam techniques, robust models are essential. 2.By lowering spam-related hazards worldwide, spam filters are a contribution to global cyber hygiene. 3. Spam filtering eliminates unnecessary and potentially disruptive emails, reducing email distraction.	The study aimed to address email spam filtering challenges by utilising machine learning algorithms like Neural Networks, Support Vector Machines, and Naïve Bayes, evaluating their precision and recall, and suggesting future research directions based on the findings.	1.The dataset includes various spam archives, Spambase, Lingspam, PU1, Spamassassin, PU2, PU3, PUA, Zh1, Gen Spam, Trec 2005, Biggio, Phishing Corpus, Enron-Spam, Trec 2006, Trec 2007, Princeton Spam Image Benchmark, and Hunter.	1.Utilize various techniques like clustering, Naïve Bayes, Neural Networks, Firefly Algorithm, Rough Set Classifiers, SVM, Decision Trees, and Ensemble Classifiers for spam email classification.	1.The paper identifies research gaps and issues in email spam filtering, guiding future research and encouraging the exploration of innovative solutions to address existing challenges. 2.The paper provides a thorough analysis of machine learning-based email spam filtering methods, detailing their key concepts, historical advancements, and techniques used in this field.	1.The paper reviews email spam filtering from 2004-2018, but its relevance may be limited due to the rapid advancements in technology and machine learning. 2.The paper may overemphasize machine learning techniques, neglecting other hybrid systems combining rule-based methods, requiring a balanced approach for effective spam filtering mechanisms.	NIL
4	S. Gibson, B. Issac, L. Zhang and S. M. Jacob, "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms," in *IEEE Access*, vol. 8, pp.	1.Gather a variety of datasets to classify spam. 2.Create an enhanced grey wolf optimization algorithm (EGOA) optimized Multi Layer Perceptron (MLP) model. 3.Determine whether an email is spam.	1.Email spam classification is crucial for removing unwanted and potentially harmful messages so they don't end up in users' inboxes and cause unneeded disruption. 2. Effective email spam classification is essential for spotting phishing scams and dangerous content, shielding consumers from	The study uses a literature review, seven email datasets, and machine learning models like Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree, and Multi-Layer Perceptron. Bio-inspired algorithms like Particle Swarm Optimization and Genetic Algorithm	1.The paper discusses the use of Ling-Spam, Enron, PUA, and SpamAssassin datasets for spam email classification, incorporating individual emails as text	1.The study uses machine learning algorithms like Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree, and Multi-Layer Perceptron, along with bio-inspired optimization techniques like	1.Through performance optimization, bio-inspired algorithms such as Particle Swarm Optimization and Genetic Algorithm enhance the efficacy and precision of machine learning models utilized for spam identification. 2.The inclusion of a	1.Unique numerical representations in datasets like PUA and PU3 may hinder model interpretability and feature extraction, making it difficult to comprehend decision-making processes. 2. Bio-inspired optimization algorithms	Accuracy: SpamAssasin Dataset: 99.35% Enron Spam Dataset: 99.12% Ling Spam Dataset: 97.82% PU3: 97.87%

	187914-187932, 2020, doi: 10.1109/ACCESS.2020.3030751.	4.Adjust the model's parameters to increase spam detection accuracy.	security risks and unwanted access to private data.	are integrated for improved classifier performance. Experiments, parameter tuning, and results analysis are conducted, followed by a Python-based spam detection framework.	files with varying content and characteristics.	Particle Swarm Optimization and Genetic Algorithm.	variety of datasets, including Enron and Ling-Spam, enables a thorough assessment and guarantees the model's resilience to changes in email content and attributes.	may cause computational overhead, affecting the scalability and speed of the spam detection model, depending on the size and complexity of the datasets.	
