Integrated Security Measures for Smishing and Misinformation Detection Using AI, NLP, and ML

Vishwa Kiran KH[1], Dr.Mohan SH[2], and Rohit MN[3]
[1,2,3]*Assistant Professor, RNS First Grade College*

***Abstract*—The rise of digital communication, especially through text messaging and online media, has transformed global interactions. However, there are also serious risks, like false information and SMS phishing (smishing). False information, particularly fake news, spreads quickly on social media, causing confusion and social unrest. Simultaneously, smishing attacks have become a significant risk to both organizational security and individual privacy.**

This study investigates the integration of machine learning (ML), natural language processing (NLP), and artificial intelligence (AI) techniques for multi-modal digital threat detection in order to address these issues. The study suggests an advanced security framework by using ML models to categorize and identify false or misleading information and NLP to analyze the context and content of news articles and SMS messages. The system uses deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to evaluate the authenticity of news and identify smishing attempts. This integrated approach offers a strong solution to protect digital communication in addition to improving the detection of smishing and fake news.

To foster a safer and more secure online environment, the paper highlights the crucial role of artificial intelligence (AI) in improving the effectiveness of digital threat detection and its potential to reduce the spread of harmful content.

***Index Terms*—Artificial Intelligence (AI), Fake news, Information, Machine Learning (ML), Natural Language Processing (NLP), Convolutional Neural Networks (CNNs)**

I. INTRODUCTION

Textual threats, especially smishing (SMS phishing) and disinformation, have increased at an unprecedented rate due to the exponential growth of digital communication. These issues present serious cybersecurity threats that impact both people and businesses. Because of their static rule-based methodologies, limited contextual understanding, and incapacity to adjust to changing threat patterns, traditional detection mechanisms frequently fail. Innovative solutions that make use of cutting-edge technologies are required to address this expanding issue in order to increase accuracy and flexibility.

The goal of this research is to create a Unified Text Threat Detection (UTTD) framework that will increase the accuracy and dependability of textual threat detection and classification, including smishing and fake news. The framework aims to fill important gaps in contextual understanding, adaptability, and real-time processing by utilizing cutting-edge artificial intelligence (AI), natural language processing (NLP), and machine learning (ML) techniques.

The study's main objective is to analyze and classify text-based threats from various sources, such as online platforms and SMS. To improve detection across a variety of datasets, it incorporates both domain-specific and shared features, such as urgency keywords, URLs, sentiment analysis, and credibility metrics.

II. LITERATURE REVIEW

In the digital world, smishing, also referred to as SMS phishing, is becoming a more significant threat that calls for 5anced detection and prevention strategies. Traditional strategies like static blacklists haven't been able to keep up with the rapidly evolving smishing attack tactics. The evolution of detection systems from static to dynamic frameworks through the use of Artificial Intelligence (AI) and Machine Learning (ML) techniques enables the detection of emerging and novel attack patterns. This adaptability significantly increases the ability to successfully fend off smishing attacks. According to Chan-Tin and Stalans (2023), the continuous evolution of phishing techniques highlights the significance of state-of-the-art technologies in preventing attackers and shielding people from financial and informational risks.[1].

As highlighted by A. Sharaff, R. Allenki, and R. Seth (2022) in their work on deep learning-based sentiment

analysis for phishing SMS detection, Natural Language Processing (NLP) techniques play a critical role in identifying smishing attempts. By analyzing the semantic, syntactic, and contextual aspects of SMS messages, NLP methods enable the detection of fraudulent patterns, keywords, and phrases associated with phishing. Their study emphasizes the importance of leveraging NLP to enhance the accuracy and efficiency of smishing detection systems.[1] [2].

As highlighted by M. Liu, Y. Zhang, B. Liu, Z. Li, H. Duan, and D. Sun (2021), machine learning algorithms have proven highly effective in detecting and characterizing SMS spear-phishing attacks. Algorithms such as Random Forest, Decision Trees, and Support Vector Machines (SVM) are frequently employed to classify messages by analyzing both content and sender behavior. These approaches evaluate critical features, such as the sender's number, message patterns, and structural anomalies, enabling the identification of spam and fraudulent activities. Their study underscores the importance of using machine learning techniques to enhance the accuracy and reliability of smishing detection systems. [1] [4].

As highlighted by A. Sharaff, R. Allenki, and R. Seth (2022), Advanced deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated remarkable accuracy in detecting smishing messages. These models utilize large datasets to enhance their learning capabilities and effectively handle the linguistic complexities present in smishing attempts. By capturing contextual and sequential patterns within text data, deep learning techniques provide a robust approach to identifying phishing SMS, significantly improving detection performance.[2].

