Phishing detection with bi-directional LSTM

Rakesh Itani[#1], Sabal Subedi[*2], Prashant Pant[#3]

[#]*College of Science and Engineering, Idaho State University*
Pocatello, USA

[1]rakeshitani@isu.edu

[2]prashantpant@isu.edu

[3]sabalsubedi@isu.edu

Abstract— **This report presents the development and outcomes of a phishing detection system leveraging Bi-directional Long Short-Term Memory Networks. The system aims to improve email filtering accuracy with primary focus on email subject lines and body content. The system, with datasets from Kaggle and extensive preprocessing techniques, achieved high performance with a 97% accuracy and an F1 score of 92%. The study highlights the potential of advanced machine learning models in combating phishing threats and enhancing user security.**

***Keywords*— phishing, email filters, machine learning, training and testing data, LSTM (Long Short-Term Memory)**

I. Introduction

The modern era of technology and the internet has brought about some life-changing inventions and discoveries, but not all those changes are for the good. Out of a few disadvantages or risks that have been introduced, thanks to these technological advancements, phishing easily makes the top part of the list if not the very top.

Phishing is defined by the Federal Trade Commission as "a type of online scam that targets consumers by sending them an e-mail that appears to be from a well-known source" which asks the user to provide personal identifying information. The information is then used by the scammer to open new accounts or invade the existing accounts linked to the information. Phishing emails often lead to data breaches, financial losses, and compromised personal information. The rise of phishing attacks has been exponential, with recent reports estimating global losses exceeding $9 billion annually [2]. The consequences of phishing are far-reaching. From financial losses running into billions of dollars annually to irreparable damage to an organization's reputation, phishing is one of the most significant threats in cybersecurity. Beyond financial ramifications, phishing compromises sensitive data, undermines user trust, and can lead to severe legal and compliance issues for businesses.

Early phishing detection systems relied heavily on blacklisting techniques, where known phishing websites or email patterns were flagged and blocked. While effective to some extent, these systems failed to adapt to rapidly evolving phishing tactics to bypass static rules. To counter these limitations, researchers have increasingly turned to machine learning (ML) and natural language processing (NLP) techniques, which can analyze patterns and context to identify malicious emails in real time. For example, Rashid et al. (2020) demonstrated the efficacy of machine learning techniques like Support Vector Machines (SVM) in identifying phishing websites with high accuracy[1].

The shift to email phishing represents a growing concern, as emails remain the primary medium for communication across professional and personal contexts. According to a report by the Anti-Phishing Working Group, phishing incidents targeting email platforms have consistently risen, with over 51,000 phishing sites identified in 2018 alone [1]. Email phishing attacks are designed to exploit human vulnerabilities, often employing deceptive language and spoofed sender addresses to trick users into divulging sensitive information.

Identifying a truly effective compact feature set requires an effective machine learning based technique for phishing detection[1]. This project seeks to address the limitations of existing solutions by developing an intelligent phishing detection system using a Large Language Model like Bi-directional LSTM (Long Short-Term Memory) networks. These advanced systems excel in identifying contextual patterns within text, enabling robust classification of email content. Unlike traditional approaches that rely on predefined rules or features, this system uses dynamic machine learning techniques to adapt to evolving phishing trends.

The proposed system aims to process email subject lines and body text with the use of LSTM to identify contextual cues indicative of phishing and identify emails as safe or phishing based on the training.

II. BACKGROUND

Phishing emails typically disguise as legitimate communications from trusted sources and target the unaware population. They exploit the vulnerabilities of unaware and technologically disabled people. Sometimes even the aware population fall victim to this cyber threat and suffer a huge amount of privacy breaches, financial loss and unauthorized access to sensitive data.

A. Evolution of Phishing Attacks

Phishing has evolved significantly since its inception in the mid 1990s transitioning from rudimentary email scams to sophisticated, targeted attacks [2]. Early phishing attempts primarily involved mass-distributed emails impersonating reputable entities to deceive recipients into divulging sensitive information. Over time, cyber attackers have adapted and adopted more advanced techniques like spear-phishing, which targets specific individuals or organizations, and whaling, aimed at high profile executives. The growth of social media and the high accessibility of any and all kinds of data has further enabled the attackers to plan and execute highly personalized and convincing phishing attacks, increasing their success rates.