5	N. Saidani, K. Adi, M.S. Allili, "A semantic-based classification approach for an enhanced spam detection," Computers & Security, Volume 94, July 2020, 101716. DOI: 10.1016/j.cose.2020.101716.	1. Create a semantic-based method for spam detection.. 2. Classify the email as spam or not using the model using categories. 3. Refine the model for improved accuracy.	1.The increasing frequency of identity theft incidents needs the development of advanced email spam filters to safeguard personal data and prevent financial fraud. 2. The rise in remote work has heightened the necessity of email filters to safeguard business communications and confidential corporate data. 3. The increasing risk of falling victim to financial scams through deceptive spam emails underscores the need for improved filtering.	The methodology involves email categorization by domain using various algorithms, followed by domain-specific semantic feature extraction through manual rule creation and automatic rule induction using the CN2-SD algorithm. The classifiers are trained, evaluated, and tested on datasets, demonstrating improved spam detection through semantic features.	1.The paper evaluates the approach on the CSDMC2010 SPAM dataset, consisting of 4327 labelled emails across various categories, using a test dataset from Enron, Ling-spam, and specialised forums.	The paper presents a two-stage email spam detection algorithm, categorising emails by subject domains and building domain-specific semantic features using Naive Bayes and SVM classification algorithms.	1.The paper introduces a domain-specific approach for spam detection, enhancing the precision of spam identification in specific contexts. 2. The approach enhances spam detection by combining manually-specified rules with automatically generated semantic features, leveraging expert knowledge and data-driven insights.	1. The approach's reliance on rule-based semantic features may hinder its adaptability to evolving spam tactics and may struggle to detect previously unseen patterns. 2. The paper's discussion on the generalisation of the proposed approach to diverse datasets may restrict its applicability to email corpora beyond the training data's scope.	Naive Bayes: Accuracy: 98.92% Recall: 98.83% Precision: 98.97% F measure: 98.9% Decision Tree: Accuracy: 97.67% Recall: 97.66% Precision: 97.92% F measure: 97.79%
6	U. Murugavel, R. Santhi, "Detection of spam and threads identification in E-mail spam corpus using content-based text analytics method," Materials Today:	1.Create a MSSCA algorithm model for spam thread classification.. 2. Classify the email spam categories. 3. Refine the model for improved	1.Email exploitation for malware spread necessitates advanced filters that can identify and neutralize potential threats in messages. 2.Email communication channels must combat the rise in deceptive messages to ensure secure digital interactions and maintain	The methodology uses MATLAB to tokenize and eliminate stop words from a large email dataset, then uses the Multi-Split Spam Corpus Algorithm to filter and classify emails into spam and ham, identifying spam threads for improved performance.	1. The paper explores the spam corpus dataset with 1040 instances with 15 distinct email attributes.	The Multi-Split Spam Corpus Algorithm (MSSCA) is a tool that enhances spam corpus performance by preprocessing, filtering, classifying email datasets, identifying spam	1.The Multi-Split Spam Corpus Algorithm (MSSCA) enhances spam detection by identifying and categorising spam threads, thereby improving filtering and reducing unwanted emails. 2. Content-based text	1.The paper lacks transparency about the experimentation dataset, potentially limiting reproducibility and making it difficult to assess the algorithm's performance in diverse settings. 2. The text lacks specific evaluation metrics for	High

	Proceedings, Volume 33, Part 7, 2020, Pages 3319-3323. DOI: 10.1016/j.matpr.2020.04.742.	accuracy.	trust. 3.The increasing issue of spam underscores the significance of dependable email experiences, which can be achieved through advanced filters.			threads, and extracting frequent occurrences.	analytics enhances spam detection by categorising spam threads, identifying frequent spam keywords, and distinguishing between legitimate and spam emails.	the algorithm's performance, making it challenging to assess its effectiveness and its real-world applicability.	
