**Phishing Attacks: Leveraging Natural Language Processing (NLP) Algorithms for Ensemble Phishing Detection Model in Facebook**

A Thesis Presented to the faculty of the College of

Information Technology and Computer Studies

Pamantasan ng Lungsod ng Muntinlupa

Muntinlupa City

In Partial Fulfillment of the Requirements of the Degree

Bachelor of Science in Computer Science

**Sanchez, Aaliyah Felicity J.**

**Gayta, Jhon Vincent L.**

**Tarroza. John Angelo S.**

**MAY 2025**

**TABLE OF CONTENT**

**PRELIMINARIES Page(s)**

Title Page

Approval Sheet

Acknowledgement

Dedication

Executive Summary

Table of Content

List of Figures

List of Tables

**CHAPTER**

1. **INTRODUCTION**

Background of the Study 2

Statement of the Problem 4

Objective of the Study 6

Significance of the Study 7

Scope and Limitation 9

1. **REVIEW OF THE RELATED LITERATURE AND STUDIES**

Foundation of the Study 11

Related Literature 12

Related Studies 15

Synthesis 20

Conceptual Framework 21

Definition of Terms 23

1. **METHODOLOGY**

Materials 25

Hardware 25

Software 25

Data 26

Methods 27

Research design 27

Research Question 28

Population 31

Methods of collecting data 32

Data Analysis 32

Context Diagram 34

Algorithmic Structure 35

Algorithm Flowchart 39

Mathematical model 41

References 43

**LIST OF FIGURES**

**Figure Title Page**

Figure 1: Conceptual Framework 22

Figure 2: AI-Driven Phishing Attacks: Leveraging Natural Language

Processing (NLP) Algorithms for Phishing Detection in Facebook

Context Diagram 34

Figure 3: Flowchart of the design of the ensemble NLP phishing detection

model for URL and text features 39

Figure 4: Algorithm flowchart of the ensemble NLP phishing detection

model for URL and text features 40

Figure 5: Mathematical model of getting the word’s Term Frequency 41

Figure 6: Mathematical model representation of the Inverse Document Frequency 42

Figure 7: Mathematical model of Term Frequency-Inverse Document Frequency 42

Figure 8: The mathematical model for the binary logistic regression algorithm 42

**Chapter 1**

**INTRODUCTION**

This study proposes designing and developing an AI-powered phishing detection system tailored explicitly for the Facebook platform, focusing on phishing text and URLs. It integrates Natural Language Processing (NLP) algorithms to analyze and classify suspicious content that may appear in Facebook posts, comments, or private communications. The increasing frequency of phishing operations against Facebook users also threatens serious risks like identity theft, data breaches, and financial loss. Existing detection techniques mostly rely on user complaints and keyword matching, which are ineffective against more sophisticated phishing tactics. This study addresses that limitation using machine learning to examine linguistic patterns and deceptive structures in phishing text and URLs.

The system will gather phishing-related data from public Facebook content, phishing databases, and simulated text. NLP preprocessing techniques including tokenization, lemmatization, sentiment analysis, and named entity recognition will be applied to extract essential features. These features will learn machine learning models such as Logistic Regression, Random Forest, and Neural Networks to distinguish the data between phishing and non-phishing classes. The system will be trained, validated, and tested in order to make it extremely accurate and reliable.

The proposed detection model is designed to operate as a browser plugin or a backend integration, enabling real-time flagging of suspicious text and URLs. Through the automation of phishing detection by intelligent systems, the research intends to enhance user security and lower phishing occurrences on Facebook. In addition, it provides language composition insights of phishing attacks and delivers a scalable model for social media sites.

**Background of the Study**

With the accessibility of technology and digital communication, people are so reliant on social media to connect that platforms like Facebook have become a part of everyday life, indispensable tools for friendship building, business promotion, and information sharing. Facebook has become the dwelling place for billions of users' personal and professional data globally and, simultaneously, the most desired place for cybercriminals to carry out their actions. Phishing attacks are the most common and one of the most severe dangers the platform is experiencing nowadays. These attacks mainly involve people who are fraudulent and try to get information that is considered secret by pretending to be those they can trust. Most of the time, the attacks take the form of trusted e-mails that ask the victims for their sensitive data or to click the link to fill in their Facebook credentials. Through social engineering, the essence of the attacks has changed from a one-dimensional spam method to a highly sophisticated, targeted attack, which is much harder for the user to spot. The security patches that are in place are increasingly failing to stop modern phishing techniques.

Conventional phishing detection systems, rule-based and static blocklists, fail to address the challenges raised by linguistic exploitation and slight design changes when the phishing texts can successfully evade filters. Moved by this state of affairs, many researchers have begun immersing themselves in the potential of artificial intelligence (AI) approaches in Natural language processing (NLP) applications, which are much more dynamic and adaptable solutions that can be used to overcome such problems. With NLP, the machine can learn to process and understand human language, which then allows the detection of hidden signals, such as textual patterns, emotional cues, and intent in the written word. On top of that, NLP makes it more effective to find phishing text, which usually contains manipulative language and expresses urgency or requests for confidential information, that a well-trained AI may flaunt.

This research addresses the rising challenge of AI-driven phishing, especially on platforms like Facebook, whereby cybercriminals now cleverly opt to employ artificial means to lure users and bypass the conventional security measures these sites offer. A fundamental problem is that the existing detection system does not recognize very advanced and context-aware phishing messages created by artificial language models to look like regular communications. Most of these systems pay attention to static rules regarding matching keywords. This means that they become useless against the dynamic and linguistically rich methods that current attackers find relevant. Moreover, there is a noticeable absence of an AI-based NLP model that could be dedicated to the Facebook environment, constituting a broad range of user-generated content across posts, messages, and comments. The gap is being filled by creating an intelligent model to detect phishing content accurately through the combination of a linguistic indicator and ultimately facilitate phishing detection on Facebook vis-a-vis quickly changing mechanisms used by cyberspace threats.

**Statement of the problem**

Phishing attacks have gotten much more sophisticated with the increase in artificial intelligence (AI), with the attacks being smarter and harder to detect. Social network platforms like Facebook are most vulnerable due to the massive amount of user-generated content in the guise of false posts, malicious links, and texturized content to trick users into revealing confidential information. Traditional rule-based and signature-based detection methods rarely pick up on phishing attacks, as attackers continuously modify their methods with Natural Language Processing (NLP) techniques to mimic authentic communications. One of the most critical among these challenges is the detection of malicious captions and text. Phishing attacks on users often take the form of apparently authentic messages and captions that lead them into providing valuable information unwittingly. Traditional detection mechanisms are not sufficient to scan for contextual and semantic nuances, particularly if phishing attackers utilize AI-generated or well-crafted sentences. In such situations, these mechanisms have a tendency to overlook the slightest linguistic detail that indicates a phishing attack. Detection of phishing URLs is another crucial issue. Threats typically conceal dangerous links within posts by making them appear as legitimate. Although certain detection technologies rely on URLs, they frequently overlook dynamically generated and obfuscated URLs, common in AI-based attacks. Such a limitation subjects users to adaptive threats that slip beyond conventional security measures. Besides, existing phishing detection technologies are drastically hindered by lack of flexibility and interoperability.

