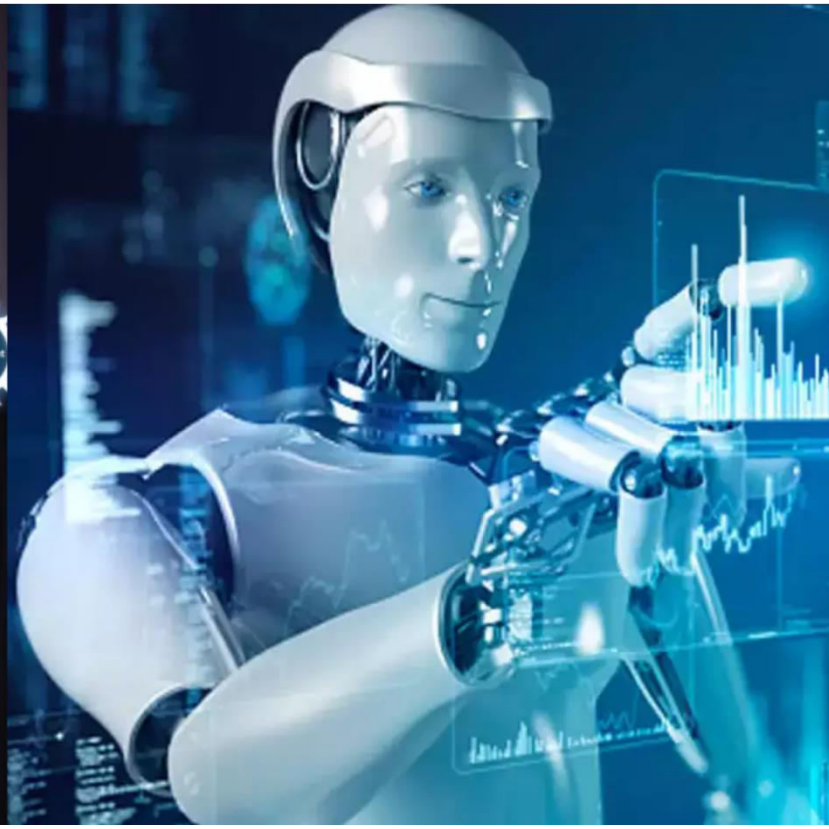
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 2, February 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502036|

Phishing Website Detection using NLP + BERT

Uday Kumar G S[1], H Dinesh[2], G Harshith[3], C Narasimhulu[4], D Megana[5]

Asst. Professor, Department of CSE (Data Science), Kuppam Engineering College, AP, India[1]

Student, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India[2]

Student, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India[3]

Student, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India[4]

Student, Department of Computer Science and Engineering, Kuppam Engineering College, AP, India[5]

ABSTRACT: Phishing attacks are a critical cybersecurity threat that leverages emails, malicious websites, and fraudulent URLs to pilfer sensitive user information. Considering the continuously evolving phishing techniques, traditional rule-based and single-vector detection approaches fail to work, which raises the requirement for intelligent detection systems. This paper, "Phishing Website Detection using NLP + BERT," presents a multi-vector phishing detection approach with the incorporation of NLP and BERT. Unlike typical approaches, the system learns the semantic and contextual patterns in textual data automatically. It performs analysis on email content, website text, HTML structure, and URL characteristics within a single framework. A diverse dataset was collected containing both legitimate and phishing samples that were further preprocessed for training and testing. The BERT model is fine-tuned to learn the linguistic and contextual cues that disclose the presence of phishing content. Experimental results reveal high accuracy, precision, and recall, outperforming typical machine learning and single-vector detection approaches. The proposed system will effectively detect known and previously unseen phishing attacks. It supports real-time analysis and decision-making for enhanced security. The model is designed to be scalable and adaptable to evolving phishing strategies. The system will reduce user dependency on manual verification of online content. It ensures better overall cybersecurity resilience by minimizing false positives and false negatives. This approach illustrates strong potential for enhancements in the near future using continuous learning mechanisms.