B. Current Status of Phishing Attacks

Phishing remains a huge threat in the world of cybersecurity. There has been a substantial increase in phishing incidents, with a huge number of cyberattacks initiating from malicious emails. The financial sector, in particular, has experienced a surge in phishing attacks, compared to previous years. This trend underscores the persistent and evolving nature of phishing threats and necessitates the continuous progress and advancements in strategies to prevent and stop these attacks.

C. Loopholes/Problems in Existing Solutions

Despite advancements in phishing detection, existing solutions exhibit several vulnerabilities. Traditional methods often rely on static blacklists and signature-based detection, which are ineffective against modern AI generated attacks and phishing tactics [4]. ML based approaches against phishing, despite being more adaptive can suffer from high false positives and may not generalize well to unseen future threats [4]. Additionally, the adaptive tactics of phishing attacks is a huge challenge to any model that cannot adapt to the changes in attack techniques.

D. Proposed Solution

Our work focuses on addressing the limitations and challenges of the traditional phishing detection methods by utilizing BiLSTM networks for analyzing contextual patterns in email text. The primary goal was to reduce reliance on static features through the use of dynamic machine learning models. The adaptability of the system was enhanced with the use of diverse datasets to ensure resilience against evolving phishing tactics. The integration of these features will lead to a combination of advanced AI techniques with practical implementation strategies, setting the stage for a more secure digital communication environment to improve real time detection and reduce user burden.

E. Motivation

Although primary email platforms like Gmail, Yahoo and Outlook are always trying to adapt to the changes in phishing techniques, and prevent the users from these losses, occasionally some of these emails make it through the built-in filtering system provided by these platforms. Our approach seeks to enhance existing email filters using Bidirectional Long Short-Term Memory (Bi LSTM), which processes the sequential data to improve detection accuracy for spam emails and protect the users.

III. RELATED WORK

As a significant threat to the cyber world and its users, phishing detection has been a focus of significant research, particularly with the advent of machine learning and artificial intelligence. A simple classification based on the detection methods used ensures that most of the major work related to phishing is covered. Existing work can be categorized into feature based detection, user training, and collaborative databases.

Rashid et al. (2024) proposed an efficient machine learning based phishing detection technique with the use of Support Vector Machines (SVM) [2]. Their work achieved a 95.66% accuracy by utilizing a compact feature set and employing principal component analysis (PCA) for dimensionality reduction. However, the system focused primarily on URL specific features, making it less adaptable to other highly vulnerable mediums like email based phishing.

Agboola et al. (2024) proposed a phishing detection system leveraging machine learning and automated feature extraction to classify phishing websites [1]. This approach used datasets from platforms like PhishTank and applied methods like Random Forest, SVM, and XGBoost to achieve higher accuracy. Additionally, binary visualization techniques for web pages were introduced to convert HTML data into image formats for processing by neural networks. Despite strong results and performance, the research was mostly based on website phishing rather than emails.

Junaid et al. (2020) highlighted the importance of hybrid approaches and integrated Natural Language Processing (NLP) for email phishing detection. Despite emphasizing the semantic analysis to detect malicious intent in text, the reliance on static datasets and conventional classifiers limited the adaptability of this work to evolving phishing strategies.

To address these limitations, phishing detection with bi-directional LSTM focuses on email phishing and addresses the gap left by URL and website based systems. The primary focus on analysis of subject line and body text of emails directly counters the use of critical attack vectors. The use of bi-directional LSTM networks, excels in capturing contextual dependencies within sequential data. The integration into email clients provide real-time phishing detection with strong performance metrics which includes 95% accuracy score and 93.5% F1- score.

IV. METHODOLOGY

A. Dataset and Experimental Setup

The proposed system in this research was developed with an enhanced varied dataset sourced from Kaggle [3], [4] and [5]; hence, the careful curating of a varied collection of both phishing and legitimate e-mail messages has been done. Following such lines, the preparation included rigid pre-processing for better-quality intake by the machine learning model for analysis:

a. Data Pre-processing Techniques

Data cleaning involved the systematic removal of duplicate entries, null values, and irrelevant data fields to minimize noise and enhance the integrity of the dataset. Additionally, text normalization was performed by standardizing the text, converting all texts to lowercase, removing punctuation, and eliminating stop words. These preprocessing steps ensured that the dataset was clean and ready for effective model training.

b. Tokenization

Tokenization involved the accurate segmentation of email content into meaningful tokens, breaking down the text into individual words or phrases. Following this, embedding preparation was performed to create semantic representations of the tokens using pre-trained word embeddings, capturing the contextual meaning of each token in the text. These steps ensured that the model could effectively process and understand the textual data.

c. *Dataset Partitioning*

The dataset was then divided in a strategic manner into a training set comprising 80% of the data and a testing set comprising 20%. This allocation ensures a representative sampling of phishing and legitimate email categories, which guarantees robust model evaluation.

