**Ham-Spam Detection Using Machine Learning**

Jatin verma

Computer Science with Artificial Intelligence and Machine Learning Chandigarh University Mohali, Punjab

Shaveta Jain

Computer Science with AIT-CSE Department Artificial Intelligence and Chandigarh University, Machine Learning

Mohali, Punjab

***Abstract*—**Email communication is a fundamental part of our daily lives, and with its ubiquity comes the proliferation of spam emails. Spam emails are not only a nuisance but can also pose security risks and lead to a loss of productivity. In this research article, we explore the application of Machine Learning (ML)techniques for the detection of spam emails, commonly referred to as "spam" and legitimate emails, often referred to as "ham." We analyze the effectiveness of various ML algorithms in accurately classifying emails into these two categories and propose a robust model for efficient ham-spam email detection.

**Key Words: Natural Language Processing, NLTK, SpaCy, Term Frequency Inverse Document Frequency, Naïve Bayes, Corpus, Stopwords, Bag of Words**

1. INTRODUCTION

Spam emails are the most irritating enemy of privacy, sometimes called unsolicitted bulk emails are frequent in email components. These emails often include frauds, phishing, malware and others .It is a major problem to individuals and organisations because it involved money loss, system invasion and many more.

Ham emails are regular emails exchanged betwen individuals and organisation .The major and paramount issue is to discovery the difference between hamand spam emails so as achieving the goal of email exchange.

Through years of dealing with very large quantities of spam, sophisticated anti-spam techniques have evolved. Traditional anti-spam techniques, such as blacklisting and rule-based filtering, occasionally provided quite good results, but (for the same set of reasons) were not adaptive enough to address emerging spam attacks. This drove researchers to explore more advanced, more adaptive methods. One of the most promising approaches is Machine Learning.

The emerging ML, especially the supervised learning algorithms in this domain, has rendered quite promising solutions. A model targeting this can be trained using labeled datasets of ham and spam emails for automation in email filtering. This automation reduces the toil of manual effort in sorting out unwanted spam emails from a user's inbox, consequentially enhancing security and the user experience with emails.

The algorithms applied to the classification of emails using machines learning include Naive Bayes, Support Vector Machines, Decision Trees, and newer deep learning techniques such as Convolutional Neural Networks and Recurrent Neural Networks. Each algorithm has its strengths and weaknesses, where some are more effective in handling high dimensions and sparsity typical in text data as in emails.

The algorithms require great support from the methods of feature extraction that are done at the preprocessing stage for the representation of email data. They include techniques such as term frequency-inverse document frequency, bag of words, and word embeddings like Word2Vec and GloVe. These methods turn raw text into numerical vectors for machine-learning algorithms to use; as such, they enable the models to understand what the emails are about and hence classify them properly.

This research will, hence, focus on the evaluation of various machine learning algorithms' capabilities in classifying ham and spam emails effectively and on showing the efficiency of various techniques in the extraction of relevant features. An effective model would be proposed for the efficient detection of ham and spam emails. In such a way, we wish to make our own small contribution to the long-drawn campaign aimed at making email communications much more secure and quicker across the cyber environment. In particular, this research will indicate which combinations of algorithms with feature extraction methods will ensure the greatest accuracy and robustness of classification for spam versus legitimate emails.

Ultimately, these will pave the way to a strong, scalable, adaptive email classification system to sustain and match the pace of spammers, who work relentlessly to introduce innovative tactics to evade filters. It will bring efficiency in managing emails and security to the personal and organizational communications channels by reducing the spams email risks.

II. LITERATURE SURVEY

Spam and ham were detected in emails in the early epochs of information security@and electronic mail communication; machine learning techniques@have played a very focal role in mitigating this.

This paper on the literature survey makes a review of some major research works and developments in the area of ham-spam detection using machine learning over these years.