The majority are still dependent on static blacklists or simple heuristics, which are useless in fighting the rapid-evolving phishing attacks. Little emphasis is also laid on the usage of integrated approaches combining NLP-based text analysis and URL-based detection for boosting precision. Absent an integrated system, present mechanisms remain siloed, decreasing their total efficiency in catching phishing material on Facebook. With these limitations, this study aims to develop an AI-powered phishing detector model that employs NLP techniques to monitor text for malicious intent as well as test embedded URLs for phishing indicators. By filling these gaps, the study will enhance detection efficiency and present a better defense against phishing attacks on social media platforms that has URL and text captions:

1. How have phishing attacks evolved to become more sophisticated and evasive compared to traditional methods?
2. What makes Facebook a primary target for phishing attacks in the context of social media platforms?
3. Where do current phishing detection solutions fall short in analyzing the semantic and contextual subtleties of malicious text?
4. When do existing systems fail to integrate text-based and URL-based analysis, and what impact does this have on phishing detection accuracy?
5. How can an AI-based model that combines NLP and URL analysis enhance the detection of phishing content on Facebook?

**Objective of the study**

This study aims to design a phishing detection model for Facebook by leveraging Natural Language Processing (NLP) algorithms. The specific objectives are:

1. To design an intelligent phishing detection framework that integrates the following key components:
   1. Natural Language Processing (NLP)-based text analysis to identify malicious intent within Facebook posts and captions.
   2. URL phishing detection mechanisms that apply heuristic and machine learning techniques to recognize suspicious or obfuscated links.
   3. A user-centric dashboard interface for presenting detection results, issuing threat alerts, and supporting seamless user interaction.
2. To implement core functionalities using appropriate technologies and tools, including:
   1. Python for backend logic and NLP-based processing.
   2. Machine learning libraries and frameworks for phishing classification and pattern recognition.
3. To conduct rigorous system validation through structured testing approaches:
   1. Alpha Testing for internal quality assurance, algorithm refinement, and bug detection.
   2. Beta Testing to gather user feedback, assess real-world applicability, and evaluate scalability.
4. To assess system quality and performance using the ISO/IEC 25010:2011 Software Product Quality Model, with emphasis on:
   1. Functional suitability for accurate phishing content detection.
   2. Reliability under diverse and dynamic input conditions.
   3. Security against adversarial manipulation and evasion tactics.
   4. Usability for ensuring accessibility and ease of use across a broad user base.
5. To deploy the solution as a browser extension, enabling phishing detection while maintaining a smooth and non-intrusive user experience

**Significance of the Study**

The study investigates a relevant issue against an emerging phishing trend through social media, especially Facebook. Sophisticated social engineering attacks take advantage of humans whose linguistic capabilities to deceive and convince would fail. The traditional way of detecting these kinds of attacks through rule-based detection systems and blocklists seems insufficient since these systems cannot adapt to new patterns or decipher the subtlety of the language utilized by the perpetrators. This advanced research proposal presents an intelligent and dynamic mechanism incorporating Artificial Intelligence (AI) and Natural Language Processing (NLP). NLP enables machines to grasp, interpret, and evaluate the contextuality and semanticity of languages in a highly fine-grained analysis of phishing-related content, way beyond mere keyword detection. The AI-based models proposed in this research can also help mitigate phishing attempts masquerading as legitimate content on Facebook by improving detection accuracy, adaptability, and real-time interventions. Specifically, explaining how NLP works to enable phishing detection provides some robust challenges to the existing cybersecurity systems, which are often outdated, to offer the very perspectives deemed essential for dealing with modern cybersecurity issues.

However, the implications of this study are far-reaching among varied stakeholders. The research enhances the comprehension of AI-based threat detection in computer science and cybersecurity by creating a working model that can be improved, modified, and extended into many online platforms. It also serves as a link between academic research and applying it to a real-world case because it shows how the use of advanced technologies such as NLP can resolve some of the most complex security problems. Developers and engineers will be offered a flexible framework integrated into their current systems to provide real-time protection; educators and students will use this working example to apply machine learning concepts to security situations. Thus, companies dealing with social media can use this research to strengthen their security platforms and increase their trustworthiness to lessen losses resulting from phishing scams. Most importantly, this work protects end-user-average social media users from the darker threats of identity theft, data loss, and other phishing-related threats. This research furthers a more considered, proactive, and AI-centric cybersecurity model by answering questions about who gains from this study, how it promotes digital safety, and how it works towards building secure online communities.

**Scope and Delimitations**

This study aims to create an AI-based phishing detection system for Facebook that utilizes Natural Language Processing (NLP) and machine learning (ML) methods, designed specifically as a browser plugin. The research will examine Facebook posts, captions, and URLs embedded within them for phishing signals, utilizing publicly available sets of phishing posts from cybersecurity stores, synthetically created phishing messages for testing, and real Facebook posts for comparison. The platform will utilize NLP models such as TF-IDF and the Logistic Regression algorithm for textual analysis to tag malicious intent within post captions with a combination of ML-based heuristics in the analysis of URLs, looking into domain reputation as well as lexical attributes to spot malicious links. The prototype will be created with Python as the programming language for the core ML/NLP functionalities and combined into a web-based interface (HTML/CSS/JavaScript) for end-user interaction, and testing in controlled environments before looking at real-world deployment. Alongside looking at conformance to ISO/IEC 25010:2011 software quality requirements for attributes such as functional suitability and security.

Nonetheless, this research is subject to some significant limitations that need to be recognized. The study is limited by its use of available datasets and simulated phishing posts, which might not accurately reflect real-time adversarial strategies, exacerbated by Facebook's data privacy policies that limit access to private posts and constrain training data scope. The efficacy of the system can be threatened by continually evolving AI-based phishing techniques, especially zero-day exploits through hitherto unknown methods, and NLP models may produce false negatives/positives by classifying sarcasm, humor, or aberrant language patterns incorrectly. Technically, the solution only supports text and URL analysis, not multimedia-based phishing (for example, image or video fraud), and is specifically crafted as a browser extension that must have active internet connectivity to operate, since the NLP models and URL analysis functions rely on cloud-based processing. This browser extension form also implies the system cannot run as an independent application or be tightly integrated with Facebook's native client without API consent. In addition, the results might not be generalizable to other social media sites such as Twitter or LinkedIn because of variations in user interactions and content architectures. These limitations serve to define crucial constraints on the system's functionality while suggesting opportunities for future development in phishing detection technology.

**Chapter 2**

**REVIEW OF THE RELATED LITERATURE AND STUDIES**

**Foundation of the Study**

Social media, particularly Facebook, has taken communication to a global range and is shared by billions of users. Among other things, the high popularity of the platforms has led them to attract cybercriminals to carry out cyber-phishing activities to obtain privacy-compromising information, generally through deceiving user behavior. Such attacks are not limited to the technical domain but also emphasize the deceitful language and social engineering practices of hacking through the judgment of human beings. With more sophistication and frequency in phishing attacks, contemporary security mechanisms such as blocklists and rule-based systems are no serious obstacle to appreciating and thwarting fresh-generation phishing. On this explicit note, the research will revolve around significant theories of AI, in particular, stemming from Natural Language Processing (NLP).

NLP goes hand in hand with AI and forces the understanding and sorting of human language upon machines. This implies access to quasi-human tools, like sentiment analyzers, grammatical definitions, information from the context, and lexical-grammatical distractors that are particularly well structured and instrumental for checking the underground activity continually fed by phishing messages. Therefore, it plays host to a well-structured mandatory phase, supervised machine learning, in accordance with which the machine will remember patterns learned in historical input and forecast the nature of the anticipated message. This theory, coupled with Social Engineering Theory, provides profound explanations for how attackers use trust and, as a result, sow deception. Given the knowledge circulated in these disciplines, this thesis will undertake the mission of developing an intelligent adaptive system through which phishing will be detected on Facebook.

