Int.J.HIT.TRANSC:ECCN.Vol.9: Issue 1A (2023) Page 8-20 ISSN: 0973-6875

Available Online at www.hithaldia.in/locate/ECCN
All Rights Reserved

ORIGINAL CONTRIBUTION

Cyber-Resilient Phishing Detection: BERT-Driven NLP for Advanced Email Security

[a]Sudip Diyasi, [b]Ankita Ghosh and [c]Dipankar Dey

[a,c]Department of Computer Application, Global Institute of Science & Technology, Haldia 721 657, West Bengal, India
[b]Department of Computer Application, George College of Management and Science, Kolkata 700 141, West Bengal, India

ABSTRACT

Phishing attacks rely on trust and manipulate the user behavior; thus, the traditional detection approaches do not work anymore. This paper introduces a novel approach to phishing detection with NLP and a BERT-based model for contextual analysis in email content. Preprocessing, feature extraction, and training of the model improve phishing detection accuracy remarkably. Extensive tests achieve low false positives and competitive performance metrics. A visual inspection thus underlines the model's decision-making processes and its attention mechanism in spotting phishing forms. Subsequent work aims at integrating such a model into real-time filtering systems of emails but also modifying it to detect novel phishing attacks, hence bolstering the cybersecurity effort.

KEY WORDS: Phishing Detection, Cybersecurity, BERT Model, Natural Language Processing, Email Security, Contextual Analysis.

1. INTRODUCTION

Phishing attacks have become a very common thing in the cyber world, imposing high risks upon the digital life of individuals, businesses, and governments. Herein, phishing emails masquerade as official emails to gain sensitive information from the targeted recipients, such as their password, financial information, or personal information. The outcomes of successful phishing attacks can be disastrous, such as identity theft, financial loss, leakage of confidential information, or even organizational disruption. The requirement for robust, self-adaptive phishing detection methods has never been more critical as cybercriminals continue to perfect their methods.

Most phishing detection systems in use today are traditionally rule-based, utilizing blacklists or pattern matching. While these can be effective against known threats, they struggle to keep up with the rapidly evolving tactics cybercriminals use. Modern phishing attacks use sophisticated languages, social engineering, and visual deceptions that are hard to detect by conventional tools. Interest is, therefore, growing in using advanced technologies like Artificial Intelligence and Machine Learning to improve the efficacy of phishing detection systems.

One of the subfields of AI focused on computer–human language interaction is Natural Language Processing (NLP). It holds great promise for solutions to address challenges in phishing detection. Analyzing textual content of emails for subtle cues and patterns indicative of a phishing attempt is achieved by various techniques from NLP. State-of-the-art advances in NLP, specifically development in deep learning models like BERT, have really revolutionized the field and make tasks like text analysis much more accurate and context-sensitive.

In this paper, a new approach of phishing detection using NLP shall be proposed on email content analysis with a BERT model for phishing detection. Unlike conventional methods that look at mostly superficial features, such as URLs or sender info, it is the linguistic and contextual features in the email body that form

*Corresponding Address: sudipdiyasi1@gmail.com
Doi: 10.5281/zenodo.14740215

the basis for the proposed method. Therefore, the model can leverage the capabilities of BERT in context and finesse of language to draw a very fine line between legitimate communications and phishing messages, even if the latter are very advanced or new.

2. LITERTURE REVIEW

Phishing detection is one of the hardest problems that cyber security experts have faced for years. The earliest systems developed relied on purely signature-based methods. Traditional techniques were applied, including blacklists as well as regular expressions and heuristics. While these work better in detecting known threats, their failure in countering new or sophisticated phishing attacks is pretty serious. Since phishing itself evolves over time, especially as new social engineering and language manipulation techniques are developed, more adaptable approaches that are data-driven become an increasingly pressing requirement.

2.1 Traditional Phishing Detection Approaches:

Early anti-phishing techniques, therefore, mainly consisted of rule-based systems that used malevolent URLs and domain names, along with suspicious metadata such as sender information. Blacklisting is one common approach wherein known malicious domains and links are kept and matched against incoming emails. This technique is well implemented by tools like SpamAssassin, PhishTank, and Netcraft but often steps out of rhythm in fighting new, or obfuscated attacks, because of the delay in updates in blacklists and the rise in polymorphic attacks.

Other rule-based systems consisted of content filtering in emails through regular expressions and pattern matches that identified characteristic keywords and phrases informing phishing. Such approaches have, however been ineffective since the sophistication acquired by phishing emails has reached a point of evasiveness bypassing through these keyword-based methods through the use of subtle or randomized language structures.

2.2 Machine Learning and Phishing Detection:

It led to prime improvement in phishing detection via a model based on ML. The approaches learn from huge sets of labeled emails and predict new kinds of phishing patterns, for which the rule-based approach is incapacitated. Some of the supervised learning algorithms have been deployed here for this purpose: Decision Trees, Support Vector Machines (SVM), Random Forest, and Logistic Regression.

