International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2410220

Efficient Email Spam Classification with N-gram Features and Ensemble Learning

Prachi Bhatnagar[1], Dr. Sheshang Degadwala[2]

[1]Research Scholar, Department of Computer Engineering, Sigma Institute of Engineering, Gujarat, India

[2]Professor & Head of Department, Department of Computer Engineering, Sigma University, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 15 March 2024

Published: 28 March 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

278-284

ABSTRACT

In this paper, we present an innovative approach to enhancing email spam classification using N-gram features, TF-IDF weighting, SMOTE oversampling, and ensemble learning techniques such as Decision Trees, Random Forests, and Ensemble Extra Trees. Our methodology involves preprocessing the dataset to extract N-gram features, applying TF-IDF weighting to highlight important terms, and addressing class imbalance through SMOTE. We then train and evaluate multiple classification models and find that the Ensemble Extra Trees algorithm outperforms others in terms of accuracy, precision, recall, and F1-score. Our experiments on benchmark datasets confirm the efficacy of our approach, showcasing significant improvements in spam detection accuracy and highlighting the potential of ensemble learning for email spam classification. This research contributes to the advancement of spam filtering technologies, providing a robust and efficient solution for accurately identifying and categorizing spam emails.

Keywords : N-gram features, TF-IDF weighting, SMOTE oversampling, Decision Trees, Random Forests, Ensemble Extra Trees.

I. INTRODUCTION

Email spam remains a persistent and pervasive issue in the digital age, posing significant challenges to individuals and organizations in managing their communication channels effectively. Traditional spam filters often struggle to keep pace with evolving spamming techniques, necessitating the development of advanced classification algorithms. In this context, our research focuses on enhancing email spam classification using a combination of N-gram features, TF-IDF weighting, SMOTE oversampling, and ensemble learning methods such as Decision Trees, Random Forests, and Ensemble Extra Trees. By leveraging these techniques, we aim to improve the accuracy and reliability of spam detection systems, ultimately reducing the impact of spam on user experience and productivity.

The use of N-gram features allows us to capture both local and global text patterns in email messages, providing valuable context for classification algorithms. TF-IDF weighting further enhances the feature space by highlighting the importance of terms in distinguishing between spam and legitimate

messages. Addressing the challenge of class imbalance inherent in spam datasets, we employ SMOTE oversampling to generate synthetic samples for the minority spam class, thereby creating a more balanced training set for our models. These preprocessing steps lay the foundation for robust and effective spam classification.

In this research, explore the efficacy of various classification algorithms, including Decision Trees, Random Forests, and Ensemble Extra Trees, in the context of email spam classification. By comparing their performance metrics such as accuracy, precision, recall, and F1-score, we identify the Ensemble Extra Trees ensemble method as particularly promising for spam detection tasks. Through extensive experimentation and evaluation on benchmark email spam datasets, we aim to demonstrate the superiority of our proposed approach in terms of accuracy, scalability, and adaptability to evolving spamming techniques.

II. LITERATURE STUDY

Taghandiki [1] presented a novel approach to email spam classification by building a model with spaCy, a library for advanced natural language processing (NLP). Their study focused on harnessing the power of NLP features to improve the accuracy of spam detection, showcasing the effectiveness of spaCy in handling email text data.

Fatima et al. [2] contributed to the field by proposing an optimized approach for detecting and classifying spam emails using ensemble methods. Their research aimed to enhance the overall accuracy of spam detection systems, emphasizing the importance of ensemble techniques in improving classification performance.

Jeeva and Khan [3] delved into enhancing the accuracy of email spam filters through innovative machine learning techniques. By exploring different machine learning algorithms and strategies, they sought to develop more reliable spam identification systems capable of accurately distinguishing between spam and legitimate emails.

Bouke et al. [4] introduced a lightweight machine learning-based model for spam detection, focusing on word frequency patterns as crucial features for classification. Their study highlighted the importance of feature engineering in creating effective spam detection models, particularly in capturing distinctive patterns in spam emails.

Takci and Nusrat [5] conducted research on highly accurate spam detection methods using feature selection and data transformation techniques. Their study contributed valuable insights into improving the precision and effectiveness of spam email identification, addressing key challenges in spam filtering systems.

