iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55052

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Spam Spoiler for Public and Private E-Mail Services

Manikandan[1], Sreerambabu[2], Kalidasan[3], Mohammed Riyaz[4]
[1]PG Scholar, [2]Head of the Department, [3, 4]Assistant Professor Dept of MCA

Abstract: ***In recent years, cyber security incidents have occurred frequently. In most of these incidents, attackers have used different types of spam email as a knock-on to successfully invade government systems, well-known companies, and websites of politicians and social organizations in many countries. The detection of spam mail from big email data has been paid public attention.***
However, the camouflage technology of spam mail is becoming more and more complex, and the existing detection methods are unable to confront the increasingly complex deception methods and the growing number of emails. In this project, we proposed to design a novel efficient approach named Spam Spoiler for big email data classification into four different classes: Normal, Fraudulent, Harassment, and Suspicious E-mails by using LSTM-based GRU. The new method includes two important stages, the sample expansion stage and the testing stage under sufficient samples. This project the LSTM-based GRU efficiently captures meaningful information from E-mails that can be used for forensic analysis as evidence. Experimental results revealed that Spam Spoiler performed better than existing ML algorithms and achieved a classification accuracy of 98% using the novel technique of LSTM with recurrent gradient units. As different types of topics are discussed in E-mail content analysis. Spam Spoiler effectively outperforms existing methods while keeping the classification process robust and reliable.
Keywords: Spam spoiler, spam, machine learning, deleting

I. INTRODUCTION

Dispatch stands for Electronic Correspondence. It's a system to shoot dispatches from one computer to another computer through the Internet. It's substantially used in business, education, specialized communication, and document relations. It allows communicating with people all over the world without bothering them. In 1971, a test dispatch was transferred to Ray Tomlinson to himself containing the textbook. E-mail dispatches are conveyed through dispatch waiters; it uses multiple protocols within the TCP/IP suite. For illustration, SMTP is a protocol, that stands for simple correspondence transfer protocol and is used to shoot dispatches whereas other protocols IMAP or POP are used to recoup dispatches from a correspondence server.

However, you just need to enter a valid dispatch address, word, If you want to log in to your correspondence account. Although utmost of the webmail waiters automatically configures your correspondence account, thus, you're only needed to enter your dispatch address and word. Still, you may need to manually configure each account if you use a dispatch customer like Microsoft Outlook or Apple Mail. In addition, to enter the dispatch address and word, you may also need to enter incoming and gregarious correspondence waiters and the correct harborage figures for each one.Dispatch dispatches include three factors, which are as follows

1) Communication envelop It depicts the dispatch'selectronic format.
2) Communication title It contains the dispatch subjectline and sender/ philanthropist information.
3) Communication body It comprises images,textbooks, and other train attachments.

II. EXISTING SYSTEM

The paragraph discusses various techniques used for email spam filtering. Algorithms analyze the content of emails, including words, occurrences, and distributions, to categorize them as spam or non-spam. Different approaches such as case-based filtering, heuristic or rule-based filtering, previous likeness-based filtering, and adaptive filtering are employed.

Machine learning classifiers like Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (DT), and AdaBoost are utilized for classification.

However, these approaches have certain disadvantages, including the possibility of misclassifying important emails, the time-consuming nature of block listing, and the need for manual feature engineering. Forensic tools relying on keyword searches may also produce irrelevant results.

©IJRASET: All Rights are Reserved

III. PROPOSED SYSTEM

The proposed approach for email classification involves using LSTM-GRU models and dividing email datasets into different classes. The emails are segmented into word sequences and transformed into vectors using embedding techniques. LSTM networks address long-term dependency issues, while GRU networks mitigate the vanishing gradient problem. The advantages of this approach include effective email content classification, reliable bracketing process, pattern identification, and elimination of manual labeling.

