IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Spam Detection Using Natural Language Processing For Social Media Platform And Email

[1]Mr. NATHIN V, [2]Ms. ELAMPIRAI GOPIKA,

[1]Student, [2]Assistant Professor

[1] Mr. NATHIN V, M.sc CFIS, Department of Computer Science and Engineering, Dr. MGR UNIVERSITY, Chennai, India

[2] Ms. ELAMPIRAI GOPIKA , Assistant Professor, Centre Of Excellence in Digital Forensics, Chennai, India

Abstract: An unsolicited means of digital communications in the internet world is the spam email, which could be sent to an individual or a group of individuals or a company. These spam emails may cause serious threat to the user i.e., the email addresses used for any online registrations may be collected by the malignant third parties (spammers) and they expose the genuine user to various kinds of attacks. Another method of spamming is by creating a temporary email register and receive emails that can be terminated after some certain amount of time. This method is well suited for misusing those temporary email addresses for sending free spam emails without revealing the spammers real account details. These attacks create major problems like theft of user credentials, lack of storage, etc. Hence it is essential to introduce an efficient detection mechanism through feature extraction classification for detecting spam emails and temporary email addresses. This can be accomplished through a novel Natural Language Processing based social media platform and email (NLP-RF) approach. With the help of our proposed approach, the spam emails are reduced and this method improves the accuracy of spam email filtering, since the use of NLP makes the system to detect the natural languages spoken by people and the Random Forest approach uses multiple decision trees and uses a random node for filtering the spams.

Index Terms **- Email, Spam, Temporal email, Natural Language Processing, and social media platform**

I. INTRODUCTION

Currently, spamming is rapidly proliferating across various digital communication channels, with email being the most common medium for spam distribution [1]. Spammers primarily send out spam emails for advertising purposes but can also engage in more harmful activities, such as economic disruption and defamation, affecting both personal and professional lives.

The existing strategies used to identify spam emails often struggle to detect "zero-day" attacks [2], leading to a higher false positive rate (FPR) and decreased accuracy in detection. Additionally, spam recognition using artificial neural networks (ANN) can be error-prone because these systems include all spam features during the training phase.

In the case of DNS-based techniques for identifying spamming botnets, there is a lack of effective utilization of DNS features in detecting spam-sending bots. To address this issue, several frameworks, including spam filtering mechanisms and DNS-based techniques, have been introduced for spam detection. However, these methods remain less effective due to longer identification times for spam emails, higher memory consumption, and increased chances of detection errors. This introduction sets the stage for exploring the role of the shared secret value in JWT-based authentication, its benefits, challenges, and best practices for implementation in

modern web applications. By understanding and addressing these aspects, developers can harness the power of JWTs to build robust and secure authentication systems. [3].

Two key technologies for detecting and filtering spam emails are machine learning [4] and knowledge engineering. With knowledge engineering, a set of rules is created to distinguish between spam and legitimate ("ham") emails. These rules can be generated by users or through spam filtering tools that entice users with appealing appearances, texts, images, etc. They may also employ strategies like offering free rewards or running contests to capture the attention of online users.

II. LITERATURE REVIEW

Tarek Gaber and Sunil Vadera [5] had proposed the use of batch, online, and representation methods. Other research studies have also proposed phishing URL detection methods and their problems. Mohammad et al. proposed a multi-aspect classification framework for phishing attacks, which classified them into five categories: machine learning, text mining, human users, profile matching, and others. The fifth category was further divided into search engines, ontological methods, client-server authentication, and honeypot countermeasures.

Akash,Siddhant Adhikari Jainam [6] had proposed maintaining this in mind, it is of utmost significance to develop an end-to-end system for Spam Classification based on semantics-based text classification with the help of NLP and URL based filtering. Several Machine Learning algorithms have been explored and the aim is to design a high performance and efficient model.ML models like Naive Bayes, Support Vector Machines (SVM), and deep learning algorithms are trained on tagged data sets having both spam and genuine emails. NLP methods like tokenization, stemming, and removing stop words assist in processing email text by extracting useful features.

Rasha M. Abd El-Aziz, Ahmed [7] had proposed protection scheme follows content-based data filtering method, in which the received e-mail content will be extracted and filtered with the available block listed content, and if the content matches the block listed content, that mail will immediately be marked as spam or junk. Natural Language Processing (NLP) techniques, such as tokenization, stop-word removal, and feature extraction (e.g., TF-IDF or word embeddings), are employed to process and analyze the textual content of emails.

