Grenze International Journal of Engineering and Technology, Jan Issue

Enhancing Spam Detection Accuracy using Genetic Algorithm

Akshitha Singireddy[1], Harika Nallapati[2] and Dr. E. Padmalatha[3]
[1-3]Dept of CSE, Chaitanya Bharathi Institute of Technology, Hyderabad, India
Email: akshitha1302@gmail.com, nharikachowdary04@gmail.com, epadmalatha_cse@cbit.ac.in

***Abstract*— The technology advancement has resulted in increased usage of internet which in turn has provided certain groups a chance to misuse the resources and services. Due to this exploitation, spam messages are sent to the customers because of which they are missing out on the important information. Spam has become a critical problem in online social networks. Spam in social networks refers to the unwanted, malicious, unsolicited content or behavior, fake friends, fraudulent reviews. Algorithms used for spam detection can take a lot of time to be trained as it includes irrelevant and redundant features. This results in poor predictions and high computational overhead. The proposed method mainly focuses on selecting the best features to increase the accuracy. Thus, selecting relevant feature subsets can help in reducing the computational cost and speed up the learning process. An extensive research was done to implement Apriori technique as well as genetic algorithm on different email datasets, along with feature extraction and pre-processing. The comparison of our results show the best suitable model is also discussed.**

***Index Terms*— Spam Messages, Malicious content, computational cost**

I. INTRODUCTION

Spam messages are a significant problem in today's digital world, as they can harm regular email communication. It is imperative to identify and eliminate harmful spam emails to protect users from potential harm. In the past, several methods, such as blacklisting, whitelisting, content-based filtering, and machine learning-based strategies, were used for spam detection.

However, these methods have limitations and cannot provide 100% accurate spam detection. Blacklisting involves maintaining a database of known spam email IDs and checking incoming messages against these IDs. If an email is found to be coming from a blacklisted ID, the system immediately recognizes it and reports it. Similarly, some advanced systems use IP address based data and mail filtering to block messages from suspected IP addresses, enhancing system security. Whitelisting, on the other hand, identifies known and trusted email IDs, allowing users to block other IDs and protect the system from future emails. The whitelisting system also stores known-party mail IDs, acting as a data store for efficient email handling system security.

Content-based filtering is another method where the content of received emails is extracted and filtered against predefined blocked content. If the content matches the blocked content, the email is marked as spam or junk. Despite these existing methods, there is currently no foolproof solution for spam detection. Each method has its limitations and may not provide complete spam detection ability. Therefore, further research and development are needed to enhance spam detection techniques and effectively combat the issue of spam messages. Spam messages

Grenze ID: 01.GIJET.10.1.8

are a significant problem that needs to be addressed in the digital world. While various methods such as blacklisting, whitelisting, and content-based filtering have been used in the past, there is no perfect solution for spam detection. Continued efforts are required to improve spam detection techniques and protect users from the harmful effects of spam emails in the present digital scenario.

II. Related Work

Numerous methodologies have been proposed to tackle the issue of spam, which includes four primary categories of detection techniques: 1) syntax analysis based detection, 2) feature analysis based detection, 3) hybrid techniques, 4) blacklist based detection. Syntax analysis-based detection techniques involve analyzing the syntax of tweets, such as URLs and tweet content.[1] Since Twitter limits the number of characters in each tweet and URLs often contain many characters, URL-based detection techniques often focus on shortened URLs. However, this presents a challenge as shortened URLs can hide direct lexical meanings and dangerous links. To address this, Thomas et al. proposed a shortener identification method to identify spammers by calculating the percentage of descriptions that were shortened by shortening services like bit.ly. The discrimination used to spot spam tweets is calculated as the division product of the two ratios.

Spam has been addressed through the development of spam detection models. These models are categorized as non statistical and statistical approaches. Non statistical approaches lack a learning component, which may result in legitimate messages being classified as spam, leading to undetected spam and increased automated response messages. To overcome these limitations, machine learning algorithms have been employed. Support Vector Machines, Artificial Neural Networks, and other popular learning algorithms have also been used in spam detection. However, these approaches may suffer from high computation overheads and low detection rates due to the inclusion of irrelevant features. To address these issues, various techniques such as feature selection have been put forward in spam detection research. Researchers have turned to machine learning algorithms, such as Support Vector Machines, Boosting, and Artificial Neural Networks, to improve the performance of spam detection models. These techniques aim to select relevant features to enhance the accuracy and efficiency of spam detection. By utilizing these techniques, researchers aim to overcome the barriers of non-statistical approaches and upgrade the detection rates of spam while minimizing false positives and false negatives.