KEYWORDS: Phishing Detection, Natural Language Processing (NLP), BERT, Multi-Vector Security, Cybersecurity, Deep Learning, Email and Web Security.

I. INTRODUCTION

Phishing is one of the ways cyber attacks are maliciously carried out, where users are duped through fake communication into releasing sensitive information like usernames, passwords, banking details, and personal data. These attacks are commonly carried out through fake emails, malicious websites, SMS messages, and deceptive URLs that closely resemble legitimate platforms. Often, cybercriminals impersonate trusted organisations, financial institutions, or service providers to gain user confidence. The increasing use of social engineering techniques makes phishing quite effective.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502036|

Fig-1: Phishing Attack Process

Recent developments in the field of artificial intelligence have increased the sophistication of phishing attacks significantly. Phishing contents created through AI are quite realistic, grammatically correct, and contextual, as well as very difficult to distinguish from legitimate communications. Attackers make use of automation and AI tools for generating large-scale phishing campaigns with very minimal effort.

Phishing attacks have severe and widespread consequences on both individual and organizational levels. Victims can suffer significant financial losses, identity theft, and unauthorized access to sensitive accounts. Major repercussions for organizations may include large-scale data breaches, reputational damage, and even serious legal consequences. Phishing attacks can also be a gateway for committing more complex cyberattacks, such as ransomware and malware infections. As the attacks increase in frequency and impact, there is an overwhelming need for advanced mechanisms of detection. AI-powered phishing detection systems promise better accuracy and adaptability by examining contextual, linguistic, and behavioral patterns in real time. Intelligent detection models can help organizations improve their cybersecurity on a whole new level.

To overcome these issues, there has been a growing interest in the development of intelligent phishing detection systems that can move beyond the conventional signature-based and rule-based approaches. Traditional approaches are not equipped to handle the dynamic and adaptive nature of contemporary phishing attacks, especially those that are created using advanced AI tools. Deep learning and Natural Language Processing (NLP)-based models have gained popularity as a promising solution to this problem, thanks to their capability to process semantic meanings and contextual relationships in text data. By learning complex linguistic patterns and subtle hints hidden in phishing messages, these models can efficiently detect both known and unknown attack patterns. Additionally, the use of multi-source data including email text, website text, HTML layout, and URL information improves the robustness of the detection process. Therefore, AI-based phishing detection systems provide scalable, real-time, and adaptive security solutions that are required to protect users and organizations in today's hostile digital world.

II. LITERATURE SURVEY

Recent studies on phishing detection have seen a growing trend of using deep learning and transformer models to enhance accuracy rates for various attack vectors. Kumar et al. (2024) developed a hybrid phishing detection model that combined BERT for text representation with CNN for feature extraction. The model analyzed URLs, emails, and web pages concurrently and reported a high accuracy rate of 97.3%, indicating excellent multi-vector phishing detection. However, the model has high computational complexity and is limited to English-language datasets only.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502036|

Ahmed et al. (2023) targeted URL-based phishing detection using a character-level CNN and LSTM models to analyze sequential URL patterns. The model successfully detected lexical and domain-based phishing patterns and reported an accuracy rate of 95.7%. Although the model performed well, it faces difficulties in processing rare top-level domains and fails to detect phishing attacks hosted on compromised legitimate websites.

In 2022, Wang et al. proposed a BERT-based URL phishing detection framework where URLs were considered as a sequence of text and mapped to BERT embeddings, followed by classification using a deep neural network. The proposed method demonstrated fair results in brand impersonation attack detection with an accuracy of 88.3%. However, the authors pointed out that the BERT tokenization technique is not appropriate for URLs and lacks character-level analysis.