## *1*. Early Approaches

Rule-Based Systems and Heuristics:

The first stages of email filtering were ruled by heuristic systems and rule-based systems. These systems used the simple matching of keywords, regular expressions, and manually crafted rules to trigger probable spam emails. While such approaches provided basic filtering, they were, by design, limited in their@inability to dynamically learn the tactics of evolving spams.@In most cases, such rules were soon bypassed by spammers and resulted in a lot of false positives and negatives..

Limitations of Early Approaches:

The inherent@disadvantage of the rule-based systems was their@inflexibility. They had to be updated continuously and maintained to enhance their monitoring mechanisms against new spamming techniques. More importantly, most of the systems couldn't fully understand the intricacies of the language that human beings use, which led to considerable misclassification of ham as spam and vice versa.

*2.* Transition to Machine Learning

Paradigm Shift:

The arrival of machine@learning brought a paradigm shift in email@classification. The researchers started using ML algorithms@for automation of the detection process by allowing the systems to learn from the data and hence adapt accordingly with the@change in spam patterns, which brought immense improvement in accuracy and efficiency towards spam detection.

a) Naive Bayes:

VC One of@the very early applications of ML in spam detection was done@using Naive Bayes classifiers. Rennie et al. introduced the@use of Naive Bayes for spam detection in 2003, showing that@this probabilistic model was much more effective at high@accuracy when trained on a large dataset.

They realize the conditional independence of features quite well, and at the@same time, they are computationally efficient, hence@making them quite suitable for real-time spam filtering.

b.) Support Vector Machines (SVM):

Cristianini@and Shawe-Taylor (2000) applied Support Vector Machines@on spam filtering. One major property of SVMs is the ability to deal with high-dimensional data, and this would generally@make them very appropriate for text classification of emails. The strength of SVM lies in its capacity to create a clear margin of separation between spam and ham emails within a high-dimensional feature space. This would improve classifier precision and recall.

c.) Ensemble Methods:

In 2001, Elkan studied the application of ensemble methods like AdaBoost@in spam filtering. Ensemble methods work on the technique@of combining many weak classifiers to get one strong classifier with improved accuracy. This compensates for the failures of the base classifiers and gives a robust solution to the spam detection problem by gathering the advantages of a variety of models.

*3*. Feature Extraction:

Among the most influential elements for any ML

models' performance over email classification feature extraction techniques are. The features to be provided to the

model will drastically affect the accuracy and effectiveness of the classification model.

a) Bag of Words (BoW):

Sahami et al. (1998) proposed the Bag of Words representation of e-mails, where e-mails were described as vectors of word frequencies. The basic characteristics of BoW make this method foundamental for text-based classification due to its simplicity and effectiveness in rendering text data under a format convenient for the application of machine learning algorithms.

b.) Term Frequency-Inverse Document Frequency (TF-IDF):

Forman, 2004, discussed the effectiveness of Term Frequency-Inverse Document Frequency in spam detection. TF-IDF provides the functionality of weighting the words at the level of their importance against a single document and the whole corpus combined; it acts to a great degree of prominence of the right terms for classification. This approach enriches the vector representation of text data by considering the significance of the terms against the context of the whole dataset.

c) Word Embeddings:

Mikolov et al. (2013) proposed Word2Vec, a word embeddings learning technique. This process of word embedding will base its semantic information of words by representation in continuous vector spaces. Since this was applied to spam detection, it was expected that word embeddings would elevate the capabilities of classifiers and, therefore, undertake the comprehension of content within emails.

## 4. Challenges and Future Directions

### While machine learning has obtained very good results in ham-spam detection, there are still a number of areas open to improvement. Most of these challenges must be addressed in order to come up with better and more robust spam detection systems. This section touches on the main challenges in the area and discusses some possible future ways to surmount them.

### Imbalanced Datasets:

One of the challenging issues in the detection of spam is related to class imbalance. Normally, spam emails form a very minimal portion of the total dataset, while legitimate emails form a major part. This can introduce biases to the models toward the majority class, ham, which will end up having very poor detection rates on spam emails.