In their Conference paper, Sang Lee, Dong Kim, Ji Kim, and Jong Park present an optimised spam detection model based on Random Forest. Their approach incorporates simultaneous optimization of parameters and feature selection using Random Forest, offering four main advantages. Firstly, it optimizes the parameters of RF to enhance the model's performance. Secondly, it identifies important features as numerical values, providing insights into the most relevant features for spam detection. "Ref. [2]"Lastly, it achieves spam detection with less processing overheads and higher detection rates, making it a promising solution for practical use. The anticipated detection rates are then contrasted with a predetermined threshold value, T. The phase is finished if the detection rates satisfy the design criteria. In every other case, feature selection is used to develop the spam detection model through iterative approaches. To find the ideal selection of characteristics, the researchers offer two methods for constructing the spam detection model. The first approach involves using the same parameter values as the optimal parameter values of the initial model until the rebuilding process is finished. The alternative strategy involves doing parameter optimisation each time a feature is dropped. Using these techniques, the researchers may iteratively improve the model to produce the best Random Forest spam detection solution.

In their paper, Shahad Suhail and Karim Hashim discussed the objectives of their system, which aims to segregate emails as either spam or non spam. Pre-processing, training, and testing are the three main modules of the system, each with a number of smaller ones and components. To find patterns based on the connections between item-sets, the researchers used association rule algorithm, a data mining process. "Ref. [3]" Additionally, they used the Information Gain (IG) approach to minimise the dimensionality of the feature space and the Term Frequency Inverse Term Frequency (TFIDF) method to extract features from the dataset. The suggested method in their study uses the aforementioned methods to classify emails as spam or not based on the content of the messages.

In the proposed system, association rule mining is used. However, implementing association rules algorithms can result in a large number of item-sets and require significant space, leading to lower accuracy. To tackle this problem, the dataset was divided into segments to enhance the accuracy of classification using association rule algorithms, leading to improved classification accuracy. The segmentation was performed based on the information gain algorithm, which identified features with high weight values for inclusion in the segments. It was found that the performance of classifiers improves with larger training sample sizes.

Zahra Razi and Seyyed Amir Asghari conducted research on enhancing spam detection using a Genetic Algorithm assisted Artificial Immune System, and compared it with standard AIS in their paper. In order to reduce algorithm

mistakes and runtime, the proposed method was created to cooperate with various filtering systems. The researchers modified the AIS model in a number of ways to increase its effectiveness by applying GA. “Ref. [4]”Support Vector Machine (SVM) was used for email classification, and performance was compared with decision trees, in the study. SVM was used for email classification, and performance was compared with traditional AIS.The study’s conclusions showed that when dealing with huge datasets, SVM can be computationally time- and memory-intensive.

The accuracies reported in the studies on review spam detection using various machine learning techniques vary. Logistic regression achieved an AUC of 0.78, and SVM and Na¨ıve Bayes achieved 89.8% accuracy in different studies. Using topic models and SVM, accuracies ranged from 52.0% to 64.7% for cross-domain review spam detection. Stylometric features with SVM achieved an F-measure of 84%. Na¨ıve Bayes achieved a high accuracy of 99.59% with Random Oversampling for Arabic review spam. The unsupervised approach using Semantic Language Models (SLM) had an impressive AUC of 0.9987. Semi-supervised learning with Na¨ıve Bayes reached an F-Score of 0.609 and 0.631 with co-training and agreement modification. It is important to consider the diversity of datasets and evaluation metrics when comparing these results.

III. Theoretical Foundation

A. Feature Selection

Feature selection is an important step in machine learning and data preprocessing, where the most relevant and important features (also known as variables or attributes) are selected from the original features to be given as input for a model. Feature selection is crucial for improving the performance and interpretability of machine learning models, as it can reduce the complexity of the model, mitigate the risk of overfitting.