IV. PROBLEM DESCRIPTION

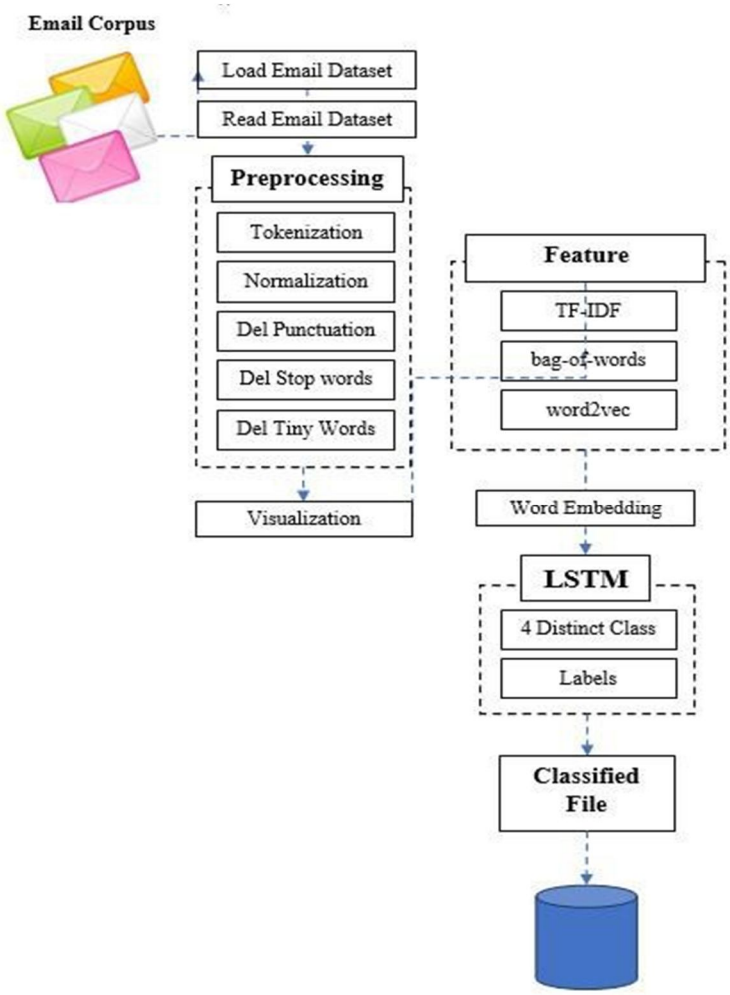
Emails have a title and a body. The title includes information like the subject, sender, and timestamps. The body contains various types of content. Before classifying emails, we preprocess them by removing URLs, HTML, CSS, JavaScript, and special symbols. We then use machine learning algorithms and tokenization to convert the email text into terms. We extract features using TF-IDF and predict using Gaussian Naïve Bayes. In deep learning, we map words to vectors using embeddings. Our dataset includes normal, fraudulent, harassment, and suspicious emails. We enhance the dataset by adding suspicious emails from various sources. The dataset consists of different classes with varying numbers of emails.

V. DATASET DESCRIPTION

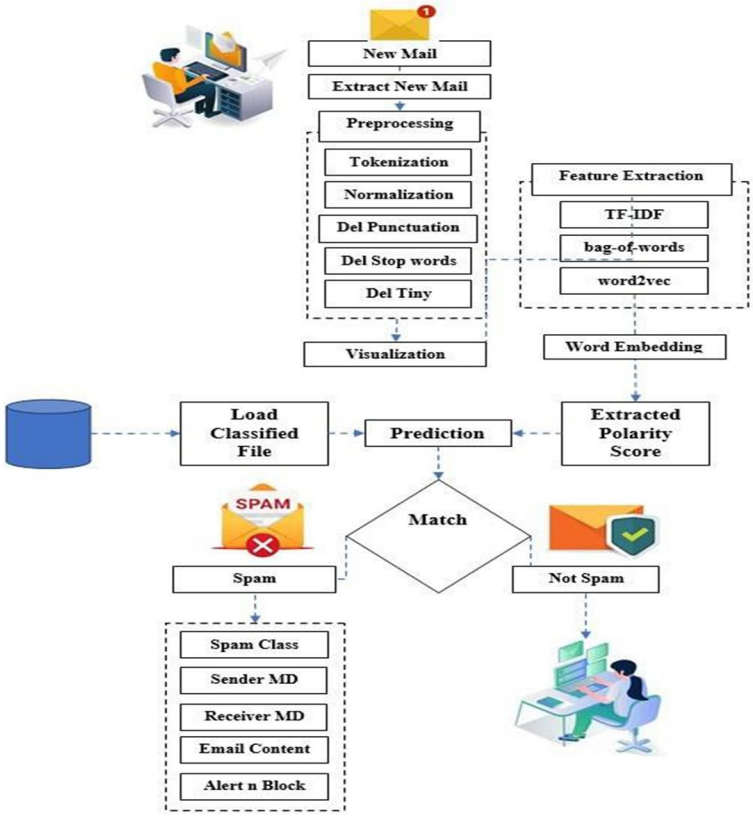
Our dataset is a combination of four datasets. It includes normal emails, fraudulent emails, harassment emails, and suspicious emails. The dataset is merged into a structured file for multiclass email classification.

