ANALYSING THE EMAIL DATA USING STYLOMETRIC METHOD AND DEEP LEARNING TO MITIGATE PHISHING ATTACK

Abstract:

The high-volume usage of email has attracted cybercriminals to the platform and criminals are aware of difficulties users often have in separating legitimate from illegitimate emails and seek to take advantage of those difficulties by impersonating staff of a trusted organisation to persuade users into divulging their private information. To help users overcome the difficulty in detecting phishing attacks, a system is proposed. Recent advancement uses: stylometric features, gender features and personality features to carry out a sender verification process. The existing approaches are more complex and if the system fails to detect bad email, and it gets to users, the possibility of becoming a victim becomes high if not detected by the user. The proposed framework adds Colour Code to Email Verification (CCEV). It conducts sender's verification at the recipients' end based on 3-features related with senders, writing pattern, gender, and header.

Keywords: Email data, phishing attack, spear-phishing, mitigation, sender verification, Colour Code

1 Introduction:

Email is a common medium for phishing attacks [1]. Attackers frequently use it to defraud users through phishing emails, which could lead to the invasion of important information systems [2]. This may result in users disclosing personal information or making unauthorized purchases on behalf of the phisher, with the consequences of falling victim to a phishing attack.

Phishing attacks predominantly utilize email as the primary channel, contributing to nearly 91% of successful cyber-attacks and security breaches through the dissemination of deceptive or spoofed emails [3]. It can manifest in various forms, including scams through email or text messages, fraudulent websites, or phone calls. These attacks frequently leverage social engineering tactics, aiming to exploit human emotions, curiosity, or trust. Upon obtaining access to the user's credentials or personal data, attackers can employ them for nefarious purposes such as identity theft, financial fraud, or initiating data breaches.

According to [4], phishing constitutes a type of cyber-attack designed to illicitly obtain sensitive information, such as login credentials and financial data. This is achieved by adopting the guise of a trustworthy entity in electronic communications. Phishing attacks frequently exploit social engineering techniques to deceive users, taking advantage of their trust in well-known brands or authoritative figures. Phishing attacks can be broadly categorised into social engineering and technical subterfuge, with phishing being considered a specific type of social engineering. In this classification, infected emails containing malicious code or malware are classified as social engineering attacks that involve phishing emails. On the other hand, embedded phishing and URL websites are identified as a subtype of phishing attacks, as illustrated in Figure 1.

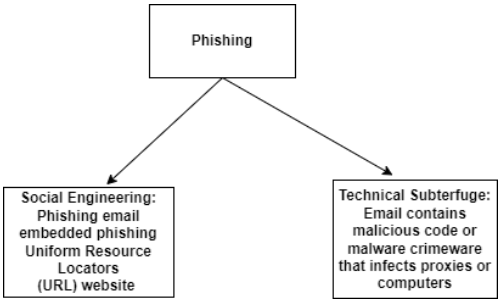

Figure 1. shows Phishing attack taxonomy. Reproduced from Figure 2 in [4].

Socially engineered messages are disseminated by phishers through digital channels to manipulate individuals into undertaking specific actions that ultimately benefit the attacker [5]. The attacks hinge on deception, aiming to deceive users into disclosing sensitive information. Phishers have refined their skills in crafting deceptive websites and emails that closely mimic legitimate ones. Employing tactics such as replicating the authentic design, layout, logos, and images of genuine websites, they enhance the credibility of their phishing emails. This

sophistication can make it challenging for users to discern phishing emails or websites, heightening the risk of falling prey to such attacks. Clicking on a link or attachment in a phishing message may discreetly install malware on a user's device, resulting in significant data loss and downtime, particularly in the case of a ransomware attack [6] (refer to Figure 2).

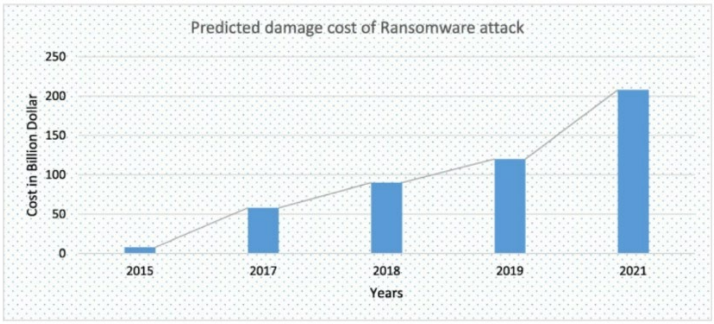

Figure 2. shows the global ransomware damage costs from year 2015 - 2021 [7].

Ransomware attacks have the potential to inflict damage or even destruction upon data, coupled with post-attack business disruptions that contribute to the overall cost of the attack [7]. Even if users refrain from entering any information into a phishing link, their devices may still fall victim to the installation of malware or spyware. The pervasive challenge in detecting phishing makes individuals susceptible to scams through these deceptive messages on a daily basis. According to the UK Government's Cyber Breaches survey in 2022, 83% of cyber-attacks were identified as phishing attacks [8]. (See Figure 3).

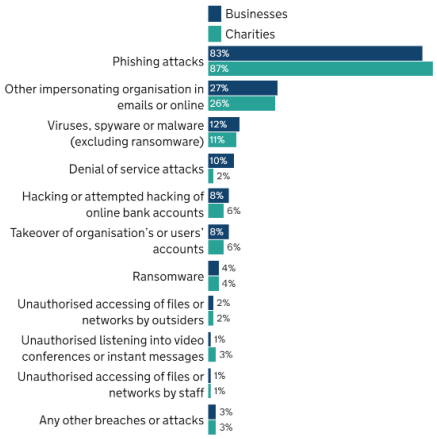

Figure 3. show the rise in phishing attack [8].

