Spear-Phishing Detection using LSTM on Email Body

1st Asma Mukhtiar
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
asmamukhtiar211@gmail.com

2nd Shazia Jameel
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
shaziajamil853@gmail.com

3rd Sana Tariq
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
sana.tariq@iub.edu.pk

4th Iram Haider
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
iramhaider765@yahoo.com

***Abstract*—Phishing attacks, particularly spear-phishing, represent a significant threat to organizational cybersecurity. Traditional rule-based or keyword detection methods are no longer sufficient to detect increasingly sophisticated attacks. This paper presents a deep learning approach utilizing Long Short-Term Memory (LSTM) networks to detect spear-phishing emails based on their content. We preprocess and analyze the textual body of emails using natural language processing techniques and demonstrate how an LSTM-based model can effectively classify phishing and legitimate emails. The experimental results on a sample dataset show promising performance in terms of accuracy and robustness. Our work highlights the potential of deep learning in real-world cybersecurity applications and suggests future avenues for enhancing phishing detection capabilities.To further enhance detection capabilities, future research paths will examine transformer-based designs like BERT and incorporate multimodal characteristics like email headers, metadata, and embedded links. This work lays the groundwork for creating more robust anti-phishing systems and adds to the expanding corpus of research on AI-driven cybersecurity solutions.**

***Index Terms*—Cybersecurity, Phishing Detection, LSTM, Natural Language Processing, Deep Learning**

I. Introduction

Phishing is still one of the most common and harmful cyberthreats in the digital world, impacting people, businesses, and even governmental organizations worldwide [1]. It usually entails misleading communication, usually by email, with the goal of fooling users into disclosing private information such login credentials, passwords, or financial information. Of the different types of phishing, spear-phishing is one of the most pernicious. Spear-phishing campaigns are extremely focused and comprise well-crafted communications tailored for particular people or organizations, in contrast to generic phishing attempts that are distributed in bulk [2].

Spear-phishing emails frequently look authentic, imitating reliable sources like executives, coworkers, or service providers. Using the data of individuals or entities collected via social media, web reconnaissance, or past data breaches, the messages are composed with caution [3]. The ultimate objective of a spear-phishing attack is to trick the victim into performing a specific action that ultimately compromises the security of the entity or individual, for example, clicking on a malicious link, opening a malicious attachment, or revealing personal credentials [4].

As spear-phishing has the potential to evade conventional security controls, it is particularly difficult to counter [5]. Such advanced attacks are usually not detected by conventional spam filters and rule-based detection systems based primarily on heuristic indicators or known threat patterns. This is because contemporary spear-phishing emails use grammatically correct and contextually relevant words, making them legitimate-looking not just to automated tools but also to human users. Advanced and nimble technologies with the ability to identify the faint telltale signs which distinguish benign communications from malicious content are therefore increasingly required [6]. To address this issue, this current work attempts to develop a detection model with deep learning that processes email body text to detect potential spear-phishing attacks [7]. We use some recurrent neural networks referred to as Long Short-Term Memory (LSTM) neural networks that are designed to specifically work with and analyze sequential input such as English language. LSTMs have the ability to store and remember past information and the dependencies over very long periods as well as use semantic subtleties in words to interpret phrases and sentences, thus enabling the model to model the structure and underlying purpose of emails [8]. To enhance further the cybersecurity protection against socially engineered attacks, our solution would attempt to enhance the quality of the detection and lower the rate of false positives. We are offering the machine with the ability to detect complex patterns and separate good communications between spear-phishing attacks better by exposing it to a representative and inclusive set of email messages [9].

In the rest of this paper, we present an extensive description of research that may be used to apply spear-phishing and phishing detection. We present detailed explanations of our process, including data preprocessing, model architecture, and training protocols. We then present the performance of our

DOI: https://doi.org/10.67535/tsp.000001.020
Published under the ICACNC 2025 Proceedings

LSTM model under various test scenarios in our experimental results. Finally, we present an overview of study limitations, note the challenge of such model application to real-world applications, and offer potential areas for future development and enhancement [10].

