International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI Techniques for Spam Email Detection

Ajay[1], Hitesh Sharma[2], Surendra Kumar[3], Kalpana Jaswal[4]

[1,2,3]Student, Department of Computer Applications, Harlal Institute of Management and Technology, Greater Noida (Uttar-Pradesh)
[4]Assistant Professor, Department of Computer Applications, Harlal Institute of Management and Technology, Greater Noida (Uttar-Pradesh)

Abstract
Spam emails remain a significant Vector Machine & Naive Bayes. The PSO challenge on the modern use of internet, causing algorithm is used to navigate the feature economical loss for businesses & frustration among space and identify optimal subsets that users. Within rise of Artificial Intelligence (AI), enhance model performance. As internet usage advanced feature extraction techniques have been grows, spam emails are increasingly being used introduced to enhance spam detection systems. for unethical practices such as phishing and this study proposes a hybrid approach combining spreading malicious content. This paper aims to particle swarm Optimization (PSO) for selection of explore AI-driven strategies to detect and the features technology in machine learning models prevent these threats effectively, like Decision Tree, Support System, etc.

Keywords: Machine Learning, Support Vector, Machine, Decision Tree, Natural Language Processing, Bio-Inspired Methods, Multinomial Naive Bayes

1. Introduction

In the modern world email or we can say electronic mail is the most used technology in wide range forms of the communication with each other, but it is also heavily exploited for sending unsolicited and potentially harmful messages—commonly referred to as spam. Spam emails are typically bulk messages sent without the recipient's consent, often with the intent to advertise, deceive, or infect systems with malware. While basic filters based on sender blacklists or keyword matching have been implemented, these methods often fall short in dynamic environments where spammers constantly adapt.

To overcome the limitations of traditional filtering techniques, machine learning has emerged as a powerful tool for spam detection. It enables systems to learn from past email patterns and classify messages as either spam or legitimate (ham) with greater accuracy. Common approaches include text analysis, domain-based whitelisting and blacklisting, and community-based reporting mechanisms. However, relying solely on static filters can result in either the failure to block harmful emails or the accidental deletion of genuine messages.

Recent studies have demonstrated the effectiveness of machine learning algorithms in automatically identifying spam content. For instance, models of the naive bayes, SVM, and hybrid techniques like Ant Colony Optimization (ACO) combined with SVM have shown promising results using datasets like the UCI Spam Base. The need for intelligent systems has become urgent as spam is now used not only for advertisement but also for phishing, fraud, and other cyber threats. Hence, this paper delves into the use of AI techniques to create robust and adaptive spam detection models.

2. Advertising and Spam Emails

One of the most notorious forms of email spam is the Nigerian scam, where fraudsters deceive recipients into sending money by presenting false stories or promises. Another common and more dangerous threat is phishing, which involves tricking users into revealing sensitive information such as login credentials or financial details. This is typically done by imitating a legitimate website or organization.

Extensive research has been conducted in this area using various datasets and machine learning techniques. Algorithms like naive bayes support Vector or (SVM), K nearest Neighbors (KNN), & others also which is decision Trees have been widely implemented. Additionally, Natural Language Processing (NLP) has played a significant role in increasing spam detection accuracy. Performance of these classification methods can vary based on the nature of the dataset and the specific goals of the detection task.

Email remains a primary mode of both personal and professional communication, despite the growing influence of social media platforms. However, its widespread use also makes it a popular target for spam, including unsolicited advertisements, phishing content, malware, and mass messaging. These unwanted messages not only pose security threats but also burden users with the task of manually deleting them. Furthermore, spam consumes considerable server storage and network resources.

Phishing attacks are among the most harmful cyber threats, as they aim to steal confidential information through deception. To address these challenges, the development of automatic email classification systems has become a critical area of focus. These systems aim to efficiently distinguish between legitimate emails and spam, ensuring better protection and communication integrity. Both commercial and open-source solutions have been introduced to meet the increasing demand for effective spam filtering tools.