As attackers perceive phishing attacks as a convenient means to bypass organisational security measures, they often exploit human vulnerabilities to infiltrate organizations. A vulnerability is known as a weakness which allows attacker to diminish the security assurance of a system [9]. Analysing email data through stylometry proves valuable in discerning the email's originator by identifying distinctive writing patterns, encompassing aspects like word choice, sentence structure, and punctuation. This analytical approach is particularly effective in unmasking the author's identity, especially in instances where the sender endeavours to stay anonymous or adopts a false identity [10]. The identification of the email author via stylometry analysis enables the implementation of suitable measures to thwart fraudulent or malicious activities.

Verizon's 2019 data breach investigation report reveals that malware delivered via email messages accounts for up to 94% of the incidents [11].

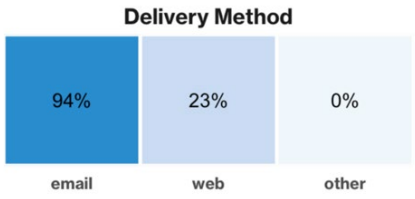

Figure 4. Malware delivery methods [11]

Understanding the origin of an email message is a crucial aspect, and stylometry plays a vital role in identifying a sender through features embedded in the sender's message. Consequently, these stylometric attributes can be considered as tools for enhancing the security of email communication channels. By uncovering the true sender of an email to a recipient, stylometric analysis serves as a means to achieve this without requiring the user's awareness, effectively identifying the original author [12]. By employing this method, the occurrence of email impersonation attack can be minimized.

The main objective of this study is to develop a phishing mitigation framework, assisting users in distinguishing between legitimate and illegitimate emails.

1. This will involve the implementation of Colour Code Email Verification (CCEV) using Natural Language Processing (NLP) and Long Short-Term Memory (LSTM) techniques.
2. Evaluate the efficacy of the proposed system.

2 Background and Related works:

Phishing is one of the most organised crimes of the twenty-first century, and it is defined as a form of malware or a term for which an attacker sends out a spoofed email to random victims, aiming to obtain private information [13].

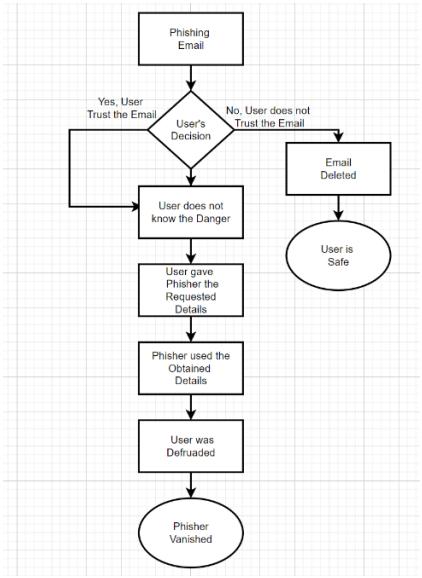

Figure 5. Overview of a phishing attack [14]

[15] presents comprehensive information on a variety of anti-phishing algorithms designed to detect phishing attempts. These algorithms integrate various technologies and features to enhance security for a wide range of users. The primary goal of the paper is to elucidate these concepts and their contributions to global cybersecurity. The authors assert that anti-phishing algorithms have shown considerable success in identifying phishing attacks. The technologies derived from these algorithms have proven effective in thwarting evolving phishing strategies, with a key aspect being the notification of users about potential attacks.

In their work, [16] introduced RAIDER, an acronym for Reinforcement AIded Spear Phishing DEtectoR, which is a system based on reinforcement learning for feature evaluation. This system autonomously identifies optimal

features for detecting various attack types. Employing a reward and penalty mechanism, RAIDER facilitates automatic feature selection, focusing solely on critical characteristics essential for identifying phishing emails and detecting spear-phishing attacks. The aim is to minimize the number of features used. Evaluating RAIDER on over 11,000 emails and across three attack scenarios revealed that using reinforcement learning for feature identification can reduce the required feature dimension by 55% compared to existing machine learning-based systems. The system achieved a 4% increase in accuracy, enhancing the detection of spoofing attacks from 90% to 94%.

In their work [17], the authors proposed a technique for detecting targeted spear-phishing attacks by identifying crucial similarities within the specific domain under attack. The methodology involves assessing the authenticity of the domain, distinguishing between genuine and counterfeit domains using multiple innovative grading algorithms. Consequently, this study contributes to addressing the challenge of targeted attacks on specific organizations by introducing a novel enterprise solution. The detection system centers around domain names, specifically registered ones that enjoy trust among victims. According to the authors, the results indicate that the detection system has proven effective in substantially mitigating email phishing attacks.

[18] presents a methodology for identifying phishing attacks through the application of machine learning, specifically Random Forest, using a set of features designed to recognise deception targeted at users. The study suggests that with slight modifications, this approach can also detect phishing websites and the emails guiding victims to them. The detection method relies on various classifiers, including Support Vector Machines, rule-based methods, conventional decision trees, and Bayesian approaches. The evaluation involves a 10-dimensional feature set, incorporating factors such as IP-based URLs, the count of links in the HTML section(s) of an email, the number of domains, non-matching URLs, HTML emails, the count of dots, the age of linked domains, links to non-modal domains, Spam-filter output, and the presence of JavaScripts [8]. For validation, the authors employed two publicly available datasets: the ham corpora from SpamAssassin, comprising the 2002 and 2003 ham collections (easy and hard), totaling around 6,950 non-phishing non-spam emails, and 860 email messages from the publicly accessible phishing corpus.