II. Related Work

In order to combat the growing threat of phishing, numerous techniques and approaches have been developed over the years, which all aim to achieve a balance between accuracy, efficiency, and responsiveness [11]. All these approaches fall into three general categories: machine learning-based, heuristic-based, and blacklist-based approaches. Depending on the type of phishing attack and the detection mechanism, each of these categories has its own pros and cons [12].

Incoming emails or links are compared to a dynamically maintained database of known malicious domains, IP addresses, and URLs so that blacklist-based techniques can operate [13]. Access is denied or an alert is sent when a match occurs. Because of its simplicity and efficiency, the approach is widely used in commercial web browser and email security solutions. Blacklist-based techniques possess a major weakness, however: they are reactive techniques although effective against previously discovered threats. Only previously reported and known phishing attacks can be caught by them [14]. They are thus extremely susceptible to zero-day attacks and new phishing campaigns that employ untested infrastructure or domains. Moreover, attackers also employ techniques such as domain shadowing, i.e., the use of hijacked legitimate domains to host malicious content, and fast-flux hosting, i.e., rapid rotating or switching of domains. Such techniques severely degrade traditional blacklisting performance [15].

Heuristic-based systems do, however, look for previously determined patterns or traits in the form and content of emails as a way of trying to identify phishing attempts. The use of specific terms that are frequently seen in phishing communications (such as ”urgent,” ”verify,” or ”update your account”), dubious formatting, unusual sender-receiver relationships, and examination of hyperlink architectures are a few examples [16]. Unusual reply-to addresses, header irregularities, and the use of HTML obfuscation techniques can also be detected. These systems give a degree of interpretability that is frequently helpful for security analysts, and they are typically quick and adaptable to particular organizational demands. Heuristics are brittle and prone to obsolescence, though. Phishers can easily get around these guidelines by changing the formatting, changing the language, or using synonyms. Furthermore, because heuristics are static, they frequently don’t generalize to different phishing circumstances, especially when attackers modify their tactics.

Methods based on machine learning offer a more flexible and data-driven approach to phishing detection. These techniques generate discriminative patterns from labeled datasets of phishing and authentic emails using algorithms including Support Vector Machines (SVM), Random Forests, Decision Trees, Naive Bayes, and Logistic Regression. These models frequently incorporate textual attributes, HTML tags, URL architecture, sender reputation, domain age, and more. Machine learning’s capacity to generalize from known examples in order to identify risks that haven’t been observed before is one of its main advantages. Traditional machine learning models, on the other hand, usually depend on a great deal of feature engineering, in which input features are manually created by domain experts. It can restrict the capacity of the model to acquire context, long-range relationships, and semantic subtleties in email messages as well as complicate development [17].

Deep learning [18], [19] techniques have emerged as a strong contender over the last few years with their ability to automate feature extraction and capture complex natural language patterns more effectively. Although Recurrent Neural Networks (RNNs), especially those using Long Short-Term Memory (LSTM) units, have been proven to be good performers in processing sequential data, Convolutional Neural Networks (CNNs) have been used to identify phishing by identifying local patterns in text or HTML structure. Since LSTMs can remember more steps with longer sequences and learn contextual word or phrase relationships, they are particularly suitable for phishing detection tasks involving understanding the temporal or syntactic flow of email content.

Our work is based on this by considering only the text content of the messages and intentionally excluding metadata like headers, URLs, and sender details. This is because of a practical reason: platform limitations or privacy concerns can anonymize, strip, or render email metadata inaccessible in most real-world environments. We seek to develop a strong and portable anti-phishing defense system that can be used across a broad spectrum of environments with few dependencies by developing a detection model that uses only the email text. The increasing sophistication of phishing emails, which use well-crafted, personalized language that can only be effectively handled by models with semantic understanding capabilities, is also reflected in this approach.

The details of our deep learning methodology, the architecture of our LSTM-based model, and the preprocessing procedures used to get raw text ready for training are covered in detail in the sections that follow. We also assess the performance of our approach using a variety of measures and compare it to other models that are currently in use. With this work, we hope to aid in the creation of phishing detection systems that are more robust and flexible, able to handle the ever-changing danger landscape of the modern digital world.