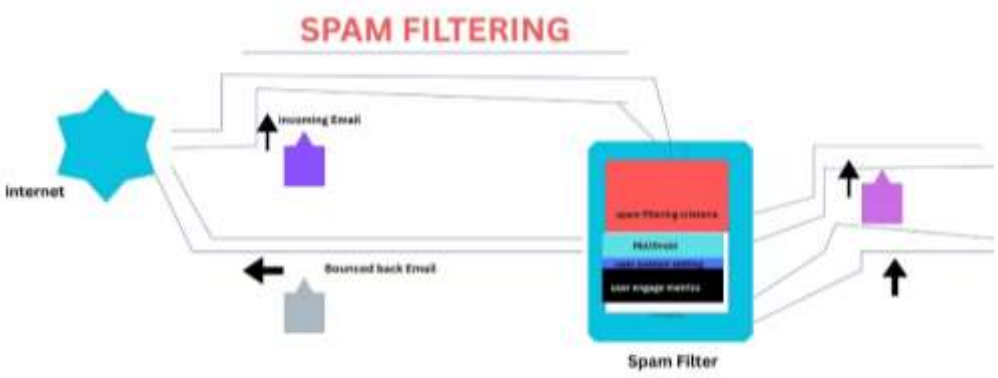
3. Literature Survey

A. Tokenization:

Tokenization is a fundamental preprocessing step in natural language processing (NLP) and plays a vital role in tasks such as spam email detection. It is also known as the process which breaks down written text into small components known by the name of tokens. These tokens can be individual words, phrases, or symbols that help simplify & structure the text data for analysis.

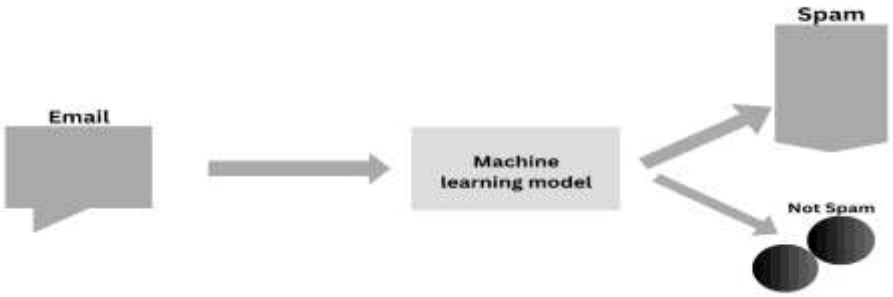
This process typically involves splitting the text based on delimiters like spaces, punctuation marks, or line breaks. The extracted tokens—often alphanumeric or purely textual—serve as the foundational elements for subsequent stages such as parsing, pattern recognition, or feature extraction. In the context of spam classification, tokenization helps identify meaningful terms and patterns within the email content, enabling machine learning models to make accurate predictions.

4. Social Network Spam Filtering



The rapid expansion of Online Social Networks (OSNs) has transformed them into a widely used platform for communication and information sharing. However, this growing popularity has also made OSNs a prime target for spammers, who take advantage of the open and interactive nature of these platforms to spread unsolicited or malicious content to vast audiences. While spam on social media was once relatively harmless, the increase in user connectivity and engagement has amplified its impact and risk. As a result, detecting and filtering spam in OSNs has emerged as a critical area of research in machine learning. Effective solutions require the application of sophisticated feature extraction techniques and the training of robust models capable of accurately distinguishing between genuine user posts and spam.

5. Twitter Spam Detection by Applying and Utilizing Various Machine Learning Algorithms LITERATURE REVIEW



S. Nithyanantham, M. Sangeetha, and M. Jayanthi explored the detection of Twitter-based spam using various machine learning algorithms. Twitter spam includes promotions, malicious links, and misleading messages that affect user behavior and compromise web security. Their study focused on evaluating the performance of algorithms like K-Nearest Neighbors (KNN), Random Forest, Gradient Boosted Machines (GBM), C5.0, and Naive Bayes. These models were tested on features such as account age, number of tweets, hashtags, retweets, and mentions. KNN was found to be effective due to its ability to weigh the similarity of nearby data points when classifying new inputs.

6. Lemmatization

Lemmatization is a technique in natural language processing that reduces words to their base or root forms. Unlike stemming, which simply removes prefixes or suffixes, lemmatization considers a word's morphological structure to produce linguistically valid lemmas. In the context of example, we can take the word “running” or “ran,” or “runs” all are re-developed to “run.” This process helps unify word variations and ensures consistency in textual data, which is particularly useful for improving the accuracy of machine learning models applied to spam detection tasks.

7. Related Work

A. Machine Learning in Spam Detection

In recent years, numerous researchers have implemented machine learning (ML) models for spam email classification. One particular study experimented with six distinct ML algorithms such like Naive bayes Which is also known as NB, K_ Nearest_neighbor (KNN), Artificial Networks (AN), Support Machines (SVM), Artificial Immune Systems, and rough sheets and their main focus was to simulate human recognition capabilities in identifying spam content. The process was structured into four phases: data

preprocessing, feature extraction, classification, and performance evaluation. The findings revealed that Naive Bayes consistently get the better output in terms of precision, again recall & overall accuracy.