The challenge has been faced head-on by researchers who have resorted to different techniques:

The challenge has been faced head-on by researchers who have resorted to different techniques:

Oversampling: This method involves duplicating spam email instances to balance the dataset. While it increases the representation of the minority class, it can also lead to overfitting.

* Undersampling: This technique reduces the number of ham email instances to balance the classes.

Although it helps in addressing imbalance, it can result in loss of valuable information from the majority class.

* **Synthetic Data Generation**: Techniques like Synthetic Minority Over-sampling Technique (SMOTE) create synthetic spam instances based on existing ones. This approach helps in balancing the dataset without duplicating data, thereby reducing the risk of overfitting.

### b) Evolving Spam Tactics

### Spammers always find a way to improve their tactics in evading a detection system. It is this constant adaptation that really challenges the Spam detection model and forces it to evolve continuously. The spammers gan by tactics like -

### Obfuscation Techniques:

### Spammers use obfuscation techniques by inserting random characters or using similar looking symbols to avoid detecting by keyword-based techniques.

### Content Variation: A treaty bit of content variation in every spam mail makes it knowledge grasping for the detection models, which look for common patterns.

### New Attack Vectors: Spammers generally keep on coming up with new kinds of spam mails as a part of a continuous cycle. They at times can be sophisticated types of phishing attacks or malware-laden messages.

### These tactics will therefore continue requiring investment in research and development in the coming future to stay ahead with robust and adaptive models that can learn from new data and, thus, manage to notify new forms of spam. In this line, continuous model retraining combining with live data feeds can help in keeping the effectiveness of spam detection systems.

C) Multimodal Content

It is trending that emails nowadays contain multimedia, such as images, audio, and video, alongside text. This shift in trend challenges traditional models of spam detection, which were based on simple text analysis and hence could hardly cope with such diversified content types.

The researchers also consider several approaches to overcome the challenge:

* Picture Processing: The body of an e-mail may contain images over which techniques like OCR can be applied to extract text.
* Audio Analysis: In a similar vein, audio attachments can also be analyzed for spam detection by transcribing their content using speech-to-text technologies.
* Multimodal Learning: This would involve training a model capable of fusing and estimating information from more modalities, including text, images, and audio, to increase the overall accuracy and resilience of spam detection systems.

d) Privacy and Ethical Issues

While the machine learning models for spam detection are increasingly sophisticated, important concerns are inevitable: user privacy and ethical considerations. There could be a number of key issues here:

* Privacy of data: This is very important in having the best level of privacy and security in treating user data used for training and deploying models of Spam Detection. Some techniques that could be used toward keeping the information of users private include data anonymization and encryption.
* Model Bias: There could be biases within the machine learning models themselves. They could then discriminate between different types of emails or senders, which is unfair. Models should be developed with no bias, and their performance over time has to be monitored continuously to ensure that there is no discriminative behavior at all.

The long-term key research direction has to focus on developing privacy-preserving machine learning and considering the ethical standards while creating and deploying the spam detection system. Basically, this involves adhering to the data protection regulation, frequent audits in the performance of the model, and building fairness checks to prevent biased outcomes.

By pitting ourselves against the challenges and finding novel ways to creatively solve them, the development of ham-spam detection techniques will make it possible to have more complete and reliable email communication systems in the near future.

III. METHODOLOGY

3.1. Data Preprocessing

Before feeding data to machine learning algorithms, a number of critical preprocessing steps should be performed on the data to cleanse the text data for consistency and into a form ready for analysis.

They are:

Text Cleaning:

The raw email text is cleaned of unwanted elements: removing all HTML tags, special characters, and excessive whitespace. For example, an email with <html><body>Hello, user</body></html> HTML content is cleaned to simply "Hello, user!". This would ensure minimal noise and allow focusing on relevant content.