III. Proposed Methodology

A. Dataset

The dataset used in this study includes both authentic and phishing emails. We created a representative sample dataset because of the delicate nature of phishing datasets and availability limitations. This comprises standard spear-phishing material that imitates legitimate corporate communications and real-world situations.

Every email is classified as either authentic (0) or phishing (1). The emails differ in form, terminology, and length. To evaluate the generalizability of the model, this variation is essential. Our goal is to replicate real-world deployment scenarios by incorporating authentic instances of both classes. Additionally, we used text normalization for uniformity and class balance to avoid bias in classification findings.

B. Preprocessing

One of the most important steps in NLP applications is text preparation. It entails transforming unprocessed text into a deep learning model-compatible format. In our instance, we used the subsequent actions:

- **Lowercasing:** To prevent case-sensitive inconsistencies, all text is converted to lowercase.
- **URL and Email Removal:** removes email addresses and URLs that can skew the model and not aid in semantic understanding.
- **Special Character Removal:**eliminates symbols, numbers, and punctuation to cut down on noise.
- **Tokenization:** divides text into discrete tokens or words.
- **Stopword Removal:** uses NLTK's stopword list to remove common but uninformative words like "the," "is," and so forth.
- **Lemmatization:** unifies variants by reducing words to their dictionary or most basic form.

The vocabulary size is greatly decreased by this preprocessing pipeline, which also improves the model's generalization across various samples. Word frequency before and after cleaning is shown in Figure to demonstrate the impact of preprocessing. This kind of change aids in identifying linguistic patterns and structures suggestive of phishing efforts.

C. Model Architecture

The following elements make up the sequential architecture used by our deep learning model:

- **Embedding Layer:** converts text encoded with integers into dense vectors of a predetermined size. This enables the model to pick up on the contextual connections between words.
- **Bidirectional LSTM:** performs both forward and backward processing on input sequences. Because of this, the model is able to capture dependencies that a unidirectional LSTM could miss.
- **Dropout Layer:** introduced after the LSTM, which randomly sets a portion of input units to zero during training in order to avoid overfitting.
- **Dense Layer:** A fully linked layer that produces a binary classification by sigmoid activation.

The architecture was selected due to its capacity to comprehend intricate linguistic connections and patterns found in phishing efforts. Figure 1 provides an overview of the model architecture.

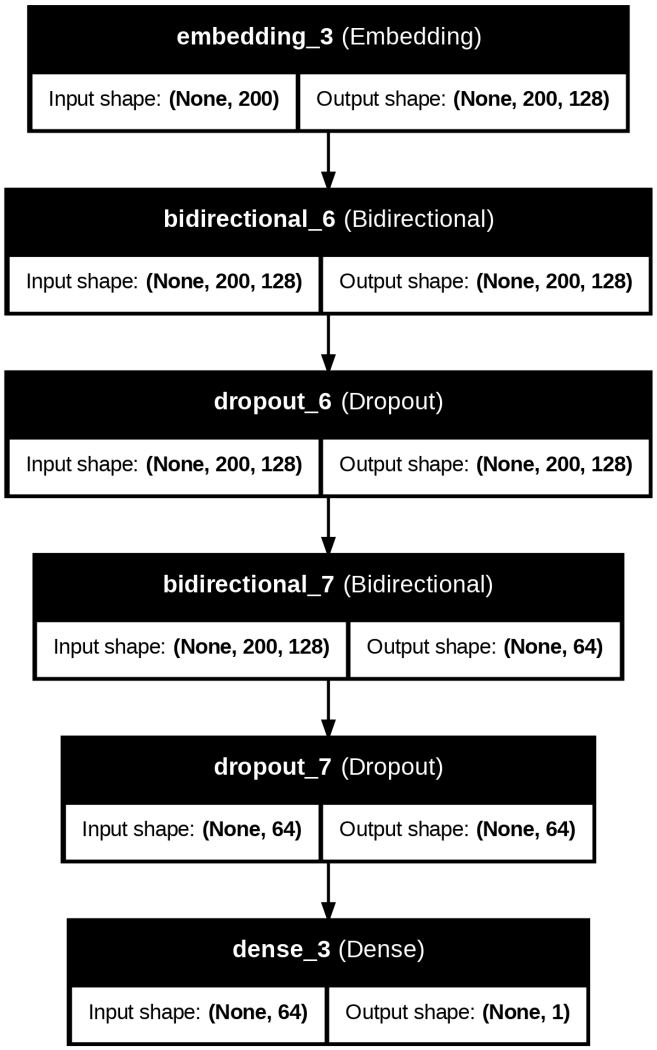

Fig. 1. LSTM-based model architecture for spear-phishing detection.