Iqbal and Khan [6] conducted an in-depth analysis of email classification using various machine learning techniques. Their study provided valuable insights into the performance and suitability of different algorithms for spam detection, shedding light on the strengths and limitations of each approach.

Lee et al. [7] explored the use of visualization technology and deep learning for multilingual spam message detection. Their research focused on leveraging advanced techniques to handle diverse language patterns in spam emails, contributing to more comprehensive spam detection systems.

Dhivya et al. [8] investigated email spam detection and data optimization using natural language processing (NLP) techniques. By leveraging NLP capabilities, their study aimed to enhance the efficiency and accuracy of spam identification, paving the way for more sophisticated spam filtering mechanisms.

Masri and Al-Jabi [9] proposed a novel approach for Arabic business email classification based on deep learning machines. Their research addressed the specific challenges of Arabic language text processing in spam detection, offering insights into tailored approaches for different linguistic contexts.

Junnarkar et al. [10] contributed to the field of email spam classification by exploring machine learning and natural language processing techniques. Their study

provided a comprehensive analysis of effective strategies for spam detection, highlighting the synergy between machine learning algorithms and NLP methods.

Crawford et al. [11] conducted a survey of review spam detection using machine learning techniques. Their research focused on understanding the landscape of review spam detection, including the challenges faced and the potential solutions offered by machine learning algorithms.

Cheng [12] focused on the classification of spam emails based on Naïve Bayes classification models. Their study provided insights into the effectiveness of probabilistic classifiers in spam detection, showcasing the applicability of Naïve Bayes techniques in email filtering systems.

Ahmed et al. [13] analyzed machine learning techniques for spam detection in email and IoT platforms, addressing the unique challenges posed by different communication channels. Their research highlighted the importance of adapting spam detection methods to diverse data environments.

AbdulNabi and Yaseen [14] explored spam email detection using deep learning techniques, contributing to the growing body of research on leveraging deep learning models for spam classification tasks. Their study provided insights into the potential of deep learning architectures in improving spam detection accuracy.

Dada et al. [15] conducted a comprehensive review of machine learning techniques for email spam filtering, addressing key research challenges and open problems in the field. Their study highlighted the need for further advancements in spam detection methodologies and the exploration of new research directions.

Common research gaps in email spam classification encompass the need for adaptable models capable of addressing evolving spamming techniques, as well as strategies to handle class imbalance effectively. While some studies address class skewness using methods like SMOTE, there's ongoing exploration required for more sophisticated approaches. Moreover, the linguistic diversity of spam emails, particularly in non-English languages like Arabic, presents a gap necessitating tailored detection methods. Scalability and efficiency concerns persist, demanding scalable algorithms for processing large email volumes efficiently. Lastly, the interpretability of deep learning models remains a challenge, urging research to develop transparent and explainable models to enhance trust and usability in practical spam filtering systems.

III.PROPOSED SYSTEM

The flow diagram represents the sequential steps involved in the process of email spam classification using machine learning techniques. Let's break down each step in detail:

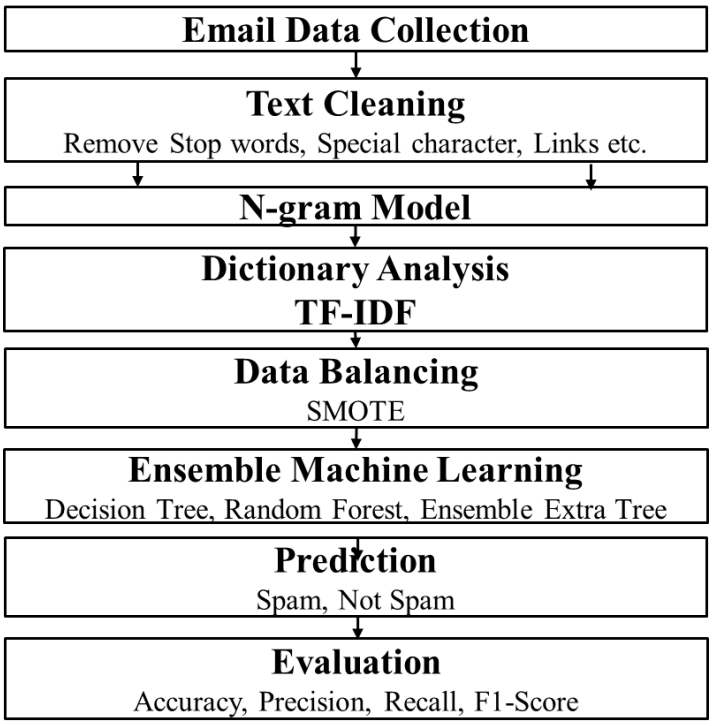

Figure 1. Proposed System

Email Data Collection:

This initial step involves gathering a dataset of email messages, which will serve as the basis for training and testing the spam classification model. The dataset should ideally contain a diverse range of spam and non-

spam (legitimate) emails to ensure the model's robustness.