Tokenization:

Tokenization is when cleaned text is broken down

into words or tokens. For example, the sentence "Spam detection is essential" would be tokenized into ["Spam", "detection", "is", "essential"]. This step is very important in changing text into structured format for analysis by machine learning algorithms.

Stopwords removal:

Common words like "the," "and," "in" are usually

removed because, in the context of classification, they add little value to the meaning of the text. For example, the sentence "This is an example of an email" has a removal of the stop words which result in: ["This", "example", "email"]. Such stop-words are filtered to provide dimensionality reduction and bring out the best performance from classification algorithms.

Stemming or Lemmatization:

We further normalize the text by applying stemming or lemmatization. Stemming removes the suffixes from words and reduces them to their root form, and lemmatization reduces a word to its base form using vocabulary and morphological analysis. For example, "running" and "runner" would both be reduced to "run," which enables us to group similar words and increase consistency in our text data.

3.2. Feature Extraction

Feature extraction transforms the preprocessed text into numerical representations that can be fed into machine learning models. Several techniques are commonly used for this purpose:

Bag of Words (BoW):

#### The Bag of Words@model represents each email as a vector of word frequencies. For example, if the dataset contains the words ["spam", "email", "filter"], and an email has the text "spam email filter spam", the BoW representation might be [2, 1, 1], indicating the@frequency of each word in the email. This@technique captures the presence of important words.

#### Term Frequency-Inverse Document Frequency (TF-IDF):

#### TF-IDF enhances the Bag of Words model by assigning weights@to@words based on their importance in the document and@the entire corpus. For example, the word "spam" might appear frequently in all emails, while a unique word like "investment" appears only in some. TF-IDF assigns a higher weight to "investment" in an email where it appears, indicating its significance.

#### Word Embeddings:

#### We explore the use of pre-trained word embeddings such as Word2Vec or GloVe to capture semantic relationships between words.@For example, in Word2Vec, words like "email''@and "message" might be close together in the vector space, indicating their semantic similarity. These embeddings are@trained on large corpora and capture intricate semantic and syntactic relationships between words.

### 3.3. Model Selection and Training

### Selecting@and training the right machine learning model is crucial for effective spam detection. We experiment with a range of algorithms, each with its strengths:

#### Naive Bayes:

#### Naive@Bayes is a probabilistic classifier that applies Bayes' theorem with strong independence assumptions. For instance if the word "free" often appears in spam emails, Naive Bayes will consider an email with this word more likely to be spam. It is particularly effective for text classification tasks and is computationally efficient.

#### Support Vector Machines (SVM):

#### SVMs are powerful classifiers that find the optimal hyperplane separating different classes in a high-dimensional space. For example, if spam emails are often short and contain specific keywords, while legitimate emails are longer and more varied, SVMs can use these characteristics to classify emails. They are effective for handling large feature sets typical of text data.

#### Random Forest:

#### Random@Forest is an ensemble method that builds multiple decision trees and@merges their results to improve accuracy and robustness. For@example, one tree might classify an@email based on the presence of certain words, while another tree might use the length of the email. This algorithm is@particularly useful for handling complex datasets with many features.

#### Gradient Boosting:

#### Gradient Boosting builds models sequentially, with each new model correcting@errors made by the previous ones. For instance, if an initial model misclassifies some spam emails as legitimate, the next@model will focus more on those misclassified examples. This method is@highly effective in improving the performance of weak classifiers and can achieve high accuracy by focusing on difficult-to-classify cases.

For each algorithm, we split the dataset into training and testing sets@to evaluate their performance. We employ a variety of@evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to comprehensively assess the effectiveness@of each model.

### 3.4. Model Evaluation

To@ensure the robustness and generalizability of our models, we use@cross-validation techniques. Cross-validation involves dividing the dataset into multiple folds and training the model on@different subsets while testing on the remaining folds. For example, with 5-fold cross-validation, the dataset@is split into five parts, and the model is trained and tested five times,@each time using a different part as the test set. This process@helps in avoiding overfitting and provides a more reliable estimate of model performance.