7	Alhogail, A. and Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. Computers & Security, Volume 110, 102414. https://doi.org/10.1016/j.cose.2021.102414	1.Gather dataset for spam classification. 2.Create a phishing email classifier based on machine learning that improves accuracy by utilizing graph convolutional networks (GCN) and natural language processing (NLP) on email body text. 3. Classify email as spam or not.	1.Digital security is becoming increasingly threatened by the rise in phishing assaults. 2.Strong spam and phishing email filtering systems are required to protect both personal and corporate accounts from potential cyber dangers. 3.Spam categorization that prioritizes user privacy balances effectiveness and security, safeguarding private information during classification.	The study proposes a three-phase method for constructing a phishing detection classifier using deep learning GCN algorithms. The first phase involves data pre-processing, which involves cleaning and preparing email data. The second phase involves graph construction, which transforms the pre-processed dataset into a large graph representing words and emails. The final phase involves the GCN classifier, which is trained on a publicly available fraud dataset.	1.This study uses a curated dataset of 8,579 emails, including 3,685 phishing and 4,894 legitimate emails, from the fraud dataset to train and evaluate phishing detection models.	1.The research uses a Graph Convolutional Network (GCN) algorithm for phishing email detection, enhancing accuracy through data pre-processing, graph construction, and GCN classifier.	1.The proposed GCN-based phishing detection model outperforms multiple existing techniques with high accuracy, precision, and recall. 2.The effectiveness of the model comes from the removal of the necessity for manual feature extraction and domain expert participation.	1.The computing resource needs of the model could be high during training. 2.Evaluation measures are mostly contrasted with similar works, and the selection of standards for comparison may be biased.	Accuracy: 98.2% Recall: 98.3% Precision: 98.5% F measure: 98.5%
8	S. A. A. Ghaleb, M. Mohamad, S. A. Fadzli and W. A. H. M. Ghanem, "Training Neural Networks by Enhance Grasshopper Optimization	1.Gather dataset for spam classification. 2.Create a MLP model for EGOA algorithm. 3. Classify the email as spam or	1. Adaptive spam nature leads to inaccurate detection, increasing false positives/negatives. 2.Traditional MLP model training has difficulties with overfitting in spam detection, local minima, and sluggish convergence.	This study develops a Spam Detection System (SDS) using an Artificial Neural Network (ANN) and Enhanced Grasshopper Optimization Algorithm (EGOA). The SDS consists of three modules: SD, ANN, and	1.The SpamBase dataset, sourced from the University of California at Irvine, has 57 features and class labels of 39% spam and	1.The EGOA is utilized to train a Multilayer Perceptron model for spam detection, initializing the grasshopper population and processing the	1.The EGOA-MLP-SDS model, using EGOA and MLP, improves spam detection accuracy, enhancing the effectiveness and reliability of a spam filtering system. 2. The model effectively	1. The suggested model's scalability and adaptation to huge and diverse datasets have not been fully explored, potentially impacting its real-world applicability. 2. An in-depth examination of the	SpamBase: Classification Accuracy: 96.9% Detection Rate: 97.2% FAR: 0.037 SpamAssasin: Classification Accuracy: 98.1%

	Algorithm for Spam Detection System," in *IEEE Access*, vol. 9, pp. 116768-116813, 2021, doi: 10.1109/ACCESS.2021.3105914.	not. 4. Refine the model for better accuracy.	3. Spam classification boosts productivity by removing irrelevant emails for focused communication.	EGOA. EGOA optimizes the ANN's structure and weights, aiming for superior classification accuracy and robustness in identifying spam emails.	61% non-spam. The SpamAssassin dataset, sourced from Kaggle, has 31.4% spam and 68.6% non-spam. The UK-2011 Webspam dataset has 53% spam and 47% non-spam.	training dataset	identifies spam emails, outperforming benchmark datasets. Its use of EGOA and MLP enhances its generalization and adaptability across various contexts.	EGOA-MLP-SDS model's computational complexity and resource requirements, which is essential for practical implementation concerns, might be lacking in this article.	Detection Rate: 97.8% FAR: 0.012 WebSpam: Classification Accuracy: 95.6% Detection Rate: 96.6% FAR: 0.053
9	Zeeshan Bin Siddique et al. "Machine Learning-Based Detection of Spam Emails." AI-enabled Decision Support System: Methodologies, Applications, and Advancements 2021. Research Article, Open Access, Volume 2021, Article ID 6508784. https://doi.org/10.1155/2021/6508784.	1. Gather dataset from kaggle. 2. Convert it to urdu using Googletrans. 3. Train model using the dataset. 4. Classify email as spam or not.	1.Prioritizing legitimate messages in spam filters reduces email overload and boosts communication effectiveness. 2.Effective spam filters reduce inbox congestion, resulting in streamlined communication and a better user experience. 3.Phishing assaults can be avoided by identifying and blocking phishing emails through spam classification.	The study uses machine learning algorithms and deep learning to detect and categorize spam emails in Urdu. The dataset is created by translating English emails into Urdu and manually correcting translations. Data preprocessing steps include tokenization, stop word removal, stemming, and feature extraction.	1. The dataset includes 5000 spam and ham emails collected from kaggle and translated to urdu using Googletrans.	1.The research paper employs various machine learning algorithms such as Naive Bayes, Support Vector Machine, Convolutional Neural Network, and Long Short-Term Memory.	1. The study achieves a high accuracy of 98.4% with the Long Short-Term Memory (LSTM) model. 2. It utilizes a unique dataset using urdu, expanding the scope of email spam classification to other languages.	1. The LSTM model has a longer training time and may use high computations. 2. There is limited transparency regarding the source of the dataset as it is not specifically mentioned.	Accuracy: LSTM: 98.4% Naive Bayes: 98% CNN: 96.2% SVM: 97.5%