1) *Apriori Algorithm*: This technique that focuses on establishing association rules among items, which represent relationships between two or more objects. It specifically examines whether customers who purchase one product, such as Product A, tend to also purchase another product, such as Product B. The main goal of this algorithm is to generate association rules that describe the connections between different items. Another commonly used term to refer to the Apriori algorithm is ”frequent pattern mining.”

The Apriori algorithm consists of three key components: Support, Lift, and Confidence. Support refers to the frequency or occurrence of a particular item or itemset in the dataset being analyzed. Confidence measures the likelihood that an association rule holds true, given the support of the antecedent (input) and consequent (output) items. Lift, on the other hand, quantifies the degree of dependence or correlation between the antecedent and consequent items, indicating how much more likely the consequent item is to be purchased when the antecedent item is purchased compared to when it is not. It consists of three components, namely Support, Confidence, and Lift, are used to evaluate the strength and relevance of the generated association rules.

$$\text{Support}(X \rightarrow Y) = P(XUY) \quad (1)$$

$$Confidence(X \rightarrow Y) = \frac{P(XUY)}{P(X)} \quad (2)$$

$$Lift(X \rightarrow Y) = \frac{P(XUY)}{P(X)xP(Y)}$$

2) *Genetic Algorithm*: It is a type of optimization algorithm inspired by the process of natural selection in biological evolution. It is a search algorithm used in machine learning and computational optimization to find approximate solutions to complex problems where traditional methods may be inefficient or impractical. A population of potential solutions to a problem is evolved over several generations in a genetic algorithm to find an ideal or nearly ideal answer. The answers are shown as distinct entities known as chromosomes, which are commonly encoded as strings of binary or realvalued values. The genetic algorithm replicates the process of genetic recombination and mutation in biological organisms by applying genetic operators such as crossover (recombination) and mutation to the chromosomes to produce new progeny.

A fitness function that gauges how well each person in the population solves problems is used to assess each person’s fitness. Higher fitness values increase the likelihood that an individual will be chosen for reproduction and pass on their genetic makeup to the following generation. Over successive generations, the population evolves, with better solutions typically emerging over time. Genetic algorithms are particularly well-suited for solving complex optimization problems where the search space is large, multi-dimensional, or poorly understood. It involves phases which are given as below:

- Initialization

- Fitness Assignment
- Selection
- Reproduction
- Crossover
- Mutation
- Termination

IV. METHODOLOGY

A. Proposed System

Many organisations and people now have more convenient contact alternatives because of electronic mail. Spammers who send malicious emails make use of this technique to make fraudulent gains. This paper seeks to introduce a technique for spam email detection using machine learning algorithms that are enhanced using bio-inspired techniques. A survey of the literature is conducted to investigate effective techniques used on various datasets to get good outcomes. To apply the Apriori technique and genetic algorithm to various email datasets, along with feature extraction and pre-processing, substantial study was conducted. Our results are compared, and the optimal model for the situation is also considered.

B. Pre-Processing

To enable further analysis and learning in a model, text data needs to be transformed into a clear and consistent format. This is achieved through text preprocessing methods, which can be generic and applicable to various applications, or tailored for a specific purpose. The text preprocessing steps used in the proposed method are as follows.

1) *Cleaning the data*: Cleaning text data, such as text messages, involves additional considerations due to the unstructured nature of text data. Text normalization involves converting text to a consistent format to reduce inconsistencies. This can include converting all text to lowercase or uppercase, removing punctuation, and removing or replacing special characters or symbols. Text messages may contain special characters, emojis, or other symbols that can impact the analysis. Removing or replacing these special characters and symbols can ensure that the text data is processed correctly.

2) *Stopword removal*: Stopwords are the most frequent words that are unlikely to aid text mining, including prepositions, articles, and pronouns. Since these words are present in every text document, text mining can be used without their usage. These words are all omitted. Any word group will work for this purpose. Additionally, it reduces the amount of text data and enhances system speed.

3) *Stemming*: Stemming is a technique commonly used in text data pre-processing to reduce words to their root form. It removes suffixes from words to obtain the core part or root, which can help in reduction the dimensionality of the text data and grouping similar words together. Word stemming involves applying stemming algorithms, such as Porter, Snowball, or Lancaster, to reduce words to their root form. These algorithms use various rules and heuristics to identify and remove common suffixes from words. For example, ”running”, ”runs”, and ”ran” would all be stemmed to ”run”.