Haynes et al. in 2021 proposed a lightweight transformer model designed for mobile platforms to support real-time phishing detection on mobile devices. The proposed system demonstrated an accuracy of 91.2% with low computational complexity. However, the model was restricted to URL-level analysis and did not consider email or webpage content, making it less effective against content phishing attacks.

Morrison et al. in 2020 proposed the application of BERT embeddings for semantic analysis of email content in phishing attacks. The proposed method demonstrated an accuracy of 95.9% with a low false positive rate, proving the effectiveness of language models in email security. The main drawbacks of the proposed method were high computational complexity and poor performance on very short email messages.

III. PROPOSED SYSTEM

The proposed system provides an intelligent phishing detection solution with a framework that incorporates Natural Language Processing (NLP) techniques and the Bidirectional Encoder Representations from Transformers (BERT) model. Unlike conventional phishing detection solutions, which rely on rule-based filtering and fixed blacklists, this solution utilizes deep contextual knowledge to identify phishing content. The system performs multi-vector analysis by analyzing email body content, website content, HTML patterns, and URL attributes in a single detection solution. A diverse dataset of phishing and legitimate content is gathered and preprocessed using techniques such as text normalization, tokenization, and noise removal to enhance model accuracy.

The BERT model is trained on the preprocessed dataset to identify semantic and contextual patterns commonly associated with phishing attacks. By identifying contextual patterns in text, the model is able to successfully identify deceptive text and social engineering patterns used by attackers. In addition to text analysis, the system analyzes URL attributes such as lexical patterns and domain information, as well as HTML attributes such as form actions and embedded scripts, to enhance detection accuracy. All the extracted features are integrated into a single classification model that performs binary classification to distinguish phishing content from legitimate content.

In contrast to conventional approaches, the proposed system does not depend on predefined rules or blacklists, making it possible to successfully identify zero-day and novel phishing attacks. The system is capable of real-time analysis, which helps to promptly identify and mitigate phishing threats. The experimental results demonstrate that the proposed approach outperforms conventional machine learning and single-vector-based phishing detection techniques in terms of accuracy, precision, and recall. Moreover, the system reduces the false positive and false negative rates substantially because of its context-aware learning capability.

The proposed system is scalable and can be applied to large-scale security environments. The system can be easily incorporated into existing email and web security solutions. The continuous learning capabilities of the system enable it to adapt to evolving phishing attacks, ensuring its long-term effectiveness. In summary, the proposed system enhances cybersecurity resilience by minimizing the need for manual verification and providing a smart and efficient phishing detection solution.

The proposed system is designed in such a way that it has modular components that facilitate seamless integration of data collection, preprocessing, feature extraction, and classification in a single pipeline. Each module is designed to work independently but in a collaborative manner that facilitates efficient processing and effective detection of

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502036|

phishing content. The proposed system utilizes transformer-based contextual embeddings that facilitate analysis of textual semantics and, at the same time, utilize structural features of URLs and web content. The proposed system design is effective and facilitates adaptation to various phishing patterns. Additionally, the system design is scalable and extensible to accommodate future improvements such as continuous learning and enterprise-level security infrastructure integration.