This study will also offer a pool of verifiable undertakings intrinsic to deploying NLP and AI to detect phishing attacks in real-time on social networks. It paves the way for a richer understanding of AI models in the context of language patterns, contextual data, and profiling for the safe integration of an intelligent and more dynamic defense mechanism against phishing vices. By tapping into this area, the study significantly contributes to ongoing computer science and cybersecurity endeavors to furnish smart and safe digital areas through intelligent automation.

**Related Literature**

**Phishing Attack Incidents and Trends**

Over the years, phishing attacks have become more crucial to people and organizations due to their latest trend and increased incident reports. According to [1], phishing attacks are evolving using artificial intelligence (AI) and the phishing-as-a-service method by cyber criminals. The [2] stated that in the past half of 2024, credential phishing had increased by 703%, where the attackers employed zero-day malicious links to exploit the common detection systems. In addition, in the second half of 2024,  phishing attacks through email also increased by 202%.

**Utilization of Artificial Intelligence for Phishing Attacks**

In the age of information, technological advancement is seen everywhere, as AI is utilized for social engineering attacks, especially for phishing. [3], also stated that phishing detection is more challenging due to the possible exploitation and utilization of trusted platforms such as Google Drive and OneDrive for malicious content hosting. Meanwhile, [4[ demonstrated in their study that human-crafted emails can be generated by AI. Specifically, large language models (LLMs) can generate phishing emails that have a click-through rate of 54%. In addition, [5] tackled how Microsoft Copilot can be the generative AI for personalized phishing email generation, which concerns that AI can be applied for the phishing-as-a-service model development. Similarly, [6] introduced a framework that outlines how social engineering attacks are done by attackers, especially on automation, personalization, and phishing campaigns. The framework is called the Generative AI Social Engineering (GenAI-SE) Framework.

**Natural Language Processing (NLP) Applications and Challenges**

Natural Language Processing (NLP), according to the Britannica page, written by [7] is designed and utilized to enable computers to understand, process, and generate both spoken and written languages that meet humans’ capabilities—Achieved when statistics, machine learning, deep learning, and computational linguistics models are applied. Based on the comprehensive survey of [8], through fine-tuning and pre-training approaches, models such as BERT have the capability of transforming NLP applications in the generation of text, data augmentation and prompting.

Despite language-based capability advancements, NLP model deployment also encounters issues and challenges. According to [9], NLP models often lack robustness during adversarial input exposure, which leads to degradation in performance in real-world applications, despite their high benchmark dataset results. It suggests the design of models that are resistant to adversarial attacks, have multi-modal robustness, and can maintain reliability across different applications. Similarly, [10] stated that NLP models are susceptible to inherent societal biases present in the dataset they were trained on, which can lead to unfair and biased outcomes in diverse scenarios. In addition, [11] stated that NLP models often struggle in less-spoken languages than Spanish, English, and Chinese due to related resource limitations. Even though mBERT offers a potential solution for the said challenge, further research and study are still advised for embedding language-specific embeddings.

**Logistic Regression and Its Phishing Detection Capabilities**

Logistic Regression, based on [12] is a statistical method used for estimating the relationship between the dependent binary variable and independent variables through the logistic function for modeling the probability of the event or class existence. According to [13], logistic regression is a supervised machine learning algorithm that can predict the categorical dependent variable output through the fitting of data into a logistic function, which generates probabilities used to map two or more discrete classes. In addition, logistic regression consists of different types, namely: binomial, multinomial, and ordinal.

Logistic regression has been applied for phishing detection based on extracted features from the URL. [14] used logistic regression to detect phishing URLs through the incorporation of feature selection based on Mutual Information (MI), which has a 99.97% accuracy rate. In addition, [15] performed a comparative study of various machine learning algorithms' effectiveness for phishing website detection. It is found in their study that even though BERT has the best performance with an accuracy of 99%, logistic regression, on the other hand, offers balance in performance and efficient computation, which makes it an ideal algorithm for lightweight real-time URL phishing detection applications.

**Related Studies**

**Local Studies**

[16] stated in their study that Logistic Regression is like any machine learning algorithm that contains limitations and possible issues like overfitting, class imbalance, and large data sets, which impact both its accuracy and efficiency. So, they improved the Logistic Regression algorithm by introducing the utilization of recursive Feature Elimination for huge data sets, Principal Component Analysis to handle overfitting issues, and Term Frequency-Inverse Document Frequency (TF-IDF) to deal with class imbalance. Furthermore, the improved Logistic Regression model performed significantly better in spam detection with an accuracy rate of 98%, with the help of TF-IDF. Similarly, the RFE enables the model’s robustness in handling large data sets while maintaining its performance, and the PCA reduces the risks of overfitting due to enhanced model generalization. The authors of this study then claim that their enhanced Logistic Regression model is highly applicable to email security matters in the real world.

Moving forward, [17] conducted qualitative research about grammatical deviations within phishing emails in the Philippines and explored how grammatical issues in emails can be potentially deceptive. The study included fifteen emails that were considered phishing to be analyzed, and it was found that all of them consisted of grammar issues and usage errors. According to the researcher, the most frequent grammar issues are errors in word forms, capitalization, and punctuation. In addition, based on the findings, the researcher inferred that phishers in the Philippines lack proficiency with grammar due to the common errors in phishing emails that are expected to be taught across the education systems in the Philippines. Also, it leads the researcher to conclude that people who do not have strong foundations in grammar are more susceptible to phishing attacks.

On the other hand, the study of [18], delves into the effectiveness of machine learning (ML) techniques that can be utilized for the improvement of the detection of phishing in Sulu, Philippines. The gathered dataset for their study was from legitimate websites, which incorporate relevant features within Sulu’s context, In banking institutions, e-commerce platforms, and government services. Machine learning algorithms such as Naive Bayes, Support Vector Machine, and Random Forest were utilized and trained from the gathered dataset, which achieved high accuracy in phishing website detection. The Random Forest algorithm outperforms the other two algorithms with an accuracy rate of 98.7%, then the Support Vector Machine has 96.5%, and the Naive Bayes has with 94.2% accuracy rate. In addition, through feature-importance analysis, it is found that the factors that significantly affect the accuracy of phishing detection include login form presence, URL structure, and domain age.

**Foreign Studies**

Several international studies have shown the utilization of Natural Language Processing (NLP) algorithms to identify phishing attacks across the internet's web pages. [19], proposed a model for detecting phishing attacks that focuses on the text content from suspicious web pages, not the URL address, using Natural Language Processing (NLP) and Deep Learning (DL) algorithms. It was found in their study that the best-performing DL algorithm for the model is Bidirectional GRU (BiGru), with an accuracy of 97.39%. However, their study is limited to the content of a text and is not multimodal. It suggests the necessity of performing certain algorithms that detect phishing attacks using ensemble methods with at least two DL algorithms on multiple levels—Web content level, URL, and third-party information.