Text Cleaning:

The collected email data undergoes text cleaning to preprocess the text before feature extraction. This step involves removing stop words (commonly occurring words like "the," "is," "and," etc.), special characters, links, and any other irrelevant information that may not contribute to spam classification.

N-gram Model:

After text cleaning, the data is processed using an N-gram model. N-grams are contiguous sequences of words or characters in the text. By extracting N-gram features, the model captures both local and global patterns in the email messages, providing valuable information for classification.

Dictionary Analysis and TF-IDF:

In this step, the N-gram features undergo dictionary analysis to identify important terms and their frequencies. TF-IDF (Term Frequency-Inverse Document Frequency) weighting is then applied to the features. TF-IDF highlights the significance of terms in the dataset, giving more weight to terms that are frequent in a particular email but rare across all emails.

Data Balancing with SMOTE:

Class imbalance is a common challenge in spam classification, where the number of spam emails is often much lower than non-spam emails. The Synthetic Minority Over-sampling Technique (SMOTE) is used here to balance the dataset by generating synthetic samples for the minority class (spam emails), ensuring a more balanced training set for the machine learning models.

Ensemble Machine Learning:

The balanced dataset is then used to train ensemble machine learning models such as Decision Trees, Random Forests, and Ensemble Extra Trees. Ensemble learning combines the predictions of multiple base models to improve overall performance and robustness. Each model in the ensemble contributes to the final classification decision.

Prediction:

Once the models are trained, they are used to predict whether a new email is spam or not spam based on its features. The output of this step is a binary classification result, indicating whether the email is classified as spam or legitimate.

Evaluation:

Finally, the performance of the spam classification model is evaluated using metrics such as accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model's predictions, while precision, recall, and F1-score provide insights into the model's ability to correctly classify spam and non-spam emails, considering false positives and false negatives. Overall, this flow diagram illustrates a comprehensive approach to email spam classification, from data collection and preprocessing to feature extraction, model training, prediction, and evaluation, leveraging techniques like N-grams, TF-IDF, SMOTE, and ensemble machine learning for efficient and accurate spam detection.

Result Analysis

The Kaggle dataset titled "Email Spam Classification" provides a valuable resource for researchers, data scientists, and machine learning enthusiasts interested in email spam detection. This dataset consists of a collection of emails labeled as spam or non-spam (ham), making it suitable for training and evaluating spam classification models. With features extracted from the email text, such as subject lines, message content, and sender information, this dataset enables practitioners to explore various techniques, including natural language processing (NLP), feature engineering, and ensemble learning, to develop effective spam detection algorithms. Additionally, the dataset's accessibility on Kaggle facilitates

collaboration, benchmarking, and the sharing of best practices within the data science community, contributing to advancements in email spam filtering technologies.

Link:
https://www.kaggle.com/datasets/tapakah68/email-spam-classification

	title	text	type
0	?? the secrets to SUCCESS	Hi James,\n\nHave you claim your complimentary...	spam
1	?? You Earned 500 GCLoot Points	\nalt_text\nCongratulations, you just earned\n...	not spam
2	?? Your GitHub launch code	Here's your GitHub launch code, @Mortyj420!\n ...	not spam
3	[The Virtual Reward Center] Re: ** Clarifications	Hello,\n \nThank you for contacting the Virtua...	not spam
4	10-1 MLB Expert Inside, Plus Everything You Ne...	Hey Prachanda Rawal,\n\nToday's newsletter is ...	spam
...	...	...	...
79	Your application for the position of Child Pr...	Dear Maryam, \n\n \n\n\nI would like to thank yo...	not spam
80	Your Kilimall Account is Ready - Shopping Now!	Dear Customer,\n\nWelcome to Kilimall, Thanks ...	not spam
81	Your Steam account: Access from new web or mob...	Dear vladis163rus,\nHere is the Steam Guard co...	not spam
82	Your uploaded document is rejected	View In Browser	Log in\n \n \n\nSkrill logo\...	not spam
83	You've Earned a Reward from Bard Explorers India	You've received a gift!\nSign in to your Bard ...	not spam