VI. SYSTEM ARCHITECTURE

System Architecture – Training Phase



System Architecture – Testing Phase



VII. CONCLUSION

With the growing trend of cybercrime and accidents resulting from vulnerabilities, proactive monitoring and post-incident analysis of email data is crucial for organizations. Cybercrimes like hacking, spoofing, phishing, E-mail bombing, whaling, and spamming are being performed through E-mails. The existing email classification approaches lead toward irrelevant E-mails and/or loss of valuable information. Keeping in sight these limitations, we designed a novel effective approach named E-MailSinkAI for E-mail bracket into four different classes Normal, Fraudulent, Hanging, and Suspicious E-mails by using LSTM- grounded GRU that not only deals with short sequences as well long dependences of 1000Ccharacters.We estimated the proposed E-MailSinkAI model using evaluation criteria similar as perfection, recall, delicacy, and f- score. Experimental results revealed that E-MailSinkAI performed better than being ML algorithms and achieved a bracket delicacy of 95 using the new fashion of LSTM with intermittent grade units.

VIII. FUTURE ENHANCEMENT

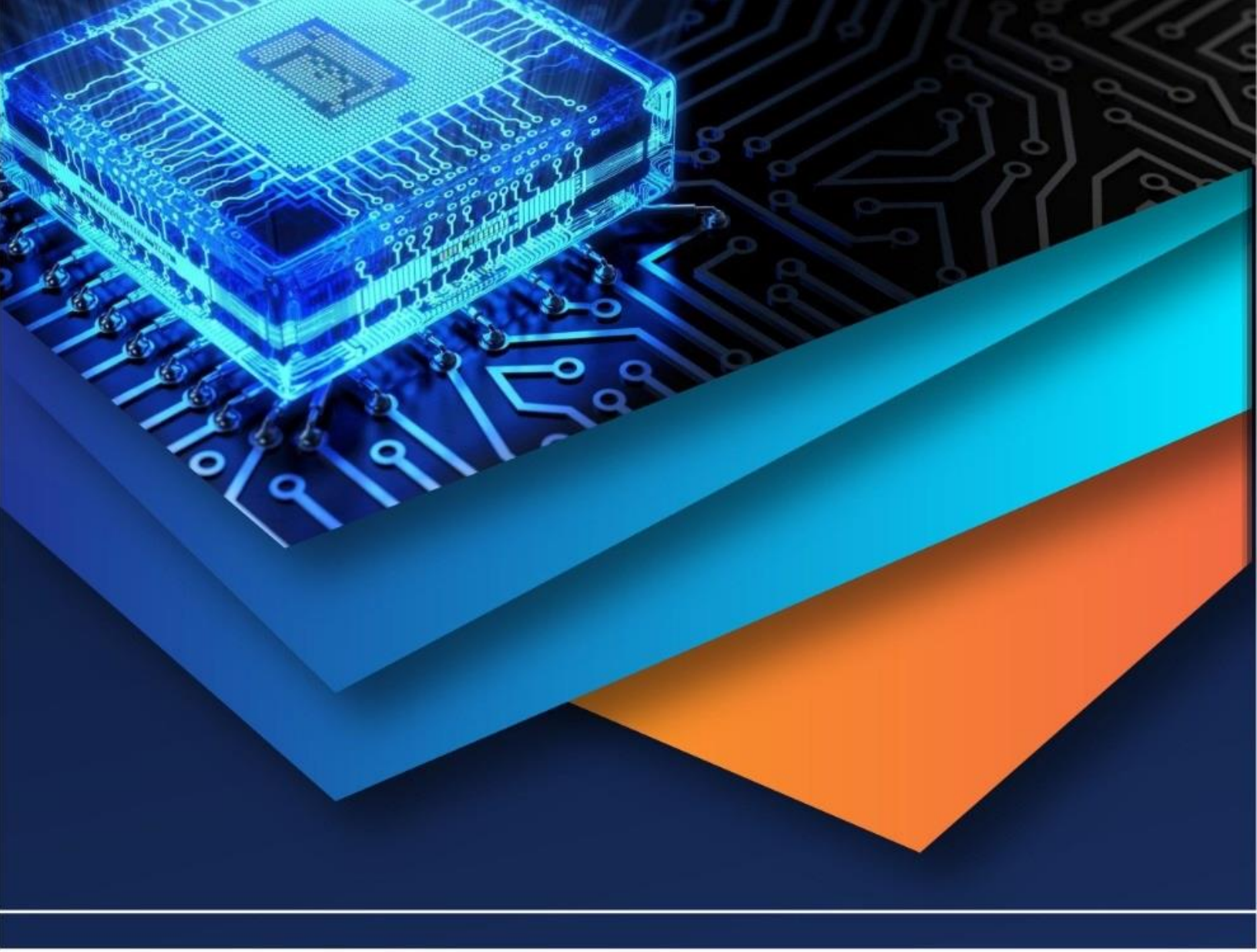
Unborn improvement For now, we're considering e-mail classes similar as normal, importunity, fraud, and suspicious; still, numerous other classes can be added to this work in the presence of a massive quantum of-mail data.

REFERENCES

