iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67874

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 13 Issue III Mar 2025- Available at www.ijraset.com

LSTM Based Phishing Detection for Big Email Data

Dwarakasai Y, Dhanush R, Bharath N, Dr. Ramshankar N, Dr. Ezhumalai P
Department of Computer Science and Engineering, R.M.D. Engineering College (Autonomous), Chennai, Tamil Nadu, India

Abstract: ***Phishing attacks remain a significant concern for email security and typically utilize fraudulent messages that appear legitimate. In this study, we propose a new method that appropriately identifies phishing attacks based on email traffic using advanced NLP methods coupled with LSTM networks. The approach begins by applying NLP techniques to extract and analyze email content to identify specific linguistic cues indicative of phishing as well as other features in the text. The identified features will serve as inputs to the LSTM model which is suited to sequences of data and will help the system recognize patterns to discriminate a phishing attack from a non-phishing email. With deep learning systems and advanced NLP techniques, the proposed system has a high degree of accuracy when processing an abundance of emails. One of the major strengths of the proposed system is that it minimizes and attempts to eliminate false positives which can cause legitimate emails to be flagged as phishing attempts. The objective of this research is to improve email security by employing an effective and efficient mechanism for detecting Detecting possible phishing threats instantly.***

Index Terms: Phishing Detection, Email Security, Natural Language Processing (NLP), Long Short-Term Memory (LSTM), Deep Learning, Spam Detection, Cybersecurity, Pattern Recognition

I. INTRODUCTION

Phishing attacks are becoming more dangerous: The fre- quency and sophistication of phishing attacks are rising quickly, endangering both individuals and organizations.

Cybercriminals are continuously advancing their tactics, causing receivers to reveal sensitive information by allocating bootleg emails that seem to be from trusted sources. These cyber threats can cause data leaks, identity fraud, and major monetary damages. The growing threat illustrates the urgency of the need for intelligent, flexible security solutions to protect vulnerable targets in an increasingly digital world while at the same time dynamically, counter this ever-changing phishing method.

Leveraging Advanced NLP and LSTM Methods: This approach utilizes cutting-edge Natural Language Processing (NLP) techniques to extract valuable insights from email text, effectively combating sophisticated phishing tactics. This involves analyzing the textual characteristics related to the language format, grammatical structures, and contextual nu- ances. By integrating NLP techniques with LSTM networks, this model can detect subtle differences between genuine and fraudulent email text, leveraging LSTM's strength in identifying temporal and sequential patterns in data. This combination isolates hidden patterns that could otherwise be discarded while also improving the detection of direct phishing metrics.

High Precision for Proactive Security The LSTM-based model is a good directional choice in phishing detection, as it was rigorously designed to lessen False alarms and undetected anomalies, which correlates to a enhanced degree of accuracy. That degree of precision affords the system the ability to recognize and thwart potential threats before they escalate into data breaches or other security events. Accurate detection allows organizations to maintain trust from their consumers in their email communications systems, protect sensitive data, and maintain legitimate operations. Scalable Risk Mitigation: The massive volumes of email traffic that modern organizations handle pose hurdles in terms of both volume and complexity. With its scalable design, the recommended approach can process and evaluation massive email a number swiftly and efficiently. This is made possible by deep learning architectures' built-in parallel processing capabilities, strong feature extraction methods, and simplified data preliminary treatment pipelines. The technique decreases the risk of phishing attacks by offering consistent, predictable security as the threat environment changes by efficiently handling large data streams and sophisticated email formats. Smooth Integration and Realistic Implementation: The in- tended phishing detection system is made to be applicable in the actual world. It can be incorporated into already-existing email security systems to offer automatic threat responses and real-time monitoring. This simplicity of connectivity improves overall cybersecurity and reduces interference with ongoing activities. The system also allows for regular updates and retraining, which guarantees that it will continue to be effective against new phishing techniques. Both business settings and individual users find the solution appealing due to its useful deployment framework.