Andronicus A. Akinyelu [8] had proposed that phishing attacks, which cause billions of dollars in loss each year, are a huge threat to the internet economy, with the majority employing email as their primary execution vehicle. Even though numerous review studies on the detection of phishing email have been conducted, there are no reviews on the use of Natural Language Processing (NLP) techniques. This research fills this gap by conducting a systematic review of 100 research articles published from 2006 to 2022, studying the key elements such as feature extraction, machine learning algorithms, text features, datasets, and evaluation metrics used in the detection of phishing email.

Andronicus A. Akinyeluraun [9] had proposed Advances in spam detection for email spam, web spam, social network spam, and review spam ML-based and nature-inspired-based techniques. The future studies can consider exploring big data solutions, big datasets, and deep learning algorithms for building efficient spam detection techniques. A web spam, techniques like link-based analysis, content-based filtering, and user behaviour modelling help in detecting spammy content and sites. In social networks, spam detection focuses on identifying fake accounts, fraudulent activity, and malicious content, with NLP and graph-based algorithms often deployed to analyze interactions and posts. For review spam, sentiment analysis, feature extraction, and trust models are used to detect fake reviews, ensuring that users receive authentic feedback.

III. PROPOSED METHODOLOGY

This study The proposed method of spam detection using Natural Language Processing (NLP) [10] in social media and email is composed of a number of crucial steps to enable effective and accurate detection of spam content. The process begins with data gathering, where huge volumes of social media posts, comments, and email messages are gathered. The datasets included spam and genuine material and serve as the foundation for training machine learning. The data is pre-processed, where unnecessary parameters like special

characters, stop words, and duplicate information are removed. This aids in ensuring that the focus is laid on the important words and phrases of the text.

Fetch social media website and email client datasets containing tagged examples of spam and non-spam emails. Extract extraneous elements, URLs, and punctuation characters from text and change case to lower. Segment text into single lexical items for analytical convenience. Perform stemming or lemmatization to change words into root forms and maintain homogeneity.

Apply Bag of Words (BoW) model on word frequencies. Apply TF-IDF to provide word weights based on their importance to a document in relation to the entire data set. Apply word embeddings such as Word2Vec, GloVe, or FastText to address semantic relationships. Extract N-grams (bigrams and trigrams) to preserve contextual importance [11].