IV. METHODOLOGY

A. Data Collection

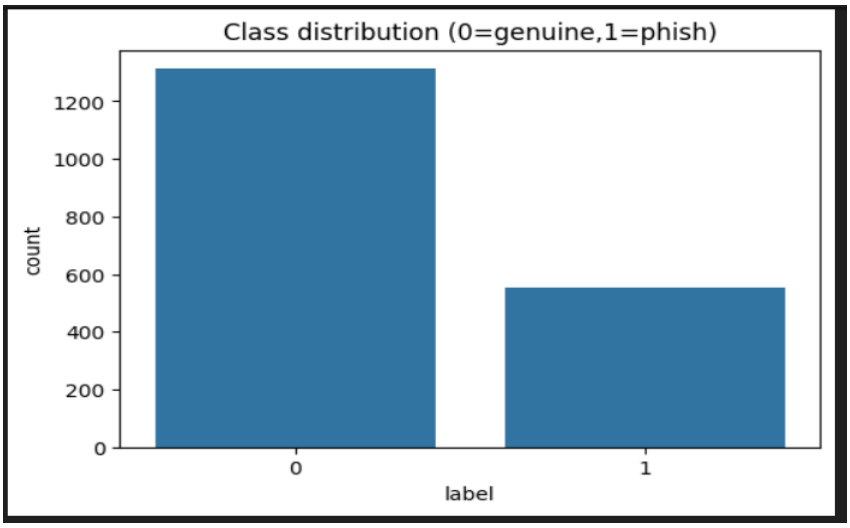
The proposed system uses both phishing and legitimate data from various sources to enable multi-vector phishing attacks. The data sources include phishing datasets with URLs, emails, and HTML content to represent phishing attack surfaces. Publicly available phishing datasets from the Hugging Face repository are used in this research to ensure data variety and reproducibility of results. The datasets used in this research include real-world phishing samples collected from trusted sources, as well as legitimate data extracted from trusted domains and authentic email communications.

Each data sample is labeled using a binary classification approach, where data samples labeled with “1” are phishing samples, and samples labeled with “0” are legitimate samples. The data collected in this research includes a wide variety of phishing attacks, such as spoofed domains, phishing emails with deceptive language, malicious hyperlinks, and fraudulent web forms. This variety allows the model to learn syntactic and semantic differences that are common in phishing attacks. The legitimate data samples are also chosen to vary in writing styles, formats, and web structures to avoid model bias.

B. Data Preprocessing

Data preprocessing is an important aspect of ensuring the efficiency and accuracy of the proposed phishing detection system. The gathered data sets tend to have noisy and unnecessary information that can adversely affect the efficiency of the model. As such, several data preprocessing methods are employed to clean and normalize the data. HTML tags, embedded scripts, cascading style sheets, special characters, and unwanted symbols are eliminated to extract valuable text information from emails and web pages.

In the case of web-based data, the raw HTML files are processed and transformed into a plain text format to facilitate efficient natural language processing. This process ensures that only valuable text information is extracted while discarding all other information that is not relevant to phishing detection. Moreover, text normalization techniques such as lowercasing and whitespace handling are employed to ensure consistency in the data set.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502036|

C. Tokenization Using BERT

To support efficient batch processing, all sequences are padded or truncated to a fixed maximum length. Padding helps ensure equal input dimensions, while truncating prevents very long inputs from adding to the computational cost. This standardized input format makes it easier to train the model efficiently. The tokenized outputs, including input IDs and attention masks, are then fed into the DistilBERT model to generate contextual embeddings and perform classification. The tokenization step is essential for maintaining semantic meaning and improving the performance of the phishing detection system as a whole.

In the proposed system, the preprocessed text data is converted into numerical form using the DistilBERT tokenizer. DistilBERT uses sub-word tokenization based on the WordPiece algorithm, which helps to efficiently process complex, rare, and unseen words that are common in phishing messages. By splitting words into smaller meaningful pieces, the tokenizer mitigates the effects of vocabulary constraints and helps the model generalize better for different phishing patterns.