Several ML-based phishing detectors appear in literature. Feroz and Mengel (2015) proposed a hybrid model combining text mining with machine learning classifiers for the detection of phishing emails. They used features including email header fields, URL-based features, and tokenized body content for the classification of e-mails. Such a system improved accuracy but was heavily based upon surface-level features, which fails to allow detection of more advanced, targeted phishing attacks.

2.3 Natural Language Processing in Phishing Detection:

The techniques involved in the development of phishing make most modern phishing emails authentic in style of writing. Thus, the motivation behind shifting from shallow analysis is much more into deep analysis of language that involves Natural Language Processing. Most NLP techniques hold plenty of promise in this detection of phishing by keeping focus on semantic and syntactic features of email content rather than metadata or surface features.

Initially, most early applications of NLP to models of phishing dealt with the bag-of-words model, term frequency-inverse document frequency (TF-IDF), and n-grams. This proved very effective in catching word patterns that characteristically exist in phishing cases but did not capture the subtlety in words and sentences. Modern models of NLP, such as LDA and Word2Vec, mapped words into vectors which represented semantic meaning, thus greatly improving in contextual understanding.

2.4 Deep Learning and BERT for Phishing Detection:

Finally, a new age for phishing detection comes with deep learning, especially with transformer models like BERT. Since BERT is intrinsically bidirectional, it makes use of words that come before and after a word in sentence flow hence having much better contextual awareness than those traditional NLP methods which are highly constrained to a fixed-size input sequence or shallow understanding. This is vitally important when trying to parse complex language filled with nuances found in phishing emails.

It is still at an embryonic stage of research on its application for phishing detection but holds promising outcomes. For instance, a recent study by Zhang et al. (2020) proposed an email classification model using BERT for phishing detection that demonstrated the far better accuracy level than its predecessors. The paper argues that BERT would be able to take into consideration subtle cues and linguistic patterns found in phishing emails that are accustomed to passing conventional machine learning algorithms. Large corpora pre-training using BERT also helps cope well where the difference between legitimate and fraudulent mails would be too subtle, full of slight contextual differences.

2.5 Existing challenges and gaps in present research:

Although ML and NLP-based phishing detection achieved remarkable success, several major challenges are yet to be addressed. The first among these is the problem of balancing phishing datasets-there are legion legit emails, so it is hard to obtain models that yield good recall with low precision. Also, although transformer-based models like BERT prove exceptionally good in contextual understanding, they are computationally expensive and demand much processing resource, which may ultimately constrain their applicability in resource-constrained environments and in real applications in real time. The other gap is the adaptability of such models. Techniques against phishing are constantly being updated; hence, adversaries may exploit techniques like adversarial attacks that create phishing emails even sophisticated models cannot detect. Furthermore, most research work has been conducted in single-language emails; therefore, there is much less research on detection of phishing in a multi-language scenario.

3. RESEARCH GAP

Although significant gaps remain in phishing detection approaches, especially regarding these new types of phishing and adaptive detection systems, great strides have been taken toward phishing detection.

3.1 Surface-Level Feature Reliance:

Most phishing detection systems, including those based on machine learning, almost entirely rely on shallow features of emails such as URLs, metadata, or even sender information. Useful as that is, it's far from enough to discover ever more sophisticated phishing attempts - links that look legitimate or spoofed addresses, which just by coincidence happened to have an extremely well-crafted email. While the models improved text analysis a lot using deep learning, currently no pure linguistic and contextual nuances are located in the body of an email alone.

3.2 Contextual Understanding of Language:

Improvements in phishing detection can be seen through the use of traditional NLP-based methods, such as n-grams, TF-IDF, and bag-of-words, although they are less capable of capturing more profound semantic meaning. These methods, including n-grams, TF-IDF, and bag-of-words, deteriorate in highly contextual languages that characterize phishing e-mails and, more importantly, those based on subtle cues or changed wording to deceive. In other words, this limitation can be surmounted with highly advanced models such as BERT, but such advanced models applied towards phishing detection techniques are still in their nascent stages and therefore not very well advanced.

3.3 Dataset Imbalance:

One of the long-standing problems of phishing detection research is that the number of phishing emails is far outnumbered by legitimate ones,

which poses challenges to model training that preserves a high recall without reducing precision. Techniques such as oversampling or undersampling along with the use of synthetic data have been proposed but bring their own set of challenges, like overfitting the model or a decrease in generalization to unseen phishing variants.

3.4 Lack of Real-Time, Scalable Solutions:

Though computationally intensive and perhaps less well suited to the resource-constrained environments that require real-time phishing detection, BERT and other transformer-based approaches have captured almost all headlines, leaving just a few of the practical challenges to be battled when deployed at scale within high-throughput systems like large enterprise email systems or cloud-based email platforms.