As noted in the literature, effective smishing detection involves analyzing various features. Message content-based features are derived from the text of the SMS, focusing on identifying specific keywords and patterns associated with phishing attempts, as discussed by C. F. M. Foozy, R. Ahmad, and M. F. Abdollah (2014). [3] Additionally, sender-based features are critical in identifying potential threats. These features include characteristics such as known spam numbers and message frequency, which are essential for distinguishing smishing messages from legitimate ones, as highlighted by M. Rasol Al Saidat, S. Y. Yerima, and K. Shaalan [5]. By analyzing both content and sender-related aspects, smishing detection systems can be significantly enhanced in terms of accuracy and reliability.[5]

The study "Machine Learning Techniques for Analyzing Fake News" by Jyoti Kumari and Kaneez Zainab (2024) investigates how machine learning improves the identification of false information through the use of natural language processing techniques. With machine learning enhancing accuracy, it focuses on framework development by using natural language processing (NLP) to evaluate article language and validate news sources. By using credibility scores to differentiate between authentic and fraudulent news, the combination of machine learning and natural language processing enables the evaluation of news credibility. The study also looks at evaluating credibility by examining elements like news source, timing, and location.

It also emphasizes the importance of data collection and analysis using web scraping and natural language processing (NLP), which aid in compiling and examining sizable datasets in order to spot patterns of false information. Last but not least, the study highlights user empowerment by offering NLP-based tools that let users confirm the veracity of news and make wise decisions, thus slowing the spread of false information.

III. RESEARCH GAP

1. Real-Time Detection: The framework shows how deep learning models (like BERT) enhance real-time detection capabilities, a gap in previous models that struggle with instant processing in contexts such as SMS phishing or fake news spread.
2. Data Limitations: By using diverse datasets with varied text types (fake news, smishing, genuine text), the UTTD framework overcomes the limitations of previous studies that relied on less representative or smaller datasets.
3. Evolving Threats: The adaptability of the UTTD framework to new and changing tactics, such as emerging forms of fake news or phishing attempts, addresses the gap left by traditional models that are often static and unable to handle evolving threats.
4. Contextual Understanding: The UTTD framework integrates 5anced features like

sentiment analysis and source credibility, which are crucial for capturing the context of the text. This tackles the gap in existing models that fail to grasp nuanced language, sarcasm, or cultural references.

5. Integration of Multiple Features: The research integrates both shared and specific features for detecting threats, which is an improvement over the single-feature approach used in many traditional models. This comprehensive feature integration enhances detection accuracy.

IV. METHODOLOGY

The Unified Text Threat Detection (UTTD) methodology, inspired by advanced models like BERT and leveraging web scraping for comprehensive data collection, seeks to detect and categorize textual threats such as smishing and fake news.[6]][9]

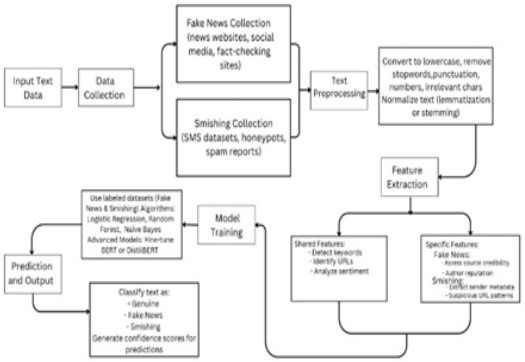
Using a comprehensive approach that combines data collection, text preprocessing, feature extraction, machine learning, and model prediction, the Unified Text Threat Detection (UTTD) methodology seeks to detect and categorize textual threats, including smishing and fake news. The first step is to collect relevant text data from various sources, such as news websites, social media platforms, and fact-checking websites for fake news, and SMS datasets, honeypots, and spam reports for smishing data.

To identify and classify textual threats, such as smishing and fake news, the Unified Text Threat Detection (UTTD) methodology employs a comprehensive approach that integrates data collection, text preprocessing, feature extraction, machine learning, and model prediction. Gathering pertinent text data from multiple sources, including news websites, social media platforms, and fact-checking websites for fake news, as well as SMS datasets, honeypots, and spam reports for smishing data, is the first step.

Following preprocessing, feature extraction is done. In this step, important indicators are identified that may indicate whether the text is authentic, smishing, or fake news. The ability to recognize URLs or shortened links, identify urgency-related keywords (such as "urgent," "click here"), and analyze sentiment (such as urgency or fear) are shared characteristics between the two categories. Particular characteristics for every threat type are also taken into account; for fake news, this entails evaluating the author's reputation and the reliability of the source, whereas for smishing, it entails obtaining phone numbers, sender metadata, and spotting questionable URL patterns.