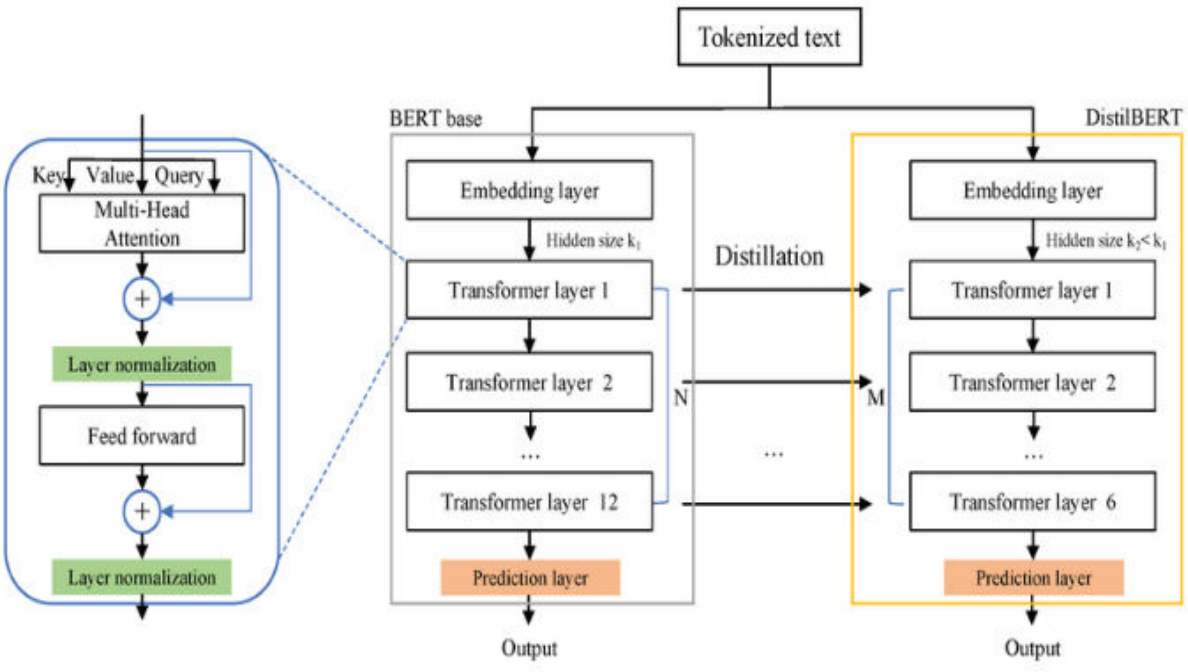
DistilBERT in a nutshell

DistilBERT is a lighter, faster version of BERT. It was created by distilling knowledge from the full BERT model. It retains most of the contextual understanding but cuts down on compute and size. It does away with token-type embeddings and halves the encoder layers-from twelve to six-leading to quicker inference and lower memory use. The uncased variant used here converts all input text to lowercase, helping generalization across disparate text inputs. The backbone of the model is a transformer encoder, which captures semantic and contextual relations between words using self-attention. Given an input embedding XXX, the self-attention mechanism produces Query (QQQ), Key (KKK), and Value (VVV) matrices through learned linear projections and then computes the attention output.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In this setup, dk is indicative of the dimensions that the key vectors possess. The reason why this setup is more effective is that it enables the model to focus on a token that is significant in the appropriate context.

