Email Spam Detection Using Machine Learning Algorithms

Abuzar
School of Computer Science and Engineering
Galgotias University
Greater Noida, India
abuzarrajpoot02@gmail.com

Amit Verma
School of Computer Science and Engineering
Galgotias University
Greater Noida, India
amitv28242@gmail.com

Noman Ali
School of Computer Science and Engineering
Galgotias University
Greater Noida, India
nomanaliko@gmail.com

Abstract – Email spam is becoming a significant issue. They are being used for fraud, phishing, and other illicit and immoral activities. sending harmful links via spam emails, which can compromise both your and our systems. Spammers find it very easy to create a phony email account and profile. They pose as real persons in their spam emails, and they prey on those who are unaware of these scams. Therefore, it's important to recognize spam emails that are fraudulent. This research will use machine learning approaches to identify those spam.

Keywords: (Machine Learning, Naive Bayes, SVM, Decision Tree, Random Forest, Boosting, Neural Network.

1. INRODUCTION

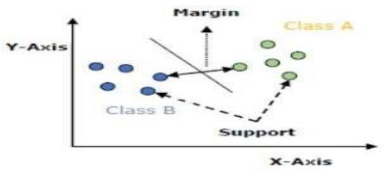
Email or electronic mail spam has been defined as "the use of email to send unsolicited emails or advertising emails to a group of recipients. The method of "using email to send unsolicited emails or advertising emails to a group of users" is known as email spam, or electronic mail. When an email is sent that is not requested, its sender has not given permission to receive it. On the internet, spam has grown to be really unfortunate. Spam is a time, storage, and message speed waste of time. Although automatic email filtering is likely the best way to identify spam, spammers may now easily get through all of these spam filtering technologies. A while backwards, the majority of spam that came from specific email addresses could be manually banned. A machine learning approach will be used to detect spam. "Text analysis, white and blacklists of domain names, and community-primarily based techniques" are some of the major ways used for junk mail screening. Textual analysis of email content is a common method for detecting spam. There are numerous ways that can be deployed on the server and by the user. Naive Bayes is one of them.

Machine learning is more effective since they utilize a pre-classified set of emails as training data samples. Multiple algorithms from machine learning techniques can be applied to email testing. "Naïve Bayes, support vector machines, neural networks, K-nearest neighbor, random forests, etc." are some of these algorithms.

2. MACHINE LEARNING CLASSIFICATION ALGORITHMS

Several machine learning classification techniques will be tested for email spam detection based on TF-IDF properties. The algorithms under consideration include.

- **Support Vector Machines (SVM):** SVM is a powerful algorithm that seeks to find an optimal hyperplane to separate spam and non-spam emails based on the TF-IDF features.



- **Random Forest:** Random Forest constructs an ensemble of decision trees to make predictions. It can effectively handle high-dimensional feature spaces and provide accurate email spam classification.
- **Decision Tree:** Decision trees use a hierarchical structure of nodes to make decisions. They can capture key aspects for email spam classification using TF-IDF values.
- **Multinomial Naive Bayes (MultinomialNB):** MultinomialNB is a probabilistic algorithm that models the conditional probability distribution of the TF-IDF features given the class labels. It can handle text-based data efficiently.

By using these machine learning classification algorithms to the TF-IDF features extracted from the email dataset, we aim to build a robust and accurate email spam detection system capable of differentiating between spam and legitimate emails, thereby improving email security and user experience.

3. OBJECTIVES OF THE PROJECT

- Create a machine learning model for email spam detection.
- Achieve high accuracy in classifying emails as spam or non-spam.
- Improve the accuracy of classification by reducing false positives and negatives.
- Increase the system efficiency for real-time spam detection.

4. METHODOLOGY

Describe the general process for detecting email spam with machine learning.

- Data Source: This component represents the source of email data, such as the Kaggle dataset or a realtime email feed.
- Data preprocessing: When the data is considered, always a very large data sets with large no. of rows and columns will be noted. However, this is not always the case; data can take many formats, including images, audio, and video files. Structured tables, etc.
  Machine doesn't understand images or video, text data as it is, Machine only understand 1s and 0s.
- Feature extraction: Utilize the TF-IDF (Term Frequency-Inverse Document Frequency) technique to convert the email text into numerical feature vectors.
- Model training and evaluation: Train a machine learning model (e.g., Naive Bayes, SVM, Random Forest) using the labeled dataset and evaluate its performance using metrics such as precision, recall, and F1-score.
- Real-time spam detection: Deploy the trained model to classify incoming emails as spam or non-spam in real-time.