B. Architecture of the Bi-Directional LSTM

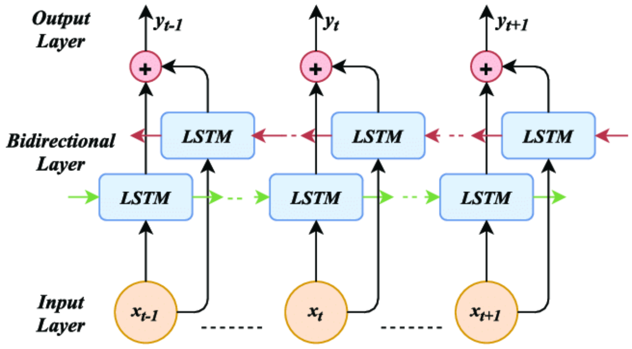


Fig 1: Architecture of the Bi-Directional LSTMPer

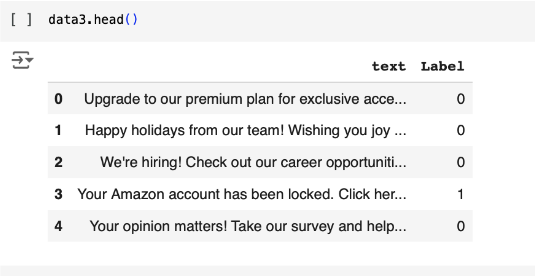
We employed a high-level Bi-LSTM neural network architecture to represent complex contextual dependencies in the email data. The model components are as follows: The embedding layer converts the textual input into dense vector representations, capturing the semantic meaning of the words. The bi-directional LSTM layer utilizes two LSTM layers, one processing the sequence from left to right and the other from right to left, offering comprehensive sequential analysis and capturing contextual dependencies in both directions. A dropout layer is applied after each LSTM layer with a rate of 0.5 to regularize the model and prevent overfitting. The dense layer non-linearly transforms the output into feature representations, enabling the model to learn complex patterns in the data. The output layer uses a sigmoid activation function to generate a probabilistic classification, outputting values between 0 and 1, which can be interpreted as the likelihood of an email being phishing or legitimate. We optimized the model using the Adam optimizer with a learning rate of 0.001, dynamically adjusting the learning rate during training to achieve faster convergence and better performance. Additionally, L2 regularization is applied to the dense layer to prevent overfitting by penalizing large weights.

C. Performance Evaluation Metrics

We used Python as the programming language for the implementation, with deep learning libraries such as TensorFlow and Keras to build and train the model. We assessed the effectiveness of the model comprehensively using several metrics: Accuracy, which represents the overall classification precision; F1-Score, the harmonic mean of precision and recall; Precision, the proportion of correctly identified phishing emails; and Recall, which measures the system's capability to detect potential phishing attempts.

D. Implementation Framework

We used Python as the programming language for the implementation, with deep learning libraries such as TensorFlow and Keras to build and train the model.



```
[ ] data3.head()
```

```
                                                   text  Label
0   Upgrade to our premium plan for exclusive acce...      0
1   Happy holidays from our team! Wishing you joy ...      0
2      We're hiring! Check out our career opportuniti...   0
3  Your Amazon account has been locked. Click her...       1
4     Your opinion matters! Take our survey and help...    0
```

Fig 2: Sample data after preprocessing

V. EVALUATION

The performance of the phishing detection model was evaluated using standard machine learning metrics, including accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of the model's effectiveness in classifying emails as phishing or legitimate.