The number of features was expanded to 47 by [19], encompassing various aspects such as bodydearword, bodyform, bodyhtml, bodymultipart, bodynumchars, bodynumfunctionwords, bodynumuniqwords, bodynumwords, bodyrichness, bodysuspensionword, bodyverifyyouraccountphrase, externalsabinary, externalsascore, scriptjavascript, scriptonclick, scriptpopup, scriptstatuschange, scriptunmodalload, senddiffreplyto, sendnumwords, sendunmodaldomain, subjectbankword, subjectdebitword, subjectfwdword, subjectnumchars, subjectnumwords, subjectreplyword, subjectrichness, subjectverifyword, urlatchar, urlbaglink, urlip, urlnumdomains, urlnumexternallink, urlnumimagelink, urlnuminternallink, urlnumip, urlnumlink, urlnumperiods, urlnumport, urlport, urltwodoains, urlunmodalbaglink, urlwordclicklink, urlwordherelink, Urlwordloginlink, urlwordupdatelink. They proposed a novel framework named ASCAI [20, 21], based on document authorship techniques. This approach detects potential mismatches between received emails and trusted authors by analysing the email body to identify the sender's writing style. In [22], the authors presented a framework that addresses phishing attacks on email users, designed as a mental model for users and adaptable for implementation in email clients. They emphasised the importance of visual indicators of security status in email clients, asserting that user awareness is crucial for successfully mitigating scams like phishing.

The authors in [23] developed algorithms and explored diverse classifiers to assess the genuineness of social network posts, averaging around 206 words, extracted from Facebook. They examined the feasibility of employing standard machine learning methods to ascertain if a specified user is the author of a written message. To classify 9259 Facebook posts as authentic or non-authentic, the authors devised various stylometry and ad hoc social networking features. The research introduces an algorithm utilising machine-learning classifiers for investigating this issue, along with exploring a voting algorithm that combines three classifiers. Challenges in applying traditional stylometry techniques to short messages, such as those on social network sites, are discussed. The authors highlight that this study represents one of the initial endeavours concentrating on authorship authentication in short messages. Experimental results, based on 30 users, indicate an average accuracy rate of 79.6% for authorship authentication.

As part of their research, the authors, as outlined in [24], curated emails from Enron, eBay comments, Java forums, and Cyberpath chats. They compiled a comprehensive feature set and conducted experiments to identify and detect similarities in the datasets [25]. Their findings indicated that the accuracy of the Support Vector Machine (SVM) in authorship identification diminished as the number of authors in a dataset increased.

In the realm of cybercrime authorship identification, [26] applied stylometry in their work. They delved into the study of obfuscated writing, where an author mimics another [27]. Another facet of their research involved the examination of doppelganger accounts posted in underground forums, encompassing English, German, and Russian texts [28]. The authors suggested that a hybrid method incorporating both stylometry and underground-specific features, such as ICQ numbers, could effectively identify some doppelgängers [28]. Additionally, [29] conducted experiments to test whether the author identification framework known as Anonymouth could handle manually anonymized texts. While Anonymouth focuses on author privacy.

[30] investigated author identification using an extensive database of forum posts. The study revealed that the number of features demonstrated a correlation with the accuracy of various classification methods. Notably, all these authorship identification studies developed their corpora using word counts exceeding 500 or character counts surpassing 7500. In a study utilising single character frequencies, [31] successfully identified Shakespeare and Fletcher as the authors of acts in Two Noble Kinsmen. The authors emphasised that samples of 500 words or less lack the accuracy required for reliable authorship attribution. An examination of email authorship challenges was carried out by De Vel, who narrowed the email topics to movies, food, and travel due to the brevity of email documents. Authors were classified based on the structural characteristics and linguistic patterns of emails [32].

Nizamani and Memon employed cluster-based classification in their research to identify authors in emails [33]. Similarly, Afroz's study, as conducted by Iqbal, explored the feasibility of forensic investigation into cybercrimes using stylometric features of emails [34]. This analysis was carried out on the Enron email dataset, which includes folder information for each of the 151 employees, encompassing details like sender and recipient email addresses, date and time, subject, body, and text. The study incorporated the majority of stylometric features outlined in Writeprints, a work by Abbasi and Chen [35].

Brocardo delved into authorship verification and introduced a method for verifying short message authors through n-gram analysis [36]. Identity Mailer, developed by the authors, validates email authorship by learning the typical email-sending behavior of an individual over time. The system then compares subsequent emails against this model. In experiments with real-world email datasets, Identity Mailer demonstrated its effectiveness in blocking advanced email attacks originating from genuine email accounts, which conventional protection systems may fail to detect. According to the authors, it is the first system capable of identifying spear-phishing emails sent from a compromised email account within an organisation [37]. IdentityMailer, the authors' new system, aims to protect corporate users' identity by verifying if an email is authored by the legitimate owner of an email account. The system learns the behavior of the email account owner and checks subsequent emails against the known profile to prevent the transmission of spear-phishing emails from a compromised account or machine before being forwarded to Simple Mail Transfer Protocol (SMTP) servers.

The authors in [38] introduced a comparable system known as "profilers," which focuses on inspecting spear-phishing emails after users have received them. This methodology incorporates features from both stylometric analysis and email metadata information. Initially, the authors construct probabilistic models for both the email metadata and stylometric features of the email content. Subsequently, in order to identify distinctive markers of spear-phishing attacks, incoming emails are compared against these models. The evaluation, conducted on a real dataset derived from 20 email users, demonstrates the effectiveness of this approach in distinguishing spear-phishing attacks from legitimate emails.

[39] introduces the Spear Phishing Email Detection Approach based on Authentication (SPBA). This technique verifies the sender through the analysis of stylometric, gender, and character features extracted from emails. The method integrates stylometric, gender, and personality features to enhance phishing detection. Experimental results indicate an impressive detection accuracy of 95.05%, surpassing single stylometric feature usage by approximately 10%. This underscores the algorithm's efficacy in phishing detection. In a related effort to counter phishing attacks [40], the focus is on distinguishing spear phishing emails from non-spear phishing ones. The authors employed 27 features extracted from email data and third-party threat intelligence platforms. To balance the data, KM-SMOTE was employed. Four machine learning algorithms (Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM)) were applied for the identification of spear phishing emails.