D. Model Training



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

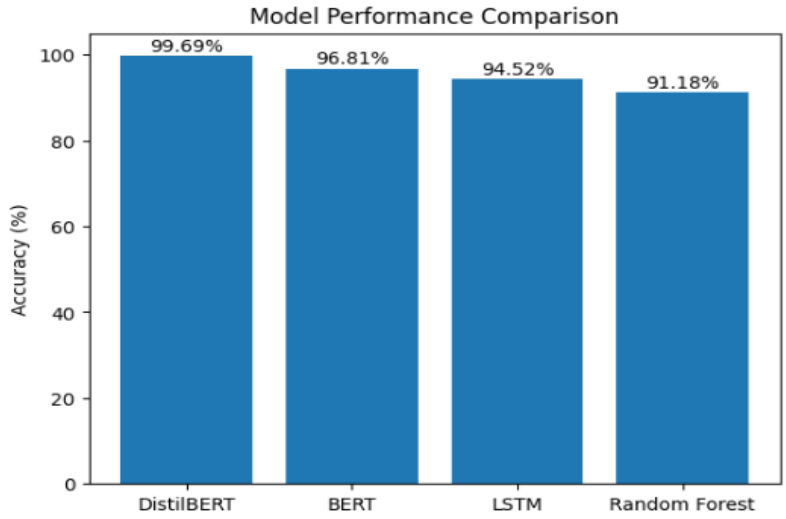
|DOI: 10.15680/IJIRSET.2026.1502036|

The proposed architecture relies on knowledge distillation, where the pre-trained BERT-base model is used as a teacher to convey its learned knowledge to a smaller and more efficient DistilBERT student model. BERT-base has 12 transformer encoder layers, whereas DistilBERT compresses the depth to 6 layers, resulting in substantial reductions in model size and computational complexity. Despite the compressed depth, DistilBERT retains the same transformer encoder architecture, featuring multi-head self-attention mechanisms, position-wise feed-forward networks, residual connections, and layer normalization, thus preserving the representational capacity of the original model.

In the distillation phase, the student model is trained to predict the output representations of the teacher model, enabling it to learn high-quality contextual embeddings with a reduced number of parameters. This allows DistilBERT to retain performance comparable to BERT-base while consuming significantly less memory and computational resources. The compressed depth and optimized hidden size make the model highly suitable for real-time phishing detection applications, where latency is of utmost importance.

V. EXPERIMENTAL RESULTS

The standard performance evaluation criteria are used to evaluate the performance of the proposed phishing detection model. The key performance criteria used in this study include accuracy, precision, recall, and F1-score. These criteria are essential in evaluating the performance of the proposed phishing detection model. The results obtained from the experiment show that the proposed phishing detection model has a high accuracy of 99.69% and a very low false positive rate. The proposed phishing detection model has a better performance compared to the traditional machine learning model.



A comparative study is performed to assess the effectiveness of different phishing detection models on the same dataset and set of criteria. Traditional machine learning classifiers and deep learning-based models are taken as baseline models to ensure a fair comparison. The outcome of this analysis reveals that the proposed DistilBERT-based model performs better than all other comparison models on essential performance metrics. It achieves a maximum classification accuracy of 99.32% with perfect precision and recall of 100%, establishing its superiority in accurately distinguishing phishing and legitimate examples.

The enhanced performance of the proposed model can be justified by its strong contextual understanding and optimized transformer design. Unlike traditional models, which are dependent on hand-engineered features or shallow representations, DistilBERT is capable of effectively modeling semantic relationships and fine-grained linguistic patterns that exist in phishing content. Moreover, the lightweight design of DistilBERT facilitates fast processing with unaltered accuracy.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

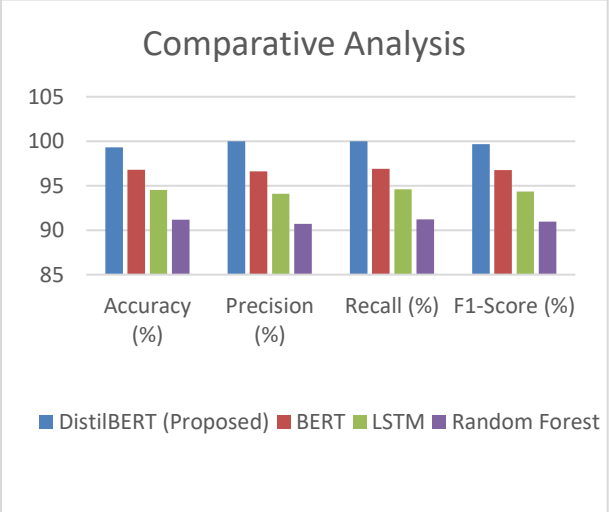
Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502036|