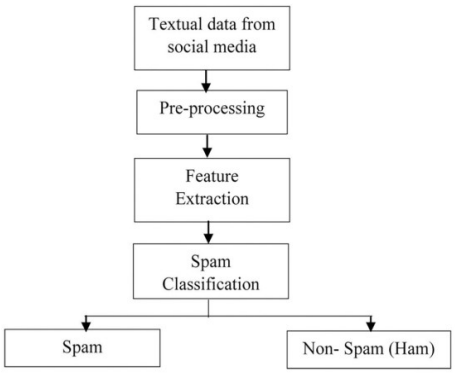

Fig. 1. Flow of the Proposed Model

4) *Term Frequency*- Inverse Document Frequency (TFIDF) technique: TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be described as the evaluation of a word's significance to a text within a corpus or series. The frequency of a word in the corpus more than makes up for the meaning increasing proportionally to how many times it appears in the text (data-set).

C. Feature Selection

The data is processed and by using TF-IDF vectorization the data is converted into mathematical vectors. Then the feature selection comes into picture. One of the Association rule mining algorithms, "Apriori Algorithm" is used to find the frequent item sets in the dataset upon which Genetic Algorithm is applied to select the features that are of most important in the dataset.

D. Algorithms Used

Upon the dataset on which Genetic Algorithm is applied is used to train and test the model. The models are built using different Machine Learning algorithms like Logistic Regression, Decision Tree, KNN, Random Forest, Voting Classifier (LR + RF+ DT).

1) *Logistic Regression*: Logistic regression is used for binary classification in machine learning. It is one of the types in supervised learning algorithm that is used in predicting the probability of an outcome belonging to a certain class, based on input features. In this, the aim is to model the relationship between a set of input features (also known as independent

2) *Decision Tree*: It is used for both classification and regression tasks. It can handle various kinds of data such as numerical data, categorical data. It is flowchart kind of representation where internal node represents a decision based on a feature and leaf node represents predicted outcome

3) *KNN*: It works by identifying the k nearest data points to a given test data point in the feature space, and then making predictions based on the majority class or average value of these k neighbors. K-Nearest Neighbor is simple, intuitive, and can handle multi-class classification problems, but it can be computationally expensive for large datasets.

4) *Random Forest*: It is an ensemble machine learning algorithm that combines various or multiple decision trees to make predictions. It uses random feature selection and bootstrapped sampling to build a "forest" of trees, and aggregates their predictions for improved accuracy and robustness. It is widely used for classification and regression tasks in various domains.

5) Voting Classifier: A voting classifier in machine

learning is an ensemble learning technique that combines the predictions of multiple base classifiers to make a final prediction. Each base classifier is trained on the same dataset using a different algorithm, and their predictions are aggregated using a majority vote or weighted vote. The voting classifier can improve overall prediction accuracy and robustness, and it is commonly used in machine learning for classification tasks.

The Whole Methodology in a brief is that initially textual data is taken and is pre-processed like Cleaning the data, Stopword Removal, Stemming, Term Frequency- Inverse Document Frequency (TF-IDF). Then the vectorized data is feeded into the apriori algorithm yealding frequent dataset items, then genetic algorithm is applied on the resulted data. Now, the dataset is then trained using Machine Learning Models and finally the algorithm with highest accuracy is chosen.

V. Experimental Setup & Results

A. Dataset

All the required packages are imported. And the dataset is loaded into the system.

Spam Mail Classifier: Spam - ham dataset This dataset is taken from Kaggle, it contains the Enron-Spam datasets, as described in the paper: V. Metsis, I. Androutsopoulos and G. Paliouras, "Spam Filtering with Naive Bayes - Which Naive Bayes?". Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006), Mountain View, CA, USA, 2006. The dataset contains 5170 messages (rows) with 4 columns namely index (order of the message arrived), label(spam or ham), text, label num (if spam it's 1, or else it's 0). The dataset has 4993 unique values and 5170 messages out of which 71% are ham and 29% are spam.