Having reviewed the above papers, these three papers below are quite related and close to my research project.

Table 1. Related papers on author identification using style analysis to detect phishing and spear phishing attacks:

Author(s)	Methods	Year	Accuracy	Limitation
[37]	IdentityMailer prevents transmission of spear phishing emails from compromised machines before they are forwarded to SMTP server. The solution focused on insiders' attack	2014	90%	It requires an email history of 1,000 or more to be effective.
[38]	EmailProfiler identifies potential spear phishing attacks on the email recipient side of the communication, checks for spear phishing emails after they are received by the user. It addresses situation where recipient do not have enough email from a given author to train any authorship profile on the receiving side.	2016	93%	The user will do the comparison of the incoming message against the profile generated by the inbox profiler and not the system.
[39]	Spear-phishing Emails Based on Authentication (SPBA). This method conducts sender verification based on three kinds of characteristic related with senders writing stylometric, gender and character extracted from emails. To detect spear-phishing emails, these three features are combined to carry out sender authentication process.	2019	95.05%	This system classifiers between two number, 1 and 2. What happens when the system is not certain about the author is not stated. If the system fails to detect the bad email, and it gets to the users, they fall for it

The IdentityMailer solution is effective in preventing the transmission of spear-phishing emails from compromised machines to the SMTP server, but it requires an email history of 1,000 or more for optimal functionality. This solution specifically addresses insider attacks. On the recipient side, EmailProfiler is employed to identify potential spear-phishing attacks by checking incoming messages against a profile generated by the inbox profiler. However, a limitation of this method is that it relies on user-driven comparisons, not system-driven ones.

Another approach, Spear-phishing Emails Based on Authentication (SPBA), conducts sender verification based on three characteristics: stylometric, gender, and character features extracted from emails. The system combines these features to authenticate senders and classify them into two categories, 1 and 2. However, the response when the system is uncertain about the author is not clearly stated.

It's noted that existing technical approaches may fail to detect malicious emails, and if they reach users undetected, users are vulnerable. Spam filters are not infallible, occasionally flagging legitimate emails as spam and allowing fraudulent emails to pass through. Consequently, user assistance is crucial for identifying phishing and spear-phishing emails and avoiding potential attacks.

Therefore, this work contributes to existing approaches by implementing sender verification at the recipients' end based on three types of features related to senders: writing pattern (stylometric analysis), gender, and email headers extracted from emails. It analyses users' emails to determine if the email originates from the claimed sender. It uses Deep learning techniques [41], Long Short-Term Memory (LSTM) which is a type of recurrent neural network [42, 43]. The system employs three colours: Green indicates safety, Amber denotes suspicion, and Red signifies high threat. The colour code assists users in promptly identifying emails. While this work shares similarities with the approach presented in [39], a distinction is made as their system classifies between two numbers, 1 and 2. The response when the system is uncertain about the author is not specified. Furthermore, their implementation is based on traditional machine learning. Also, If the system fails to detect a malicious email, and it reaches the users, they become susceptible to falling for it.

3 Methodology:

The research is based on users' verification to mitigate phishing attacks. The goal is to develop a phishing mitigation framework that helps users identify legitimate emails from illegitimate ones.

1. **Implementation of Colour Code Email Verification (CCEV):**
    - Techniques Used: Natural Language Processing (NLP) [44] and Long Short-Term Memory (LSTM).
    - The system is developed using the specified techniques.
2. **Evaluating System Efficacy:**
    - The proposed system is evaluated for its effectiveness and performance.
3. **Dataset Utilisation:**
    - Dataset: Enron email dataset.
    - Purpose: Used for model development.
4. **Feature Extraction:**
    - Features are extracted from the Enron email dataset.
    - Extraction Method: These features are used to verify and classify authors.
5. **Model Development:**
    - Utilising extracted features, a model is developed for author verification.
6. **Testing and Evaluation:**
    - The developed model is tested to assess its performance and effectiveness.

4 System Framework:

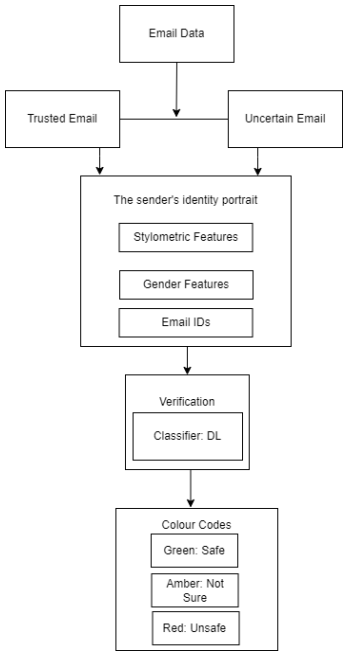

Figure 6. shows the CCEV Framework

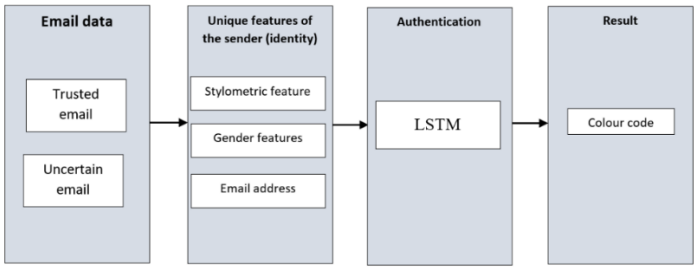

Figure 7. Framework of CCEV

Stylometric Features:
Having a similar writing style even in different works allows distinguishing the author's identity through different writing styles. Stylometric characteristics define the writing style of an author [35].

Gender Features:
There is always a need to identify gender, as gender can be misused in a variety of situations, such as email forgeries, online communities‟, forensic matters, marketing, etc.