5. PROPOSED SYSTEM

The proposed system leverages the power of machine learning algorithms to classify emails as spam or non-spam based on their content. The TF-IDF technique is used to turn email content into numerical characteristics, which capture the significance of specific terms throughout the message. These features are then fed into machine learning models for training and prediction.

In this project, a proposed email spam detection system will be designed applying an SVM algorithmSVM is a popular and effective machine learning algorithm for binary classification tasks. The system will be trained on a Kaggle dataset including labeled spam and non-spam emails. The trained SVM model will then be used to classify incoming emails as either spam or non-spam based on their content.

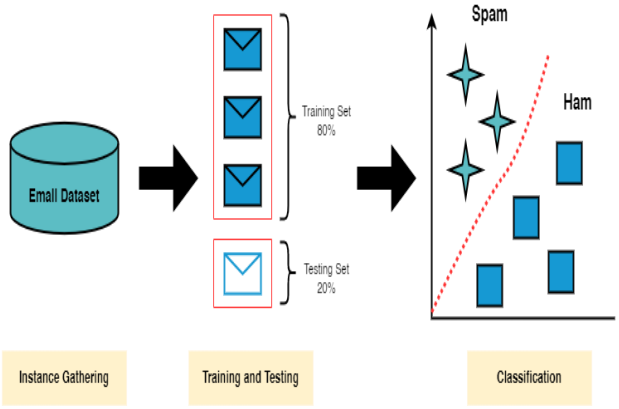
The workflow begins with preprocessing steps, including tokenization, stop word removal, and stemming, to enhance the accuracy and efficiency of the TF-IDF calculations. Next, TF-IDF vectorization approach is used to represent each email as vector of numerical values, emphasizing the significance of terms within the massage. These vectors serve as input to popular machine learning algorithms, such as Naive Bayes, Support Vector Machines (SVM), or Random Forest, which learn from labeled training data to build effective spam classification models.

6. PROJECT ARCHITECRURE DIAGRAM

The project architecture diagram provides a visual representation of the system's components and their interactions. It shows how the various components of the email spam detection system are organized and linked for achieving the desired functionality. The diagram serves as a blueprint for understanding the system's structure and flow of data and helps in the implementation and communication of the project.

- Data Preprocessing: This component involves various preprocessing steps, such as tokenization, stop word removal, and stemming, to clean and normalize the email text before feature extraction.
- Feature Extraction: This component utilizes the TF-IDF technique to convert the preprocessed email text into numerical feature vectors.
- Machine Learning Model: This component includes the selected machine learning algorithm, such as SVM, Random Forest, k-NN, or Naive Bayes. The model is trained on the labeled dataset to learn the patterns and characteristics of spam and non-spam emails.
- Model Training: This component represents the process of training the machine learning model using the preprocessed and feature-extracted email data. The model learns to classify emails based on their features and labels.
- Model Evaluation: This component assesses the performance of the trained model using evaluation metrics like accuracy, precision, recall, and F1-score. It helps in understanding the model's effectiveness in email spam detection.
- Real-time Email Classification: This component represents the application of the trained model to classify incoming emails in real-time. The system predicts whether an email is spam or non-spam based on its features and assigns the appropriate label.

- Output: This component displays the system's output, which may contain grouping results, statistical data, and visualizations for additional analysis and interpretation.



The project architecture diagram provides a comprehensive overview of how the different components of the email spam detection system interact and contribute to its overall functionality. It helps in understanding the flow of data and the role of each component in the process of email spam detection using machine learning.

7. ENSEMBLE LEARNING METHODS

1. BAGGING

Bagging classifier is an ensemble classifier that fits base classifiers each on random sub sets of the original data sets and then combined their individual calculations by voting or by averaging) to form a final prediction. "Bagging is a mixture of bootstrapping and aggregating.