Additionally, we perform hyperparameter tuning to optimize the models further. Hyperparameter tuning involves adjusting the@parameters of the machine learning algorithms to find the best combination that yields the highest performance. Techniques such as grid search or random search are commonly used to explore a range of hyperparameter values systematically. For instance, for an SVM, we might tune parameters like the kernel type and regularization parameter to improve its performance.

By following these meticulous steps in data preprocessing, feature extraction, model selection, and evaluation, we aim to develop a robust and accurate spam detection system capable of effectively distinguishing between ham and spam emails.

4. Results

Our experiments reveal that certain algorithms, such as Support Vector Machines and Gradient Boosting, outperform others in accurately classifying ham and spam emails. Additionally, using TF-IDF as a feature extraction technique tends to yield better results compared to BoW.

The performance metrics achieved by our best model are as follows:

|  |  |
| --- | --- |
| Accuracy | 98% |
| Precision | 97% |
| Recall | 99% |
| F1-score | 98% |
| ROC-AUC | 0.99 |

IV. CONCLUSION

This study demonstrates how machine learning algorithms can effectively discriminate between spam, or unsolicited emails, and ham, or valid communications. Our model's remarkable 98% accuracy rate indicates that it has the ability to significantly decrease the amount of manual labour needed to filter out unsolicited spam emails. By integrating such a system into email clients or servers, spam can be automatically removed, improving user productivity and email,,security.  
  
The findings show that using a variety of characteristics taken from the email content, machine learning is quite successful at recognizing spam emails. This automatic method lowers the dangers of harmful emails, including phishing and malware attacks, while also saving time.

Prospective Courses: Even though our model performs exceptionally well, there are a few more research avenues that could improve spam detection systems even more:  
  
Advanced Methods for Natural Language Processing (NLP)  
Investigating more complex NLP methods, such as sentiment analysis and contextual understanding, may aid in identifying sophisticated spam that evades conventional filters with

devious wording.

Models of Deep Learning Examining the application of deep learning models—like recurrent neural networks (RNNs) and convolutional neural networks (CNNs)—may help identify intricate patterns in email content that more straightforward algorithms might overlook. These models provide a more thorough comprehension of the text and are particularly helpful with big datasets.

Instantaneous Email Categorization Creating real-time email classification systems would allow for instantaneous identification of spam as soon as emails arrive, ensuring that inboxes are kept tidy and orderly. This calls for effective models that can process emails fast without compromising precision.  
  
Adjusting to Changing Spam methods continuous study is necessary to stay up to date with spammers' continually shifting strategies. Creating adaptable models that take in fresh information and identify spam trends is one way to do this.  
  
Managing,,Multimodal,,Content:   
 emails contain multimedia material (such as images, audio, and videos) more frequently, future research should concentrate on developing models that can efficiently handle and analyse a variety of content types in order to identify spam.

Ethics,and,Privacy:  
As machine learning models advance in complexity, it is imperative to tackle privacy and ethical issues. In order to preserve user confidence and comply with regulations, it is imperative that these models do not create biases or violate user,privacy.  
  
To sum up, this study highlights how machine learning may greatly increase email security and efficiency. We can create even more reliable and efficient spam detection systems by pursuing cutting-edge methods and adjusting to novel obstacles, ultimately fostering a more secure and efficient digital communication environment.