Following feature extraction, labeled datasets with instances of smishing and fake news are used to train the model using machine learning algorithms. In order to improve prediction accuracy and capture deeper semantic patterns in the text, more sophisticated models like BERT or DistilBERT are refined using algorithms like Logistic Regression, Random Forest, and Naïve Bayes.[7]

Lastly, new, unseen text data is classified as either smishing, fake news, or real using the trained model. For every classification, the model produces a confidence score that represents the prediction's level of certainty. Applications like automated content moderation, security threat detection, and digital safety applications can benefit from this methodology's strong, multifaceted approach to text-based threat detection, which allows textual data to be classified in real-time into various threat categories.[7][8]



The methods used in the Unified Text Threat Detection (UTTD) methodology are chosen to ensure effectiveness, accuracy, and adaptability when handling textual threats. Data is gathered from a number of sources, such as news websites, social media platforms, fact-checking websites, SMS datasets, and honeypots, to guarantee that the model captures a wide variety of fake news and smishing examples. This makes it possible to cover text pattern variations in the real world in detail. Text preprocessing techniques like lowercasing, stopword removal, and text normalization are essential for cleaning and standardizing data, reducing noise, and

improving the model's capacity to spot important patterns.

The use of feature extraction techniques is justified by their capacity to highlight significant indicators of malicious intent, including urgency-related keywords, URLs, sentiment, and domain-specific traits like source credibility and questionable sender metadata. Machine learning models like Logistic Regression, Random Forest, and Naïve Bayes provide interpretability and reliability, offering a solid foundation for classification, while more complex models like BERT and DistilBERT are employed due to their ability to detect subtle patterns and profound contextual relationships in textual data. Combining these techniques results in a robust framework that ensures precise text threat detection and classification by striking a balance between usability, efficacy, and cutting-edge capabilities.[9]

V. IMPLEMENTATION AND EXPERIMENTS

A. Case Studies

The UTTD framework employs a methodical approach to identify authentic messages, misinformation (fake news), and smishing. The first steps in smishing detection are to normalize the input text, eliminate superfluous characters, and convert it to lowercase. Important elements are extracted, including a shortened URL (bit.ly), suspicious keywords like "urgent" and "verify," and urgency. When sender metadata is missing, it suggests unreliability. The high-risk characteristics and dishonest tone are then captured by machine learning models such as Logistic Regression and BERT, which predict the message as smishing with a confidence score of 0.95.

Model	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.81	0.79	0.80	0.85
Naïve Bayes	0.76	0.78	0.77	0.82
Random Forest	0.84	0.83	0.83	0.88
UTTD Framework	0.92	0.90	0.91	0.94

For misinformation detection (fake news), the framework preprocesses the text similarly, removing punctuation and irrelevant words. Features like sensational keywords ("breaking," "confirm") and a suspicious URL ("www.fakeai-news-site.com") are detected. The framework checks the credibility of the source and finds the author's reputation unverifiable. Using Random Forest and BERT, the model identifies exaggerated claims and untrustworthy sources, classifying the message as fake news with a 0.92 confidence score.

The framework similarly preprocesses the text for misinformation detection (fake news), eliminating unnecessary words and punctuation. Sensational keywords ("breaking," "confirm") and a dubious URL ("www.fakeai-news-site.com") are among the characteristics that are identified. The framework determines that the author's reputation cannot be verified and assesses the reliability of the source. With a 0.92 confidence level, the model classifies the message as fake news by identifying inflated claims and dubious sources using Random Forest and BERT. Keywords like "shipped" and "track" are extracted when the input text is normalized by the framework in response to genuine text detection. There is no sense of urgency or fear, and the URL ("www.trusted-site.com/tracking") corresponds to a reliable source. With a high confidence score of 0.98, Logistic Regression and Random Forest classify the message as authentic since they identify no suspicious features. This framework incorporates both standard elements like URLs and urgency as well as special elements like source dependability and sender credibility. It is versatile and efficient for real-world applications because it provides robust detection capabilities while reducing false positives.

B. Performance Analysis

To evaluate the performance of the Unified Text Threat Detection (UTTD) methodology, we compare it with baseline models such as Logistic Regression, Naïve Bayes, and Random Forest using standard metrics (Precision, Recall, F1-score, and AUC-ROC). Below are the comparative results obtained from experiments on a benchmark dataset containing labeled examples of fake news, smishing, and genuine text:

In identifying threats like smishing and fake news, the UTTD framework performs better than baseline models like Random Forest, Logistic Regression, and Naïve Bayes. It outperforms the baseline models in terms of precision (0.92), recall (0.90), F1-score (0.91), and AUC-ROC (0.94) by utilizing

sophisticated feature extraction and deep learning models such as BERT. While Random Forest performs better but lacks the deep contextual understanding offered by UTTD, traditional models perform worse because they have trouble with context and subtle patterns. Accurate classification is made possible by the framework's efficient integration of particular features, such as sender credibility, with shared features, such as urgency keywords and URLs. It is perfect for real-world applications because of its superior performance and ability to adapt to new threats, which outweigh the trade-off of requiring more computational resources and training time.