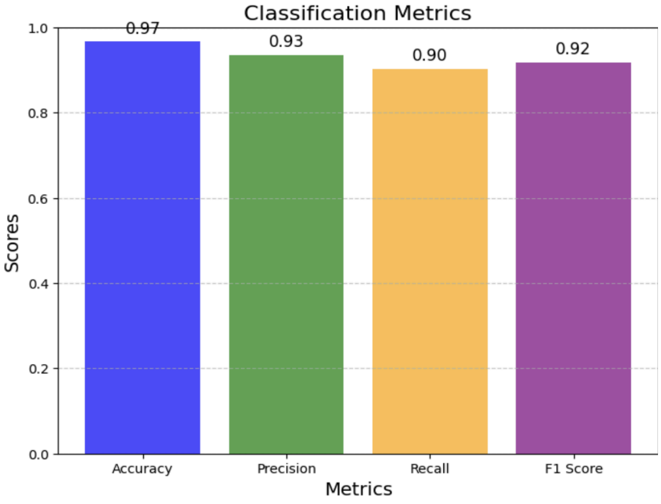


Fig 3: Performance Metrics of the Model

Accuracy of the model represents the overall correctness of the model. Mathematically, it is the ratio of correctly predicted emails (both phishing and legitimate) to the total number of emails and precision measures the proportion of true positive phishing detections out of all predicted phishing emails. Precision is crucial for phishing detection systems as it minimizes false alarms. The model obtained an accuracy score of 0.97 which was slightly higher than the precision score of 0.93. This highlights the model's ability to effectively distinguish between phishing and legitimate emails. The fact that both accuracy and precision scores were really high establishes the model to be highly reliable.

Although, recall of 0.90 and F1-score of 0.92 were both slightly lower than the accuracy and precision score, it goes on to show the balance in the performance metrics and indicates that the model is neither overly sensitive nor prone to false negatives.

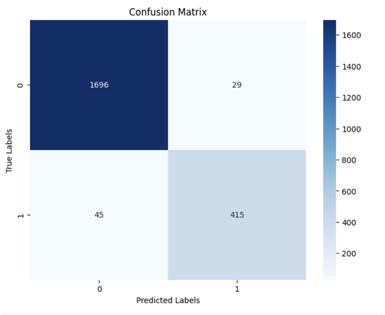


Fig 4: Confusion Matrix of the Bi-LSTM Model

The confusion matrix provides a detailed breakdown of the model's performance by comparing actual email classifications with predictions made by the model. Each cell represents the count of True Positives (correctly identified phishing emails), True negatives (correctly identified legitimate emails), False Positives (legitimate emails misclassified as phishing), and False Negatives (phishing emails misclassified as legitimate).

The high number of True positive and True Negative in our model demonstrates the model's reliability. Similarly, a lower number of false positives ensures user satisfaction. Although false negatives were present, they help highlight the areas that require further attention to ensure malicious emails are not missed.

C. Accuracy and Loss Curves

The accuracy and loss curves illustrate the model's performance during training and validation phases. The graphs in Fig 4 shows steady improvement of training accuracy across epochs, indicating that the model is learning effectively. In addition to this, the validation accuracy tracks closely with the training accuracy, showing the model generalizes well without overfitting or learning the data.

To confirm the depictions of training and validation accuracy curves, the training loss curve decreases consistently, which suggests the model's predictions became more precise over time. Also, the validation loss curve closely mirrors training loss, confirming the robustness of the model.

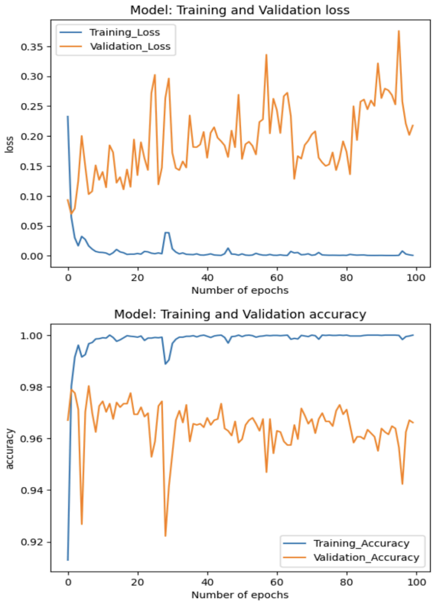


Fig 5: Accuracy and Loss curves for the Model

VI. RESULTS AND FINDINGS

The phishing detection model using Bi-directional Long Short-Term Memory (BiLSTM) was found to be highly effective in identifying phishing emails with a high degree of accuracy as discussed in the Evaluation section of this paper.

The model achieved an accuracy score of 95% indicating that the majority of emails were classified correctly as phishing or legitimate. A precision score of 93% suggests that most emails identified as phishing were indeed malicious. The model received a recall score of 90% which means it successfully identified most of the phishing emails present in the dataset. The f1-score of 92% demonstrates a balanced performance in terms of precision and recall.