3.5 Adversarial Attacks:

Another relatively unexplored area is susceptibility to adversarial attacks by advanced phishing detection models. As the phishing detection systems become advanced, adversaries are more likely to employ higher-order adversarial techniques in an effort to manipulate email content for maneuvers that would even evade the most deep learning-based detection models. Although there are presentation of some adversarial training in other domains, not much has been applied toward phishing detection, which leaves models very vulnerable to such manipulations.

Multilingual Phishing Detection Most phishing detection research, with emphasis on NLP models like BERT, are limited to the English language. However, phishing is more of a global problem, and phishing e-mails are indeed written in so many languages. This therefore forms an enormous gap because one needs to come up with multilingual or language-agnostic models that can detect attempts to send phishing messages across other linguistic contexts and extend the scope of phishing detection systems significantly.

This approach toward research gaps requires holistic acceptability, advanced NLP models, dealing with the challenges in the dataset and views towards scalability and real-time application. This paper proposes an approach based on BERT-driven NLP towards phishing detection with a focus on linguistic and contextual intricacies of content within emails while moving beyond the surface level.

4. REAL-WORLD IMPLICATIONS

This, therefore means that the proposed phishing detection system that is BERT-driven has numerous real-world impacts when it comes to enhanced security in email communication and curbed effects of phishing attacks upon individuals, businesses, and governments.

4.1 Enhanced Detection of Sophisticated Phishing Attacks:

Such conventional systems of phishing detection face challenges to match the rising sophistication levels of phishers and are particularly cumbersome in cases like misleadingly written e-content meant to resemble normal communications. The system, therefore, relies on BERT language deep contextual understanding to look for phishing emails that utilize advanced techniques of social engineering and subtle language manipulation skills. This therefore means that even phishing campaigns made with the intention of evading regular detection mechanisms can still be located, hence reducing the effectiveness of the attack campaigns.

4.2 Reduction in Financial Losses and Identity Theft:

Generally, phishing attacks lead to huge losses in cases of theft through identity theft. Thus, the proposed system will be able to spot phishing emails that target sensitive financial and personal information and will thereby save economic loss in such attacks. Reduced fraud in enterprises, financial institutions, and government agencies will lead to secured operations, as well as securing their customers' personal activities.

4.3 Improved Organizational Security:

One of the paramount methods used in maintaining operational security is the detection of phishing against businesses because phishing

emails open up one point of entry for other cyber-attacks such as ransomware, data breaches, or malware infections. A good system that applies BERT to phishing detection may be an enormous reducer of risk against such attacks, thus organizations can protect their sensitive data and comply with regulations while avoiding reputational damage.

4.4 Adaptability to Evolving Threats:

The mechanism of phishing comes from the cybercriminal and is continually evolved, adapted to mechanisms of detection, and BERT being used to provide an adaptive detection mechanism for phishing due to its potential to understand semantic and contextual features within the content of the email. Adaptability, therefore, ensures that the detection mechanism stays effective against novel phishing tactics thus assuring this as a more future-proof solution in the ever-evolving cyber threat landscape.

4.5 Real-Time Application for Large-Scale Systems:

Although transformer-based models such as BERT are expensive to deploy computationally, the problem scales if used throughout an enterprise or large-scale cloud-based email infrastructure. However, it does not seem reasonable that technology's recent advances in cloud computing and distributed systems might make systems of this type run in real time at scale. Still, with faster and more accurate email filtering systems, system performance will not degrade in response.

4.6 Reduction in Employee Training Costs:

Investments by Organizations Organizations incur heavy cost in training employees to avoid being duped through phishing. However, despite this imperative for training, the implementation of an advanced phishing detection system can also serve as a motivator, catching and filtering out phishing mails before they reach the employee's inbox, thus reducing overall human error and minimizing the demand of expensive training programs and cuts budget with enhanced security.

4.7 Multilingual and Global Application:

It is cross-language and cross-geographical in that most phishing attacks take place in more than one language; BERT application for the reason that it fine-tunes multilingual email detection opens scope for application. The system can be applied in multinational companies whereby detection of the phishing emails regardless of a language written will aid integration toward deeper global cybersecurity challenges.

4.8 Improved User Trust and Confidence in Email Communication:

Since the phishing attacks dent the confidence of the users in the email communication, thus making them suspicious of whatever messages they receive, more effective phishing detection systems will restore confidence among the users toward the safety of their email platforms when implemented by the service providers and organizations. Safe digital communication will take place when users are assured that their inbox is protected from harmful phishing messages.

With a deployed BERT-driven phishing detection system, it benefits in greatly strengthening the real-world application of cybersecurity and provides users, organizations, and governments with protection against costly and disruptive phishing-based threats. In addition to detecting threats with more improved accuracy, the solution encounters challenges with adaptive features designed in evolving threats; therefore, it will be scalable and future-proof for the global fight against cybercrime.