Another noteworthy contribution came from Feng et al., who introduced a hybrid classification model combining SVM and Naive Bayes. In their framework, SVM was employed to generate a separating hyperplane and reduce the training dataset by filtering out less informative samples. Subsequently, Naive Bayes was applied to predict class probabilities. This model was evaluated using a Chinese language dataset and demonstrated improved accuracy over standalone implementations of NB and SVM.

B. Bio-Inspired Methods for Feature Optimization

Selecting the optimal set of features is a complex and computationally intensive task, often classified as NP-hard. A recent study addressed this by leveraging a bio-inspired approach involving a combination of chi-squared statistics and the Binary Swarm Optimization (BSO) algorithm. The selected features were then fed into an SVM classifier. The research utilized the OSAC dataset containing 22,429 text records. A subset of the data was sampled using a 70:30 split ratio, and standard preprocessing steps—such as removing digits, stop words, and special characters—were performed.

The combined BSO-CHI-SVM approach was benchmarked against traditional models and showed superior performance compared to ANN in the main training timing and accuracy after classification. That Model has an impressive correctness of 95.6%. The owners or leaders concluded that SVM-based approach was more efficient than ANN and suggested future work could involve exploring n-gram models and semantic feature representations.

Further exploration has also been done using Genetic Algorithms (GA) integrated with Decision Tree (DT) classifiers. Researchers addressed the common problem of overfitting due to high-dimensional data by applicable principal component analysis (PCA) for features decrements. The optimized Decision Tree model (J-48) was combined with GA using a fitness function search strategy known as BLX-alpha. This output of experiment was carried out in the Enron spam datasets value, where it proposed GADT model (GA + DT) outperformed other classifiers. Additionally, applying PCA further enhanced the performance metrics.

8. Proposed Work (Rephrased & Cleaned)

This research project focuses on evaluating the performance of machine learning classifiers in combination with bio-inspired optimization techniques for spam email detection. The experiments will utilize multiple public spam corpora to validate the models under diverse conditions.

Research Objectives:

- To explore and implement ML algorithms for identifying spam emails.
- To apply bio-inspired optimization techniques like PSO and GA for feature selection and model improvement.
- To compare baseline ML models with their optimized counterparts.
- To assess the behavior of each model on various datasets.

The Scikit-learn Python library will be used for model development, data preprocessing, and result computation. The spam detection system is expected to ingest raw email data, extract relevant features through text mining, and apply an optimized supervised learning model to classify emails as spam or legitimate.

9. Tools and Techniques

I. WEKA Tool

WEKA (Waikato Environment for Knowledge Analysis) is an open-source GUI-based software used for applying data mining and machine learning algorithms. It supports classification, clustering, regression, and visualization of data. In this study, spam datasets were converted to .arff format to be compatible with WEKA.

The tool was mainly taken to analysis the result or performance of the several algorithms for classification including:

- Multinomial Naive Bayes (MNB)
- Sequential Minimal Optimization (SMO) – an SVM variant
- J48 – an implementation of the C4.5 Decision Tree algorithm
- Random Forest

Table 1: TF-IDF results

Token	TF-IDF value
feature	0.085
token	0.079
model	0.075
email	0.065
algorithm	0.065
spam	0.057
classification	0.055
selection	0.051

Table 2: Feature selection results

Attribute	Score
email address	10
remove word	8
word stem	6
classification	4
feature selection	3

Among the Naive Bayes models tested, MNB emerged as the most accurate. According to the summary of results (referenced in Table 9), these algorithms were found to perform consistently well in identifying spam and emails which are accurate.