II. LITERATURE SURVEY

As researchers and practitioners have worked to keep up with the changing strategies of cybercriminals, the field of phishing detection has seen tremendous change over the last ten years. Heuristic and rule-based methods were the main- stays of early detection systems. These early techniques con- centrated on spotting distinct, predetermined signs of phish- ing, like discrepancies between sender addresses and URLs, irregularities in email layout, and other readily apparent ab- normalities. These systems were intrinsically rigid and rapidly became outdated as attackers developed more sophisticated and subtle ways to avoid detection, despite being effective against established attack patterns.

After ascertaining the limitations of heuristics, the scientific community began to leverage conventional machine learning techniques. The emergence of methods like random forests, decision trees, and SVM represented a shift toward data-driven detection techniques. Unlike heuristics, these Unlike heuris- tics, these models were able to discover patterns in the data without the need for establishing rules by hand, and, by and large, it was recognized that they outperformed the previous technical methods. Notably, the success of these models was dependent on the quality of the feature engineering process. In order to detect aspects of email that may be suggestive of phishing, experts hand crafted and selected features, such as lexical features, header anomalies, and URL features, etc. Despite the improvements over heuristics, handcrafted features often limited the models' ability to learn about the sequential and complex structure of the email body content.

The emergence of deep learning has fundamentally altered phishing detection. The innovation of neural network topolo- gies, specifically LSTM networks and RNNs, has allowed researchers towards autonomously extract complex patterns from text that has not been preprocessed. LSTM networks can analyze the temporal dynamics of email content and related sequential data due to the implicit sequential data nature. LSTM networks learn the long-range dependencies in text, effectively detecting subtle phishing indicators and symptoms that traditional models would overlook. Furthermore, LSTM- based phishing detection models, which learn hierarchical representations of text data from the input itself, have shown substantially less reliance on thorough feature engineering.

The adoption of hybrid techniques, which combine deep learning with conventional machine learning methods, has also evolved. Researchers have examined the use of ensemble approaches to capitalize on the advantages of multiple models, enhancing detection precision and reducing false positives. By using strategies like the Synthetic Minority Over-sampling Technique (SMOTE), researchers have also tackled real-world issues like class imbalance, when phishing efforts are greatly outnumbered by legitimate emails. This ensures a more fair and balanced representation during training while also improv- ing the model's ability to adapt across different datasets.

A crucial element of current studies is their emphasis on adversarial robustness. Even though deep learning models perform remarkably well, they are susceptible to adversarial attacks, which are tiny, purposefully created perturbations meant to trick the system. To strengthen models against such attacks and make sure they continue to function reliably even when confronted with complex evasion attempts, studies in this field have integrated adversarial training techniques.

Transformer-based designs, such BERT, are now being investigated by emerging trends in the area because they have the ability to capture bidirectional context and sophisticated language representations that LSTM networks cannot. Initial findings are encouraging and imply that these models may further improve the accuracy and resilience of detection sys- tems, even though their use in phishing detection is still in its infancy.

This thorough analysis of the literature shows how phishing detection algorithms have evolved and become more complex over time, ranging from early rule-based from early rule-based methods to modern deep learning and ensemble techniques. The suggested LSTM-based system is built on the knowledge gathered from these studies and seeks to use these cutting- edge methods to offer a more flexible, reliable, and effective way to protect email correspondence from phishing attacks.

At present, many different forms of cybercrime are perpe- trated through the online world. Hence, this study emphasizes phishing attack analysis. While phishing first appeared in 1996, it has since grown into one of the gravest and most perilous cyber threats on the web. Identity theft scam relies on the email distortion of tricky correspondences and mock sites as its overarching mechanism for stealing the required data from its victims. Various studies have shared their efforts regarding prevention, detection, and awareness of phishing attacks, however there appears to be no adequate, compre- hensive prevention solution available at this time. For this reason, machine learning is a critical component for defending against phishing attacks as a form of cybercrime. The proposed research makes use of a phishing URL dataset, obtained from a widely recognized data repository. This dataset com- prises characteristics of both phishing and legitimate URL elements, gathered from 11,000 datasets of websites in a vectorized format. After undergoing preprocessing, several machine learning techniques were implemented and enhanced, specifically to raise awareness and strengthen defenses against URL-based phishing attacks while ensuring user security.