Email IDs.
An email spoofing occurs when the origin details of an email are altered so they appear as if they originated from a different source. This is usually done in order to fool the recipient so that the real sender remains unknown.

5 System Design:

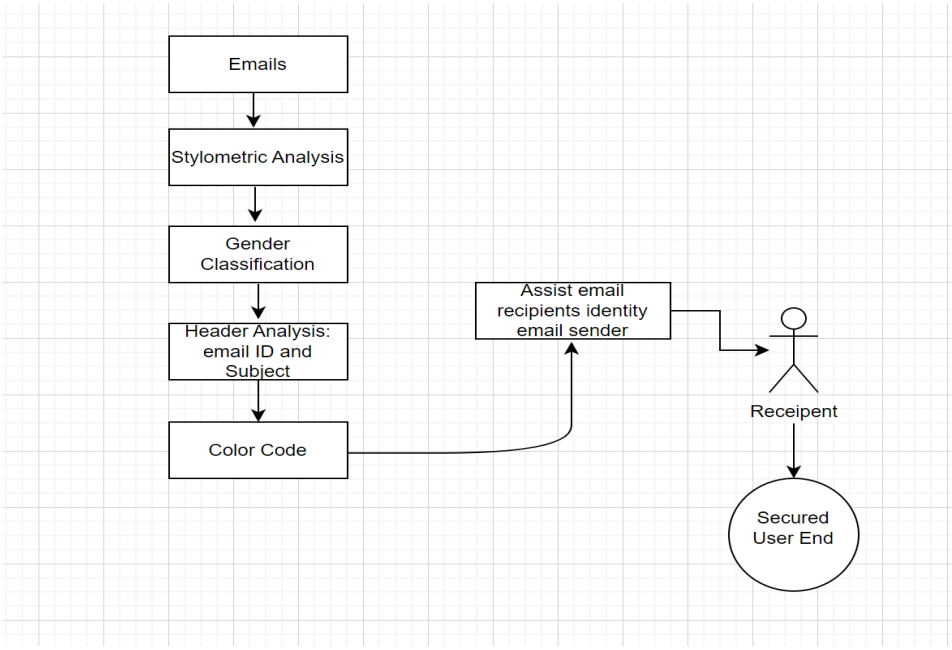

Figure 8. shows the system design flow diagram.

The system is designed to analyse all features and determine whether the sender is legitimate or illegitimate. It informs the recipient whether the sender is who they claim to be.

6 The mathematical formula for the proposed LSTM model for CCEV:

1. $x=(x_1,x_2,\ldots,x_T)$ be the input email message, where x_t represents the t-th word in the message.
2. $y=(y_1,y_2,\ldots,y_T)$ be the corresponding output sequence, where y_t is a binary value indicating whether the t-th word is part of the sender's writing pattern or not.
3. h_t be the hidden state of the LSTM at time step t, and c_t be the cell state.
4. The LSTM model consists of three gates: an input gate, a forget gate, and an output gate. These gates control the flow of information into and out of the cell state, allowing the model to selectively remember or forget information as needed.
5. The input gate determines which information from the current input and the previous hidden state should be added to the cell state. It is calculated as follows:

   $i_t=\sigma(W_i \cdot x_t+U_i \cdot h_{t-1}+b_i)$
6. The forget gate is responsible for identifying the information that needs to be discarded from the cell state. It is calculated as follows:

   $f_t=\sigma(W_f \cdot x_t+U_f \cdot h_{t-1}+b_f)$

7. The output gate determines which information should be output from the cell state to the hidden state. It is calculated as follows:

$$Ot=\sigma(Wo\cdot xt+Uo\cdot ht-1+bo)$$

8. The new cell state is calculated as follows:

$$Ct=ft\cdot ct-1+it\cdot\tanh(Wc\cdot xt+Uc\cdot ht-1+bc)$$

9. The hidden state at time step t is calculated as follows:

$$ht=Ot\cdot\tanh(Ct)$$

10. The output of the LSTM model is a sequence of probabilities, indicating the likelihood that each word in the email message belongs to the sender's writing pattern. These probabilities are mapped to a Colour Code for Email Verification, as described above.
11. If $pt \geq 0.5$, then the word is part of the sender's writing pattern, gender, and email ID, and is assigned a green code. Otherwise, is assigned an amber code or a red code.

7 Evaluation Results:

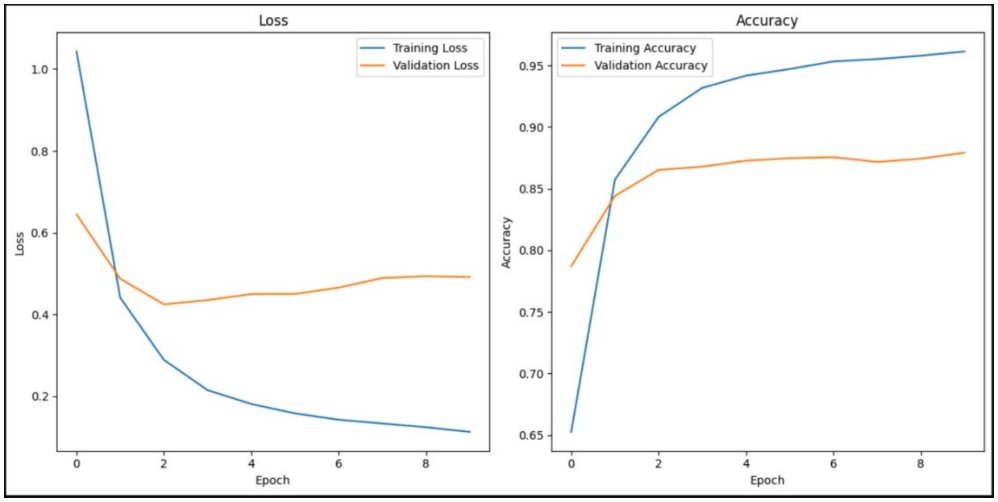

Figure 9. shows the email text classification results.

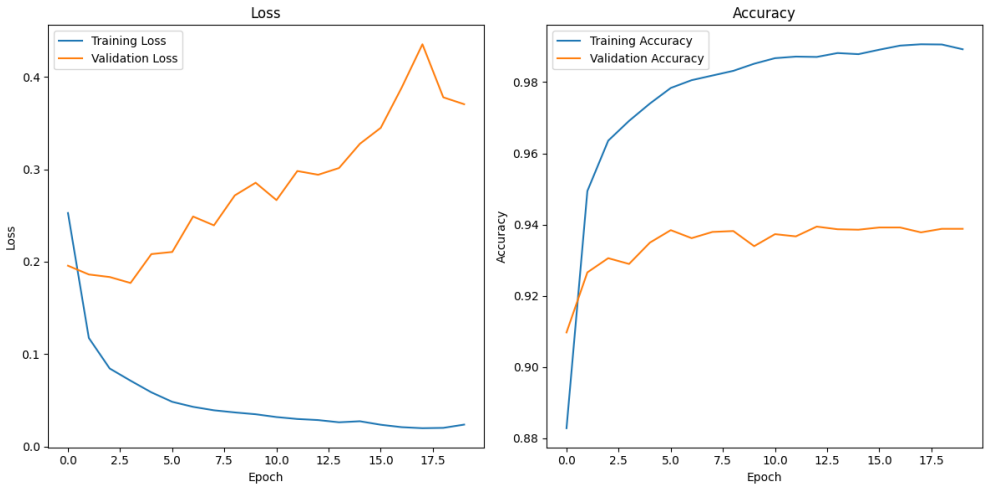

Figure 10. shows the gender classification results.

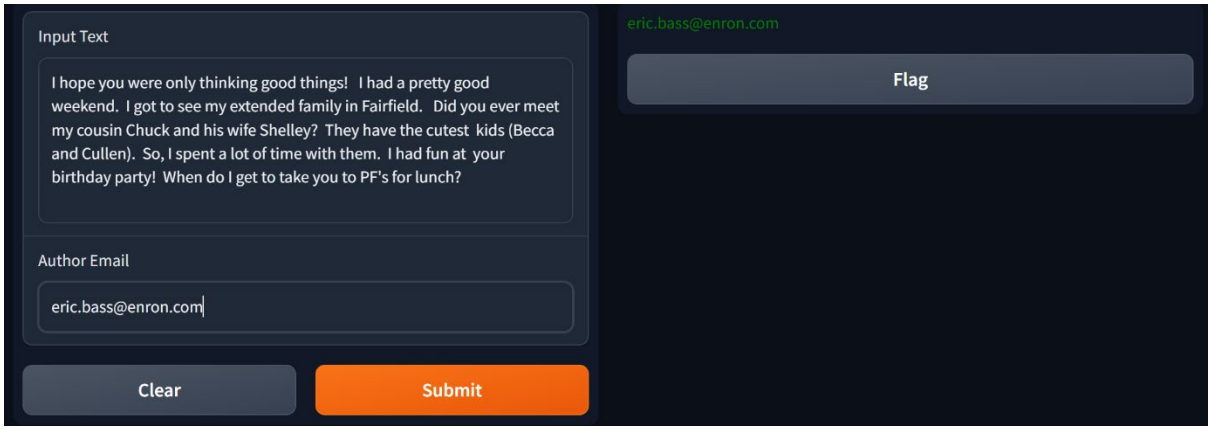

Figure 11. shows green as the user is the legitimate sender.

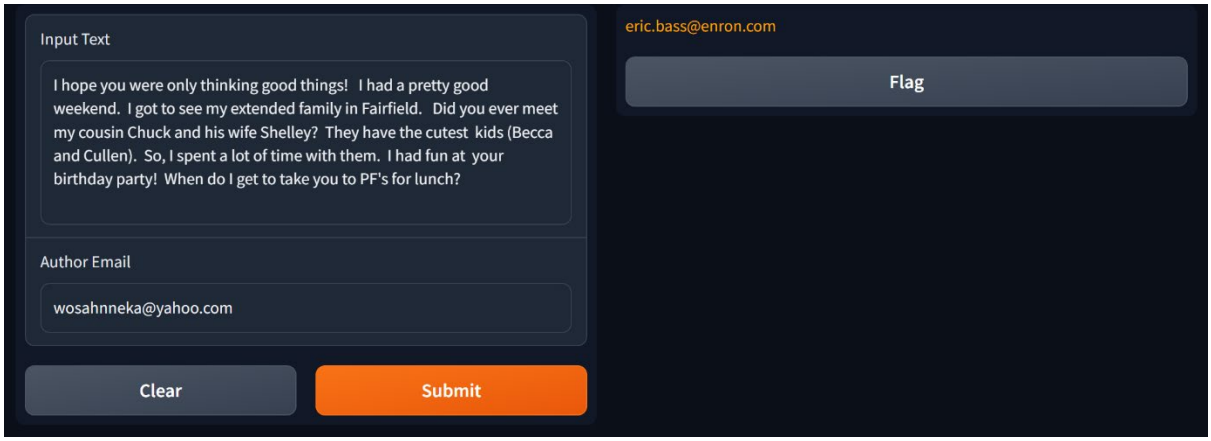

Figure 12. shows suspicious.

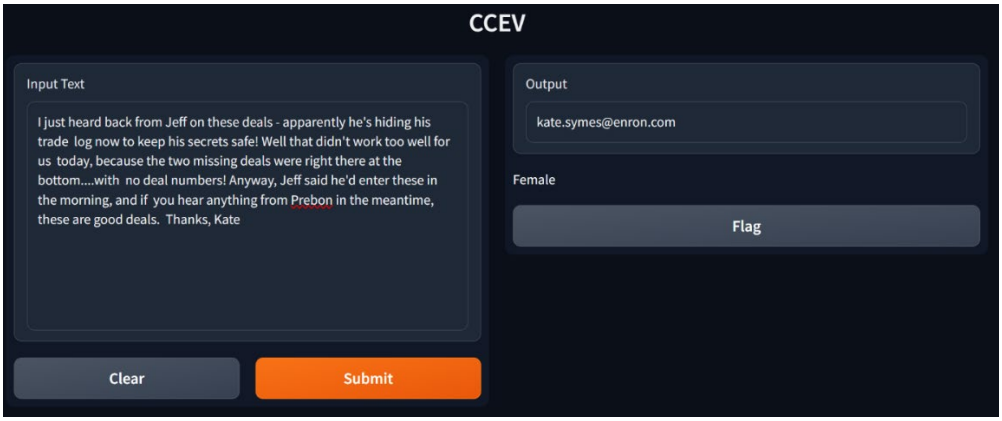

Figure 13. shows the gender and ID of the sender.

8 Discussion:

Assisting users in identifying phishing attacks is an important part to take in mitigating phishing and spear phishing attacks. knowing that humans are the weakest link in cyber security, the system considered to assist both organisations and individual users and proffered a solution to keep them safe online. Phishing is not about financial losses as it also entails risk to organisational reputation. A compromised organisational account is sometimes used by attackers to attack someone from another organisation, by sending a payment invoice to contacts that often exchange such invoices. And not all phishing attacks go with a collection of sensitive information, it can involve more than that as some phishing comes with malware and opening an attachment or clicking on a link that can lead to a successful phishing attack. Most of the attack that organisations fall for are the spear phishing impersonation attack, which is a targeted attack, and detecting this attack seems difficult for users to do.

Therefore, to mitigate these attacks, a detection system that can help users identify a dangerous email that pretend to come from a true send is needed for the security of non-technical users, as well as technical users. The proposed approach conducts sender verification at the recipients' end based on 3 kinds of features related with senders, writing pattern which is stylometric analysis, gender features and email header extracted from emails. It notifies the user of an email sender with colour code to ensure the security of the recipients. A received email is displayed with green colour if its sender's identity portrait matches that of the sender's identity trust emails. And where the email does not match the identity of the legitimate sender it displays the email with red colour. If the system is not sure whether the email is true or false it displays the result to the recipient with amber colour. With this system, users can receive emails and understand immediately the authenticity of the received emails. It reduces the time it takes users to hover on email and it curtails the fear of opening an email. Also, it ensures the email users security.

9 Conclusion and Future work:

Cyber attackers exploit emails to execute cyberattacks on both companies and individuals, aiming for financial benefits. Safeguarding user accounts against such attacks stands as a paramount priority [45]. This system is mainly designed for the security of email users and making it safer for organisations and non- technical users. Focusing on detection that assists users to identify fake email will chase phishers away from email channel.

Future Work, this implementation can be extended to other social media platforms like Facebook, Twitter, and Instagram, where identifying authors and their characteristics can also be of importance. However, like any machine learning model, there is always room for improvement. Fine-tuning hyperparameters, exploring more complex model architectures, and increasing the diversity of the training data are potential avenues for future work. Also, continuous evaluation and adaptation of the models as new data becomes available will be crucial in maintaining their accuracy and effectiveness over time.

Data Availability Statement

The dataset utilised was obtained from the web page of William Cohen, which can be accessed at http://www-2.cs.cmu.edu/~enron/

Conflict of interest declaration:

There are no conflicts of interest to disclose as all views presented in this paper belong to the author alone, and not any institution. I declare that I have no competing interests.

References:

[1] Petelka J, Zou Y, Schaub F. Put your warning where your link is: Improving and evaluating email phishing warnings. In Proceedings of the 2019 CHI conference on human factors in computing systems 2019 May 2 (pp. 1-15).

[2] Li Q, Cheng M, Wang J, Sun B. LSTM based phishing detection for big email data. IEEE transactions on big data. 2020 Mar 12;8(1):278-88.

[3] Halgaš L, Agrafiotis I, Nurse JR. Catching the phish: Detecting phishing attacks using recurrent neural networks (rnns). InInformation Security Applications: 20th International Conference, WISA 2019, Jeju Island, South Korea, August 21–24, 2019, Revised Selected Papers 20 2020 (pp. 219-233). Springer International Publishing.

[4] Rastenis J, Ramanauskaitė S, Janulevičius J, Čenys A, Slotkienė A, Pakrijauskas K. E-mail-based phishing attack taxonomy. Applied Sciences. 2020 Mar 30;10(7):2363.

[5] Nurse JR. Cybercrime and you: How criminals attack and the human factors that they seek to exploit. arXiv preprint arXiv:1811.06624. 2018 Nov 15.

[6] Alkhalil, Z., Hewage, C., Nawaf, L. and Khan, I., 2021. Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science*, *3*, p.563060.

[7] Humayun M, Jhanjhi NZ, Alsayat A, Ponnusamy V. Internet of things and ransomware: Evolution, mitigation and prevention. Egyptian Informatics Journal. 2021 Mar 1;22(1):105-17.

[8] GOV.UK. Cyber security breaches survey 2022. [Online] Available from: https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2022/cyber-security-breaches-survey-2022 [cited 2023 May 22].

[9] Anjana, Singh A. Security concerns and countermeasures in cloud computing: a qualitative analysis. International Journal of Information Technology. 2019 Dec;11:683-90.