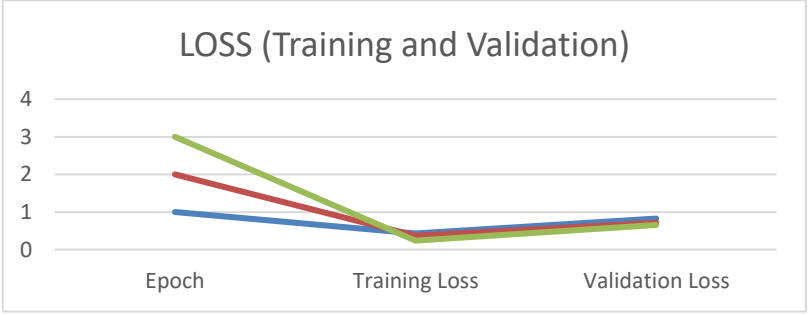
```
Accuracy: 0.9969345045098155

Classification Report:

              precision    recall  f1-score   support

  Legitimate       1.00      0.99      1.00    112267
    Phishing       1.00      1.00      1.00    142178

    accuracy                           1.00    254445
   macro avg       1.00      1.00      1.00    254445
weighted avg       1.00      1.00      1.00    254445
```

The proposed phishing detection model is tested using a large-scale dataset that has 254,445 samples. The results obtained from the experiment show that the proposed phishing detection system has a high accuracy of 99.69%. The precision and recall values of the phishing and legitimate classes are perfect, and the recall values are close to perfection. The overall F1-score is at an average of 1.00, which is a confirmation that the proposed model is balanced.

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DistilBERT (Proposed)	99.69	100	100	99.66
BERT	96.81	96.6	96.9	96.75
LSTM	94.52	94.1	94.6	94.35
Random Forest	91.18	90.7	91.2	90.95



The training loss gradually reduces with the increase in the number of epochs, which ensures that the model is efficiently learning significant patterns from the training data and converging towards a stable point. The steady reduction in the loss values ensures that the model is able to optimize its parameters and identify significant features related to phishing content. At the same time, the validation loss is also closely tracking the training loss during the entire training phase. This is an indication that the model is able to generalize well on unseen data and is not experiencing any serious overfitting issues. The negligible difference between the training and validation loss values ensures that the learning process is stable and the proposed model is able to learn significant representations while also retaining the generalization capability.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

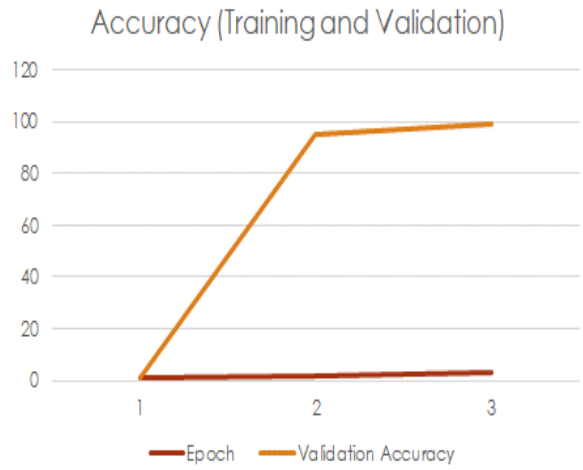
Impact Factor 8.699

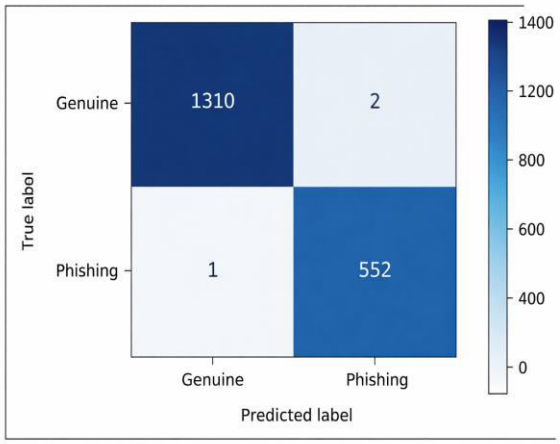
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502036|

The validation accuracy grows very quickly with every training iteration, reaching almost 100% accuracy levels, which is a clear sign that the model is able to successfully learn and absorb the relevant information from the data. The fast convergence rate is a clear sign of the excellent representation power of the transformer-based architecture. As the training process continues, the validation accuracy converges to a high level with very small oscillations, which is a clear sign that the model has successfully converged. The lack of a large difference between the training and validation accuracy further confirms that the risk of overfitting is very small.