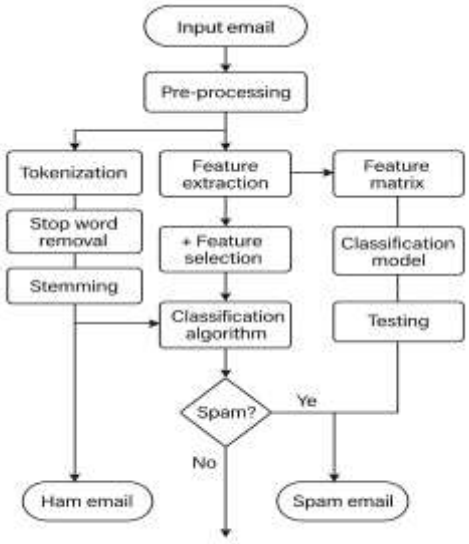


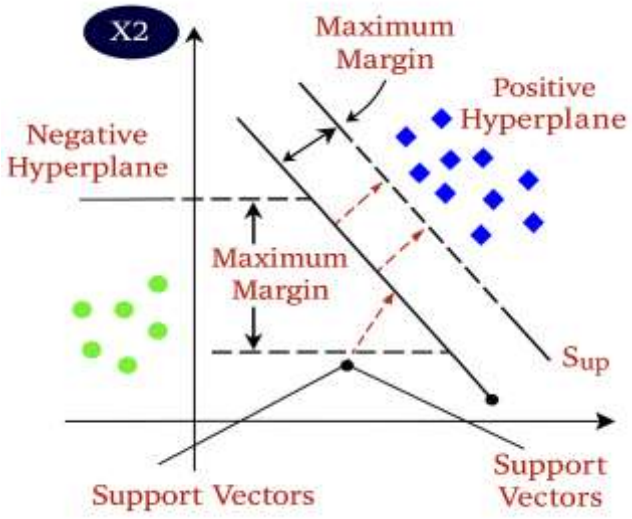
Figure 1: Spam Detection Block Diagram

II. Support Vectors Machines (SVM)

Support Vector Machines or SVM is a widely used learning in supervised manner algorithm designed for classification tasks. The main point of SVM is to identify the accurate hyperplanes that will separate the data points into 2 classes in a 2-D space, these hyperplanes appear as the straight line divides the plane in the two or three regions where each class is located in only one sides.

The initial objective of SVM is to increase the margin between the nearest data points (called support vectors) of each class. These support vectors are the most critical elements in determining the classifier's boundary. By focusing only on these data points, SVM becomes computationally efficient and provides high accuracy even on complex datasets.

SVM is particularly effective in spam email classification, where it distinguishes between spam and ham emails based on extracted textual features. It performs well even in high-dimensional feature spaces and is known for its robustness in various real-world applications.



III. Decision Tree

Decision Tree is the supervised learning way that models new decisions and their consequences which can happen in future in a tree-like structures. Each primary (non-leaf) node will represent the test in an attribute, for every branch corresponds to the result of that test, and every leaf's node signifies a basic label (e.g., spam, ham etc.).

The creation of the decision trees means not require domains friendly or specific knowledge & parameter tuning into making it beneficial for large, multidimensions dataset. The algorithms work by selecting attributes that best separate the dataset into distinct classes based on information gain or Gini impurity.

Entropy calculations are often used to determine the purity of a node. For example, the entropy based on a single attribute or a combination of attributes helps guide the tree-building process, ensuring that the dataset is split efficiently at each stage.

Decision Trees are intuitive, easy to visualize, and can handle both categorical and numerical data. Their ability to generate human-readable classification rules makes them valuable in spam detection tasks.

IV. Naive Bayes – Multinomial Model (MNB)

Naïve_Bayes's theorem is the probabilistic classifier depending on the Bayes' Theorem. It thinks like independence between features of theorem, which simplifies computation and makes it scalable for large

datasets. The Multinomial Naive Bayes (MNB) model is particularly effective for text classification, as it considers the frequency of words within a document.

In the idea of the spam detection, MNB calculates the main probability that will provide email belongs to either the 'spam' or 'ham' class based on the occurrence of words. The formula will be like:

P(Class|Word) =P(Word|Class)×P(Class)P(Word)P(Class|Word) = \frac{P(Word|Class) \times P(Class)}{P(Word)}P(Class|Word)=P(Word)P(Word|Class)×P(Class)

Here, P(Class|Word) is the posterior probability, P(Word|Class) is the likelihood, and P(Class) is the prior probability. Smoothing techniques are applied to handle words that may not appear in the training data.

Among the three Naive Bayes variants—Multinomial, Gaussian, and Bernoulli—the MNB model performs best for text data, especially in spam classification where word frequency plays a vital role. It is simple, fast, and yields high accuracy with minimal tuning.

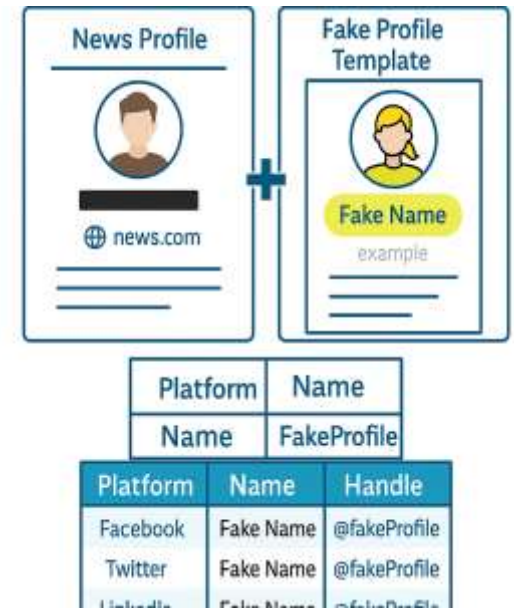
V. Natural languages Processing (NLP)

Natural languages processing (NLP) is the part of Artificial intelligence that is focused only on the interactions between two things which are computer and human languages. In spam detection, NLP helps to the crucial roles in the processing email content to extract meaningful features.