Machine learning techniques, including decision trees (DT), linear regression (LR), random forests (RF), naive Bayes (NB), gradient boosting models (GBM), K-nearest neighbors (KNN), and support vector classifiers (SVC). Additionally, a novel hybrid LSD model is introduced, integrating logistic regression. To enhance protection against phishing threats while ensuring high precision and efficiency, we employed logistic regression, Support vector machine, and decision trees (LR, SVM, and DT) machine learning models, in both soft and hard voting.In the suggested LSD model, a canopy-based method selection approach was utilized alongside stratified cross-validation and grid search optimization methods for fine- tuning hyperparameters. To assess the suggested approach, we utilized various evaluation metrics such as precision, accuracy, recall, F1-score, and specificity to showcase the model's effectiveness, influence, and performance. The findings of the comparative assessments indicate that the recommended technique outperformed alternative approaches and achieved the highest overall evaluation.

III. PROPOSED WORK

By leveraging Long Short-Term Memory (LSTM) networks in addition to Natural Language Processing (NLP), the pro- posed system seeks to revolutionize email security with an advanced email phishing detection system. This system will automatically preprocess and analyze incoming email content, allowing important features to be obtained and ultimately leading to the identification of phishing attempts, with high accuracy levels. During system development, large datasets will be used to train the LSTM model to allow it to accu- rately discern between legitimate and illegitimate emails. The proposed system will also incorporate a real-time monitoring capability so that users can be alerted of any phishing attack at- tempt immediately. In all, the overarching approach, enhances trust in email communication by providing a scalable, efficient defense against such attacks.

IV. IMPLEMENTATION

The phishing detection system uses Long Short-Term Memory (LSTM) networks combined with Natural Language Pro- cessing (NLP) to ascertain Whether an email is a phishing attempt or a legitimate message.The organized process of an a phishing detection system consists of several stages, such as data gathering, preprocessing, feature extraction, model training, and real-time classification. The use of deep learning algorithms facilitates effective and precise phishing detection. Information Gathering: The collection is collected from simulated email exchanges, organizational email traffic, and publicly accessible phishing email repositories. A supervised learning method is ensured by classifying emails as either authentic or phishing. To improve classification, additional metadata is included, such as email attachments, subject lines, and sender information.

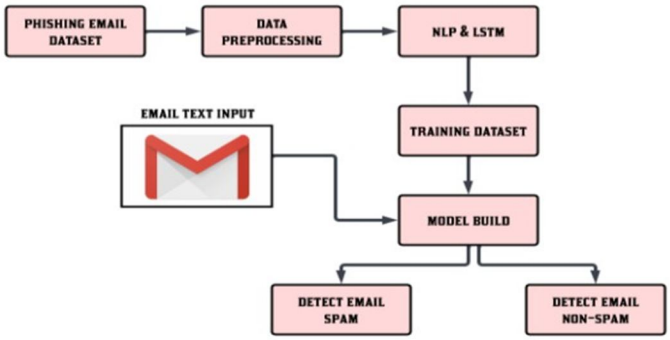


Fig. 1. Output screenshots for the Application.

1) *Preprocessing of Data:* Several preparation procedures are used to transform unstructured email text into a structured format. Unwanted characters, HTML tags, and special symbols are eliminated by text cleaning. While stopword removal removes common words that don't help detect phishing, tokenization separates the email content into distinct words. Lemmatization ensures consistency by normalizing words to their base forms. Lastly, TF-IDF, Word2Vec, or Doc2Vec methods are used to transform text data into numerical representations.
2) *Extraction of Features:* To increase the accuracy of classification, pertinent features are extracted. Lexical character- istics emphasize n-grams, suspicious phrases, and keyword frequency. Email length, attachments, and URLs are examples of structural elements. NLP approaches are used to create se- mantic characteristics, which analyze sentiment and contextual meaning to identify phishing intent.