[20] Introduced a two-phased stack generalized model, namely AntiPhishStack, which has the role of detecting malicious URLs. AntiPhishStack is a combination of term frequency-inverse document frequency (TF-IDF) for character-level features with machine learning and deep learning algorithms as classifiers. It consists of two Phases: Phase 1 is where the features are trained with the machine learning algorithms; Phase II is where a two-layered LSTM network, together with five optimizers, is used, followed by the premiere prediction of the features. In Phase II, the two-layered deep learning algorithm, which is the long short-term memory (LSTM), eliminates sole dependence on feature knowledge for phishing detection. Lastly,  the simultaneous detection from the two phases was optimized and used for training a classifier algorithm, meta-XGBoost, for a final, accurate, and robust detection. This two-phase model accomplished an accuracy rate of 96.04% in detecting benign phishing (malicious) in URLs.

On the other hand, [21] tackled fake and fraudulent social media profile detection with the utilization of multiple machine learning algorithms, such as Logistic Regression, Random Forest, Support Vector Classifier (SVC), Decision Tree, MultiLayer Perceptron (MLP), Naive Bayes classifier, and Dummy classifier. Each of the algorithms was evaluated by the use of diverse sampling techniques. This study mainly stresses the significance of model evaluation and feature selection for the detection performance of diverse detection of fake profile detection. In addition, it serves as the baseline for future research and improvements, such as the suggestion for the use of both Recurrent and Convolution Neural Networks to offer enhanced text and image classification of profiles. Also, the study suggests the addition of a Natural Language Processing (NLP) approach for analyzing and identifying deceptive language descriptions and messages, which can be applied as a factor for fake profile detection.

Moving over, [22] conducted a study about the collinearity and feature selection impacts on the Logistic Regression models’ predictive accuracy for detecting phishing attacks. In addition, several algorithms, such as K Nearest Neighbors, Decision Tree classifier, Linear Discriminant Analysis, Logistic Regression, and Gaussian Naive Bayes, were also considered in the testing to study the relationships between their predictor variables and the likelihood of being attacked through phishing, together with the predictive accuracy of each algorithm. After the experiment and comparison, it shows that considering collinearity issues and integration of feature selection improved the Logistic Regression’s predictive accuracy, compared to other basic machine learning algorithms. By following the feature engineering process, which concerns collinearity among predictors, the researcher accomplished a significant reduction of 35% false negative rate for a Logistic Regression. With that, removing the strategic batch, especially with highly correlated features, can lead to insights that are helpful for frameworks that handle multicollinearity in Logistic Regression, which is beneficial for a reliable and accurate prediction performance for phishing.

On the other hand, [23] performed a study comparing two classifiers, Random Forest and Logistic Regression, for the detection of phishing websites. Correlation-based feature selection (CFS) method was used to improve the classification performance of both algorithms in selecting the significant influencing attributes in web phishing detection. The test result shows that the application of Random Forest classification and Logistic Regression for web phishing classification resulted in an accuracy rate of 96.834% and 93.035%, respectively. However, after the CFS feature selection, the accuracy became 97.015% and 92.718%, respectively. Random Forest has an increase of 0.181% after the CFS feature selection. On the contrary, Logistic Regression decreased in an insignificant manner. Therefore, the result in this study states that after the CFS feature selection, Random Forest classification slightly outperformed Logistic Regression in terms of accuracy.

**Synthesis**

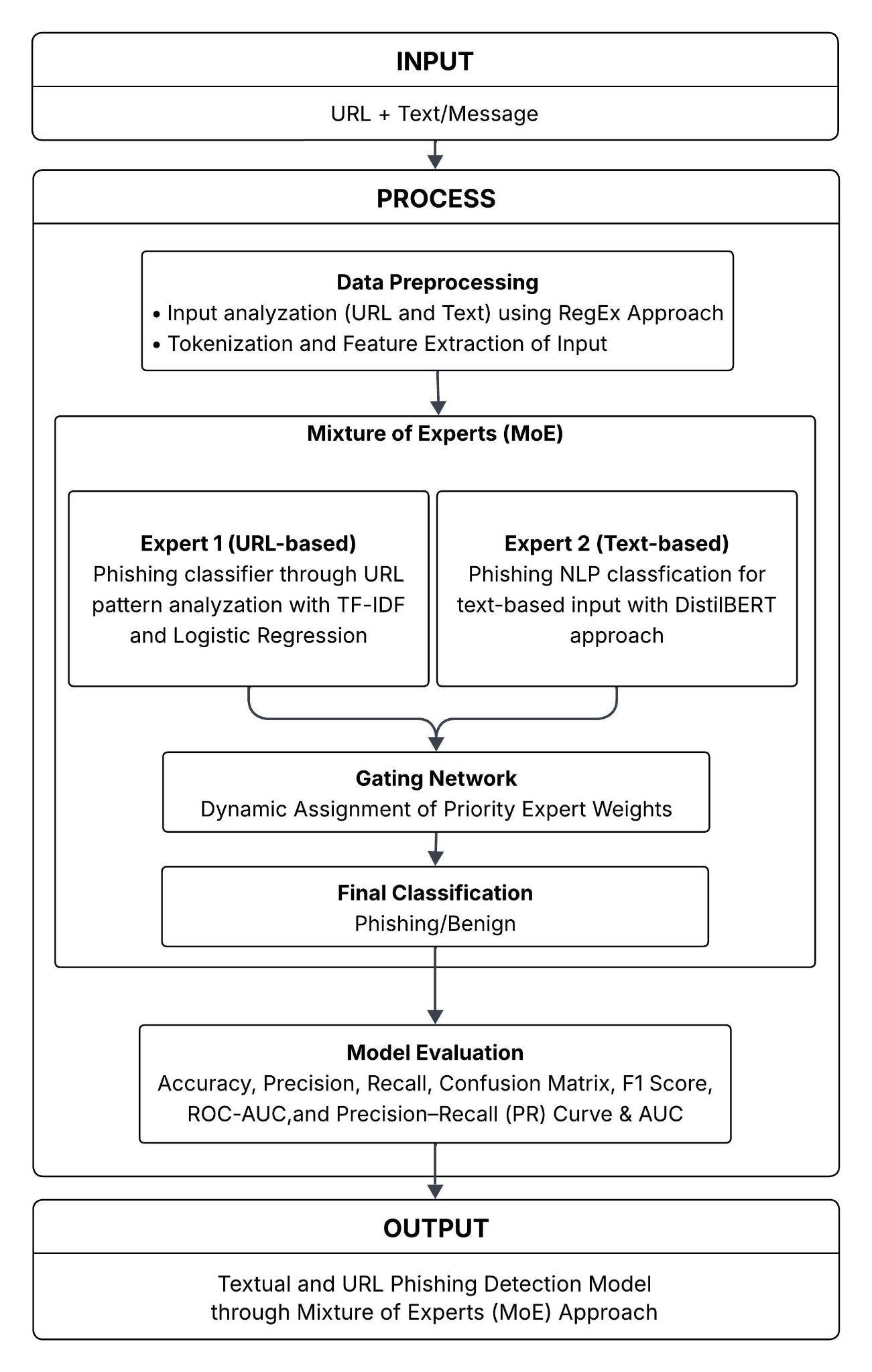
Existing literature and studies highlight how artificial intelligence and machine learning techniques play crucial roles both in developing and defending against phishing attacks. From the literature review section, it can be concluded that phishing attacks are still increasing significantly nowadays, where AI advancement influences improved phishing implementation. Due to the capabilities of AI in generating phishing contexts, several researchers have raised concerns about this matter. However, there is literature and studies that introduce machine learning techniques in combating phishing attempts through URL detection and email analysis. Specifically, Natural Language Processing (NLP) is commonly used in this matter since phishing is commonly in the form of language/text. Specifically, Term Frequency-Inverse Document Frequency (TF-IDF) is commonly used for feature extraction in phishing detection. Also, character-level feature extraction is used to further analyze the text, especially when detecting suspicious URLs, just in the case of [20].