Bagging= Bootstrap AGGregatING

Bootstrapping helps to lessening the variance of the classifier and it also decline the overfitting by just resampling the data from the training data with same cardinality as in original data set. High variance is not good for the model. Bagging is very effective method for limited data, and by just using samples you are able to get estimate by aggregating the scores.

2. BOOSTING AND ADABOOST CLASSIFIER

Boosting is a ensemble method that is use to create a strong classifier using a number of weak classifier. Boosting is completed by creating a model training data set, followed by the building of second model that corrects the original model's errors. Boosting Models are added until the training set is accurately predicted.

AdaBoost= Adaptive Boosting

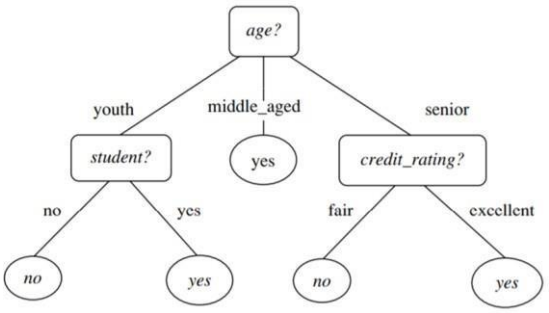


Fig.3. Decision Tree Structure

Decision tree Induction:

The building of "decision tree classifiers" doesn't need "any domain knowledge or parameter setting that is suitable for examining knowledge. "It handles multidimensional information. the learning and classification phases of decision

tree induction is simple and fast. Characteristic choice events are utilized to choose the characteristic that top parcel the tuple into particular classes. At the point when choice tree ismanufactured a significant number of the branches may result may reflect commotion and anomalies in the preparation information. tree pruning endeavors to recognize and evacuate such branches, with the objective of improving classifier precision on an inconspicuous information.

8. SCOPE OF STUDY

The project's suggested system will successfully detect spam emails, extract spam emails using machine learning techniques, and provide more accurate results. and good performance. This project required a coordinated scope of work.

These project scopes will help focus the project. The scopes are:

- Modifying existing machine learning algorithm.
- Use and classify data sets, including data preparation, classification, and visualization.
- Score the data to determine the accuracy of spam detection.
- This proposed system will detect the credibility of the mail and it will filter spam messages.
- This proposed system will save the time of the user and it will eliminate the risk of spam mails.