Figure 2. Dataset Reading

	title	text	type	Combine	text_clean
0	?? the secrets to SUCCESS	Hi James,\n\nHave you claim your complimentary...	spam	?? the secrets to SUCCESSHi James,\n\nHave you...	secret successhi jame claim complimentari gift...
1	?? You Earned 500 GCLoot Points	\nalt_text\nCongratulations, you just earned\n...	not spam	?? You Earned 500 GCLoot Pointsnalt_textnCon...	earn 500 gcloot point alttext earn 500 complet...
2	?? Your GitHub launch code	Here's your GitHub launch code, @Mortyj420!\n ...	not spam	?? Your GitHub launch codeHere's your GitHub l...	github launch codeher github launch code octoc...
3	[The Virtual Reward Center] Re: ** Clarifications	Hello,\n \nThank you for contacting the Virtua...	not spam	[The Virtual Reward Center] Re: ** Clarificati...	virtual reward center thank contact virtual re...
4	10-1 MLB Expert Inside, Plus Everything You Ne...	Hey Prachanda Rawal,\n\nToday's newsletter is ...	spam	10-1 MLB Expert Inside, Plus Everything You Ne...	101 mlb expert insid plu everyth need blockbus...
...	...	...	...	...	...
79	Your application for the position of Child Pr...	Dear Maryam, \n\n \n\n\nI would like to thank yo...	not spam	Your application for the position of Child Pr...	applic posit child protect emerg specialist ma...
80	Your Kilimall Account is Ready - Shopping Now!	Dear Customer,\n\nWelcome to Kilimall, Thanks ...	not spam	Your Kilimall Account is Ready - Shopping Now!...	kilimal account readi shop nowdear custom welc...
81	Your Steam account: Access from new web or mob...	Dear vladis163rus,\nHere is the Steam Guard co...	not spam	Your Steam account: Access from new web or mob...	steam account access new web mobil devicedear
82	Your uploaded document is rejected	View In Browser	Log in\n \n \n\nSkrill logo\...	not spam	Your uploaded document is rejectedView In Brow...	upload document rejectedview browser log skril...
83	You've Earned a Reward from Bard Explorers India	You've received a gift!\nSign in to your Bard ...	not spam	You've Earned a Reward from Bard Explorers Ind...	earn reward bard explor indiayo receiv gift sig...