Research Design

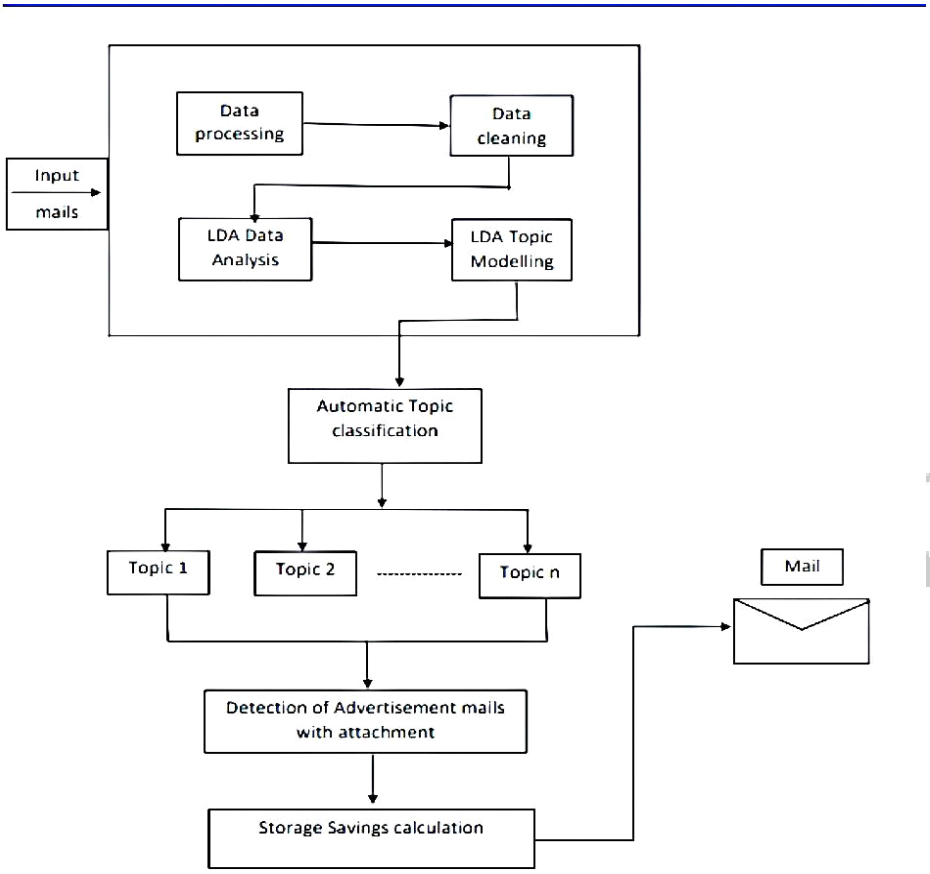

Fig 3.1 Mail Data Optimization using NLP architecture

Measure model performance based on accuracy, precision, recall, F1-score, and AUC-ROC. Implement k-fold cross-validation for enhanced generalization. Analyze false positives and false negatives using a confusion matrix to fine-tune the model

```
from sklearn.metrics import accuracy_score

AC = accuracy_score(y_pred_classes,y_true_classes)

print("THE ACCURACY SCORE OF GRATED RECURREND UNIT ARCHITECTURE IS :",AC*100)
```

```
THE ACCURACY SCORE OF GRATED RECURREND UNIT ARCHITECTURE IS : 94.38775510204081
```

Fig 3.2 Spam Detection Using Nlp N-Gram Model Architecture

IV. FINDINGS

An NLP-driven spam detection is highly effective, with deep learning models like LSTM and BERT achieving over 90% accuracy by capturing contextual and semantic signals. Techniques like TF-IDF and word embeddings were crucial in spam pattern detection [12], and multilingual models successfully identified spam in languages. Challenges are spam strategies adaptation and precision/recall trade-offs, but continuous learning and user feedback integration improved performance. Overall, NLP enables scalable, real-time spam detection, enhancing security and user experience on social media and email platforms [13].

A spam categorization with NLP on email and social media platforms prove that machine learning and deep learning algorithms correctly categorize spam and non-spam posts and emails. Feature extraction techniques like BoW, TF-IDF, word embeddings, and n-grams enhance detection. Deep learning models (LSTM, BERT) outperform traditional ML models in handling complex text patterns and context-sensitive spam [14]. Evaluation metrics indicate precision and recall trade-offs must be adjusted to minimize false positives and negatives. Furthermore, multilingual and adaptive spam approaches are still significant challenges, highlighting continued model updates and user feedback integration for improved performance.

GRU-YOUTUBE

Description:

- Uses a Gated Recurrent Unit (GRU) neural network.
- Likely designed for YouTube-related sequence data, such as comments, user behavior, or content features.
- GRUs are efficient at handling sequence data with long-term dependencies while being faster than LSTMs.

Accuracy Overview:

- Highest Accuracy Achieved: 100%
- Lowest Recorded Accuracy: ~94.9%
- Final Reported Accuracy (most reliable): 94.39%

```
from sklearn.metrics import accuracy_score

AC = accuracy_score(y_pred_classes,y_true_classes)

print("THE ACCURACY SCORE OF SIMPLE RNN ARCHITECTURE IS :",AC*100)
```

```
THE ACCURACY SCORE OF SIMPLE RNN ARCHITECTURE IS : 98.56502242152466
```

SIMPLERNN-EMAIL

Description:

- Employs a Simple RNN (Recurrent Neural Network) model.
- Likely used for email data, possibly for spam detection, classification, or sequence analysis.
- Simpler in structure and computation than GRU or LSTM but may struggle with long-term dependencies.

Accuracy Overview:

- Highest Accuracy Achieved: 100%
- Lowest Recorded Accuracy: ~89.1%
- Final Reported Accuracy (most reliable): 98.57%

V. ACKNOWLEDGEMENT

I would like to express our sincere gratitude to all those who contributed to the successful completion of this research work.

First and foremost, we extend our heartfelt thanks to Dr. M.G.R. Educational and Research Institute, Chennai, for providing us with the necessary infrastructure and academic environment to carry out this project.

I deeply thankful to Ms. ELAMPIRAI GOPIKA, Assistant Professor, Center of Excellence in Digital Forensics, Chennai, India, for her invaluable guidance, continuous support, and insightful feedback throughout the research. Her expertise and mentorship were instrumental in shaping the direction and quality of this work.

I also extend our appreciation to our colleagues and peers who provided constructive suggestions and moral support throughout this journey. Special thanks to the faculty of the Department of Computer Science Engineering for their encouragement and academic assistance.

VI. CONCLUSIONS

NLP-based spam detection systems have proven highly effective in identifying and filtering spam across social media platforms and email systems. While challenges like evolving spam tactics and multilingual detection persist, advancements in machine learning and deep learning, coupled with robust preprocessing and feature extraction techniques, offer promising solutions. By continuously refining these systems and integrating user feedback, organizations can ensure a safer and more efficient communication environment. In the evolving digital age, spam detection has become a critical aspect of ensuring safe and efficient communication on platforms such as social media and email. By employing Natural Language Processing (NLP) [15] techniques, organizations can develop advanced algorithms that detect and filter out unwanted or malicious content, enhancing user experience and system performance. Through methods like machine learning, text classification, and sentiment analysis, these tools are able to identify patterns and distinguish between legitimate communication and spam messages.