Some of the literature and studies state that Logistic Regression is a good machine learning algorithm classifier for real-time use due to its lightweight performance. However, according to the study of [23], the accuracy of Logistic Regression might be reduced after the application of the correlation-based feature selection (CFS) method. There are also local studies that utilized an enhanced model of Logistic Regression for spam detection in email. By this, it offers an idea in which aspects the Logistic Regression can be improved and what algorithms might be utilized. Lastly, the study of [21] tackles the detection of fake and fraudulent accounts on social media using Natural Language Processing (NLP).

However, most of the studies and literature rely on URLs, text, and email when detecting phishing. With that, the gap should be in the context of sentiment analysis of the phishing text (such as a caption or description). The researchers and developers of this study consider conducting further research on how sentiment analysis can influence phishing detection, together with the URL and/or email. In addition, by applying the idea of the presence of fraudulent accounts based on the study of [21], thealgorithm model would most likely be implemented in the Facebook platform.

**Conceptual Framework**

The entire process of the study is directed by the conceptual framework diagram, which represents various process flows. The conceptual framework diagram contains input, process, and output to visualize the study’s process sequence. Specifically, the conceptual framework introduces the model that will be used for this study—the hybrid/ensemble model detection using Term Frequency-Inverse Document Frequency for feature vectorization, Logistic Regression, and Mixture of Experts (MoE) for the meta-fusion of two experts/algorithms. The conceptual framework presented below is designed to become adaptable, whether the input contains only text, only a URL, or both. The Mixture of Experts (MoE) approach includes the two experts/models, the first model is a URL-based expert/model, which uses Logistic Regression and TF-IDF algorithms. Whereas the second is a text-based expert/model which utilizes DistilBERT.



*Figure 1: Conceptual Framework*

The figure above represents the conceptual framework of an ensemble NLP model for phishing detection through text and URL. The input section involves URL, text, or a mixture of both. In addition, the datasets for both URL and text are labeled as phishing and benign. The datasets are crucial for training the model and assessing its performance accuracy. Then, the input will be passed through the preprocessing phase to determine whether the URL is present, the text is present, or both. In addition, URL and text distinction will be done with the application of regular expression (RegEx) filtering. Once identified, the model will pass them on to their respective experts/models. The identified URL input will be passed to the URL-based expert. Similar to the identified text/message input, which will be passed to the text-based expert. The two models will have different approaches, where the URL-based expert will use Logistic Regression and TF-IDF. The text-based expert will use DistilBERT for NLP classification. The experts will only be activated depending on the input (URL and/or text). After each classification of two distinct experts, it will undergo the gating network, which makes the entire approach a Mixture of Experts (MoE), where it will dynamically adjust the expert weights. Specifically, the researchers will employ a Tiny Feedforward Neural Network (MLP Gate) for balanced accuracy and adaptiveness. Lastly, the entire model will undergo several tests and evaluations to assess its overall efficiency and effectiveness.

**Definition of Terms**

1. Artificial Intelligence (AI)
2. The simulation of human intelligence in machines to perform tasks like decision-making and language processing (Burgess, 2024).
3. BERT (Bidirectional Encoder Representations from Transformers)
4. An NLP model pre-trained to understand context in text (Min et al., 2021).
5. Deep Learning (DL)
6. A subset of ML using neural networks to model complex data patterns (Benavides-Astudillo et al., 2023).
7. Generative AI
8. AI systems that create text, images, or other media mimicking human output (Burgess, 2024).
9. Machine Learning (ML)
10. A branch of AI where systems learn from data patterns without explicit programming (GeeksforGeeks, 2025).
11. Natural Language Processing (NLP)
12. A subfield of AI that enables computers to understand, interpret, and generate human language (Ramathan, 2025).
13. Phishing
14. A cybercrime where attackers deceive users into revealing sensitive information by impersonating legitimate entities (Baker, 2025).
15. Social Engineering
16. Psychological manipulation to trick individuals into divulging confidential information (Schmitt & Flechais, 2023).
17. Zero-Day Exploits
18. Attacks targeting undisclosed software vulnerabilities before patches are available (Security Magazine, 2025).

**Chapter 3**

**METHODOLOGY**

**Materials**

**Software**

* The software tools and technologies used in this research include:
  + *Transformers (Hugging Face)* – For fine-tuning NLP models (e.g., BERT, RoBERTa) for text analysis.
  + *PhishTank or OpenPhish* – For accessing real-time phishing URL databases.
  + *Pandas/NumPy* – For dataset preprocessing and feature extraction.
* Development Tools:
  + *Visual Studio Code* – For coding and debugging.
  + *Jupyter Notebook* – For prototyping ML models.
  + *Git/GitHub* – For version control and collaboration.
* Testing and Deployment:
  + *Postman* – For API testing.
  + *Docker* – For containerizing the system for deployment.
  + *Chrome Extension Manifest V3* – For packaging the browser extension.

**Data**

* The datasets and data sources used in this study include:
* Phishing Text Datasets:
  + *Publicly Available Corpora*:
    - PhishTank (for phishing URL examples).
    - CERT Synthetic Phishing Emails (adapted for Facebook post text).
  + *Synthetic Data*: Generated phishing posts mimicking AI-driven attacks (e.g., using GPT-3 for adversarial samples).
* Legitimate Facebook Posts:
  + Collected from public Facebook pages (with ethical approval) to create a balanced dataset for model training.
* URL Reputation Data:
  + *Phishing-labeled URL:* Kaggle
  + *Blacklists*: URLs from OpenPhish, Google Safe Browsing.
  + *Lexical Features*: WHOIS data, domain age, and SSL certificate details.
* Evaluation Data:
  + Labeled test sets (phishing vs. legitimate) for benchmarking.
  + User feedback from beta testers to assess false positives/negatives.

**Methods**

**Research Design**

This research uses a quantitative descriptive research design to examine the application of artificial intelligence (AI) and natural language processing (NLP) in phishing content detection on Facebook. The main objective is to quantitatively describe and assess the performance of AI-based methods—specifically NLP and URL analysis—in detecting malicious messages, posts, and links shared on the site. The study is applied in nature, seeking to yield measurable and practical results that can feed into actual-world cybersecurity solutions. This is consistent with the objectives of computer science, where the development of systems to solve sophisticated problems in virtual environments is a top priority.

The research involves gathering quantifiable data and employing statistical analysis to evaluate the performance of NLP-based phishing detection methods. Data for the experiment is obtained from publicly posted Facebook content, open-source phishing data sets, or artificially generated messages that mimic known phishing attack patterns. Ethical standards are adhered to rigorously to abide by privacy policies and platform regulations. The gathered data comprise phishing and genuine posts, which are labeled as such for analysis. A preprocessing phase of structured data is applied to normalize and clean the text. This involves noise removal (e.g.,, emojis, and special characters), tokenization of words, and converting text into numerical form using techniques like TF-IDF for contextual embeddings like BERT. These operations make the dataset ready for descriptive analysis and pattern discovery. The study then entails the use of predefined machine learning models on labeled data and quantitative measures of accuracy, precision, recall, F1 score, and ROC-AUC to characterize and summarize model performance.