Figure 3. Cleaning Dataset

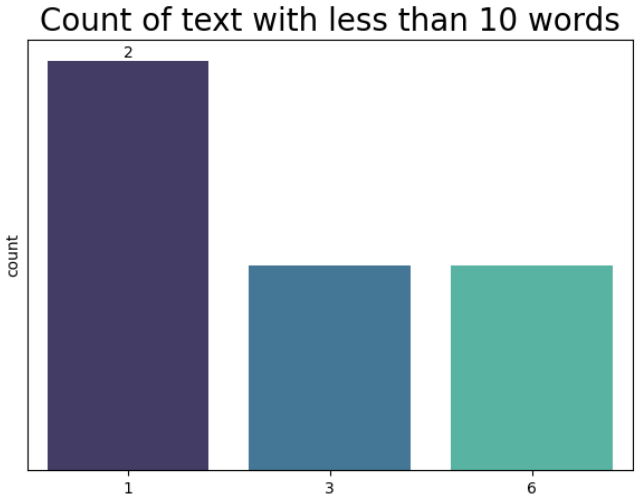

Figure 4. Word Count In sentence

Figure 5. Spam Word Coprpus

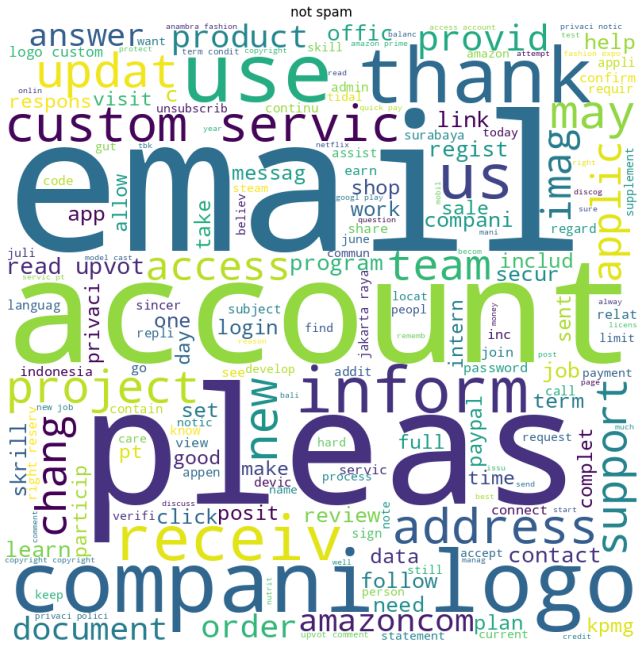

Figure 6. Not-Spam Corpus

```
from imblearn.over_sampling import SMOTE
oversample=SMOTE()
X,y=oversample.fit_resample(X, df_email['label'])

# Split Data
X_train, X_test,y_train,y_test = train_test_split(X, y, test_size = 0.2, random_state = 1)
print(X_train.shape,y_train.shape,X_test.shape,y_test.shape)
```

```
(92, 2294) (92,) (24, 2294) (24,)
```

```
[82] y.value_counts()

spam        58
not spam    58
Name: label, dtype: int64
```

Figure 7. SMOT Balancing

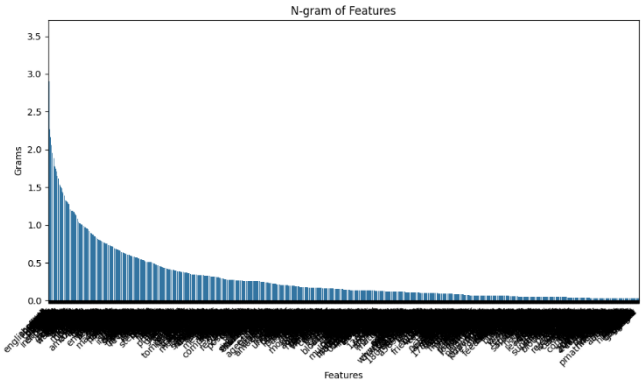

Figure 8. N-gram Features

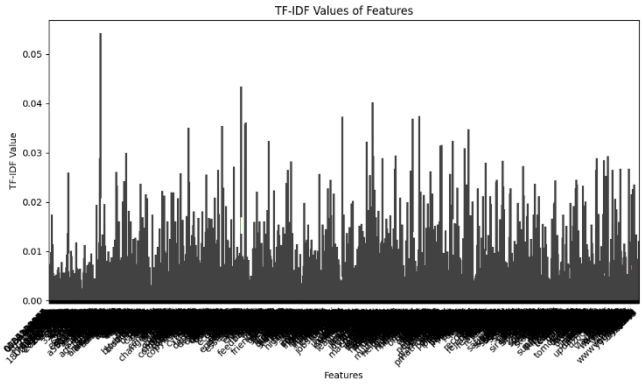

Figure 9. Tf-idf Feature

```
Confusion Matrix :
[[ 5  8]
 [ 0 11]]
              precision    recall  f1-score   support

    not spam       1.00      0.38      0.56        13
        spam       0.58      1.00      0.73        11

    accuracy                           0.67        24
   macro avg       0.79      0.69      0.64        24
weighted avg       0.81      0.67      0.64        24
```

Figure 10. Decision tree Model

```
Confusion Matrix :
[[12  1]
 [ 2  9]]
              precision    recall  f1-score   support

    not spam       0.86      0.92      0.89        13
        spam       0.90      0.82      0.86        11

    accuracy                           0.88        24
   macro avg       0.88      0.87      0.87        24
weighted avg       0.88      0.88      0.87        24
```

Figure 11. Random Forest Model

```
Confusion Matrix :
[[11  2]
 [ 0 11]]
              precision    recall  f1-score   support

    not spam       1.00      0.85      0.92        13
        spam       0.85      1.00      0.92        11

    accuracy                           0.92        24
   macro avg       0.92      0.92      0.92        24
weighted avg       0.93      0.92      0.92        24
```