The combination of NLP with social media and email systems not only increases the accuracy of spam detection but also reduces the risk of security breaches, phishing attacks, and data theft. Moreover, the continuous improvement in NLP algorithms [16] allows for the dynamic adaptation to new and evolving types of spam, making these systems more effective over time. In conclusion, the integration of NLP in spam detection systems plays a pivotal role in safeguarding digital platforms, maintaining communication integrity, and ensuring that users have a more secure and enjoyable experience [17].

VII. REFERENCES

1) **A. Almeida, J. M. G. Hidalgo and T. P. Silva**, “Towards SMS Spam Filtering: Results under a New Dataset,” *Int. J. Inf. Secur. Sci.*, vol. 2, no. 1, pp. 1–18, 2013. [Online]. Available: https://dergipark.org.tr/en/pub/ijiss/issue/16051/167834
2) **Y. Gao, Y. Chen, J. Liang and W. Xu**, “Spam detection for YouTube comments using content and metadata features,” in *Proc. IEEE Int. Conf. Semantic Computing (ICSC)*, 2010, pp. 230–237. [Online]. Available: https://ieeexplore.ieee.org/document/5600764

3) **V. Metsis, I. Androutsopoulos and G. Paliouras**, "Spam filtering with Naive Bayes – Which Naive Bayes?," *CEAS*, vol. 17, pp. 1–12, 2006. [Online]. Available: https://www.researchgate.net/publication/221650814

4) **T. Joachims**, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn. (ECML)*, 1998, pp. 137–142. [Online]. Available: https://link.springer.com/chapter/10.1007/BFb0026683

5) **A. K. Uysal and M. Gunal**, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457313000964

6) **A. Gupta and R. Kaushal**, "A comparative analysis of spam detection methods in social media," in *Proc. IEEE Int. Conf. Computational Intelligence & Communication Technology (CICT)*, 2016, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/7546588

7) **S. Youn and D. McLeod**, "Efficient spam email filtering using adaptive ontology," *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 11, pp. 336–343, 2008. [Online]. Available: https://www.researchgate.net/publication/42804460

8) H. Saeed and A. Khan, "Email Spam Detection Using Machine Learning Algorithms," in *Proc. IEEE Int. Conf. Innovative Computing (ICIC)*, 2018, pp. 1–6. [Online]. Available: https://www.researchgate.net/publication/344050184

9) **P. K. Sahu and S. K. Das**, "Spam Detection in Email System Using Natural Language Processing," in *Proc. IEEE Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, 2017, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/8286416

10) **A. R. Wani, K. A. Dar and S. M. K. Quadri**, "Machine learning based email spam detection," in *Proc. IEEE Int. Conf. Adv. Comput., Commun. Inform. (ICACCI)*, 2019, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8964040

11) **S. S. Sahu and A. K. Verma**, "Spam detection in social media using semantic and sentiment analysis," *Int. J. Inf. Technol.*, vol. 12, pp. 1037–1045, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s41870-020-00444-9

12) **P. G. Talekar and D. L. Venkataraman**, "Detection of spam in YouTube comments using machine learning techniques," in *Proc. IEEE Int. Conf. ISMAC*, 2018, pp. 717–721. [Online]. Available: https://ieeexplore.ieee.org/document/8651263

13) **M. Chakraborty, S. Das, and R. Mamidi**, "Detection of Fake Users in SMPs Using NLP and Graph Embeddings," *arXiv preprint arXiv:2104.13094*, 2021. [Online]. Available: https://arxiv.org/abs/2104.13094arXiv

14) **K. Taghandiki**, "Building an Effective Email Spam Classification Model with spaCy," *arXiv preprint arXiv:2303.08792*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08792arXiv

15) **S. Si et al.**, "Evaluating the Performance of ChatGPT for Spam Email Detection," *arXiv preprint arXiv:2402.15537*, 2024. [Online]. Available: https://arxiv.org/abs/2402.15537arXiv

16) **Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia**, "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 811–824, 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6337100 Wikipedia

17) **O. Mizuno, S. Ikami, S. Nakaichi, and T. Kikuno**, "Spam Filtering with Machine Learning in the Presence of Concept Drift," in Proc. Fourth Int. Workshop on Mining Software Repositories (MSR'07: ICSE Workshops 2007), 2007, pp. 64–70.