10	Anitha, P. U., Rao, C. V. Guru, & Babu, D. Suresh. (2021). "Email Spam Filtering Using Machine Learning Based Xgboost Classifier Method." Turkish Journal of Computer and Mathematics Education, 12(11), 2182-2190. Trabzon. doi: https://www.proquest.com/openview/9b40fcf6cc3c713db6112dc54ce40e84/1?pq-origsite=gscholar&cbl=2045096.	1. Create a machine learning model utilising the XGBoost classifier. 2. Classify the email as spam or not using the model. 3. Refine the model for improved accuracy.	1.Efficient email spam filters are crucial for maintaining a secure online environment and protecting users from evolving cyber threats. 2. An efficient email spam classification system is crucial in preventing fraudulent schemes and safeguarding users from scams. 3.The growing sophistication of social engineering tactics needs intelligent spam filters to counteract manipulative email communication attempts.	The paper presents an email spam detection method using the Extreme Gradient Boosting (XGBoost) classifier. It involves word count algorithm feature extraction, XGBoost training, spam probability detection, and email classification. The approach is evaluated on a comprehensive dataset, proving its effectiveness in enhancing spam detection accuracy.	1.The paper "spam_ham_dataset" dataset, obtained from Kaggle, contains 5,172 records. The dataset categorises emails as spam or not spam, representing undesirable business emails. The experimental work divides the dataset into a 70% training set and a 30% testing set for comprehensive evaluation.	The paper uses the XGBoost algorithm for email spam classification, implementing file selection, word count, training, detecting spam probability, testing, and classifying emails based on probability.	1.The proposed method significantly enhances spam detection accuracy, achieving 95%, proving to be a reliable and robust approach that outperforms other classifiers in comparative analysis. 2. XGBoost classifier is a popular machine learning tool due to its scalability, efficiency, and high accuracy, making it a valuable choice for large datasets.	1.The paper lacks information on the method's generalizability across different datasets, necessitating a performance assessment to ensure its applicability in real-world scenarios. 2. The paper discusses optimised hyperparameters but lacks transparency about the tuning process or chosen values, potentially affecting reproducibility and understanding of the model's configuration.	XGBoost algorithm: Accuracy: 95% Specificity: 97.73% Precision: 96.47% F1-score: 96.03% Sensitivity: 95.59%
11	R. Abdulraheem, A. Odeh, M. Al Fayoumi and I. Keshta, "Efficient Email phishing detection using Machine learning," *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2022, pp. 0354-0358, doi:	1.Collect dataset for spam emails. 2.Classify emails as either phishing or not using machine learning methods 3. Reduce the processing time to make it efficient.	1.The amount of phishing attacks has been on the rise. 2. It is necessary to filter spam and phishing emails to secure your personal and organisational accounts. 3.Rapidly identifying and fighting zero-day spam threats is crucial for proactive email protection.	The paper presents a method for detecting email phishing using machine learning techniques, focusing on Logistic Model Tree classifiers. It uses Weka environment, Principal Component Analysis, and various algorithms like Multilayer Perceptron, Decision Tree, and LMT. The study evaluates accuracy, sensitivity, specificity, and Kappa statistics, emphasising the importance of	1.The dataset, compiled using the Phish Tank archive, includes 11,000 phishing and non-phished websites, with 2,500 being phishing URLs, and includes 2,456 items and 30 features.	1.The model uses various machine learning algorithms including Logistic Model Tree(LMT), Multilayer Perceptron(MLP), and Decision Tree algorithms to create multiple linear regression models.	1.The paper uses machine learning algorithms like Logistic Model Tree and Multilayer Perceptron to effectively classify and identify phishing emails, thereby enhancing cybersecurity measures. 2. The study evaluates machine learning algorithms like MLP, J48, and LMT using PCA techniques, comparing their performance for detecting phishing	1.The study's generalizability may be limited due to its reliance on a specific dataset, as the effectiveness of machine learning models in real-world scenarios depends on diverse datasets. 2. The paper lacks detailed discussion on interpretability of machine learning models, potentially hindering transparency and accountability due	Recall and Precision: MLP: 0.968 J48: 0.959 LMT: 0.969 Accuracy: MLP: 96.77% J48: 95.87% LMT: 96.92%

	10.1109/CCWC54503.2022.9720818.			effective detection methods in cyber threats.			emails.	to challenges in understanding decision-making processes.	