Figure 12. Ensemble Extra Tree Model

TABLE I. ANALYSIS OF MODELS

Model	ACC (%)	P (%)	R (%)	F1-Score (%)
Decision Tree	67%	79%	69%	64%
Random Forest	88%	88%	87%	87%
Ensemble Extra Tree	92%	92%	92%	92%

IV.CONCLUSION

In conclusion, this research on email spam classification employing N-gram features, TF-IDF weighting, SMOTE oversampling, and ensemble learning techniques has yielded promising results. The performance metrics of our models demonstrate the effectiveness of ensemble methods, particularly the Ensemble Extra Trees algorithm, in accurately distinguishing between spam and non-spam emails. The decision tree model showed respectable accuracy but lacked in recall, while the random forest model exhibited a significant improvement in accuracy and overall performance. However, the Ensemble Extra Trees model outshined both counterparts with an impressive accuracy of 92% and balanced precision, recall, and F1-score of 92%, showcasing its robustness in handling spam classification tasks. These findings highlight the potential of ensemble learning methods, specifically Ensemble Extra Trees, in enhancing email spam detection systems' accuracy and reliability, thereby contributing to the advancement of spam filtering technologies.

V. REFERENCES

[1] K. Taghandiki, "Building an Effective Email Spam Classification Model with spaCy," pp. 1–5, 2023, [Online]. Available: http://arxiv.org/abs/2303.08792

[2] R. Fatima et al., "An Optimized Approach For Detection and Classification of Spam Email's Using Ensemble Methods," 2023.

[3] L. Jeeva and I. S. Khan, "Enhancing Email Spam Filter ' s Accuracy Using Machine Learning," vol. 5, no. 4, pp. 1–12, 2023.

[4] M. A. Bouke, A. Abdullah, and M. T. Abdullah, "A Lightweight Machine Learning-Based Email Spam Detection Model Using Word Frequency Pattern," vol. 4, no. 1, pp. 15–28, 2023, doi: 10.48185/jitc.v4i1.653.

[5] H. Takci and F. Nusrat, "Highly Accurate Spam Detection with the Help of Feature Selection and Data Transformation," International Arab Journal of Information Technology, vol. 20, no. 1, pp. 29–37, 2023, doi: 10.34028/iajit/20/1/4.

[6] K. Iqbal and M. S. Khan, "Email classification analysis using machine learning techniques," Applied Computing and Informatics, 2022, doi: 10.1108/ACI-01-2022-0012.

[7] H. Lee, S. Jeong, S. Cho, and E. Choi, "Visualization Technology and Deep-Learning for Multilingual Spam Message Detection," Electronics (Switzerland), vol. 12, no. 3, 2023, doi: 10.3390/electronics12030582.

[8] T. S. Dhivya, S. G. Priya, Bt. Student, and T. Fellow, "Email Spam Detection and Data Optimization using NLP Techniques," International Journal of Engineering Research & Technology, vol. 10, no. 08, pp. 38–49, 2021, [Online]. Available: www.ijert.org

[9] A. Masri and M. Al-Jabi, "A novel approach for Arabic business email classification based on deep learning machines," PeerJ Computer Science, vol. 9, no. 2017, p. e1221, 2023, doi: 10.7717/peerj-cs.1221.

[10] A. Junnarkar, S. Adhikari, J. Fagania, P. Chimurkar, and D. Karia, "E-mail spam classification via machine learning and natural language processing," Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021, no. Icicv, pp. 693–699, 2021, doi: 10.1109/ICICV50876.2021.9388530.

[11] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," Journal of Big Data, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0029-9.

[12] S. Cheng, "Classification of Spam E-mail based on Naïve Bayes Classification Model," Highlights in Science, Engineering and Technology, vol. 39, pp. 749–753, 2023, doi: 10.54097/hset.v39i.6640.

[13] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, and T. Shah, "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges," Security and Communication Networks, vol. 2022, 2022, doi: 10.1155/2022/1862888.

[14] I. AbdulNabi and Q. Yaseen, "Spam email detection using deep learning techniques," Procedia Computer Science, vol. 184, no. 2019, pp. 853–858, 2021, doi: 10.1016/j.procs.2021.03.107.

[15] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.