These measures form the foundation for comparison of the efficacy with which different methods identify phishing content under different scenarios. The findings of the analysis are reported in the form of statistical summaries, charts, and tables, providing a clear and organized representation of the detection system's performance. No interventions or manipulations are performed; rather, the emphasis is on objectively quantifying available capabilities and detecting patterns or constraints in the data. This descriptive method is appropriate for the research goals since it supports an organized study of phishing detection performance without experimental intervention. By being based only on quantifiable outputs and systematic evaluation, the research guarantees objectivity, replicability, and pertinence within the scope of AI-based social media security.

**Research Question**

Section A: Demographic Profile

1. Age:  
     ☐ 18–25 ☐ 26–35 ☐ 36–45 ☐ 46–55 ☐ 56 and above
2. Gender:  
     ☐ Male ☐ Female ☐ Prefer not to say
3. Occupation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. Do you use Facebook regularly?  
     ☐ Yes ☐ No
5. Average time spent on Facebook daily:  
     ☐ Less than 1 hour ☐ 1–2 hours ☐ 3–4 hours ☐ 5 hours or more

Section B: Awareness and Experience with Phishing

| Statements | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| 6. I have received suspicious messages or links on Facebook. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7. I know someone who has been a victim of phishing on Facebook. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8. I can confidently identify phishing messages. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9. Phishing scams are getting harder to spot due to realistic language. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10. I always verify links and sender information before clicking. | ☐ | ☐ | ☐ | ☐ | ☐ |

Section C: Opinions on AI and NLP for Phishing Detection

| Statements | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| 11. AI can improve phishing detection on Facebook. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 12. NLP is useful in detecting patterns in phishing content. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 13. I would trust an AI system to warn me about phishing. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 14. Facebook should use AI tools to prevent phishing attacks. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 15. I feel safer knowing phishing can be automatically detected. | ☐ | ☐ | ☐ | ☐ | ☐ |

Section D: Preferred Security Features on Facebook

| Features | Not Important | Slightly Important | Moderately Important | Very Important | Extremely Important |
| --- | --- | --- | --- | --- | --- |
| 16. Auto-flagging suspicious content. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 17. Warning messages before clicking links. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 18. Real-time analysis of messages and posts. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 19. Allowing users to report phishing attempts. | ☐ | ☐ | ☐ | ☐ | ☐ |
| 20. Educational tips about phishing attacks. | ☐ | ☐ | ☐ | ☐ | ☐ |

**Population**

The intended population for this study consists of adult Facebook users aged 18 and above who actively gain information, communicate, or socially engage with others through the medium. The phishing attempts likely to target this population usually thrive on human vulnerabilities and the trust relationships placed by adults upon the site. Different kinds of phishing formats can be misleading links, fraudulent messages, or impersonated accounts, which again make for an excellent case study to gain insights into the various aspects of AI-based phishing detection methods in this context.

This study will likely engage behavioral responses from working professionals, teachers, entrepreneurs, and other practicing adults who use Facebook for either personal or work-related accounts. Targeting this population would provide insights into users' experiences, awareness, and level of trust in these AI security interventions. Such responses would be critical in evaluating how practically applicable and accepted NLP algorithms would be in identifying phishing attacks; further, they will likely inform much tougher cybersecurity measures appropriate for adult Facebook users.

Specifically, the research population will include five (5) faculty members from the institution or experienced cybersecurity practitioners and fifteen (15) end users who regularly use Facebook for personal or professional purposes. To build a balanced composition between both experts and users on the deeper exploration of how AI-powered phishing detection tools are understood, interpreted, and optimized based on feedback from real-life experiences.

**Methods of Collecting Data**

After the implementation of the proposed algorithm, the researchers and developers will utilize survey questionnaires for the beneficiaries and IT professionals. Specifically, ISO/IEC 25010:2011 Software Characteristics will be the content of the said survey, which will allow both the beneficiaries and IT professionals to evaluate our algorithm, to be implemented through the system, according to functional suitability, reliability, security, and usability.

To maximize technological advancements nowadays, researchers will utilize Google Forms. It is where the ISO/IEC 25010:2011 Software Characteristics will be asked, together with respective email addresses. Names and other personal information of the beneficiaries and users are optional to answer by. The entire process of data gathering will strongly consider the Data Privacy Act of 2012 (DPA), legally known as Republic Act No. 10173, to ensure the confidentiality of the beneficiaries’ identity of identity.

**Data Analysis**

The research design used in this study is quantitative descriptive research, highlighting statistical description of text and URL-based phishing markers occurring in Facebook posts. The study centers on the quantitative description of features with phishing and legitimate content, without involving experimental manipulation or causation. Descriptive statistics including frequency distributions, mean word frequencies, lexical richness, and presence of keywords related to phishing are initially calculated to detect patterns in the data. The data is profiled to ascertain the distribution of phishing and non-phishing posts, identifying imbalances and common traits between classes. Quantitative feature extraction is performed to transform raw text and URLs into quantifiable variables. This involves calculating frequency of urgency-related words, suspicious punctuation, domain age of URLs and other lexical or structural features traditionally related to phishing. Each example is then represented using numerical features, e.g., TF-IDF values, word embeddings, and URL-based heuristics. These organized features allow the classification models to be used not for experimentation but for descriptive modeling that uncovers common phishing traits throughout the dataset. Statistical methods like cross-tabulation and correlation analysis are employed to discover associations between features  between occurrences of specific keywords and phishing labels.

Metrics like accuracy, precision, recall, F1 score, and ROC-AUC are presented to summarize the performance of pre-defined machine learning models like logistic regression, decision trees, and support vector machines in detecting phishing posts. These metrics give a complete picture of how accurately each model matches the underlying data patterns, giving insights into which patterns most reliably predict phishing behavior. The results are reported in the shape of tables, graphs, and charts, summarizing the performance of different text-based and URL-based features for phishing identification visually. In doing so, the research determines general phishing indicators and explains the statistical patterns observed in the sampled Facebook data. The outcomes provide quantitative grounding for understanding phishing as it appears linguistically and structurally on social media, specifically Facebook's content environment.

**Context Diagram**

A diagram of a security system

AI-generated content may be incorrect.

*Figure 2: Phishing Attacks: Leveraging Natural Language Processing (NLP) Algorithms for Ensemble Phishing Detection Model in Facebook Context Diagram*

The context diagram, Figure 1, depicts the top-level interaction among different components of the NLP-Based Phishing Detection System for processing user-generated content on Facebook. The system has at its center an intelligent detection module utilizing Natural Language Processing (NLP) algorithms to detect typical language patterns related to phishing attempts. The system's main input comes from Facebook, where real-time or batch-mode harvested messages and posts are fed for processing.

The detection system is fueled by a trained AI model, which has been created using a carefully curated dataset of both phishing and legitimate messages. The model is regularly updated and improved using a feedback loop involving security analysts. After a Facebook message or post is detected by the system as suspicious, it is directed to the Security Analysis. The feedback is cycled back into the system to improve future model training, becoming more accurate and less prone to false positives with time. Bidirectional communication between the detection system, the AI trainer, and security analysts creates a learning environment conducive to adapting, enabling the system to adapt itself to new phishing strategies. In addition to making the system a strong real-time detector, such integration enables it to support a proactive security approach that can withstand advanced, dynamically changing phishing on social media networks such as Facebook.