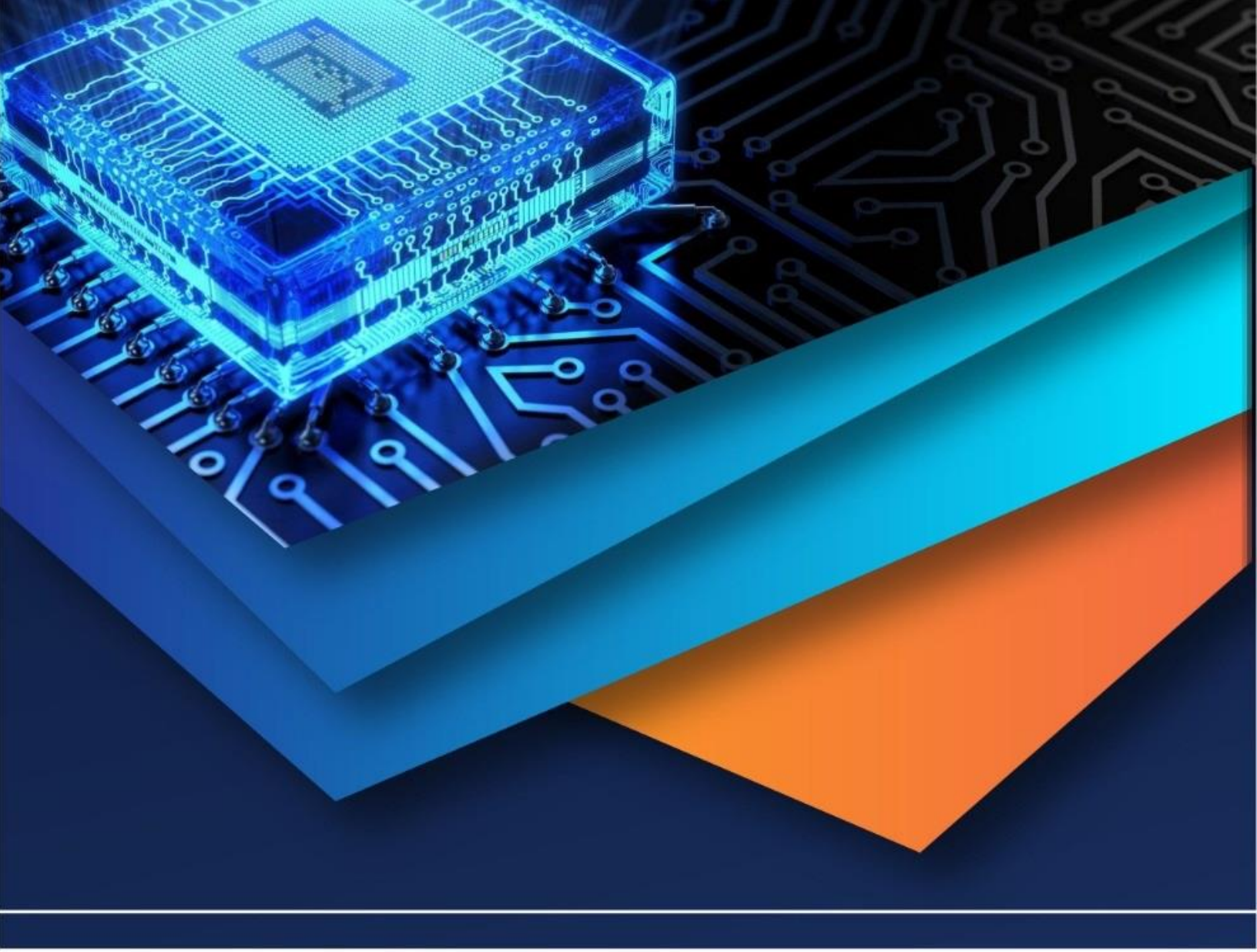
3) *Model Training through LSTM:* The LSTM model is trained in accordance with the preprocessed dataset and can recognize sequential text email text patterns. The model is trained on Binary Cross- Entropy loss, in binary optimized with the Adam optimizer classification. L2 regularization is included and dropout layers are used to prevent overfitting. By capturing contextual dependencies in email text, the LSTM network allows detection of phishing attempts with greater accuracy. After training, accuracy, precision, recall, and F1-score performance metrics are computed to measure the model.
4) *Classification & Real-Time Detection:* After training, the LSTMmodel is used for real-time email classification detec- tion. As emails are received, they are pre-processed, vectorized, and run through the model, which assigns a probability score of whether the email is phishing or legitimate. The system generates instant alerts for phishing emails, allowing detection and manner prevention to occur in real-time. An interface (CLI/Web-based) is included for easy access and monitoring by an employee.
5) *Technologies and Tools Employed:* With the help of Python and libraries like TensorFlow, Keras, Scikit-learn, NLTK, Pandas, and NumPy, the project is developed. Web-based deployment is done with Flask or Django, and integration with current email systems is made possible by TensorFlow Serving and REST APIs.
6) *Equipment and Technology Utilized:* The project is implemented in Python utilizing libraries such as NumPy, NLTK, Pandas, Scikit-learn, TensorFlow, and Keras. Web-based deployment is done with Flask or Django, and current email systems can be integrated with TensorFlow Serving and REST APIs.