B. Parameters used in Genetic Algorithm

- *Population Size*: The population size determines the number of individuals (solutions) in each generation. A larger population can potentially explore more diverse solutions, but it also increases the computational cost.
- *Number of Generations*: The number of generations represents how many iterations the genetic algorithm will run. It is essential to set this parameter to a sufficient value to allow the algorithm to converge. Here for the dataset taking 5 number of generations gave good results. • Mutation Rate: The mutation rate defines the probability that a selected individual's gene (parameter) will be subject to mutation. Mutation introduces new genetic information and can prevent the algorithm from getting stuck in local optima. A typical mutation rate

is in the range of 0.01 to 0.1. Here as the dataset contains more than 5000 features 0.2 is an optimal choice for mutation rate

- *Crossover Rate*: The crossover rate determines the probability that two selected individuals will undergo crossover to produce new offspring. A high crossover rate allows for more exploration of the search space, but too high a value might lead to premature convergence. 0.5 value of Crossover Rate is taken for better results.

C. Evaluation Metrics

TABLE I EVALUATION METRICS

Algorithm Used	Accuracy	Precision	Recall	F-Score
Logistic Regression	98.9	99	98	99
Random Forest	98.4	98	98	98
Decision Tree	92.9	92	91	92
KNeighbours Classifier	96.3	97	94	95
Voting Classifier	98.7	98	99	98

Finally, Logistic Regression is chosen which is giving the highest accuracy of 98.9 and Spam Detection can be done.

VI. CONCLUSIONS

This paper gives a novel approach for e-mail spam detection called Genetic Processing with Apriori technique, which achieves a high accuracy of 98.9%. The proposed methodology incorporates feature selection to eliminate irrelevant features from the textual messages, resulting in a significant reduction in the number of selected features. This method is effective in improving the efficiency of spam detection by removing unnecessary features from the set. This is a unique approach of applying Apriori algorithm which filters the frequent item sets, upon which Genetic Algorithm is applied to filter out the most important features and excluding the features which are of no use.

The processed dataset is used to train multiple machine learning algorithms for classifying messages as spam or ham. Additionally, this algorithm incorporates Natural Language Processing techniques to accurately identify spam emails. The proposed algorithm produces highly accurate outcomes in terms of classification accuracy, as demonstrated in the Results and Discussion section, indicating its intelligent performance.

VII. FUTURE SCOPE

As technology continues to evolve, the future scope for spam detection methods is expected to be promising. Some potential areas of improvement can be:

- Explainable AI, which allows users to understand how and why a model makes certain decisions, could be incorporated into spam detection methods to provide transparency and accountability. This could help users understand the basis for spam classification and build trust in the spam detection system.
- Human judgement can play a crucial role in spam detection, as spammers often employ social engineering techniques to bypass automated filters. Future spam detection methods may involve human-in-the-loop approaches, where human feedback is actively incorporated into the system to improve spam detection accuracy. This could involve leveraging crowdsourcing or incorporating feedback from end-users to continuously improve the spam detection system.
- With the proliferation of multiple communication channels such as email, messaging apps, social media, and voice calls, future spam detection methods may need to evolve to detect spam across different platforms. This could involve developing unified spam detection frameworks that can handle multiple communication channels and provide consistent and accurate spam detection results.

REFERENCES

[1] Zeynab Fallah Sokhangoee , Abdoreza Rezapour, “A novel approach for spam detection based on association rule mining and genetic algorithm”, 2022;

[2] Sang Min Lee, Dong Seong Kim, Ji Ho Kim, Jong Sou Park, “Spam Detection Using Feature Selection and Parameters Optimization”,2019; https://www.researchgate.net/publication/221328414
[3] Shahad Suhail Najam , Karim Hashim AL-Saedi, “Spam classification by using association rule algorithm based on segmentation”, 2018; https://www.researchgate.net/publication/327906366
[4] Zahra Razi, Seyyed Amir Asghari, “Providing an Improved Feature Extraction Method for Spam Detection Based on Genetic Algorithm in an Immune System”, 2017;
[5] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race,” in Proc. 26th Int. Conf. World Wide Web Companion, 2017, pp. 963–972.
[6] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, “BotOrNot: A system to evaluate social bots,” in Proc. 25th Int. Conf. Companion World Wide, 2016, pp. 273–274.
[7] A. Davoudi, A. Z. Klein, A. Sarker, and G. Gonzalez-Hernandez, “Towards automatic bot detection in twitter for health-related tasks,” AMIA Summits Transl. Sci. Proc., vol. 2020, p. 136, May 2020.