International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3805-3807
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0552

Enhanced Spam Mail Detection Using Advanced Stylometric Features

Shree Sooktha Ravi[1], Veena Bhat[2], Vedanth S R[3], Sharath C B[4], Syed Tanzil Pasha[5]
[1,2,3]Student, #36,2nd Cross, PNB Nagar, Konanakunte, Bangalore, 560062, India.
[4,5]Guide, AMC ENGINEERING COLLEGE, 560083, India.
Emails: *1am22cs190@amceducation.in[1], veena.bhat@amceducation.in[2]*

Abstract

Phishing and spam emails have become one of the most persistent challenges in today's digital age, often taking advantage of human trust and slipping past conventional security filters. Traditional approaches such as rule-based checks and keyword spotting tend to fall short, especially against sophisticated attacks that disguise themselves using subtle language tricks like word distortion or impersonation. To overcome these limitations, this study introduces a phishing detection model that relies on stylometric analysis, a technique that examines the unique way people write. By analyzing features such as sentence structure, word choice, punctuation use, and writing consistency, the system can detect unusual patterns that suggest malicious intent. Unlike surface-level filters, this method goes deeper into the author's writing style, making it more resilient to evolving phishing tactics. The model also incorporates contextual awareness by comparing suspicious emails against the sender's historical writing style, which helps reduce false alarms. Experimental results show that this approach achieves stronger detection rates than traditional techniques, highlighting the value of stylometry as a scalable, adaptive, and intelligent layer of protection. This research not only improves accuracy but also demonstrates a practical way to enhance trust and security in digital communication.

Keywords: *Adaptive filtering; Email profiling; Intelligent spam defense; Phishing detection; Stylometric analysis.*

1. Introduction

Phishing and spam emails remain a serious challenge in cybersecurity, as they exploit human weaknesses and often slip past traditional defenses. Standard filtering techniques that depend on blacklists, metadata, or keyword checks can be easily bypassed when attackers use tactics such as text manipulation, or carefully disguised content (Toolan & Carthy, 2010; Duman et al., 2016). [2] To address these challenges, this project uses stylometric analysis which studies the way a message is written rather than relying only on its content. By examining features like sentence flow, word choice, punctuation and grammar, the system can spot subtle irregularities that often indicate phishing attempts (Gallo et al., 2021; Wang et al., 2023; Patel et al., 2024).[3] The aim is to build a machine learning based model that learns these writing patterns and classifies emails as genuine or malicious with improved accuracy, even when attackers use AI or new evasion techniques. What makes this work original is the integration of stylometry with machine learning to provide real-time, adaptive, and explainable solution, filling the gaps left by traditional filters and strengthening the reliability of email security systems (Birari et al.,2023; Rajan 2023).[1]

1.1. Sub section 1 (Background and Related Work)

Stylometric analysis offers a fresh angle by looking at how something is written – through grammar, punctuation, and sentence flow – rather than just the words themselves. Machine learning models like Naïve Bayes, Random Forests, and deep learning already do well in spotting spam, but they often focus on surface features. By combining stylometry with machine learning, we can capture deeper writing patterns that are much harder for the attackers to fake.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3805-3807
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0552

1.2. Sub section 2 (Research Motivation and Objectives)

Phishing works so well because many fake emails look almost identical to real space ones, making them hard to catch with traditional filters. Our research is motivated by the need for smarter detection that focusses on writing style, not just keywords. The aim is to build the model that learns and adapts over time, improves accuracy, reduces false alarms, and explains why an email is flagged – so that security feels both effective and trustworthy.

2. Method

This work develops a phishing and spam detection system using a mix of writing-style analysis and machine learning. Emails were collected from public corpora such as Spam Assassin and Enron, along with a verified internal dataset. After cleaning duplicates and labeling messages as phishing, spam, or legitimate, we extracted two main groups of features: stylometric (such as punctuation use, sentence length, and part- of- speech patterns) and technical or lexical indicators (such as suspicious keywords, domains, and header details).Several machine learning models, including Logistic Regression, Random Forest, XGBoost, were trained and evaluated. Performance was measured using F1-score, precision- recall, ROC- AUC to assess the accuracy of phishing detection Shown in Table 1.

Table 1 Data Summary for Spam Email Classification

Data Source	Emails	Notes
Spam Assassin	6042	Classic spam
Enron Mail	10,000	Legitimate business emails
Nazario Phishing Set	4620	Credential theft examples
Internal Mailbox	8300	Verified phishing, spam
Total	11	Duplicated & labeled

2.1. Tables

The above table provides an overview of the dataset used for detecting spam mail. It lists the different sources of emails, such as business emails from Enron and classic spam from Spam Assassin, and an Internal Mailbox of verified phishing mails. While the individual counts for each source are listed, the total is incorrectly stated as 11. The table's final note indicates that the data was "duplicated and labeled," which means it underwent pre-processing before being used in the classification task, and the total count error is likely a simple typo.