Flowchart of the model

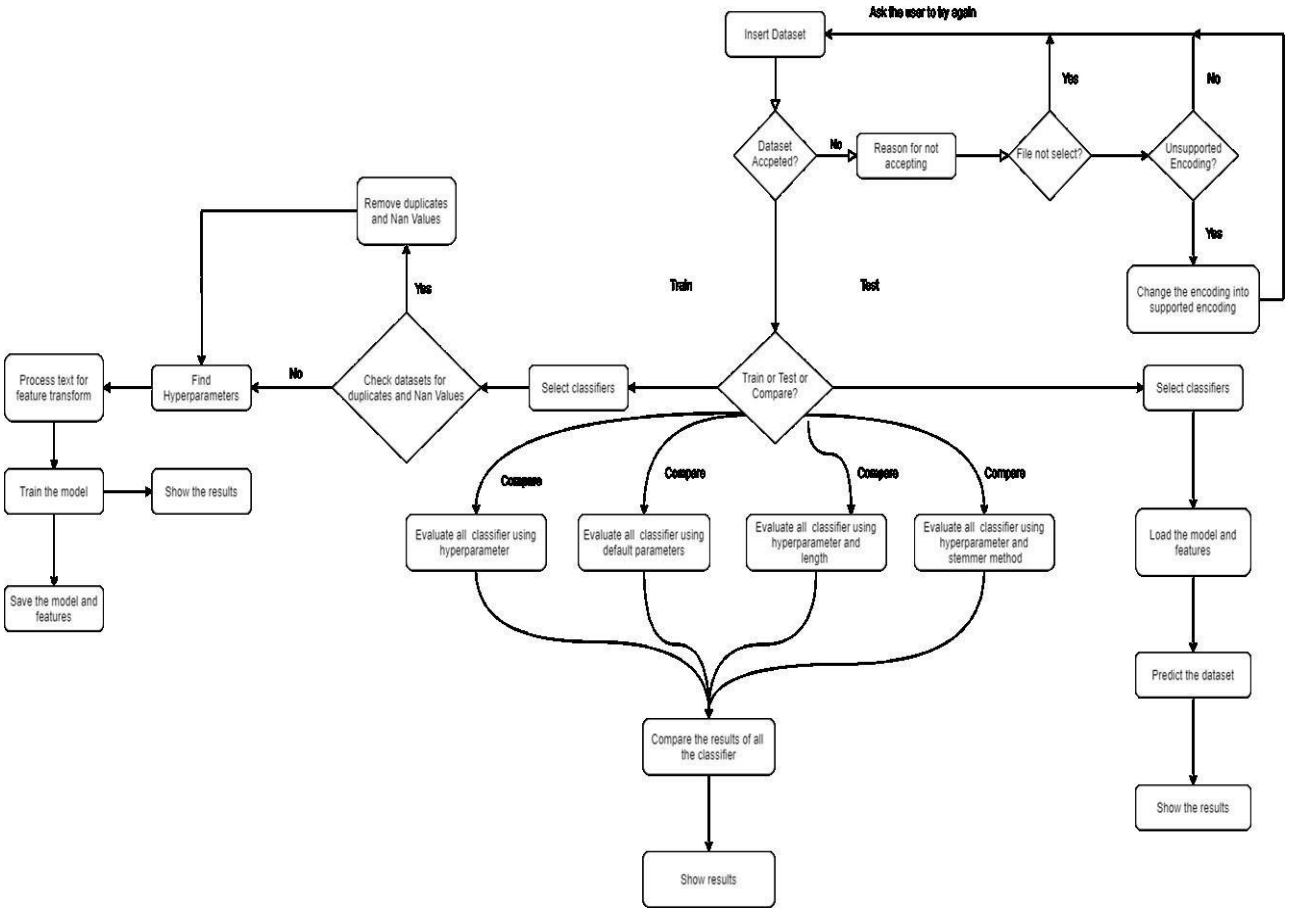


Fig.4. Flow Chart of Model

9. RESULT

To improve accuracy, we trained our model with various classifiers and compared the results. Each classifier will give its evaluated results to the user. After all the classifiers return its result to the user; then the user can compare it with other results to see whether the data is “spam” or “ham”. Each classifier result will be shown in graphs and tables for better understanding. The dataset originated from the "Kaggle" website for training. The name of the dataset used is “spam.csv”. To test the trained machine, a different CSV file is developed with unseen data i.e. data which is not used for the training of the machine; named “emails.csv”. After the text has been modified, the paper is ready for the template. Duplicate the template file using the Save As command, then name your paper according to the conventions established by your conference. Highlight the entire contents of this newly created file and then import the prepared text file. You are now ready to style your work; use the scroll down window to the left of the MS Word Formatting Toolbar.

10. ACKNOWLEDGEMENT

We would like to express our sincere gratitude and appreciation to Mr. Akhilesh Kumar Singh for his invaluable guidance, support, and expertise throughout the duration of this project. His extensive knowledge in the field of machine learning and email spam detection has been instrumental in shaping our understanding and approach.

We would also like to extend our heartfelt thanks to our fellow team members, Abuzar, Amit Verma, and Noman Ali, for their diligent efforts and collaboration. Their contributions, insights, and dedication have been crucial in the successful completion of this project.

This project would not have been possible without the collective effort and support of all those mentioned above.

Thank you for being an important part of this journey and making it rewarding and enriching.

TABLE I. COMP ARISION TABLE

	Classifiers	*Score 1*	*Score 2*	*Score 3*	*Score 4*
1	Support Vector Classifier	0.81	0.92	0.95	0.92
2	K-Nearest Neighbour	0.92	0.88	0.87	0.88
3	Naïve Bayes	0.87	0.98	0.98	0.98
4	Decision Tree	0.94	0.95	0.93	0.95
5	Random Forest	0.90	0.92	0.92	0.92
6	AdaBoost Classifier	0.95	0.94	0.95	0.94
7	Bagging Classifier	0.94	0.94	0.95	0.94