[1] S. Sinha, I. Ghosh, and S. C. Satapathy, ``A study for ANN model for spam classification," in Intelligent Data Engineering and Analytics. Singapore: Springer, 2021, pp. 331-343.
[2] Q. Li, M. Cheng, J. Wang, and B. Sun, ``LSTM based phishing detection for big email data," IEEE Trans. Big Data, early access, Mar. 12, 2020, doi: v10.1109/TBDATA.2020.2978915.
[3] T. Gangavarapu, C. D. Jaidhar, and B. Shanduka,
[4] ``Applicability of machine learning in spam and phishing email filtering: Review and approaches," Artif. Intell. Rev., vol. 53, no. 7, pp. 5019-5081, Oct. 2020, doi: 10.1007/s10462-020-09814-9.
[5] E. Bauer. 15 Outrageous Email Spam Statistics That Still Ring True in 2018, RSS. Accessed: Oct. 10, 2020. [Online]. Available: https://www.propellercrm.com/blog/email-spam- ststatistics.
[6] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and

[7] M. Alazab, ``A comprehensive survey for intelligent spam email detection," IEEE Access, vol. 7, pp. 168261-168295, 2019.
[8] K. Singh, S. Bhushan, and S. Viji, ``Filtering spam messages and emails using fuzzy C means algorithm," inProc. 4th Int. Conf. Internet Things, Smart Innov. Usages (IoT-SIU), Apr. 2019, pp. 1-5.
[9] R. S. H. Ali and N. E. Gayer, ``Sentiment analysis using unlabelled email data," in Proc. Int. Conf. Compute. Intell. Knowl. Economy (ICCIKE), Dec. 2019, pp. 328-333.
[10] K. Agarwal and T. Kumar, ``Email spam detection usingan integrated approach of naïve Bayes and particle swarm optimization," in Proc. 2ndInt. Conf. Intell. Compute. ControlSyst. (ICICCS), Jun. 2018, pp. 685-690.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)