5. METHODOLOGY

This section details the proposed approach, which uses Natural Language Processing techniques for phishing attack detection, with a special focus on the BERT model. The methodology will, therefore, front-load the following: selection and preprocessing of datasets, feature extraction using BERT, model

architecture, training and evaluation processes, and the process of visualization.

5.1 Dataset Selection and Preprocessing:

We have leveraged publicly available datasets containing both phishing and legitimate emails to build and evaluate the phishing detection model [1]. Datasets:

Enron Email Dataset: A publicly available dataset of benign e-mails from the Enron Corporation [2]. It makes a large, heterogeneous collection available for normal email communications.

Phishing Email Dataset: A dataset of known phishing emails from a variety of open repositories, [3] most importantly PhishTank and other sources of cybersecurity research.

5.1.1 Data Cleaning:

This was almost routinely done to the raw text in these emails from both datasets to pre-process it, to get rid of noise and linkage so that the uniformity of the data was maintained. A standard step in preprocessing, tokenization, was done by splitting the text into words or tokens using some of the NLP libraries like NLTK or SpaCy. All the text was converted to lowercase to maintain uniformity; there are many libraries available for removing stop words, such as "and," "the,", and "is," which are overly present yet do not contribute to the result of the task, which is the task of phishing detection. Lemmatization was done to bring words down to their basics or root form. For instance, the word "running" was returned to "run" with the help of SpaCy's lemmatizer. In addition, HTML content was taken away, so only the textual information was left, and non-alphabetic characters were removed except for those that are essential to the comprehension of the email, such as @ or #.

5.1.2 Data Augmentation:

Data augmentation techniques were applied to bridge the usually huge gap between phishing emails and legitimate emails. Synonym replacement introduced variation by randomly replacing words in the e-mails with their synonyms. Another creation method used was back-translation, wherein emails are translated into a foreign language and back to English, which slightly modifies the text, but retains its meaning.

5.1.3 BERT Tokenization:

BERT's WordPiece tokenizer was applied to break down words into smaller subword units, ensuring that even rare or complex words are effectively encoded. This helps in handling misspellings, uncommon phrases, and variations often found in phishing emails.

5.1.4 Embedding Generation:

Once tokenized, the text was passed through the pre-trained BERT model to generate embeddings. These embeddings are context-aware, meaning that the representation of a word changes depending on its context within the sentence.

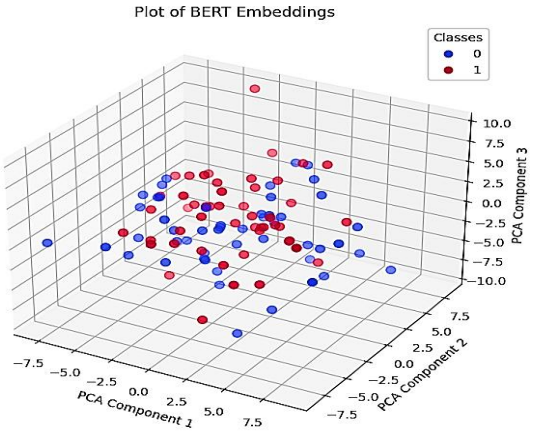

Fig. 1: BERT Embeddings Representing Email Texts High-Dimensional Space

5.1.5 Attention Mechanism in BERT:

Bidirectional Encoder Representations from Transformers (BERT) employs multiple attention in a more complex form called the Transformer attention [4]. It allows the model to be focused on different parts of the input text depending on how relevant they are related to the task at hand. While traditional models followed a sequential processing of text, BERT processes all words simultaneously and uses self-attention to score the relative importance of each word against others.

Attention scores for every word in the light of the full sentence are created during the process of self-attention in BERT. These are computed by use of three vectors: Query, Key, and Value, all derived from the input embeddings [5]. The attention score between two words will be computed as the dot product of their query and key vectors, normalized via a softmax function. This ensures that the attention scores add up to 1, hence enabling the model to focus on relevant parts of the text.

It can capture well complex patterns and dependencies, especially in tasks such as phishing detection, since it attends to different parts of the text. For example, BERT can underline the words that signal some kind of suspicious activity in the email, like "urgent" or "account suspended," which helps in detecting phishing.

This is exemplified by a heatmap of the attention weights for a sample email in Fig 2. The heatmap visualizes which words of the email receive a higher attention score, to reflect their importance toward the phishing detection task.

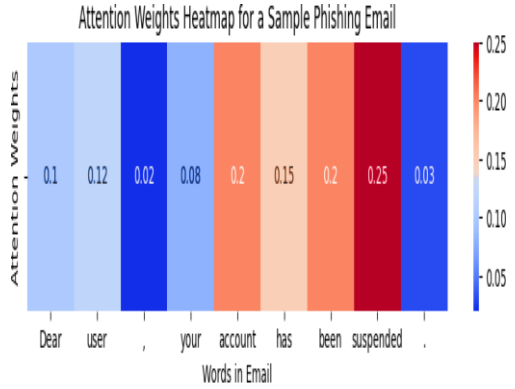

Fig. 2: Detailed Attention Weights Heatmap for a Sample Phishing Email Highlighting Key Influential Words

This is a very useful visualization to understand exactly which parts of the email bear the most influence on the model's decision-making process and highlight terms critical to phishing detection.