12	S. A. A. Ghaleb *et al*., "Feature Selection by Multiobjective Optimization: Application to Spam Detection System by Neural Networks and Grasshopper Optimization Algorithm," in *IEEE Access*, vol. 10, pp. 98475-98489, 2022, doi: 10.1109/ACCESS.2022.3204593.	1.Gather dataset for spam classification. 2.Create a MLP model for EGOA algorithm. 3. Classify the email as spam or not. 4. Refine the model for better accuracy.	1.Traditional email spam detection systems face a significant challenge due to the ongoing growth of spamming strategies. 2.Conventional spam classifiers frequently rely too heavily on particular characteristics, including email content or sender information, which limits their capacity to adapt to different spam patterns. 3.Identifying deepfake material in emails neutralizes deceptive strategies and upholds security and trust.	The methodology uses an Enhanced Grasshopper Optimization Algorithm (EGOA) and a Multilayer Perceptron (MLP) model for spam detection, preprocessing the spam dataset, training the MLP with extracted features, and iteratively optimizing the structure and weights, enhancing classification accuracy and detection rate.	1.The SpamBase dataset, sourced from the University of California at Irvine, has 57 features and class labels of 39% spam and 61% non-spam. The SpamAssassin dataset, sourced from Kaggle, has 31.4% spam and 68.6% non-spam. The UK-2011 Webspam dataset has 53% spam and 47% non-spam.	1.The paper uses EGOA to train a Multilayer Perceptron model for spam detection, optimizing structure and weights through iterative iterations.	1.The model makes use of the Multilayer Perceptron (MLP) and Enhanced Grasshopper Optimization Algorithm (EGOA), combining global optimization and neural network capabilities for better performance. 2.The suggested EGOA-MLP-SDS model has favorable findings in terms of spam detection accuracy, making it a suitable contender for efficient email filtering systems.	1.The study's reliance on certain datasets may restrict the applicability of the suggested model, and it would benefit from addressing potential bias or diversity issues in the datasets. 2. A complete understanding of the applicability of the suggested EGOA-MLP-SDS model is hampered by the paper's lack of detailed explanation of any potential restrictions or difficulties with it.	SpamBase: Classification Accuracy: 97.5% Detection Rate: 98.1% FAR: 0.033 SpamAssasin: Classification Accuracy: 98.3% Detection Rate: 98.3% FAR: 0.018 WebSpam: Classification Accuracy: 96.4% Detection Rate: 97.2% FAR: 0.043
13	Mageshkumar, N., Vijayaraj, A., Arunpriya, N., Sangeetha, A. (2022). Efficient spam filtering through intelligent text modification detection using machine learning. Materials Today:	1. Create a python model with leetspeak and diacritics. 2. Classify the email as spam or not using the model. 3. Refine the model for improved	1.Emerging spam techniques, particularly zero-day attacks, underscore the need for proactive spam filters that can swiftly identify and mitigate new threats. 2.Email spam requires innovative solutions to address evolving tactics, privacy concerns, and false positives to ensure effective	The paper presents a Python-based approach to improve the Naive Bayes Spam Filter's accuracy by incorporating semantic, keyword, and machine learning algorithms. It addresses text modifications like leetspeak and diacritics, and uses real-time testing to compare	1. There is no specific dataset as it runs on a keyword semantic approach but it was tested on Spamassasin.	The paper employs a Python-based method that integrates semantic analysis, keyword processing, and machine learning algorithms like Naive Bayes, advanced classifiers, and a	1.The paper introduces a Python-based method for spam detection, enhancing accuracy by addressing text modifications like leetspeak and diacritics. 2. A novel algorithm and advanced classifiers enhance the accuracy of the Naive Bayes Spam Filter, demonstrating	1.The paper discusses the use of semantic, keyword, and machine learning algorithms, but lacks specific details, limiting the reproducibility and understanding of the proposed methodology. 2. The paper lacks specific experimental results, such as	High

	Proceedings, Volume 64, Part 1, Pages 848-858. https://doi.org/10.1016/j.matpr.2022.05.364	accuracy.	filtering. 3.The project involves the creation of systems that can swiftly respond to emerging spam threats.	performance with Spamassassin. The framework's effectiveness is evaluated using graphical analyses.		novel text modification detection algorithm.	innovation in email classification.	accuracy or comparison with existing methods, which could hinder the assessment of the proposed approach's practical effectiveness.	