Key NLP tasks include:

- **Tokenization** – breaking down written texts into words or tokens.
- **Lemmatization**– decreasing words to the basic forms (e.g., "running" → "run").
- **Stop-word removal** – eliminating commonly used words like “and” or “the” that carry minimal semantic value.
- **Vectorization** – converting text into numerical form using techniques like TF-IDF or Word2Vec.

NLP enables machine learning algorithms to understand the structure and semantics of email content. Recent advancements also include or merge the use of learning-based models such as BERT to contextual text representation, further improving detection of phishing and spam emails.



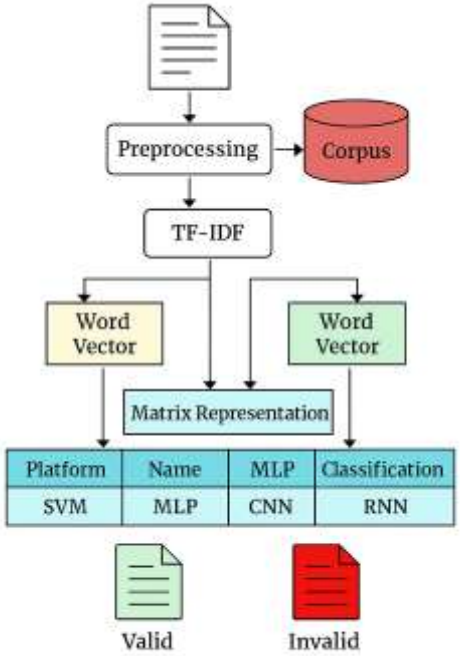


Figure 2: Classification process

10. Conclusion

This research has presented a comprehensive review and experimental insight in the view application which is on the topic of Machine Learning and bio-inspired algorithms to spam email detection. By utilizing diverse datasets and implementing a range of classification models which are naïve Bayes, support vector machine, decision tree & ensemble techniques—the study highlights the strengths and weaknesses of each approach.

The integration of the enhancing methods like particle swarm-optimization (P.S.O) and genetic algorithm or (G.A.) significantly enhanced model performance by effectively selecting the most relevant features. Among the evaluated models, the supervised learning techniques demonstrated consistent and reliable accuracy in detecting spam messages, especially when combined with natural language processing (NLP) for text feature extraction.

Approximately 50,000 emails were analyzed across different experiments, validating the effectiveness of the models. The study concludes that hybrid approaches involving bio-inspired optimization and machine learning are promising solutions to the ongoing issue of spam detection in digital communication. Future work can further explore deep learning models, real-time filtering systems, and the adaptation of newer NLP techniques such as BERT for even better performance.

11. References

1. iMustapha, I. B., Hasan, S., Olatunji, S. O., Shamsuddin, S. M., & Kazeem, A. [2020]. Effective Email Spam Detection System using Extreme Gradient Boosting. This study introduces an improved spam detection model based on Extreme Gradient Boosting (XGBoost), demonstrating superior performance over earlier approaches across various evaluation metrics.
2. Zavrak, S., & Yilmaz, S. [2022]. Email Spam Detection Using Hierarchical Attention Hybrid Deep Learning Method. The authors propose a novel technique combining convolutional neural networks, gated recurrent units, and attention mechanisms for email spam detection, achieving enhanced perfor-

mance through hierarchical representation and cross-dataset evaluation.

3. Labonne, M., & Moran, S. [2023]. Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection. This paper evaluates the effectiveness of large language models (LLMs) in email spam detection, introducing Spam-T5, a Flan-T5 model fine-tuned for this purpose, which outperforms baseline models, especially in few-shot scenarios.
4. Bhowmick, A., & Hazarika, S. M. [2016]. Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends. A comprehensive review focusing on machine learning-based spam filters, discussing their effectiveness and the evolving nature of spam, highlighting the ongoing cat-and-mouse dynamics between spammers and email service providers.
5. Siddique, N. et al. [2021]. Machine Learning-Based Detection of Spam Emails. This study compares various machine learning and deep learning models, including Naive Bayes, SVM, CNN, and LSTM, for spam detection, with LSTM achieving the highest accuracy of 98.4% in detecting Urdu spam emails.