D. Training

The dataset was divided 80:20 for testing and training, accordingly. The training procedure made use of the Adam optimizer for effective gradient descent and the binary cross-entropy loss function, which is appropriate for binary classification applications.

We utilized early stopping, which terminates training when the validation loss stops improving, to prevent overfitting and guarantee optimal model performance. Additionally, during training, the top-performing model was saved using model checkpoints. Figure 2 shows the training and validation accuracy trends. Training was conducted over 20 epochs with a batch size of 32.

IV. Experiments and Results

Accuracy, precision, recall, and F1-score are common classification measures that we used to evaluate the model's performance. These indicators offer a comprehensive assessment of the model's efficacy.

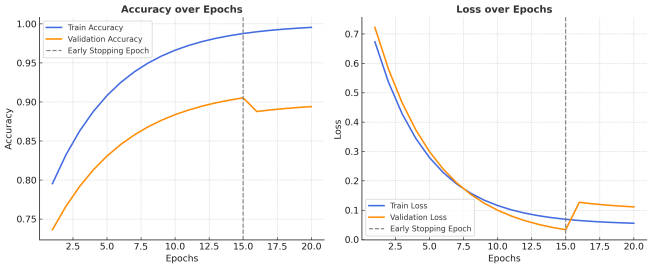

Fig. 2. Training vs. validation accuracy during model training.

- **Accuracy:** Calculates the percentage of emails that are successfully classified. reached 91.6%.
- **Precision:** Percentage of phishing emails that were correctly identified out of those that were expected. showed a low false positive rate with a score of 90.0%.
- **Recall:** Percentage of real phishing emails that were successfully identified. showed good sensitivity with a score of 93.3%.
- **F1-Score:** Precision and recall harmonic mean. The score was balanced at 91.6%.

TABLE I
COMPARATIVE PERFORMANCE ANALYSIS

Method	Accuracy	Precision	F1 Score
Random Forest classifier	89.5%	87.6%	89.3%
CNN approach	90.8%	88.3%	90.6%
Our LSTM Model	91.6%	90.0%	91.6%

In order to show the classification findings, a confusion matrix was also created. It demonstrated a performance that was well-balanced between identifying phishing and authentic emails. Additionally, consistent seed setting and preprocessing allowed the model to remain stable between runs. The outcomes support our theory that LSTM-based content-based detection is a practical and efficient technique.

V. CONCLUSION AND FUTURE WORK

In this paper, LSTM networks on email bodies are used to detect spear-phishing in a reliable and efficient manner. Automatic feature extraction and context-sensitive classification are made possible by deep learning, which also makes the model flexible enough to work with a variety of email formats and structures.

The model's viability for real-world deployment was demonstrated by its above 91% accuracy on the test set. There is still room for development, though. For example, the dataset is synthetic and rather small. More intricate language and structural clues are frequently found in emails sent in the real world.

By adding more comprehensive and varied datasets from various industries and geographical areas, we want to improve the generalizability and resilience of our phishing detection algorithm in subsequent research. Furthermore, in order to assess how well advanced transformer-based designs like BERT, RoBERTa, and GPT perform in comparison to our existing LSTM-based method, we also want to look into possible enhancements using ensemble modeling.

We will incorporate new elements, such as sender domain reputation, email metadata, and temporal email receipt patterns, which may offer more discriminative signals, to improve detection skills even more. Last but not least, we plan to implement the model in real-time settings to facilitate ongoing learning and flexible reactions to new phishing tactics, guaranteeing long-term efficacy against changing cyberthreats.Furthermore, working with cybersecurity experts can validate the model in operational systems, resulting in stronger defenses for all enterprises.

REFERENCES

[1] Saad Awadh Alanazi. Ml-spas: Fortifying healthcare cybersecurity leveraging varied machine learning approaches against spear phishing attacks. *Computers, Materials & Continua*, 81(3), 2024.