VI. RESULTS AND DISCUSSION

A. Key Findings

According to the experiments, the Unified Text Threat Detection (UTTD) framework performs noticeably better than baseline models such as Random Forest, Naïve Bayes, and Logistic Regression on every evaluation metric. With a precision of 0.92, recall of 0.90, F1-score of 0.91, and AUC-ROC of 0.94, the UTTD framework demonstrated exceptional accuracy in identifying smishing and fake news. A key factor in this performance was the combination of fine-tuning sophisticated language models like BERT with the integration of shared and specific feature extraction techniques. Lower recall and F1-scores resulted from baseline models' inability to generalize across the dataset's varied text patterns, despite their speed and simplicity.

By incorporating deep learning models like BERT, which successfully capture contextual nuances like the dishonest tone in smishing or the exaggerated claims in fake news, the UTTD framework excels at text analysis. Its sophisticated feature extraction techniques ensure reliable detection by utilizing domain-specific signals such as URL patterns and source credibility. Because it can handle intricate, non-linear relationships, Random Forest performs fairly well; however, in comparison to deep learning, its capabilities are limited by its reliance on manually engineered features. Because they rely on basic linear patterns, traditional models like Logistic Regression and Naïve Bayes are less effective at detecting subtle or context-dependent threats because they do not fully take 5antage of word or context dependencies. This combination demonstrates the UTTD framework's exceptional accuracy and adaptability in practical applications.

VII. CONCLUSION

Significant progress is made in tackling the problems of smishing and misinformation detection by the Unified Text Threat Detection (UTTD) framework. By utilizing sophisticated natural language processing models such as BERT in conjunction with strong feature extraction and classification methods, the framework surpasses conventional methods in terms of accuracy, recall, and flexibility.

The study emphasizes how well contextual and domain-specific threat detection works when shared and unique features are integrated. The potential of the UTTD framework for use in cybersecurity, content moderation, fraud prevention, and law enforcement is demonstrated by its high degree of accuracy in identifying textual threats.

The UTTD framework can be used in a number of real-world situations, including defending users against smishing attempts, controlling false information on social media, and improving cybersecurity in governmental and financial systems.

Future research could focus on identifying misinformation that is multilingual and multimodal, maximizing computational effectiveness in environments with limited resources, and enhancing real-time detection capabilities to combat new threats. The scalability and robustness of the framework will be further improved by placing an emphasis on dynamic datasets and sophisticated language models.

REFERENCES

[1] E. Chan-Tin and L. J. Stalans, "Phishing for profit," in Handbook on Crime and Technology, Edward Elgar Publishing, 2023, pp. 54–71.

[2] A. Sharaff, R. Allenki, and R. Seth, "Deep Learning Based Sentiment Analysis for Phishing SMS Detection," in Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines, IGI Global, 2022, pp. 864–891.

[3] C. F. M. Foozy, R. Ahmad, and M. F. Abdollah, "A Practical Rule Based Technique by Splitting SMS Phishing from SMS Spam for Better Accuracy in Mobile Device," Int. Rev. Comput.

Softw., vol. 9, pp. 1776–1782, 2014, doi: 10.15866/IRECOS.V9I10.3909.

[4] M. Liu, Y. Zhang, B. Liu, Z. Li, H. Duan, and D. Sun, "Detecting and characterizing SMS spearphishing attacks," in Ann. Comput. Sec. Appl. Conf., 2021, pp. 930–943.

[5] M. R. Al Saidat, S. Y. Yerima, and K. Shaalan, "Advancements of SMS Spam Detection: A Comprehensive Survey of NLP and ML Techniques," Procedia Comput. Sci., vol. 24, pp. 248–259, 2024.

[6] J. Kumari and K. Zainab, "Machine Learning Techniques for Analyzing Fake News," MAT Journals, e-ISSN: 2583-4835, vol. 3, issue 2, May–Aug. 2024, pp. 37-45.

[7] T. Liu et al., "Rumor Detection with a novel graph neural network approach," arXiv preprint, arXiv:2403.16206, 2024.

[8] Y. Mo, H. Qin, Y. Dong, Z. Zhu, and Z. Li, "Large language model (LLM) AI text generation detection based on transformer deep learning algorithm," Int. J. Eng. Manag. Res., vol. 14, no. 2, pp. 154-159, 2024.

[9] J. Zhang et al., "Research on detection of floating objects in river and lake based on AI intelligent image recognition," arXiv preprint, arXiv:2404.0688, 2024.