Ham-Spam Detection Using Machine Learning

Bhudil Mallick
Computer Science with Artificial Intelligence and Machine Learning
Chandigarh University
Mohali, Punjab
bhudil.mallick@gmail.com

Kalpana Singh
E14950
AIT-CSE Department
Chandigarh University,
Mohali, punjab

Kiruthik Ranga S.V
Computer Science with Artificial Intelligence and Machine Learning
Chandigarh University
Mohali, Punjab
kiruranga10@gmail.com

***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 toas "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

I. INTRODUCTION

Spam emails, or unsolicited bulk emails, have been a persistent problem since the early days of email communication. They often contain fraudulent offers, phishing attempts, malware, or other malicious content. On the other hand, ham emails represent legitimate communication between individuals or organizations. The challenge lies in differentiating between these two types of emails accurately.

Machine Learning, particularly supervised learning algorithms, has shown promising results in the field of email classification. By training a model on a labeled dataset of ham and spam emails, we can develop a system capable of automating the process of email filtering.

In this study, we aim to:

Investigate the performance of various machine learning algorithms for ham-spam email classification.

Evaluate the effectiveness of different feature extraction techniques. Propose a comprehensive model for efficient ham-spam email detection.

II. LITERATURE SURVEY

The detection of spam (unsolicited) and ham (legitimate) emails has been a long-standing challenge in the field of information security and email communication. Machine Learning (ML) techniques have played a pivotal role in addressing this problem.

This literature survey explores key research works and developments in ham-spam detection using machine learning over the years.

1) Early Approaches

Early approaches to ham-spam detection primarily relied on rule-based systems and heuristics. These systems often used simple keyword matching and regular expressions to identify spam emails.

While they provided basic filtering, they were limited in their ability to adapt to evolving spam tactics.

2) Transition to Machine Learning

Machine learning brought a paradigm shift to email classification. Researchers started using ML algorithms to automate the detection process, enabling systems to learn from data and adapt to changing spam patterns.

a) Naive Bayes:

- Rennie et al. (2003) introduced the idea of using Naive Bayes classifiers for spam detection. Their work demonstrated that Naive Bayes can achieve high accuracy when trained on a large dataset.

b) Support Vector Machines (SVM):

- Cristianini and Shawe-Taylor (2000) applied SVMs to spam filtering. SVMs are known for their ability to handle high-dimensional data, making them suitable for text-based email classification.

c) Ensemble Methods:

- Elkan (2001) explored the use of ensemble methods, such as AdaBoost, for spam filtering. Ensemble methods combine multiple weak classifiers to create a strong classifier, improving overall accuracy.

3) Feature Extraction

Feature extraction techniques have a significant impact on the performance of ML models in email classification.

a) Bag of Words (BoW):

- Sahami et al. (1998) proposed using the Bag of Words model, where emails are represented as vectors of word frequencies. BoW remains a foundational technique in text-based classification.

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

- Forman (2004) discussed the effectiveness of TF-IDF in spam detection. TF-IDF assigns weights to words based on their importance in a document relative to a corpus.

c) Word Embeddings:

- Mikolov et al. (2013) introduced Word2Vec, a technique for learning word embeddings. Researchers have since applied word embeddings to capture semantic information in email content.

4) Challenges and Future Directions

While significant progress has been made, challenges persist in the field of ham-spam detection using machine learning.

a) Imbalanced Datasets:

- Class imbalance remains a challenge, as spam emails often constitute a minority of the dataset. Techniques like oversampling, undersampling, and synthetic data generation are explored to address this issue.

b) Evolving Spam Tactics:

- Spammers continuously adapt their tactics. Staying ahead of these tactics requires ongoing research and the development of robust models capable of detecting new forms of spam.

c) Multimodal Content:

- With the inclusion of multimedia content in emails (e.g., images, audio), researchers are exploring approaches that can handle diverse content types.

d) Privacy and Ethical Considerations:

- As ML models become more sophisticated, concerns about user privacy and the potential for bias in classification models need to be addressed.

III. Methodology

3.1. Data Preprocessing

Before feeding the data into machine learning algorithms, we perform several preprocessing steps:

Text Cleaning: We remove HTML tags, special characters, and excessive whitespace.
Tokenization: We split the text into individual words or tokens.
Stopword Removal: Common words like "the," "and," "in," etc., are removed as they often do not contribute to classification.
Stemming or Lemmatization: We reduce words to their base or root form to normalize the text.

3.2. Feature Extraction

To represent the email content in a format suitable for machine learning, we extract features from the text. Commonly used techniques include:

Bag of Words (BoW): This technique represents each email as a vector of word frequencies.
Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF assigns weights to words based on their importance in the corpus.
Word Embeddings: We explore the use of pre-trained word embeddings like Word2Vec or GloVe to capture semantic relationships between words.

3.3. Model Selection and Training

We experiment with a range of machine learning algorithms, including but not limited to:

Naive Bayes
Support Vector Machines (SVM)
Random Forest
Gradient Boosting

For each algorithm, we split the dataset into training and testing sets and evaluate performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

3.4. Model Evaluation

We assess the models' performance using cross-validation techniques to ensure robustness and avoid overfitting. Additionally, we fine-tune hyperparameters to optimize performance further.

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

A This research demonstrates the effectiveness of machine learning algorithms in ham-spam email detection. With an accuracy of 98%, our proposed model can significantly reduce the burden of sorting through unwanted spam emails manually. Such a system can be integrated into email clients or servers to automatically filter out spam, enhancing email security and user productivity.

Future work in this area could involve exploring more advanced Natural Language Processing (NLP) techniques, deep learning models, and real-time email classification systems. Additionally, ongoing efforts are needed to adapt to evolving spam techniques and email content.

Acknowledgment

First and foremost, we extend our sincere thanks to our academic advisors Kalpana Singh, whose guidance, expertise, and continuous 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.