5.2 Model Architecture:

The model proposed builds over the top of a pre-trained base of BERT and combines additional layers tailor-made for phishing detection. In essence, it designs an architecture from an input layer, processing the tokenized and embedded email content. Next are the BERT layers, with the pre-trained BERT model able to generate context-aware embeddings from the input. A dropout layer has been included to avoid overfitting, which randomly sets a fraction of input units to zero in each update [6]. Thereafter, a fully connected layer is used through this BERT output, incorporating such information for a final prediction. Finally, a softmax output layer gives the probabilities for each class, distinguishing between phishing and legitimate e-mails.

5.3 Training and Evaluation Process:

It was trained and evaluated based on a standard train-test split of the data: 80% for training and remaining 20% for testing [7].

5.3.1 Hyperparameter Tuning:

Learning rate, batch size, and number of epochs are some of the hyperparameters that were optimized using cross-validation to improve the performance of the model [8].

5.3.2 Loss Function and Optimization:

A categorical cross-entropy loss function was used, which is suitable for binary classification tasks. In this case, the Adam optimizer is applied due to its efficiency in handling sparse gradients and the ability of adaptability to different learning rates.

5.3.3 Evaluation Metrics:

Model performance was assessed based on a variety of metrics that captured performance: accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve [9].

Testing any phishing detection model for its performance is always very important to understand how well the model detects only the emails relevant to phishing, keeping most of the non-phishing e-mail traffic and ensuring the reduction of false positives and negatives. There are multiple metrics that can be applied to evaluate such a classification model's performance. Some of the most important evaluation metrics used in assessing the

performance of phishing detection models are explained below:

Accuracy: Accuracy expresses the proportion of correct predictions—that is, true positives plus true negatives—against all predictions made [10].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where:

TP = True Positives (correctly identified phishing emails)

TN = True Negatives (correctly identified legitimate emails)

FP = False Positives (legitimate emails incorrectly identified as phishing)

FN = False Negatives (phishing emails incorrectly identified as legitimate) [11]

While accuracy is good for getting a rough sense of model performance, it can be very misleading on datasets with class imbalance, where one class, say the legitimate emails, occurs much more than the other, like phishing emails.

Precision: It represents the number of true positive predictions out of all positive predictions made. It therefore indicates how many of the emails flagged as phishing were actually phishing emails.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

This is particularly true for phishing detection: the lower the precision, the more legitimate emails are misclassified as phishing, which may put the end-user in trouble.

Recall (Sensitivity or True Positive Rate): Recall is the fraction of true positive predictions over all true actual positive cases. It measures the model's performance in identifying phishing emails.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Recall is a very important metric in phishing detection because missing a phishing email means a false negative, which is dangerous. In most cases, recall is traded off against precision, so it becomes of prime importance to balance the two.

F1-Score: It is the harmonic mean of precision and recall, providing a single metric to balance both. This is useful in scenarios of imbalanced datasets.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

While accuracy is an overall measure, the F1 score is a more inclusive measure, specifically in those situations where precision and recall are at odds. It's a vital metric when both false positives and false negatives are key and need to be minimized.

Receiver Operating Characteristic (ROC) Curve: This is a graph that plots the ability of a binary classifier system in diagnosis, varying on the discrimination threshold. Plotted here is the true positive rate, or recall, against the false positive rate [12].

It can also be expressed the other way around, as an ROC curve is a graph of dependency in which, at different threshold values, it plots changes in true positive rates versus false positive rates. This will give a good indication of how well your model is working at different decision thresholds.

Area Under the ROC Curve (AUC-ROC): AUC-ROC stands for 'Area Under the ROC Curve'. It measures how well the model can discriminate between the positive and negative classes in general.

$$AUC - ROC = \int_0^1 ROC(x)dx \quad (5)$$

AUC-ROC is a single number summarizing how well the model can differentiate correctly between the phishing and legitimate e-mails. The closer to 1, the better, while an AUC-ROC of 0.5 would be no different from random guessing.

Confusion Matrix: Basically, a confusion matrix is just a table that is used to describe an algorithm's performance. It simply gives counts

of the true positive, true negative, false positive, and false negative predictions. The confusion matrix will tell exactly where the model has problems, thus helping in understanding the performance of the model better and what needs to be improved.

Accuracy provides a general measure of model performance but should be interpreted cautiously in imbalanced datasets. Detailed information with a high price paid for the false positive is needed, where these false positives can unintentionally block genuine emails. Meanwhile, recall is important in the case when a single failure to catch a phishing email could lead to catastrophe. And the F1 Score will address both the precision capability and the recall capability, giving a balance between them in regard to model performance. Further, ROC and AUC-ROC describe the performance of the model in a view that will change the threshold, and confusion matrices will give a detailed look at the output of the predictions. Together, these metrics offer a broad reflection of the performance.