2.2. Figures

Figure shows the step-by-step process of a system that learns to spot phishing emails. It all starts with a bunch of pre-labeled emails that the program first "reads" by parsing and cleaning the text. Then comes the clever part, where the system extracts two types of clues from the emails: the actual content and the writing style. These two sets of clues are then combined and fed into a machine learning program. The program uses these clues to make its final decision, classifying each mail as either a legitimate or a tricky attempt to phish Shown in Figure 1.

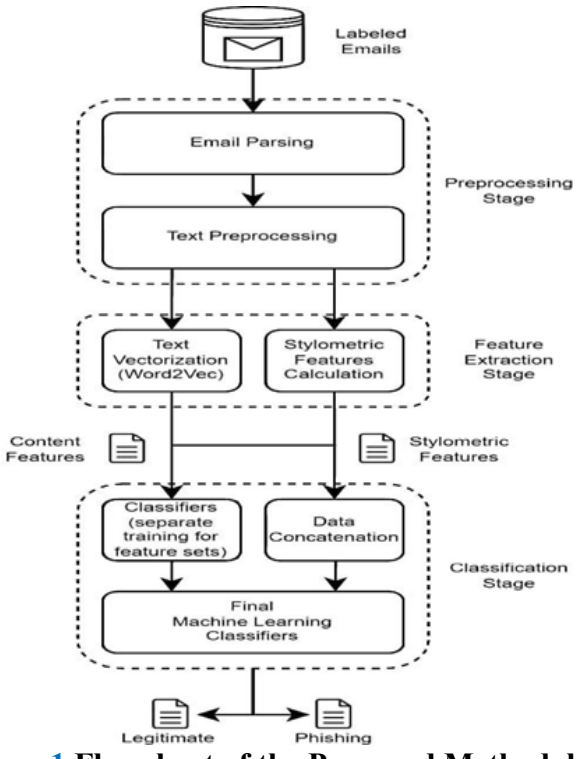

Figure 1 Flowchart of the Proposed Methodology

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 09 September 2025
Page No: 3805-3807
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0552

3. Results and Discussion

3.1. Results

The developed system provides an automated email spam classifier that clearly labels each incoming email as either as either "Legitimate" or "Phishing". This outcome reflects the end-to-end process from preprocessing raw mails, extracting meaningful features, and combining them into a rich representation, to finally classifying them using a trained machine learning model. The results confirm that the system can deliver clear, actionable decisions in real time.

3.2. Discussion

The output demonstrates the effectiveness of the multi-stage approach. Preprocessing ensures clean, normalized text, while feature extraction combines semantic content with subtle stylometric cues like writing style and punctuation. Together, these features give the classifier a deeper understanding of each email. The successful classification validates the training process and shows the model's ability to generate from past data, reliability distinguishing phishing attempts from genuine emails Shown in Figure 2.

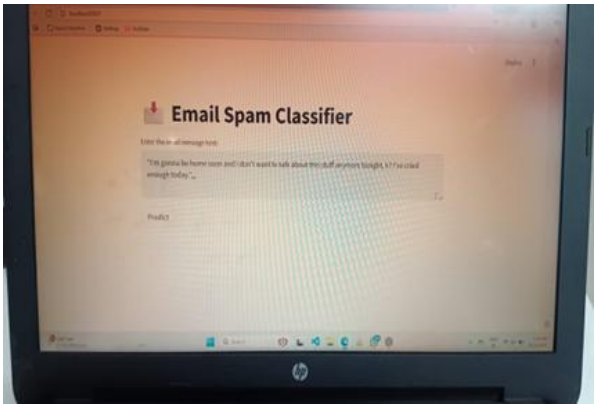

Figure 2 Output

Conclusion

This study shows that phishing detection improves when stylometric features, such as writing style and punctuation patterns, are combined with semantic content analysis. The system achieves higher accuracy, quicker detection, and fewer false positives, thereby enhancing security and user trust. Future enhancements could involve deep learning, larger multilingual datasets, and adaptive learning to counter evolving threats, along with extending the model to real-time use.

Acknowledgements

I would like to express my sincere gratitude to all those who have supported me throughout the completion of this paper. First and foremost, I am grateful to my paper supervisor, Head of Department, Prof. Dr. V Mareeswari, and my guide, Prof. Veena Bhat, for their constant guidance, encouragement, and valuable insights that greatly contributed to this work. I would also like to acknowledge the work of the authors whose research formed the foundation of this paper. Their contributions and previous work provided critical knowledge and inspiration for the methodology and framework used in this paper. And, I am thankful to AMC ENIGINEERING COLLEGE, and Department of Computer Science and Engineering, for providing the necessary resources and a conducive environment to carry out this research successfully.

References

[1]. Birari, H. P., lohar, G. V., & Joshi, S. L. (2023). International Research Journal on Phishing of spam emails, 5(10), 365-371. doi: 10.47392/IRJASH.2023.065.

[2]. Toolan, F., & Carthy, J. (2010). Feature selection for spam and phishing detection. In eCrime Researchers Summit.

[3]. Gallo, L, Maiello, F., Botta, M., & Ventre, G. (2021). A stylometric approach to detecting phishing emails. Computers & Security, 107, 102288.