V. CONCLUSION

By effectively categorizing phishing emails using LSTM networks and natural language processing (NLP) method- ologies, this proposed phishing detection system enhances privacy for email services. Unlike traditional methods, this system constantly learns from new data while reducing false positives and adapting to new threats. The inclusion of con- tinual monitoring and automatic analysis promotes an active approach to threat detection, reducing risk to individuals and businesses. The system can efficiently scale and manage large email repositories. Its seamless implementation inside cybersecurity systems helps to improve digital communication security overall. Our study supports a safer online space by protecting users from phishing attacks and developing an AI- empowered cybersecurity solution to address evolving cyber threats.

REFERENCES

[1] S. Salloum, T. Gaber, S. Vadera and K. Shaalan, quot;A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques,quot; in IEEE Access, vol. 10, pp. 65703-65727, 2022, doi: 10.1109/ACCESS.2022.3183083..
[2] A. Khalid, M. Hanif, A. Hameed, Z. Ashraf, M. M. Alnfiai and S. M.
[3] M. Alnefaie, quot;LogiTriBlend: A Novel Hybrid Stacking Approach for Enhanced Phishing Email Detection Using ML Models and Vector- ization Approach,quot; in IEEE Access, vol. 12, pp. 193807- 193821, 2024, doi: 10.1109/ACCESS.2024.3518923.
[4] J. Lee, Y. Lee, D. Lee, H. Kwon and D. Shin, quot;Classification of Attack Types and Analysis of Attack Methods for Profiling Phishing Mail Attack Groups,quot; in IEEE Access, vol. 9, pp. 80866-80872, 2021, doi: 10.1109/ACCESS.2021.3084897.
[5] R. Zieni, L. Massari and M. C. Calzarossa, quot;Phishing or Not Phishing? A Survey on the Detection of Phishing Websites,quot; in IEEE Access, vol. 11, pp. 18499-18519, 2023, doi: 10.1109/AC- CESS.2023.3247135.
[6] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari and S. R. K. Joga, quot;Phishing Detection System Through Hybrid Machine Learning Based on URL,quot; in IEEE Access, vol. 11, pp. 36805-36822, 2023, doi: 10.1109/ACCESS.2023.3252366.
[7] I. Kara, M. Ok and A. Ozaday, quot;Characteristics of Understanding URLs and Domain Names Features: The Detection of Phishing Websites With Machine Learning Methods,quot;in IEEE Access, vol. 10, pp. 124420-124428, 2022, doi: 10.1109/ACCESS.2022.3223111.
[8] A. Butnaru, A. Mylonas, and N. Pitropakis, “Towards lightweight URL- based phishing detection,” Future Internet, vol. 13, no. 6, p. 154, Jun. 2021..
[9] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si,
[10] F. Zhang, and J. S. Dong, “Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages,” in Proc. 30th USENIX Secur. Symp., Aug. 2021, pp. 3793–3810.
[11] S. Ariyadasa, S. Fernando, and S. Fernando, “Phishing web- sites dataset,” Nov. 17, 2021. Distributed by Mendeley Data, doi: 10.17632/n96ncsr5g4.1.
[12] S. D. Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, “Modeling hybrid feature-based phishing websites detection using machine learning techniques,” Ann. Data Sci., vol. 10, pp. 1–26, Mar. 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY