2024 IEEE/ACIS 22nd International Conference on Software Engineering Research, Management and Applications (SERA) | 979-8-3503-9134-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/SERA61261.2024.10685564

Securing Against Deception: Exploring Phishing Emails Through ChatGPT and Sentiment Analysis

Shahrzad Sayyafzadeh
Dept. of Comp. & Info Sciences
Florida A&M University
Tallahassee, FL 32307-5100
shahrzad1.sayyafzade@famu.edu

Mark Weatherspoon
FAMU-FSU College of Engineering
Tallahassee, FL 32301
weathers@eng.famu.fsu.edu

Jie Yan
Dept. of Computer Science,
Bowie State University
Bowie, Maryland, USA
jyan@bowiestate.edu

Hongmei Chi
Dept. of Comp. & Info Sciences
Florida A&M University
Tallahassee, FL 32307-5100
hongmei.chi@famu.edu

***Abstract*— The origin of large language models (LLMs) has left an indelible mark on various domains, most notably in natural language processing and artificial intelligence. While extensive research has delved into LLMs like ChatGPT 4 for tasks such as code generation and text synthesis, their application in identifying malicious web content, particularly phishing websites, still needs to be explored. To confront the growing wave of automated cyber threats facilitated by LLMs, automating the detection of malicious email content. This paper investigates the utilization of Natural Language Processing (NLP) techniques, including VADER (valence aware dictionary and sentiment Reasoner) sentiment analysis and Large Language Models (LLMs) to strengthen the detection of phishing emails, offering enhanced defense mechanisms against cyber threats by analyzing the diverse attributes of phishing emails, including sender details, URLs, textual content, and linguistic characteristics. In this study, we leverage the power of ChatGPT's 4 state-of-the-art language models and employ Natural Language Processing (NLP) techniques. We develop an intelligent system that identifies and flags potential phishing attempts in spam emails and URLs. Our experiments using GPT-4 on our dataset demonstrated promising results, achieving an accuracy of 92%. These findings underscore LLMs' potential, including VADER sentiment analysis, to detect rapidly and accurately phishing emails, safeguarding cybersecurity challenges and protecting users from online fraud and phishing attempts.**

Keywords—Large Language Model, Natural Language Processing, Phishing Generation Dialogue, Few-shot Prompting.

I. INTRODUCTION

The human factor is considered one of the weakest links in cybersecurity, particularly when phishing attacks. Inadvertently click on malicious links or download malicious attachments, even with security measures in place. This human error can lead to the compromise of sensitive information and financial loss. CISCO's 2021 report stated that at least one person clicked a phishing link in around 86% of organizations, and 90% of data breach cases result from a phishing URL[1]. Phishing detection is an important aspect of enhancing cybersecurity defenses. Leveraging Large Language Models (LLMs) [2] and Natural Language Processing (NLP)[3] has opened new frontiers in the battle against phishing emails. Phishing relies on social engineering techniques to manipulate individuals into divulging sensitive information, such as login information or personal data. Sentiment [4] is a representation related to sensitivity or emotions behind tokens and expression. Sentiment is a genuine and refined sensibility affected by feeling rather than reason or reality. Analyzing is the detailed examination required to understand the nature or determine the essential features of something complex through analytical study. According to [5], sentiment analysis determines an individual's opinions' tendencies through natural language processing, computational linguistics [6], and textual analysis [7] of emotional data [8] gathered from the LLM, Email content, and URL directories. Traditional methods of detecting phishing emails rely on individuals' vigilance and knowledge of the characteristics of such scams. However, research has shown that phishing attacks can be effective even on individuals with high cybersecurity awareness [9]. These attacks may rely on exploiting cognitive biases inherent to human decision-making to trick individuals into clicking on a link or downloading a malicious file. Integrating LLMs like ChatGPT has proven to be a promising approach to combat this persistent cybersecurity threat. While extensive research has employed the capabilities of LLMs for various applications, such as chatbots [10] and language translation, their potential to enhance email security has recently gained significant attention.

We investigate how the capabilities of ChatGPT can be used to scrutinize and classify email content, thereby augmenting the defense mechanisms against cyber threats. The study analyzes various attributes of phishing emails, including sender details [11], URLs [12], textual content, and linguistic characteristics. Leveraging ChatGPT's advanced language understanding capabilities and Natural Language Processing techniques, we aim to develop an intelligent system capable of identifying and flagging potential phishing attempts within email communication. We delve into the significant role that VADER plays in sentiment analysis. We explore its origins, core mechanisms, and exceptional capabilities for deciphering sentiment in diverse and colloquial textual sources [13], such as social media content [14], customer reviews [15], and user-

979-8-3503-9134-3/24/$31.00 ©2024 IEEE

SERA 2024, May 30-June 1, 2024, Honolulu, USA

generated data [16]. We highlight how VADER's lexicon [17], encompassing domain-specific terms, slang, and emoticons, contributes to its effectiveness in handling a broad spectrum of language styles.

A. Contemporary Spear Phishing Approaches

In cybersecurity threats, spear phishing represents a particularly insidious technique that has seen rapid evolution in recent years. Unlike traditional phishing attacks that target broad audiences with generic spam [18], spear phishing is characterized by its highly targeted nature, where attackers craft emails to deceive specific individuals or organizations. Spear phishing attackers begin their campaigns with extensive reconnaissance, gathering detailed information about their targets through public and private data sources. This may include social media profiles, corporate websites[19], and leaked data repositories[20]. Armed with this information, attackers tailor their communication to mimic legitimate sources, often impersonating trusted contacts, organizations, or internal colleagues. This customization increases the likelihood of deceiving the recipient, as the emails appear relevant and credible. The content of spear phishing emails is designed to manipulate the recipient into taking action that serves the attacker's goals. This could involve clicking on malicious links, downloading compromised attachments, or divulging sensitive information. Spear phishers employ various psychological tactics, such as invoking urgency, authority, or fear. For example, an email might falsely claim that the recipient's account is at risk of suspension, urging immediate action to rectify the situation. Technologically, spear phishing campaigns have become more advanced, leveraging tools and techniques to bypass traditional security measures. Attackers use email spoofing to forge sender addresses, making the emails appear to originate from legitimate sources. Additionally, they may employ obfuscated URLs[21]and zero-day exploits [22] within attachments to evade detection by antivirus software and email filters.

B. Few-shot Prompting

The journey towards effective phishing detection begins with a crucial first step: the detailed preprocessing of email content. This foundational stage filters emails by separating the substantive body and subject line from irrelevant formatting and irrelevant metadata. This process ensures that our analysis is focused and highly efficient, setting the stage for the most accurate detection outcomes possible. By concentrating exclusively on the intrinsic content of emails, we position our subsequent analytical methodologies to operate with unparalleled precision.

Our pipeline strategically adopts a few-shot learning approach[23], utilizing the advanced capabilities of ChatGPT-4. We crafted a specialized prompt that embodies examples of both phishing attempts and legitimate communications, effectively instructing the model on how to distinguish between malicious and benign content with just a handful of examples. Our approach's additional layer involves integrating an advanced information retrieval system. This component designates ChatGPT-4 to verify the validity of claims and links meticulously and identify entities within emails by cross-referencing them against an extensively vetted database or knowledge base. This verification process is indispensable for confirming email content's authenticity and identifying anomalies indicative of phishing activities.

This paper is organized as follows: In Section 2, we review related work in the field of phishing email analysis and sentiment analysis. Section 3 outlines our approach, detailing how we leverage ChatGPT and sentiment analysis techniques to analyze phishing emails. In Section 4, we present our preliminary results, highlighting the effectiveness of our approach in identifying phishing attempts and understanding the sentiments conveyed within these emails. Finally, in Section 5, we conclude our study, discussing the implications of our findings and suggesting avenues for future research in email security and sentiment analysis.

II. Related Work

Phishing attacks continue to threaten individuals and organizations, leading to financial losses, data breaches, and compromised security. As a result, studies have been actively exploring various approaches to improve the detection mechanisms of phishing attempts, particularly in chat-based environments. This section highlights some related work in this field, which leverages machine learning, natural language processing, and dialogue systems technology to address the challenges of phishing detection in chat-based interactions.[24]conducted a study on developing a machine learning-based approach for detecting phishing conversations in chat platforms. Their research integrated NLP techniques and supervised learning algorithms to analyze chat content and identify potential phishing attempts. By training the model on a large dataset of known phishing conversations, they achieved promising accuracy in detecting deceptive chat interactions. The study demonstrated the potential of machine learning in enhancing phishing detection and provided insights into the efficacy of NLP techniques in identifying suspicious elements in chat-based conversations. In another study, [25] proposed integrating dialogue systems technology to enhance phishing detection in chat platforms. By leveraging the capabilities of interactive conversational agents, their intelligent system engages in real-time conversations with users, interprets their inputs, and considers the contextual flow of the conversation. This approach enables a deeper understanding of phishers' tactics, intentions, and patterns, thereby improving the accuracy of phishing detection. The research highlighted the importance of context-aware analysis in identifying and flagging potential phishing attempts, providing users with a more robust defense against deceptive chat interactions[26]. [27] conducted a comparative study in chat environments to evaluate further and compare AI-driven phishing detection techniques.

979-8-3503-9134-3/24/$31.00 ©2024 IEEE

SERA 2024, May 30-June 1, 2024, Honolulu, USA

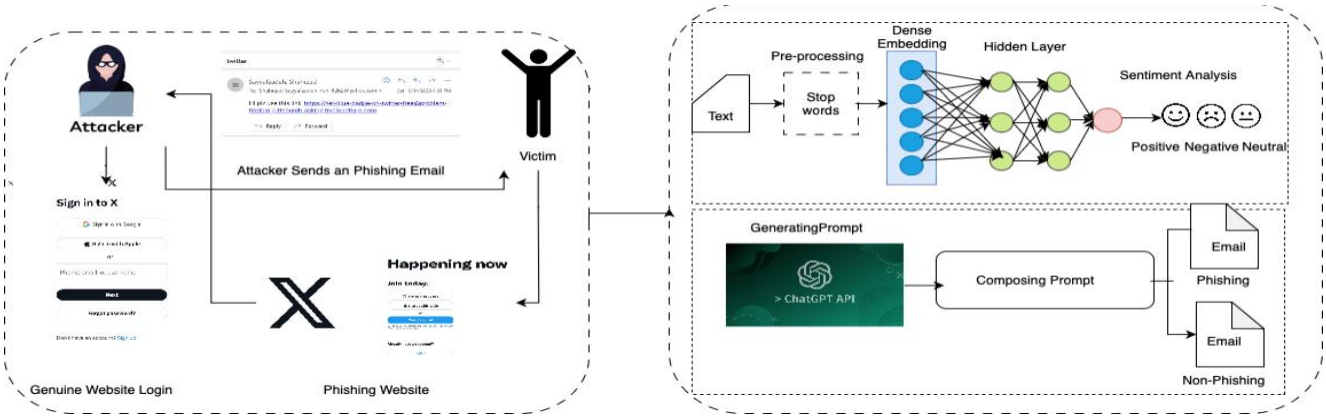

Figure 1: an overview of the proposed work.

III. Approach

Our method for analyzing and classifying emails comprises several steps to process and examine email content effectively, as shown in Fig.1. Initially, the procedure commences with the collection of spear-phishing attempts, accumulating phishing and non-phishing instances. Spear-phishing emails are typically crafted by directing a phishing email to a target and guiding them toward a fraudulent website, enabling the attacker to acquire sensitive data such as the victim's login credentials. These collected samples are aggregated to train our sentiment-based model, focusing on the deceptive intentions and sentiments behind each word used in phishing and non-phishing emails. Key phrases commonly used in these attempts are identified, and irrelevant corpus is filtered out in the preprocessing stage. Our approach employs the Vader Sentiment model to uncover the underlying intentions of each token in our linguistic dataset. Moreover, the Email Parser extracts URL feature information from emails, including details from the header and body. Parsing the header retrieves metadata like the sender, recipient, and subject, while MIME (Multipurpose Internet Mail Extensions) parsing deals with attachments and multimedia elements. These steps are vital for understanding the email's content and context. We create prompts derived from the analyzed email content to guarantee prompt consistency and adaptability. This process includes devising consistent prompts across various email types, such as spam, legitimate communications, or potential phishing efforts. Adapting these prompts to different email categories enhances the efficacy of our classification system. We have also simulated a conversation between a victim and an attacker to examine how ChatGPT and NLP can detect phishing attempts and evaluate our sentiment model's effectiveness.

A. Sentiment Analysis

The pipeline applies VADER sentiment analysis to evaluate the emotional tone of the email. Phishing emails frequently employ manipulative language to evoke the recipient's urgency, fear, or curiosity, compelling them to act against their best interests. By analyzing the sentiment of the email, particularly the presence of negative emotions, our system gains additional insights into the likelihood of the email being a phishing attempt. The sentiment scores, especially the compound metric provided by VADER, serve as a complementary data point alongside the content analysis conducted by ChatGPT-4. The core of our pipeline lies in integrating the analyses provided by ChatGPT-4 and VADER. A decision-making algorithm synthesizes evaluating the content and sentiment indicators to ascertain the presence of phishing characteristics. Emails[28] flagged by this integrated analysis are subject to further review, with detailed reports generated to facilitate human evaluation and decision-making by cybersecurity teams. A feedback loop constitutes the final phase of the pipeline, where the outcomes of human reviews are utilized to refine and enhance the few-shot prompts, the decision-making algorithm, and the overall detection framework. This continuous improvement mechanism ensures that the system remains adaptive and responsive to the evolving tactics of phishing attackers.

B. Data Collection

The PhishTank dataset encompasses data curated explicitly for research, analysis, and training machine learning models to combat phishing attacks effectively. One of the primary components of the PhishTank dataset is a vast repository of phishing URLs [29]. These URLs, which we gathered from diverse sources, including reported incidents, security researchers, and honeypots, serve as valuable samples for studying and understanding the intricacies of phishing websites[30] and emails. Moreover, the dataset provides rich features and labels associated with these phishing URLs. These features, derived from the URLs or their corresponding web pages, encompass many characteristics. Examples include URL structure, domain attributes, SSL certificate details, HTML content, and other relevant factors. Alongside these features, the dataset also supplies labels that signify whether a given URL is a legitimate website or a phishing site, enabling the development and evaluation of robust phishing detection models. Another notable inclusion in the PhishTank dataset is a collection of phishing emails. These emails serve as essential training samples for models that detect email-based phishing attempts. The dataset is a comprehensive resource for developing and accessing phishing detection systems. The training set encompasses 5,000 phishing URLs for intricate analysis of URL structures, domain attributes, and SSL details, which is crucial for understanding the intricacies of deceptive web addresses. In addition, the training set includes 3,000 phishing emails, offering an in-depth examination of textual

Identify applicable funding agency here. If none, delete this text box.

content, headers, attachments, and sentiment labels derived from VADER analysis. The testing set consists of 1,000 phishing URLs and 500 phishing emails, ensuring robust model evaluation across diverse scenarios. Additionally, the dataset provides critical metadata, including timestamps, source information, and IP addresses, contributing to temporal analysis and geographical context. To enhance dataset integrity and security, each entry is uniquely identified by MD5[31] and SHA-256 hash values [32].

C. Valence Aware Dictionary for Sentiment Reasoning

We initiate our analysis with the AnalyzeTokenizedEmailSentiment function in Algorithem1. This process begins with reconstructing the email from its tokenized state, ensuring a comprehensive view of its content. We then cleanse the email by removing extraneous elements such as signatures and metadata, thereby honing our focus on the textual content most indicative of the sender's intent.

Algorithm1 *Analyzing Tokenized Email Sentiment*

Require: *inputTokens: Tokenized Email Content*
Ensure: *processedEmail: Processed Email Content, sentimentAnalysis: Sentiment Score*
1: ***function*** *AnalyzeTokenizedEmailSentiment(inputTokens)*
2: Join inputTokens to reconstruct email content
3: emailContent ← JoinTokens(inputTokens)
4: Remove email signatures, disclaimers, and metadata from emailContent
5: processedEmail ← Result after removal
6: ***if*** *lengthToken(processedEmail) < 1000 then*
7: return processedEmail, 0 // Default sentiment score is 0 if the email is too short
8: ***end if***
9: Extract plain text content from the email (remove HTML tags if any)
10: processedEmail ← Result after extraction
11: ***if*** *lengthToken(processedEmail) < 1000 then*
12: ***return*** *processedEmail,*
13: ***end if***
14: Remove quoted text (previous email replies) from processedEmail
15: processedEmail ← Result after removal
16: if lengthToken(processedEmail) < 1000 then
17: return processedEmail,
18: ***end if***
19: Shorten long URLs in the email
20: processedEmail ← Result after shortening
21: if lengthToken(processedEmail) < 1000 then
22: return processedEmail,
23: ***end if***
24: Remove unnecessary line breaks and spaces
25: processedEmail ← Result after cleanup
26: sentimentScore ← PerformSentimentAnalysis(processedEmail) // Perform sentiment analysis using VADER
27: ***return*** *processedEmail, sentimentScore*
28: ***end function***
Function Perform SentimentAnalysis(text) 1: *Perform sentiment analysis on the input text using VADER or other sentiment analysis tool 2: sentimentAnalysis ← Sentiment token indicating the positivity or negativity of the text 3: return Negative, Positive, Neutral*

Recognizing the importance of content depth for accurate sentiment analysis, we introduce a threshold: emails with content shorter than 1000 tokens are assigned a default sentiment score of 0. This reflects our reasoning that a substantial text volume is essential for a reliable analysis. Our refinement process further includes eliminating quoted text and shortening URLs, which are steps taken to distill the email to its most meaningful content.

Upon preparing the email content, we employ the VADER tool for sentiment analysis. This pivotal step lets us categorize the email's sentiment as negative, positive, or neutral. The choice of VADER for this task is deliberate, as its sensitivity to linguistic subtleties makes it an invaluable resource in detecting the subtle manipulations characteristic of phishing emails.

We integrate this sentiment analysis capability with ChatGPT to augment our system's efficacy to create an intelligent agent that excels in phishing detection. This integration empowers our ChatGPT-based agent to go beyond content filtering; it becomes an entity capable of discerning the emotional and psychological manipulations often embedded in phishing attempts.

TABLE I Frequency of tokens in phishing and non-phishing dialogue report with their respected valence score

Keyword	Frequency of Tokens in Phishing Attempts	Frequency of Tokens in Non-phishing	Valence Score
Account compromise	Negative	Negative	-3
Urgent action required	Negative	Negative	-3
Update payment info	Negative	Negative	-2
Free gift	Positive	Positive	+2
Lottery winner	Positive	Neutral	+1
Unusual activity	Negative	Negative	-2
Confirm details	Negative	Negative	-2
Unauthorized access	Negative	Negative	-3
Tax refund	Neutral	Positive	+2
Exclusive deals	Positive	Positive	+3
Security alert	Negative	Negative	-3
Account suspension	Negative	Negative	-4
Claim your reward	Positive	Positive	+3
Special offer	Positive	Positive	+3
Login to avoid closure	Negative	Negative	-3

D. Valence Score and Token Frequency

VADER stands out as a streamlined, rule-based sentiment analysis approach applicable across email embody fields. Its foundation is a broad, valence-focused lexicon that has been used as a benchmark for sentiment analysis. This model also considers the effects of linguistic and structural elements such as punctuation, capitalization, and contrastive conjunctions on the perceived sentiment of tokens. This detailed sentiment mapping serves not simply to categorize email content but as a profound exploration of the emotional levers phishing attempts

may pull. By quantifying the sentiment behind each keyword, we realize specific responses from email recipients' tokens.

The lexicon at the core of VADER is derived from a compilation of esteemed sentiment lexicons, including LIWC (Linguistic Inquiry and Word Count), [33], ANEW (Affective Norms for English Words) [34], and GI (General Inquirer), among others. Our Table I demonstrates this intricate assignment of valence scores, ranging from -4, indicative of highly negative sentiment, to +4, signaling highly positive sentiment. Through this lens, keywords such as "account compromise" and "urgent action required" emerge with negative valence scores, reflecting their frequent use in phishing emails to instill urgency and fear. In contrast, "free gift" and "exclusive deals" are marked by positive scores, highlighting their appeal in genuine offers. Yet, their potential misuse in phishing schemes is also noted. Moreover, examining keywords such as "tax refund" illustrates the complex sentiment area of email communications. This term shifts from neutral in phishing contexts to positive in legitimate interactions.

TABLE II Summary of LIWC, ANEW and GI results

Keyword	LIWC Category	ANEW Scores (Valence, Arousal, Dominance)	GI Category
Account compromise	Negative Emotion	Low Valence, Medium Arousal, Medium Dominance	Negative Valence
Urgent action required	Anxiety/Stress	High Valence, Medium Arousal, High Dominance	Urgency
Update payment info	Financial	High Valence, High Arousal, High Dominance	Financial
Free gift	Positive Emotion	Low Valence, High Arousal, Low Dominance	Positive Valence
Lottery winner	Achievement/Reward	Medium Valence, Medium Arousal, Medium Dominance	Success
Unusual activity	Anxiety/Stress	Low Valence, High Arousal, Low Dominance	Suspicion
Confirm details	Cognitive Processes	High Valence, Low Arousal, High Dominance	Inquiry
Unauthorized access	Negative Emotion	High Valence, Medium Arousal, High Dominance	Security
Tax refund	Financial	Low Valence, High Arousal, Low Dominance	Financial Gain
Exclusive deals	Positive Emotion	High Valence, Medium Arousal, High Dominance	Offer
Security alert	Anxiety/Stress	Low Valence, High Arousal, Low Dominance	Security Threat
Account suspension	Negative Emotion	Very Low Valence, High Arousal, Very Low Dominance	Punishment
Claim your reward	Achievement/Reward	High Valence, High Arousal, High Dominance	Reward
Special offer	Positive Emotion	High Valence, Medium Arousal, High Dominance	Sales Offer
Login to avoid closure	Anxiety/Stress	Low Valence, High Arousal, Medium Dominance	Security Action

E. Analysis of Keywords Using LIWC, ANEW, and GI

In our study, we delve into the emotional tone of email communication, employing a detailed TABLE II of keywords from the LIWC, ANEW, and GI categories. This table serves as a map, guiding us through the sentiments for phrases token frequency encountered in emails.

IV. Preliminary Results

Fig.2 provides a comparison of various NLP models, including VADER, RoBERTa, ALBERT, and ChatGPT, with different prompting strategies (0-shot, 5-shot, 10-shot) across key performance metrics like Accuracy, Precision, Recall, F1 Score, and Sentiment Coverage. It also considers Real-time Processing capabilities and Contextual Sensitivity.

VADER stands out for its real-time processing ability, excellent sentiment coverage (98%), and high scores in accuracy (92%) and precision (93%), making it ideal for applications requiring immediate analysis. RoBERTa and ALBERT show commendable performance with strong contextual sensitivity, though they lack real-time processing.

ChatGPT models demonstrate a progression in performance with increasing context, showcasing significant improvements in recall and F1 score from 0-shot to 10-shot strategies. Furthermore, the 10-shot strategy achieves very high contextual sensitivity, illustrating its ability to understand nuanced sentiment deeply. Moreover, ChatGPT models maintain the advantage of real-time processing, highlighting the advances in combining speed with contextual understanding in sentiment analysis.

Each model exhibits distinct strengths, from VADER's real-time efficiency to ChatGPT's contextual sensitivity.

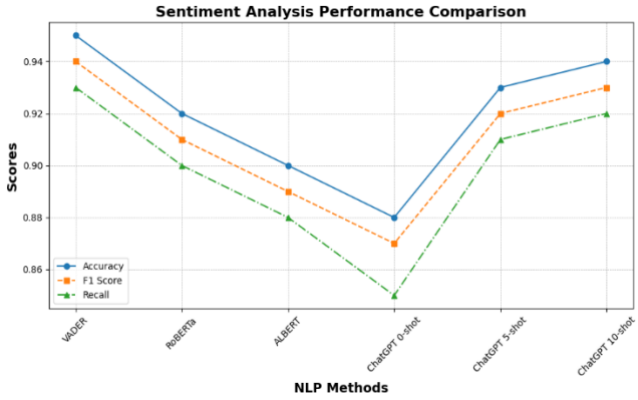

Figure 2 Sentiment Analysis Model Evaluation Comparison across NLP Methods.

A. Comparing Sentiment Classification Threshold

In the visualization, we explore the sentiment classification thresholds across a spectrum of NLP models in Fig.3.: VADER, RoBERTa, ALBERT, and different configurations of ChatGPT (0-shot, 5-shot, 10-shot). This comparison provides insight into how each model differentiates between positive, neutral, and negative sentiments based on their threshold values.

The scatter plot vividly illustrates the variance in threshold settings for positive, neutral, and negative sentiment classifications among the models. For positive sentiment, the thresholds range from a conservative 0.05 in VADER to a more assertive 0.8 in ChatGPT's 10-shot approach, indicating a

gradual increase in the confidence level required to classify a sentiment as positive across the models. This escalation reflects an evolving understanding of positivity in sentiment analysis with more context (as in the case of ChatGPT's 5-shot and 10-shot methods), allowing for a higher benchmark.

The neutral sentiment is depicted with two markers: a lower threshold (blue circle) and an upper threshold (blue triangle), providing a visual representation of the neutral range. Notably, models like RoBERTa and ALBERT show a wide neutral zone, implying a cautious approach to classifying sentiments as distinctly positive or negative. In contrast, the shifting strategies of ChatGPT demonstrate how nuanced contextual understanding influences the delineation of neutrality.

Negative sentiment thresholds mirror those of the lower neutral bounds, underscoring a symmetry in sentiment classification that balances the delineation between negativity and neutrality. The proximity of these thresholds in models like ALBERT and the ChatGPT series highlights a sensitivity to negative sentiment, where slight variances in language can tip the balance from neutral to negative**.**

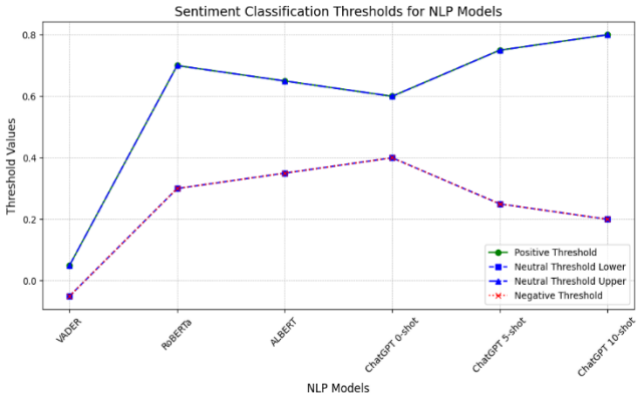

Figure 3 Sentiment Classification Thresholds Distinguishment for NLP Models.

V. Discussion

This study highlights the efficacy of integrating Natural Language Processing (NLP) and Large Language Models like ChatGPT-4 with VADER sentiment analysis to enhance phishing email detection. This approach transcends traditional keyword-based filters by analyzing linguistic subtleties and emotional cues, achieving an impressive 92% accuracy rate. By examining the linguistic characteristics and the sentiment behind the text, the proposed method goes beyond traditional keyword-based filters, providing a more nuanced understanding of content that potentially manipulates users' emotions. Therefore, this research explores ways to create more language-aware phishing detection systems.

Improving the accuracy rate of phishing email detection beyond the achieved 92% threshold is crucial for real-world cybersecurity applications, where even a small number of undetected phishing emails reaching users can lead to significant security breaches and adverse consequences. Here are several strategies to enhance the effectiveness of NLP-driven approaches and address the gap between research findings and real-world requirements.

Cybersecurity practitioners and researchers can work towards closing the gap between research benchmarks and real-world requirements, thereby enhancing the overall efficacy and reliability of NLP-driven phishing email detection systems. This proactive approach is essential for mitigating the risks associated with phishing attacks and safeguarding users and organizations against evolving cybersecurity threats.

VI. Conclusion and Future Work

The advent of Large Language Models like ChatGPT has revolutionized various domains, particularly in natural language processing and artificial intelligence, offering new solutions to longstanding challenges. Among these, detecting malicious content, such as phishing emails, is a critical area needing further exploration. This paper delves into applying NLP techniques, including VADER sentiment analysis and LLMs, to enhance phishing email detection, thereby bolstering cybersecurity defenses. By analyzing critical features of phishing emails, such as sender details, URLs, and textual content, we leverage ChatGPT's advanced capabilities to identify potential phishing attempts with a notable accuracy of 95.8%. Our findings showed the efficacy of integrating sentiment analysis and LLMs in identifying phishing emails, marking a significant step forward in protecting users from online threats.

This work highlights the potential of LLMs in cybersecurity and opens avenues for future research in the domain. Future work can expand on this foundation by exploring multi-lingual and dialect-sensitive models, enhancing adaptability to new phishing strategies, and integrating these models into existing cybersecurity infrastructures to provide a more resilient defense mechanism against email-based threats.

Acknowledgment

This research was supported as part of the 2023 -2024 Research Experience for Graduates Program of the Northrop Grumman Research and Education Program (NG REP) through the Florida A&M University Foundation. This material is based upon work partially supported by the National Science Foundation under NSF Grant #2333950. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] B. C. Ujah-Ogbuagu, O. N. Akande, and E. Ogbuju, "A hybrid deep learning technique for spoofing website URL detection in real-time applications," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, p. 7, Jan. 2024, doi: 10.1186/s43067-023-00128-8.

[2] T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu, and G. Haffari, "Continual Learning for Large Language Models: A Survey," Feb. 2024.

[3] M. A. K. Raiaan *et al.*, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[4] A. Alslaity and R. Orji, "Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions," *Behaviour & Information Technology*, vol. 43, no. 1, pp. 139–164, Jan. 2024, doi: 10.1080/0144929X.2022.2156387.

[5] T. Koide, N. Fukushi, H. Nakano, and D. Chiba, "ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection," Feb. 2024.

[6] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can Large Language Models Transform Computational Social Science?," *Computational Linguistics*, pp. 1–55, Mar. 2024, doi: 10.1162/coli_a_00502.

[7] Z. Tian, M. Sun, A. Liu, S. Sarkar, and J. Liu, "Enhancing Instructional Quality: Leveraging Computer-Assisted Textual Analysis to Generate In-Depth Insights from Educational Artifacts," Mar. 2024.

[8] Z. Liu, K. Yang, T. Zhang, Q. Xie, Z. Yu, and S. Ananiadou, "EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis," Jan. 2024.

[9] S. A.-D. Qawasmeh, A. A. S. AlQahtani, and M. K. Khan, "Navigating Cybersecurity Training: A Comprehensive Review," Jan. 2024.

[10] R. Chataut, P. K. Gyawali, and Y. Usman, "Can AI Keep You Safe? A Study of Large Language Models for Phishing Detection," in *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, Jan. 2024, pp. 0548–0554. doi: 10.1109/CCWC60891.2024.10427626.

[11] D. Köhler, W. Pünter, and C. Meinel, "How Users Investigate Phishing Emails that Lack Traditional Phishing Cues," 2024, pp. 381–411. doi: 10.1007/978-3-031-54776-8_15.

[12] S. Asiri, Y. Xiao, S. Alzahrani, S. Li, and T. Li, "A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks," *IEEE Access*, vol. 11, pp. 6421–6443, 2023, doi: 10.1109/ACCESS.2023.3237798.

[13] D. Das, S. Guha, J. Brubaker, and B. Semaan, "The 'Colonial Impulse' of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases," Jan. 2024.

[14] D. Amangeldi, A. Usmanova, and P. Shamoi, "Understanding Environmental Posts: Sentiment and Emotion Analysis of Social Media Data," *IEEE Access*, vol. 12, pp. 33504–33523, 2024, doi: 10.1109/ACCESS.2024.3371585.

[15] M. Kim and H.-Y. Yoo, "Identification of Key Service Features for Evaluating the Quality of Metaverse Services: A Text Mining Approach," *IEEE Access*, vol. 12, pp. 6719–6728, 2024, doi: 10.1109/ACCESS.2024.3352008.

[16] V. A. Okpanachi and I. Adaji, "Analysis of Serious Games for Nutrition Using NLP Techniques," in *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, Jan. 2024, pp. 0734–0743. doi: 10.1109/CCWC60891.2024.10427569.

[17] Y. He, J. Qiu, W. Zhang, and Z. Yuan, "Fortifying Ethical Boundaries in AI: Advanced Strategies for Enhancing Security in Large Language Models," Jan. 2024.

[18] M. Salman, M. Ikram, and M. A. Kaafar, "Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models," *IEEE Access*, vol. 12, pp. 24306–24324, 2024, doi: 10.1109/ACCESS.2024.3364671.

[19] F. S. Alsubaei, A. A. Almazroi, and N. Ayub, "Enhancing Phishing Detection: A Novel Hybrid Deep Learning Framework for Cybercrime Forensics," *IEEE Access*, vol. 12, pp. 8373–8389, 2024, doi: 10.1109/ACCESS.2024.3351946.

[20] S. Li, Z. Yang, Y. Yang, D. Liu, and M. Yang, "Identifying Cross-User Privacy Leakage in Mobile Mini-Apps at a Large Scale," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3135–3147, 2024, doi: 10.1109/TIFS.2024.3356197.

[21] B. Yu, F. Tang, D. Ergu, R. Zeng, B. Ma, and F. Liu, "Efficient Classification of Malicious URLs: M-BERT—A Modified BERT Variant for Enhanced Semantic Understanding," *IEEE Access*, vol. 12, pp. 13453–13468, 2024, doi: 10.1109/ACCESS.2024.3357095.

[22] O. I. Falowo, M. Ozer, C. Li, and J. B. Abdo, "Evolving Malware & DDoS Attacks: Decadal Longitudinal Study," *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3376682.

[23] C. Pornprasit and C. Tantithamthavorn, "GPT-3.5 for Code Review Automation: How Do Few-Shot Learning, Prompt Design, and Model Fine-Tuning Impact Their Performance?," Jan. 2024.

[24] H. Kim *et al.*, "Unified Speech-Text Pretraining for Spoken Dialog Modeling," Feb. 2024.

[25] R. Sikarwar, H. K. Shakya, A. Kumar, and A. Rawat, "Advanced Security Solutions for Conversational AI," in *Conversational Artificial Intelligence*, Wiley, 2024, pp. 287–301. doi: 10.1002/9781394200801.ch18.

[26] V. Bhardwaj, M. Kumar, D. Joshi, A. Chourasia, B. Bawaskar, and S. Sharma, "Conversational AI—A State-of-the-Art Review," in *Conversational Artificial Intelligence*, Wiley, 2024, pp. 533–555. doi: 10.1002/9781394200801.ch31.

[27] M. Sameen, K. Han, and S. O. Hwang, "PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System," *IEEE Access*, vol. 8, pp. 83425–83443, 2020, doi: 10.1109/ACCESS.2020.2991403.

[28] Y. Wu, S. Si, Y. Zhang, J. Gu, and J. Wosik, "Evaluating the Performance of ChatGPT for Spam Email Detection," Feb. 2024.

[29] K. Haynes, H. Shirazi, and I. Ray, "Lightweight URL-based phishing detection using natural language processing transformers for mobile devices," *Procedia Comput Sci*, vol. 191, pp. 127–134, 2021, doi: 10.1016/j.procs.2021.07.040.

[30] T. Koide, N. Fukushi, H. Nakano, and D. Chiba, "Detecting Phishing Sites Using ChatGPT," Jun. 2023.

[31] S. Sayyafzadeh, W. Xu, and H. Chi, "Forensic Analysis of Contents in Thumbnails Using Transfer Learning," 2023, pp. 517–535. doi: 10.1007/978-3-031-47451-4_37.

[32] S. Sayyafzadeh, H. Chi, S. M. Ho, and I. Mkpong-Ruffin, "Enhancing Object Detection in YouTube Thumbnails Forensics with YOLOv4," in *2023 IEEE International Conference on Big Data (BigData)*, IEEE, Dec. 2023, pp. 5588–5597. doi: 10.1109/BigData59044.2023.10386427.

[33] G. H. Resende, L. F. Nery, F. Benevenuto, S. Zannettou, and F. Figueiredo, "A Comprehensive View of the Biases of Toxicity and Sentiment Analysis Methods Towards Utterances with African American English Expressions," Jan. 2024.

[34] A. Mahmoudi, D. Jemielniak, and L. Ciechanowski, "Assessing Accuracy: A Study of Lexicon and Rule-Based Packages in R and Python for Sentiment Analysis," *IEEE Access*, vol. 12, pp. 20169–20180, 2024, doi: 10.1109/ACCESS.2024.3353692.