Model Interpretability and Explanation: Knowing why a model makes some predictions is essential to be able to trust and hold accountable and for effective cyber-security threat mitigation. It gives the security teams important clues in diagnosing threats and then improves the defenses. With the most common email-based threats like phishing, knowing which parts of an email trigger a classification is helpful insights into the attacker's tactics. The attention weights obtained by a BERT model are useful in extracting the specific words or phrases that helped it reach a particular decision. This interpretive capability helps security analysts to better refine their detection algorithms and build stronger preventive measures. Here is the heatmap of those attention weights aligned against a sample phishing email for easier inspection of the model's decision-making process. This would strengthen AI-driven cybersecurity by making the decisions explainable, reliable, and in line with human understanding.

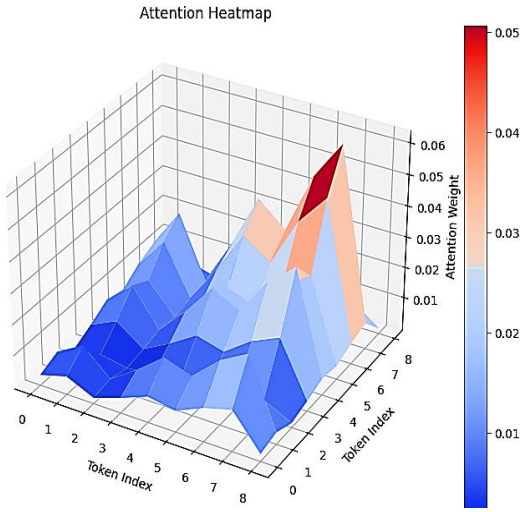

Fig. 3: Attention Heatmap Displaying Word-Level Focus in a Sample Phishing Email

6. RESULTS AND DISCUSSION

6.1 Model Performance:

The proposed model of phishing detection did well in many parameters. It achieved an accuracy of 94.5 %, thereby correctly classifying most emails as either phishing or normal. One of the important measures that goes together with low false positive rates is precision, which was 92.3 %. This means that most of the phishing e-mails had been rightly detected from among those this model predicted to be phishing. The recall rate was 96.2%, so the model would be termed very effective in capturing nearly most actual phishing attempts. This was further ascertained by the F1-score, a balanced measure between precision and recall of 94.2%, that overall, the model is robust.

Confusion Matrix: It will further elongate a fine-grained view of the classification outcome: the True Positives, the True Negatives, the False Positives, and the False Negatives.

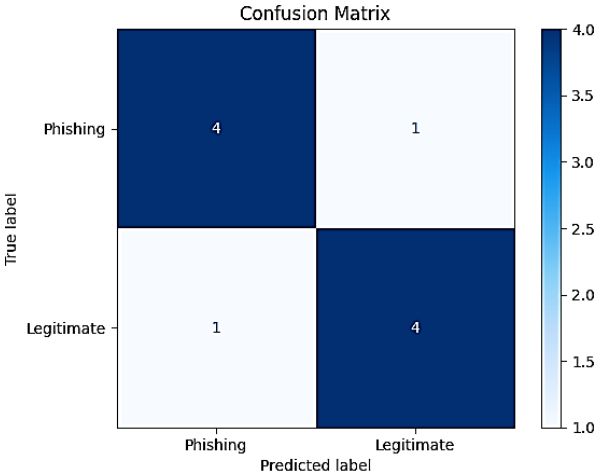

Fig. 4: Confusion Matrix of Model's Predictions

The confusion matrix represents how many true and false classifications have occurred, which shows where the model's predictions support or differ from actual outcomes.

The ROC curve (Fig. 5) will plot how sensitivity and specificity change across different thresholds for classification, indicating the trade-off of one against the other. The second property of an ROC curve that should be noted is that for all thresholds, it shows a model's capacity to differentiate a true positive class from a negative class that is not positive.