**Algorithmic Structure**

***Pseudocode***

***START***

**Step 1: Load Dataset**

* Load the dataset containing three fields: *text*, *URL*, and *label*.
* Store text messages in **T**, URLs in **U**, and labels in **L**

**Step 2: Extract URL Features** For each URL, compute the following features:

* Length of the URL.
* Number of dots (.).
* Number of at symbols (@).
* Whether the URL contains an IP address.
* Whether the URL starts with https.
* Number of slashes (/).
* Number of question marks (?).
* Number of equal signs (=).

Save these features as a vector for each URL.

**Step 3: Preprocess Features**

* Apply TF-IDF vectorization on the text data to obtain text feature vectors.
* Scale the extracted URL feature vectors so that all values are on a comparable scale.

**Step 4: Train Base Models**

* **Early Fusion:** Concatenate the text feature vectors with the URL feature vectors, and train a logistic regression model on the combined features.
* **Late Fusion:** Train two separate logistic regression models — one using text-only features and one using URL-only features.

**Step 5: Generate Probabilities**

* Use the early fusion model to generate probability scores for phishing.
* Use the text-only model to generate probability scores.
* Use the URL-only model to generate probability scores.

**Step 6: Hybrid Fusion**

* For each sample, take the average of the three probability scores.
* Apply a threshold ( 0.5) to decide whether the sample is phishing (1) or legitimate (0).

**Step 7: Evaluate the Model**

* Compare predicted labels with ground truth labels.
* Compute evaluation metrics such as accuracy, precision, recall, and F1-score.

**Step 8: Prediction for a New Sample** Given a new text and/or URL:

* If text is available, convert it into a TF-IDF feature vector.
* If a URL is available, extract its feature vector and scale it.
* If both text and URL are provided:
  + Generate probabilities from all three models (early fusion, text-only, and URL-only).
  + Average these probabilities to obtain the final phishing likelihood.
* If only text is provided:
  + Use the text-only model.
* If only a URL is provided:
  + Use the URL-only model.
* If neither is provided:
  + Default to predicting "safe."
* Return phishing if the probability exceeds the threshold, otherwise return safe.

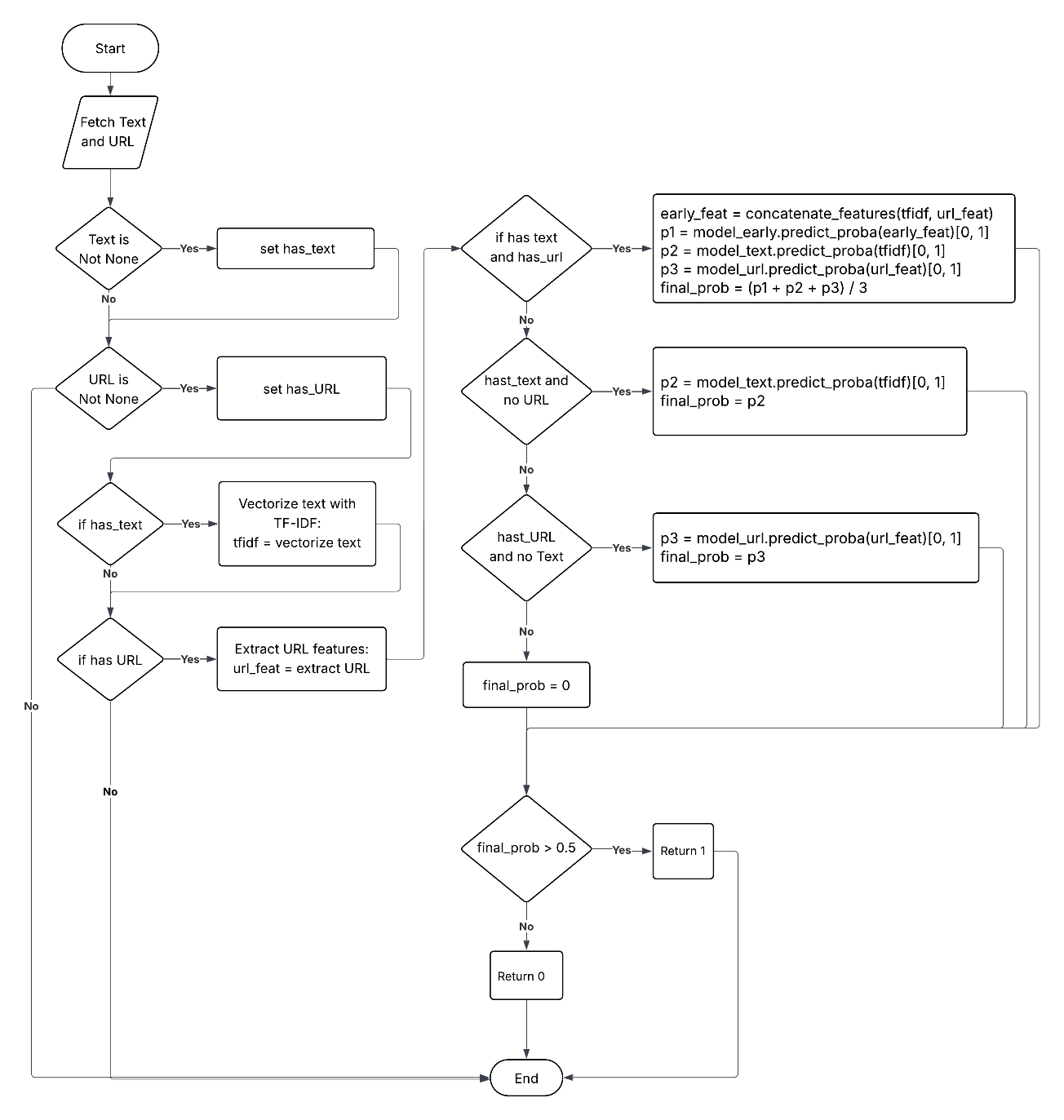
**Algorithm Flowchart**

**A flowchart of a model

AI-generated content may be incorrect.**

*Figure 3: Flowchart of the design of the ensemble NLP phishing detection model for URL and text features*

Figure 3 illustrates the flowchart for designing and training the ensemble NLP phishing detection model that can be used for both the text and URL. It does not totally represent the whole working algorithm of the model once implemented, but it shows the development process of the ensemble model using the TF-IDF vectorizer and Logistic Regression algorithm. In addition, the model will undergo an evaluation phase to display the metrics for accuracy, precision, F1 score, and Receiver Operating Characteristic- Are Under the Curve (ROC-AUC).

****

*Figure 4: Algorithm flowchart of the ensemble NLP phishing detection model for URL and text features*

The figure 4 shows the algorithm flowchart for the ensemble model phishing detection for text and URL—which will be implemented as extension for Facebook web version. The model shows the hybrid fusion of the extracted features from text and URL, which will be fetched from the Facebook web version. The flowchart implicitly shows the importation of the model that was represented in figure 3 above. Also, the threshold for the decision phase will be 0.5—the model will return an alert of phishing when the final prediction is greater than the threshold value of 0.5 (final\_prediction > 0.5).

**Mathematical Model**

This study's Natural Language Processing will utilize both Term Frequency-Inverse Document Frequency (TF-IDF) and the Logistic Regression algorithm to provide lightweight yet favorable real-time performance for phishing detection on Facebook (through the web version).