ACKNOWLEDGMENT

First and foremost, we extend our sincere thanks to our academic@advisors Kalpana Singh, whose guidance, expertise, and vcontinuous support were invaluable throughout this@research journey. Their insightful feedback and encouragement have greatly enriched our understanding of the subject matter. We are also deeply appreciative of the contributions made by the research community in the field of machine learning and email classification. The extensive body of knowledge@and research papers served as a solid foundation for our@work. We acknowledge@the pioneering researchers whose@efforts have paved@the way for advancements in spam@detection. Our gratitude@extends to Chandigarh University for providing us with access to resources, computational facilities, and a conducive

environment for research and learning. We are thankful to the library staff and IT support teams for their assistance in procuring relevant literature and data. We would like to acknowledge the efforts of our colleagues and fellow researchers who provided valuable insights and constructive feedback during discussions and presentations related to this research. Last but not least,@we wish to express our deepest appreciation to our families@and friends for their unwavering@support and understanding@during the course of this research@project. Their encouragement@and patience have been a constant source of motivation. In@conclusion, the successful completion of this research would not have been possible without the collective@efforts and support of the individuals and institutions@mentioned above. We remain indebted to them for their contributions, guidance, and encouragement.

REFERENCES

[1] Afzal and K. Mehmood, “Spam filtering of bi-lingual tweets using machine learning,” in Proceedings of the 2016 18th International Conference on Advanced Communication Technology (ICACT), pp. 710–714, IEEE, PyeongChang, Korea (South), Feb 2016.

[2] S. K. Tuteja and N. Bogiri, “Email spam filtering using bpnn classification algorithm,” in Proceedings of the 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), pp. 915–919, IEEE, Pune, India, Sep 2016.

[3] M. Mohamad and A. Selamat, “An evaluation on the efficiency of hybrid feature selection in spam email classification,” in Proceedings of the 2015 International Conference on Computer, Communications, and Control Technology (I4CT), pp. 227–231, IEEE, Kuching, Malaysia, Apr 2015.

[4] S. Suryawanshi, A. Goswami, and P. Patil, “Email spam detection:an empirical comparative study of different ml and ensemble classifiers,” in Proceedings of the 2019 IEEE 9th International.

[5] Conference on Advanced Computing (IACC), pp. 69–74, IEEE, Tiruchirappalli, India, Dec 2019.

[6] M. Mohamad and A. Selamat, “An evaluation on the efficiency of hybrid feature selection in spam email classification,” in Proceedings of the 2015 International Conference on Computer, Communications, and Control Technology (I4CT), pp. 227–231, IEEE, Kuching, Malaysia, Apr 2015.

[7] S. Suryawanshi, A. Goswami, and P. Patil, “Email spam detection:an empirical comparative study of different ml and ensemble classifiers,” in Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 69–74, IEEE, Tiruchirappalli, India, Dec 2019.

[8] Suryawanshi, Shubhangi & Goswami, Anurag &

[9] Patil, P ramod. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. 69-74.

[10] Karim, A., Azam, S., Shanmugam, B., Krishnan, K.,& Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access, 7, 168261-168295.

[11] K. Agarwal and T . Kumar, "Email Spam Det ection Using Integrated Approach of Naïve Bayes and P article Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.

[12] 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 ),

[13] 2014 Int ernational Conference on, pp.153 - 155.IEEE, 2014.

[14] Mohamad, Masurah, and Ali Selamat. "An evaluation on t he efficiency of hybrid featureselection in spam email classification."

[15] In Computer, Communications, and Control Technology (I4CT), 2015 International Conferenceon, pp. 227-231. IEEE, 2015

[16] Shradhanjali, Prof. Toran Verma “E-Mail Spam Detection and Classification Using SVM and FeatureExtraction "in International Journal Of Advance Reasearch, Ideas and Innovation In Technology,2017ISSN: 2454-132X Impact fact or: 4.295

[17] W.A, Awad & S.M, ELseuofi. (2011). Machine Learning Methods for Spam E-Mail Classification.International Journal of Computer Science & Information Technology. 3.

[18] 0.5121/ijcsit.2011.3112.

[19] K. Ameen and B. Kaya, "Spam detection in online social networks by deep learning," 2018 Internat ional Conference on Artificial Int elligenceand Data Processing (IDAP ), Malatya, Turkey, 2018, pp. 1 -4