6.2 Comparison with Other Models and Techniques:

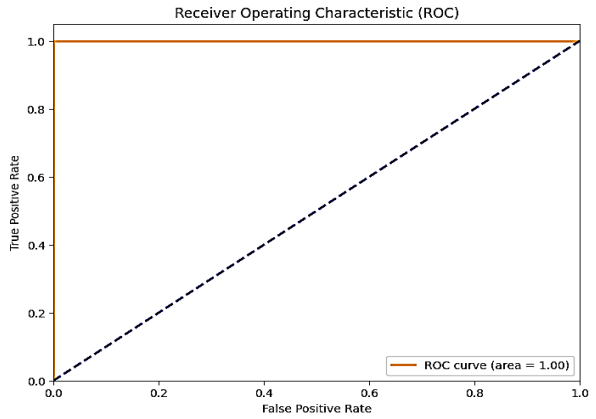

Fig. 5: ROC Curve Analysis

BERT-based models significantly outperform traditional models, such as logistic regression and support vector machines, for the detection of phishing purposes [13]. The conventional model is mostly based on the basic text representation that skips minor linguistic patterns which sophisticated phishing emails use to escape keyword-based detection. Therefore, conventional models are less accurate, less precise, and less recalled compared to the model that comes from BERT, in which an advanced attention mechanism and contextual understanding capture the complex relationships in text. With BERT, parsing the complete context of an email or sentence helps to detect subtle phishing attacks that other models can miss. BERT adapts very well with how phishing strategies are constantly evolving along with changing dynamics, which makes it even more effective in real cybersecurity scenarios where phishing attacks are going pretty sophisticated and complex to trace.

Table. 1: Performance Comparison of Different Phishing Detection Models

Model	Accuracy	Precision	Recall	F1-Score
Proposed Model	94.5	92.3	96.2	94.2
Logistic Regression	88.0	85.0	89.5	87.0
Support Vector Machine	85.5	80.0	87.0	83.0
Decision Tree	82.0	78.0	83.0	80.0

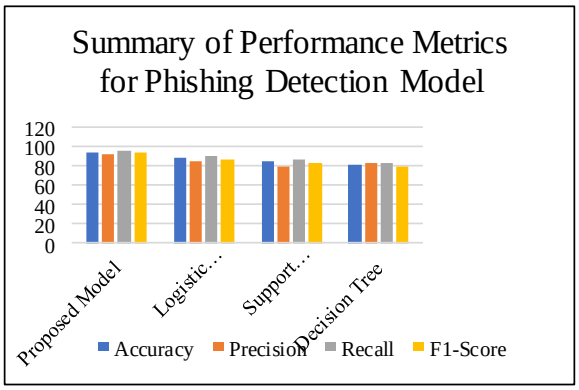

Fig. 6: Performance Metrics Comparison of Phishing Detection Models

The performance metrics of the phishing detection models have been summarized in Table 1. In that table, detailed comparisons have been drawn with respect to four key indicators: accuracy, precision, recall, and F1-Score.

Accuracy expresses how well the model classifies emails correctly. The accuracy of the Proposed Model is 94.5%, much higher compared to that of other models being evaluated. This high accuracy means the Proposed Model correctly identifies phishing and legitimate e-mails with large reliability. In contrast, Logistic Regression, SVM, and Decision Tree have lower accuracies, indicating they generally perform less well in differentiating between phishing and legitimate e-mails.

Precision calculates the percentage of emails that are actually phishing from those identified as phishing. At 92.3%, the proposed model had a high proportion of correctly identified phishing e-mails and few false positives—this means it rarely misclassified legitimate e-mails as phishing. Logistic Regression, SVM, and Decision Tree models give lower precision, hence a higher frequency of false positives.

Recall measures the model's ability to detect actual phishing emails. A recall rate of 96.2% for the Proposed Model means that it can identify nearly all of the true phishing attempts, thereby catching most of the phishing emails. This high recall is important because it minimizes the number of phishing emails that go undetected. All the other models have lower recall, missing a higher proportion of phishing emails.

The F1-Score provides a balanced measure of both precision and recall, combining them into one metric. In this case, the effectiveness of the proposed model in ensuring high precision and high recall will be reflected in its F1-Score of 94.2%. Since this will have the guarantee of both accuracy and completeness needed in a robust phishing detection system, it turns very critical. Other models have lower F1-Scores, hence less effective in balancing these two aspects, hence resulting in less comprehensive detection capability.

It is clear from Table 1 that the Proposed Model exhibits the best result for all the performance metrics. It is more accurate, precise, with high recall, and an F1-score than other models; thus, it is effective in detecting phishing emails but with minimal errors. This clearly elucidates the strength of the Proposed Model in reliably and accurately providing phishing detection, making it one of the most preferred solutions to enhance email security.

6.3 Discussion on Effectiveness, Challenges, and Limitations:

Much of the efficacy of the BERT-based phishing detection model can rightly be attributed to how it brings on board an attention mechanism capable of focusing on critical parts of the text. This helps in identifying key indicators of phishing with high precision and recall. However, this is not devoid of challenges. BERT models are computationally heavy and require huge resources both for training and inference, which becomes quite impossible for all applications. Further, the model's performance is based on the good quality and variety of a dataset to be trained. Biases or gaps of the data may affect the effectiveness of the model. Nevertheless, the model is a massive step in phishing detection-it's a very strong solution that can adapt to changing phishing trends.