a. score 1: using default parameters

b. score 2: using hy perparameter tuning

c. score 3: using stemmer and hy perparameter tuning

d. score 4: using length, stemmer and hy perparameter tuning

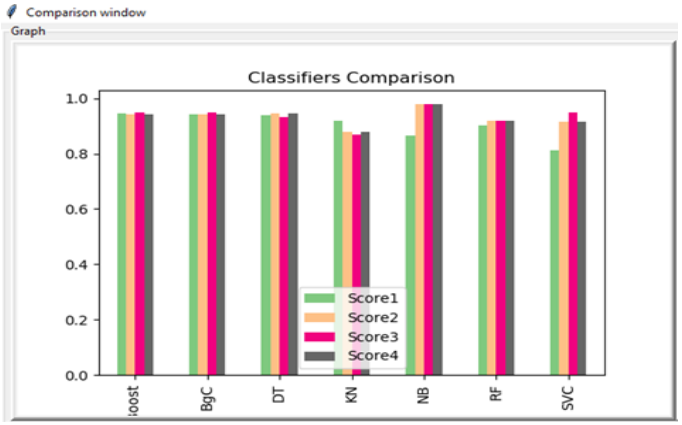


Fig.5 Comparison of all algorithms

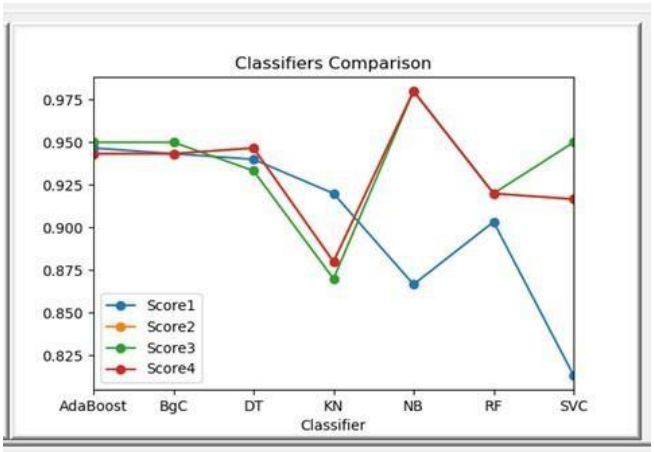


Fig.6. Comparison Graph

11. CONCLUSION

Ensemble approaches, on the other hand, have shown to be useful because they combine many classifiers for class prediction. Nowadays, a large number of emails are sent and received, making it difficult for our project to test emails with a limited corpus. Our project, therefore spam detection, is capable of screening emails based on the content of the email rather than domain names or other criteria. As a result, there is only a limited body to the email.

Our project has several potential areas for improvement. The following improvements can be made:

This strategy can be utilized by the large body to distinguish between legitimate emails and the emails they want to receive.

12. REFERENCES

1. Suryawanshi, Shubhangi & Goswami, Anurag & Patil, Pramod. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. 69-74. 10.1109/IACC48062.2019.8971582.
2. K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.
3. A. Sharaff, A. Dhadse and Naresh K. Nagwani (2016), Comparative study of classification algorithms for spam email detection, in Emerging Research in Computing, communication and applications, Information, pp. 237- 244, Springer, Berlin, Germany, DOI: 10.1007/978-81- 322-2553-9_23.
4. Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on, pp 153 -155. IEEE, 2014
5. Amin, Hossain N. & Rahman M. M., "A Bangla Spam Email Detection and Datasets Creation Approach based on Machine Learning Algorithms," 2019 3rd International Conference on Electrical, Computer & Telecommunication Engineering, Bangladesh, Rajshahi, 2019, pp. 169-172, DOI: 10.1109/ICECTE48615.2019.9303525.
6. A. Sharma & H. Kaur, Improved email spam classification method using integrated particle swarm optimization and decision tree. In Next Generation Computing Technologies 2nd International Conference on pp. 516- 521, DOI: 10.1109/NGCT.2016.7877470.
7. W.A, Awad & S.M, ELseuofi. (2011). Machine Learning Methods for Spam E-Mail Classification. International Journal of Computer Science & Information Technology. 3. 10.5121/ijcsit.2011.3112.
8. Adebayo A. Alli, Modupe Odusami, Olusola A. Alli and Sanjay Misra (2019), A review of soft techniques for SMS classification: methods, approaches and applications, Engineering Applications of Artificial Intelligence, DOI: 10.1016/j.engappai.2019.08.024.