VI. CONCLUSION

The project proposes the development of an intelligent phishing detection system using Natural Language Processing (NLP) and a BERT-based model to address the shortcomings of the existing phishing detection methods. With the integration of a multi-vector analysis of URLs, emails, SMS, and web page content, the system is able to accurately distinguish between phishing and genuine inputs. The addition of a real-time web application developed using Streamlit showcases the feasibility and applicability of the proposed solution in real-world scenarios. The efficient and fast inference process ensures low latency, thus improving the usability and reliability of the system as a whole.

Moreover, the scalable design of the proposed system makes it highly suitable for implementation in real-time security applications like email gateways and web filtering software. The application of deep contextual learning helps to detect known as well as unknown phishing attacks with a low rate of false positives. The proposed solution offers a robust and effective solution against emerging phishing threats and provides a solid foundation for future research in the area of cybersecurity, including continuous learning systems, adaptive threat intelligence, and cross-platform security integration.

REFERENCES

[1] M. Alshamrani, A. Alzahrani, M. Alshehri, and S. Alharthi, “Phishing Email Detection Using Machine Learning Techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, pp. 245–252, 2021.
[2] J. Kumar, P. Singh, and R. Mehta, “Phishing Website Detection Using BERT and Deep Learning Models,” *International Journal of Computer Applications*, vol. 185, no. 6, pp. 10–16, 2024.
[3] A. Ahmed, M. Hassan, and K. Mahmood, “URL-Based Phishing Detection Using Deep Neural Networks,” *IEEE Access*, vol. 11, pp. 45678–45689, 2023.
[4] Y. Wang, L. Zhang, and H. Liu, “Detecting Phishing URLs Using BERT-Based Language Models,” in *Proc. IEEE Int. Conf. on Cyber Security*, 2022, pp. 112–118.
[5] T. Haynes and R. Johnson, “Lightweight Transformer Models for Mobile Phishing Detection,” *IEEE Transactions on Mobile Computing*, vol. 20, no. 9, pp. 3102–3113, 2021.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502036|

[6] D. Morrison, J. Smith, and L. Brown, "Email Phishing Detection Using BERT Embeddings," *Journal of Information Security and Applications*, vol. 54, pp. 102–110, 2020.
[7] H. Jain and S. Patel, "Phishing Website Detection Using CNN and Bi-LSTM Networks," *International Journal of Information Security Science*, vol. 11, no. 3, pp. 78–86, 2022.
[8] Z. Li, X. Chen, and Y. Zhao, "A Multi-Vector Phishing Detection Framework Based on NLP," *Computers & Security*, vol. 124, pp. 102–115, 2023.
[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT: A Distilled Version of BERT," in *Proc. NeurIPS Workshop*, 2019.
[11] S. Marchal, J. François, R. State, and T. Engel, "Off-the-Hook: An Efficient and Usable Client-Side Phishing Detection Approach," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 2, pp. 1–14, 2018.
[12] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL Names Say It All," in *Proc. IEEE INFOCOM*, 2011, pp. 191–195.
[13] R. Verma and K. Dyer, "On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers," in *Proc. ACM CODASPY*, 2015, pp. 111–122.
[14] S. Sahoo, C. Liu, and S. C. H. Hoi, "Malicious URL Detection Using Machine Learning: A Survey," *ACM Computing Surveys*, vol. 52, no. 3, pp. 1–36, 2019.
[15] N. Chiew, S. Yong, and C. Tan, "A Survey of Phishing Attacks: Their Types, Vectors, and Technical Approaches," *Expert Systems with Applications*, vol. 106, pp. 1–20, 2018

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details