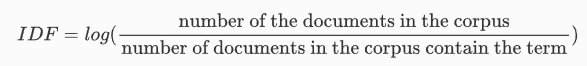
The TF-IDF will be used for information retrieval and statistical analysis. It will measure the importance of a term’s existence in a document relative to a document collection. The common term measurement (TF) and the term uniqueness measurement (IDF) are combined to achieve the term importance across the documents in a corpus. This statistical method will determine the word’s score through the Term Frequency (TF) multiplication and Inverse Document Frequency (IDF). The Term Frequency (TF) is obtained by dividing the term’s frequency in the document by the total number of terms within the document, represented as:

A black text on a white background

AI-generated content may be incorrect.

*Figure 5: Mathematical model of getting the word’s Term Frequency*

The Inverse Document Frequency (IDF) represents the term’s importance across the corpus (set of text documents) where it exists. Unique terms with a low percentage in a document are treated with higher priority than the common terms across the documents. Its mathematical model can be represented as:



*Figure 6: Mathematical model representation of the Inverse Document Frequency*

As said, the Term Frequency-Inverse Document Frequency (TF-IDF) is calculated through the multiplication of the scores by TF and IDF, represented as:



*Figure 7: Mathematical model of Term Frequency-Inverse Document Frequency*

The binary logistic regression algorithm will be used as a classifier by mapping input features from the vectors of TF-IDF to 0 and 1 probabilities. The binary logistic regression’s mathematical model is denoted as:

A black line with black text

AI-generated content may be incorrect.

*Figure 8: The mathematical model for the binary logistic regression algorithm*

Where:

*P(Y=1∣X)* denotes the probability of phishing from features X

*X1,X2,...,Xn*​ is the extracted features by the TF-IDF

*e​* represents natural logarithm base

**References**

[1] Baker, J. (2025, February 21). Phishing in 2025: Evolving tactics, AI-driven threats, and business email compromise (BEC). Social Security and IT Solutions. https://sociumsolutionsllc.com/phishing-in-2025-evolving-tactics-ai-driven-threats-and-business-email-compromise-bec/

[2] Security Magazine. (2025, February 9). Credential phishing attacks rose by 703% in H2 of 2024. https://www.securitymagazine.com/articles/101261-credential-phishing-attacks-rose-by-703-in-h2-of-2024

[3] ObserveNow. (2025, January 7). Phishing attacks surge in 2024: Netskope research highlights growing security challenges. https://observenow.com/2025/01/phishing-attacks-surge-in-2024-netskope-research-highlights-growing-security-challenges

[4] Heiding, F., Lermen, S., Kao, A., Schneier, B., & Vishwanath, A. (2024). Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. arXiv. https://doi.org/10.48550/arXiv.2412.00586

[5] Burgess, M. (2024, August 8). Microsoft’s AI can be turned into an automated phishing machine. WIRED. https://www.wired.com/story/microsoft-copilot-phishing-data-extraction/

[6] Schmitt, M., & Flechais, I. (2023). Digital deception: Generative artificial intelligence in social engineering and phishing. arXiv. https://doi.org/10.48550/arXiv.2310.13715

[7] Ramanathan, T. (2025, March 3). Natural language processing. Encyclopedia Britannica. Retrieved April 13, 2025, from https://www.britannica.com/technology/natural-language-processing-computer-science

[8] Min, B., Ross, H., Sulem, E., Pouran Ben Veyseh, A., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2111.01243

[9] Omar, M., Choi, S., Nyang, D., & Mohaisen, D. (2022). Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions. ArXiv. https://arxiv.org/abs/2201.00768

[10] i2 Group. (n.d.). The 10 biggest issues facing natural language processing. Retrieved April 13, 2025, from https://i2group.com/articles/the-10-biggest-issues-facing-natural-language-processing

[11] Krasadakis, P., Sakkopoulos, E., & Verykios, V. S. (2023). A Survey on Challenges and Advances in Natural Language Processing with a Focus on Legal Informatics and Low-Resource Languages. Electronics, 13(3), 648. https://doi.org/10.3390/electronics13030648

[12] Diniz, M. A., & Magalhães, T. M. (2022). Logistic Regression and Related Methods. In S. Piantadosi & C. L. Meinert (Eds.), Principles and Practice of Clinical Trials. Springer. https://doi.org/10.1007/978-3-319-52636-2\_122

[13] GeeksforGeeks. (2025, February 3). Logistic regression in machine learning. Retrieved April 13, 2025, from https://www.geeksforgeeks.org/understanding-logistic-regression

[14] Vajrobol, V., Gupta, B. B., & Gaurav, A. (2024). Mutual information based logistic regression for phishing URL detection. Cyber Security and Applications, 2, 100044. https://doi.org/10.1016/j.csa.2024.100044

[15] Tawil, A. A., Almazaydeh, L., Qawasmeh, D., Qawasmeh, B., Alshinwan, M., & Elleithy, K. (2024). Comparative analysis of machine learning algorithms for email phishing detection using TF-IDF, Word2Vec, and BERT. Computers, Materials & Continua, 81(2), 3395–3412. https://doi.org/10.32604/cmc.2024.057279

[16] Carlos, V. A. S., Pancho, J. C. C., & Agustin, V. A. (2024). Enhancement of logistic regression algorithm applied in email spam detection. International Journal For Multidisciplinary Research, 6(6), 1–12. https://doi.org/10.36948/ijfmr.2024.v06i06.32374

[17] Cardona, J. (2024). Grammatical Deviations in Philippine Phishing Emails. International Journal of English Language Studies, 6(2), 124–129. https://doi.org/10.32996/ijels.2024.6.2.18

[18] Benladin J. Warki, Aldam S. Ayyub, Ar-gifari A. Abdul Muktar, Sahier S. Ibrahim, Yusop S. Arbani, Ronnie E. Omar, Jurmilyn L. Muid, Jenelyn M. Mansul, Narsisa R. Ghamrasil, Nirfaisa E. Abduharim, & Shernahar K. Tahil. (2025). Enhancing Phishing Detection in Sulu, Philippines: A Machine Learning Approach to Combat Evolving Cyber Threats. Natural Sciences Engineering and Technology Journal, 5(1), 458-469. https://doi.org/10.37275/nasetjournal.v5i1.65

[19] Benavides-Astudillo, E., Fuertes, W., Sanchez-Gordon, S., Nuñez-Agurto, D., & Rodríguez-Galán, G. (2023). A phishing-attack-detection model using natural language processing and deep learning. Applied Sciences, 13(9), 5275. https://doi.org/10.3390/app13095275

[20] Aslam, S., Aslam, H., Manzoor, A., Hui, C., & Rasool, A. (2024). AntiPhishStack: LSTM-based stacked generalization model for optimized phishing URL detection. arXiv. https://doi.org/10.48550/arXiv.2401.08947

[21] Ramdas, S., & Agnes, N. N. T. (2024). Leveraging machine learning for fraudulent social media profile detection. Cybernetics and Information Technologies, 24(1), 118-136. https://doi.org/10.2478/cait-2024-0007

[22] Aghera, S., & Joshi, N. Y. (2024). Enhancing predictive accuracy in phishing attack detection: A study on the impact of collinearity and feature selection in ML-based logistic regression models. International Journal of Intelligent Systems and Applications in Engineering, 12(4), 723–728. https://ijisae.org/index.php/IJISAE/article/view/6277

[23] Farida, & Mustopa, A. (2023). Comparison of logistic regression and random forest using correlation-based feature selection for phishing website detection. Sistemasi: Jurnal Sistem Informasi, 12(1). https://doi.org/10.32520/stmsi.v12i1.1832