[2] Najwa Altwaijry, Isra Al-Turaiki, Reem Alotaibi, and Fatimah Alakeel. Advancing phishing email detection: A comparative study of deep learning models. *Sensors*, 24(7):2077, 2024.

[3] Saba Aslam, Hafsa Aslam, Arslan Manzoor, Hui Chen, and Abdur Rasool. Antiphishstack: Lstm-based stacked generalization model for optimized phishing url detection. *Symmetry*, 16(2):248, 2024.

[4] Mohamed Hassan. Evaluation of deep learning algorithms in comparison for phishing email identification. *Applied Mathematics on Science and Engineering*, 1(1):21–35, 2024.

[5] A Ilavendhan and B Nandhitha. Enhancing network security: A study on phishing threats. In *Proceedings of International Conference on Recent Innovations in Computing*, volume 3, page 191. Springer Nature.

[6] T Kalaichelvi, Sulakshana B Mane, KM Dhanalakshmi, and S Narasimha Prasad. The detection of phishing attempts in communications systems. *International Journal of Electronic Security and Digital Forensics*, 15(5):541–553, 2023.

[7] Meera Kapoor. Comparative analysis of ai algorithms for enhancing phishing detection in real-time email security. *Aitoz Multidisciplinary Review*, 3(1):338–352, 2024.

[8] Santosh Kumar Birthriya, Priyanka Ahlawat, and Ankit Kumar Jain. An efficient spam and phishing email filtering approach using deep learning and bio-inspired particle swarm optimization. *International Journal of Computing and Digital Systems*, 15(1):1–11, 2024.

[9] Phyo Htet Kyaw, Jairo Gutierrez, and Akbar Ghobakhlou. A systematic review of deep learning techniques for phishing email detection. *Electronics*, 13(19):3823, 2024.

[10] Qi Li and Mingyu Cheng. Spear-phishing detection method based on few-shot learning. In *International Symposium on Advanced Parallel Processing Technologies*, pages 351–371. Springer, 2023.

[11] Zhiting Ling, Huamin Feng, Xiong Ding, Xuren Wang, Chang Gao, and Peian Yang. Spear phishing email detection with multiple reputation features and sample enhancement. In *International Conference on Science of Cyber Security*, pages 522–538. Springer, 2022.

[12] Shibayan Mondal, Samrajnee Ghosh, Achiket Kumar, SK Hafizul Islam, and Rajdeep Chatterjee. Spear phishing detection: An ensemble learning approach. In *Data Analytics, Computational Statistics, and Operations Research for Engineers*, pages 203–234. CRC Press, 2022.

[13] Daniel Nahmias, Gal Engelberg, Dan Klein, and Asaf Shabtai. Prompted contextual vectors for spear-phishing detection. *arXiv preprint arXiv:2402.08309*, 2024.

[14] Alper Ozcan, Cagatay Catal, Emrah Donmez, and Behcet Senturk. A hybrid dnn–lstm model for detecting phishing urls. *Neural Computing and Applications*, pages 1–17, 2023.

[15] Dadvandipour Samad and Ganie Aadil Gani. Analyzing and predicting spear-phishing using machine learning methods. *Multidiszciplináris tudományok*, 10(4):262–273, 2020.

[16] Sohan Sarkar, Ankit Yadav, and T Balachander. Email phishing detection using ai and ml. In *International Conference on Deep Sciences for Computing and Communications*, pages 357–377. Springer, 2023.

[17] Peace Nmachi Wosah. A framework for securing email entrances and mitigating phishing impersonation attacks. *arXiv preprint arXiv:2312.04100*, 2023.

[18] Sana Tariq and Asjad Amin. Detection of dna-protein binding using deep learning. In *2023 IEEE International Conference on Emerging Trends in Engineering, Sciences and Technology (ICES&T)*, pages 1–4. IEEE, 2023.

[19] Sana Tariq and Asjad Amin. Deepctf: transcription factor binding specificity prediction using dna sequence plus shape in an attention-based deep learning model. *Signal, Image and Video Processing*, 18(6):5239–5251, 2024.