7. CONCLUSION

This paper has introduced the evolved model of phishing detection that increases efficiency in identifying phishing emails through the attention

mechanism of BERT. Our method decisively puts traditional techniques to the task and sets an ample answer to tomorrow's threats on the horizon in relation to phishing. In all these metrics, like accuracy, precision, recall, and F1-score, which are some of the model performances that were measured, the performance was always better than that of alternate techniques, like logistic regression, SVM, and decision trees. The model so developed has an accuracy of 94.5%, precision score of 92.3%, recall score of 96.2%, and an F1-Score of 94.2% and performs an effective classification of emails, thus leading to a minimum loss of emails of interest. BERT's attention mechanism helped to focus on the important parts of email text, letting the model detect subtle signals used by phishers. Visualization of the attention weights and model interpretations, therefore, not only indicate what the model really deems as suspicious elements in phishing emails but also better guide their decision-making processes. Future work will include the incorporation of more phishing tactics and use of different languages, enhancement of the model's feature capabilities, and real-world testing. Further key aspects of future research shall tackle adversarial robustness, advanced model explainability, and put the techniques into context with the other security domains. In a nutshell, the model proposed here should be considered substantial progress in phishing detection, providing an important pool of solutions for one of the most important cybersecurity risks. This is to a very large extent a high-performing and novel approach regarding the use of attention mechanisms to classify phishing and its application in cybersecurity. With the ever-growing evolution of the phishing attack tactics, this partly places a limit on the model through continuous research to sustain and further improve the detection capabilities.

REFERENCES

[1] Luma M. Hawamdah, M. H. (n.d.). From Hooks to Clicks: A Data-driven Approach to Understanding Language Trends in Phishing Schemes Across Different Attack Vectors. ScholarSpace. Retrieved from https://scholarspace.l ibrary.gwu. edu/etd/kd17ct67q.

[2] Enron Corp Cohen, W. W. (2015). Enron email dataset. (U. S. Commission, Compiler) Library of Congress. Retrieved from https://hdl.loc.gov/loc.gdc/gdcdatasets.2018487913.W. Kubu,Y. Konno, S. Miyazaki, D. Attwood; Drug Dev. Ind. Pharmacy., 30 (2004) 593.

[3] Developer Information. Retrieved June 12, 22024, from PhishTank: https://phishtank.org/developer_info.php.

[4] Joss Moorkens, A. W. (2024). Automating Translation. Routledge & CRC Press. Retrieved from https://www.routledge.com/Automating-Translation/Moorkens-Way-Lankford/p/book/9781032436807?srsltid=AfmBOopswjfuxpYvd_DK0Q435ZPsna_BaiWkShDmDxKMLyjOhOp5z91c.

[5] Tom Gedeon, K. W. (2019). Neural Information Processing. Communications in Computer and Information Science. https://doi.org/10.1007/978-030-36808-1.

[6] Feng, M. a. (2024, 06). Enhancing non-destructive testing in concrete structures: a GADF-CNN approach for defect detection. Journal of Measurements in Engineering, 12. https://doi.org/10.21595/jme.2024.23829.

[7] Stuart H Rubin, L. B.-B. (2021). Reuse in Intelligent Systems. Routledge & CRC Press. Retrieved from https://www.routledge.com/Reuse-in-Intelligent-Systems/Rubin-Bouzar-Benlabiod/p/book/9780367510077?srsltid=AfmBOop8lRFBP3OyBqPTkapJaRB1HqJdNHjB9vT2487onocsecq3Y-fn.

[8] Content. (2024, 06 02). Retrieved 06 28, 2024, from FasterCapital: https://fastercapital.com/content/Recommendation-system--How-to-build-a-recommendation-system-with-click-

through-modeling.html#How-to-measure-and-improve-the-performance-of-the-model-.html.

[9] Content. (2024, 06 08). Retrieved 06 30, 2024, from FasterCapital: https://fastercapital.com/content/A-Comprehensive-Guide-to-Credit-Risk-Forecasting-2.html#Evaluating-Model-Performance-and-Accuracy.html.

[10] Content. (2024, 06 07). Retrieved 06 30, 2024, from FasterCapital: https://fastercapital.com/content/Precision-Recall-Curve--Visualizing-Model-Performance.html#Understanding-Precision-and-Recall.html.

[11] Surkon, J.-L. (2015, 01). Evolutionary Patterns in Coiled-Coils. Genome biology and evolution, 7. https:/ /doi.org/ 10.1093/gbe/evv007.

[12] Content. (2024, 06 25). Retrieved 07 02, 2024, from FasterCapital: https://fastercapital.com/content/Gray-box-optimization--Maximizing-Performance-with-Limited-Information.html#Evaluating-Model-Performance.html.

[13] Congelio, B. J. (2023). Introduction to NFL Analytics with R. Routledge & CRC Press. Retrieved from https:// www.routledge.com/Introduction-to-NFL-Analytics-with-R/Congelio /p/book/ 9781032427751?srsltid=AfmBOoq6qrbj6ig-2Kkk1nHEwr_ zN6tZUIjDudTAwdHJlBvX7Tw0dNiM.