14	Jáñez-Martino, F. et al. (2023). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. Applied Soft Computing, Volume 139, 110226. https://doi.org/10.1016/j.asoc.2023.110226	1.Gather dataset for spam emails by agglomerative hierarchical clustering, classified into cybersecurity-based groups. 2.Classify emails as either phishing or not using machine learning methods 3. Refine the model for improved accuracy.	1.Digital security is under increasing threat as a result of the rise in phishing assaults. 2. Protecting individual and corporate accounts from potential cyber threats and compromises requires effective spam and phishing email screening. 3.Multilingual spam detection is required for inclusive filters to reduce false positives in a variety of linguistic contexts.	The study uses a comprehensive two-stage approach to improve spam email classification. It analyzes headers, bodies, and uses advanced techniques to identify hidden text and image-based spam. Machine learning algorithms are used to classify emails, with the TF-IDF representation being the most effective.	1.The SPEMC-15K dataset includes SPEMC-15K-E and SPEMC-15K-S, covering 11 spam categories, addressing security and privacy issues in English and Spanish emails.	1.The algorithm clusters spam emails using agglomerative methods, text processing, Bag of Words encoding, and manual category labeling, evaluating sixteen machine learning pipelines.	1.The innovative method utilizes topic-based spam classification and hierarchical clustering to enhance cybersecurity insights into spam content. 2. The SPEMC-15K-E/S datasets, categorized through hierarchical clustering, provide specialized, targeted cybersecurity services for spam analysis.	1.The proposed methodology may be specialized for specific datasets and categories, potentially facing challenges in generalizing to different spam types or evolving techniques not covered in the training data. 2. The absence of publicly available datasets hinders the reproducibility and comparison of other studies.	English Spam Classification: TF-IDF + LR: F1 Score: 0.953 Accuracy: 94.6% Spanish Spam Classification: TF-IDF + NB: F1 Score: 0.945 Accuracy: 98.5%
15	Doshi, J., Parmar, K., Sanghavi, R., Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. Computers & Security, Volume 118, 103378. https://doi.org/10.1016/j.cose.2023.103378	1. Collect dataset including both spam and phishing emails. 2. Feature selection is done on the body and content of email. 3.Construct the model using dual-layer architecture. 4. Classify the email as spam, phishing or not.	1.The evolving strategies employed by spammers pose a significant challenge to conventional email spam classifiers. 2. Current spam classifiers frequently rely on specific features like email content or sender information. 3.Classifying spam is essential for maintaining data security and preventing the introduction of dangerous content.	The research paper proposes a dual-layer architecture using deep learning techniques for spam and phishing email classification, addressing data imbalance issues and considering both email body and content features.	1.The phishing emails are sourced from Nazario's publicly accessible phishing corpus, which comprises 4,204 phishing emails. 2. The Spam Assassin project has generated a dataset of	1. Traditional machine learning algorithms like LR, SVC, KNN, NB, etc. 2. Deep learning algorithms like CNN, RNN, LSTM.	1.The dual-layer architecture consistently outperforms traditional machine learning and deep learning models, achieving high accuracy and F1-scores. 2. The dual-layer architecture is suggested as a flexible solution for text classification issues involving severe class imbalances.	1.Deep learning techniques are facing difficulties in classifying the minority class, specifically spam emails, due to a significant data imbalance. 2.The paper acknowledges the limitations in its scope, particularly in the absence of a comprehensive exploration of email attachments in feature	Accuracy: 99.51% Recall: 99.68% Precision: 99.5% F1-score: 99.52%

					1,350 spam emails and 2,664 ham emails.			engineering.	
*	My Work	1. Gather Dataset from any open source for implementation of machine learning model. 2. Create a machine learning model based on any machine learning algorithm or technique. 3. Classify email as spam or not using the machine learning algorithm. 4. Refine the model for further accuracy.	1.The evolving strategies employed by spammers pose a significant challenge to conventional email spam classifiers. 2. An efficient email spam classification system is crucial in preventing fraudulent schemes and safeguarding users from scams. 3.Prioritizing legitimate messages in spam filters reduces email overload and boosts communication effectiveness.						