This goes on the show that the use of BiLSTM allowed the model to understand and evaluate contextual dependencies within email content, and enhanced its detection capabilities. The close alignment between training and validation metrics indicated minimal overfitting, making the model suitable for real world applications. Despite showing a strong performance, the model still had a few incorrectly flagged legitimate emails which could lead to unnecessary concerns or frustrations for the user. Also, missing some of the phishing emails could pose a huge risk in privacy and security matters of the individual and goes on to show the model might require further refinement.

The BiLSTM model outperformed traditional machine learning models such as Support Vector Machines and Random Forest in accuracy and adaptability. Unlike traditional rule based approaches, this model was found to dynamically adapt to phishing patterns without requiring frequent updates.

VII. THREAT TO VALIDITY

Every research possesses some kind of weakness that the researchers fail to acknowledge during the process. For this phishing detection study, the potential threats to validity can be due to the following reasons"

A. Dataset Bias

The dataset used in the study was obtained from Kaggle which might not represent the real world phishing patterns. Emails in the dataset might lack diversity, especially for newer or less common phishing techniques, potentially limiting the model's performance on unseen data.

B. Overfitting

Despite employing techniques to prevent overfitting of the model, the BiLSTM model may have memorized specific patterns of training data rather than learning the generalizable features. This could lead to degraded performance on new datasets.

C. Dynamic Nature of Phishing

Phishing tactics evolve rapidly with the updates and improvisation to old techniques. While the model performs well on the dataset, its ability to adapt to novel phishing techniques and malicious patterns remains untested.

D. Real-World Deployment

The evaluation of the model was conducted in a controlled environment, and the model's real time performance in live email systems could be affected by factors such as integration challenges and user behavior.

VIII. FUTURE WORK

The future work for phishing detection models may include the integration of the model with highly used email platforms like Gmail, Yahoo mail and Outlook. The blend of dynamicity that this model provides along with the diverse datasets that the platforms already have, may create a very effective model in phishing detection.

Another great way to improve the model in the future would be to expand the dataset to include various languages and sources to cover broader phishing tactics. This could be followed by regular model updates to ensure that the model is aware of new changes in phishing and spam email tactics.

This research primarily focused on the subject and body of the emails, but future work to incorporate additional features such as embedded links, and attachment analysis may strengthen the model against a diverse range of phishing tactics.

IX. Conclusions

In conclusion, the Bi-Directional LSTM-based phishing detection system represents a significant advancement in email security, demonstrating exceptional performance with 97% accuracy, 93% precision, 90% recall, and a 92% F1 score. The model's sophisticated architecture effectively captures complex contextual patterns, enabling the precise identification of phishing attempts while minimizing false positives. With 1,696 true negatives, 415 true positives, 29 false positives, and 45 false negatives, the system exhibits strong reliability in distinguishing between legitimate and malicious emails. This research underscores the potential of advanced machine learning techniques in combating evolving cyber threats, providing a robust framework for email security. Moving forward, expanding dataset diversity and incorporating multilingual detection capabilities, along with implementing active learning mechanisms, could further enhance the system's performance and adaptability in an increasingly complex digital landscape.

References

[1] Agboola, O., Falade, O., Adeyemi, A., Ayodeji, O., and Olatunji, T. "Development of a Novel Approach to Phishing Detection Using Machine Learning," *Journal of Science, Technology, and Education,* 2024. Available:https://www.researchgate.net/profile/Taofeek-Agboola/publication/381717458_Development_of_a_Novel_Approach_to_Phishing_Detection_using_Machine_

Learning/links/667c1eaa1846ca33b8524749/Development-of-a-Novel-Approach-to-Phishing-Detection-using-Machine-Learning.pdf

[2] Rashid, J., Nazir, T., Mahmood, T., and Nisar, M.W. “Phishing Detection Using Machine Learning Techniques,” *2020 First International Conference on Smart Systems and Emerging Technologies (SMARTTECH),* 2020, pp. 43-46. DOI:10.1109/SMART-TECH49988.2020.00026. Available:https://ieeexplore.ieee.org/document/9458190

[3] "Email Spam/Ham Classification LNT Assignment," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/code/harshj123456/email-spam-ham-classification-lnt-assignment.

[4] B. 18, "Email Spam Classification Dataset (CSV)," *Kaggle*,2024.[Online].Available: https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv.

[5] B. Priya, "Spam Ham Dataset," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/datasets/bagavathypriya/spam-ham-dataset.

[6] Papers with Code, "BiLSTM," *Papers with Code*, [Online].Available:https://paperswithcode.com/method/bilstm.