[10] Goodman R, Hahn M, Marella M, Ojar C, Westcott S. The use of stylometry for email author identification: a feasibility study. Proc. Student/Faculty Research Day, CSIS, Pace University, White Plains, NY. 2007 May:1-7

[11] Widup S, Rudis B, Hylender D, Spitler M, Thompson K, Baker WH, Bassett G, Karambelkar B, Brannon SK, Kennedy D. Verizon Data Breach Investigations Report. 2015. URL: 1-2-DBIR-Widup (nist.gov) [accessed 2022-03-22].

[12] Alzahrani SM, Salim N, Abraham A. Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2011 May 12;42(2):133-49.
[13] Vayansky I, Kumar S. Phishing–challenges and solutions. Computer Fraud & Security. 2018 Jan 1;2018(1):15-20.
[14] Nmachi WP, Win T. Mitigating phishing attack in organisations: a literature review. InCS & IT Conference Proceedings 2021 Jan 23 (Vol. 11, No. 1). CS & IT Conference Proceedings.
[15] Sharma P, Dash B, Ansari MF. Anti-phishing techniques–a review of Cyber Defense Mechanisms. International Journal of Advanced Research in Computer and Communication Engineering ISO. 2022 Jul 31;3297:2007.
[16] Evans K, Abuadbba A, Wu T, Moore K, Ahmed M, Pogrebna G, Nepal S, Johnstone M. RAIDER: Reinforcement-aided spear phishing detector. InInternational Conference on Network and System Security 2022 Dec 7 (pp. 23-50). Cham: Springer Nature Switzerland.
[17] Al-Hamar Y, Kolivand H, Tajdini M, Saba T, Ramachandran V. Enterprise Credential Spear-phishing attack detection. Computers & Electrical Engineering. 2021 Sep 1; 94:107363.
[18] Fette I, Sadeh N, Tomasic A. Learning to detect phishing emails. InProceedings of the 16th international conference on World Wide Web 2007 May 8 (pp. 649-656).
[19] Khonji M, Iraqi Y, Jones A. Enhancing phishing e-mail classifiers: A lexical url analysis approach. International Journal for Information Security Research (IJISR). 2012 Mar;2(1/2):40.
[20] Smadi S, Aslam N, Zhang L. Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. Decision Support Systems. 2018 Mar 1;107:88-102.
[21] Hota HS, Shrivas AK, Hota R. An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique. Procedia computer science. 2018 Jan 1;132:900-7.
[22] Lötter A, Futcher L. A framework to assist email users in the identification of phishing attacks. Information & Computer Security. 2015 Oct 12;23(4):370-81.
[23] Li JS, Chen LC, Monaco JV, Singh P, Tappert CC. A comparison of classifiers and features for authorship authentication of social networking messages. Concurrency and Computation: Practice and Experience. 2017 Jul 25;29(14):e3918.
[24] Abbasi A, Chen H. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (TOIS). 2008 Apr 8;26(2):1-29.
[25] Beigi G, Liu H. A survey on privacy in social media: Identification, mitigation, and applications. ACM Transactions on Data Science. 2020 Mar 12;1(1):1-38.
[26] Afroz S, Brennan M, Greenstadt R. Detecting hoaxes, frauds, and deception in writing style online. In2012 IEEE Symposium on Security and Privacy 2012 May 20 (pp. 461-475). IEEE.
[27] Liu Y, Wu YF. Fned: a deep network for fake news early detection on social media. ACM Transactions on Information Systems (TOIS). 2020 May 5;38(3):1-33.
[28] Afroz S, Islam AC, Stolerman A, Greenstadt R, McCoy D. Doppelgänger finder: Taking stylometry to the underground. In2014 IEEE Symposium on Security and Privacy 2014 May 18 (pp. 212-226). IEEE.
[29] McDonald AW, Afroz S, Caliskan A, Stolerman A, Greenstadt R. Use fewer instances of the letter "i": Toward writing style anonymization. InPrivacy Enhancing Technologies: 12th International Symposium, PETS 2012, Vigo, Spain, July 11-13, 2012. Proceedings 12 2012 (pp. 299-318). Springer Berlin Heidelberg.
[30] Narayanan A, Paskov H, Gong NZ, Bethencourt J, Stefanov E, Shin EC, Song D. On the feasibility of internet-scale author identification. In2012 IEEE Symposium on Security and Privacy 2012 May 20 (pp. 300-314). IEEE.
[31] Ledger G, Merriam T. Shakespeare, fletcher, and the two noble kinsmen. Literary and Linguistic Computing. 1994 Jan 1;9(3):235-48.
[32] De Vel O, Anderson A, Corney M, Mohay G. Mining e-mail content for author identification forensics. ACM Sigmod Record. 2001 Dec 1;30(4):55-64.
[33] Nizamani S, Memon N. CEAI: CCM-based email authorship identification model. Egyptian Informatics Journal. 2013 Nov 1;14(3):239-49.
[34] Iqbal F, Khan LA, Fung BC, Debbabi M. E-mail authorship verification for forensic investigation. InProceedings of the 2010 ACM Symposium on Applied computing 2010 Mar 22 (pp. 1591-1598).
[35] Lin E, Aycock J, Mannan M. Lightweight client-side methods for detecting email forgery. InInformation Security Applications: 13th International Workshop, WISA 2012, Jeju Island, Korea, August 16-18, 2012, Revised Selected Papers 13 2012 (pp. 254-269). Springer Berlin Heidelberg.
[36] Brocardo ML, Traore I, Saad S, Woungang I. Authorship verification for short messages using stylometry. In2013 International Conference on Computer, Information and Telecommunication Systems (CITS) 2013 May 7 (pp. 1-6). IEEE.
[37] Stringhini G, Thonnard O. That ain't you: detecting spearphishing emails before they are sent. arXiv preprint arXiv:1410.6629. 2014 Oct 24.

[38] Duman S, Kalkan-Cakmakci K, Egele M, Robertson W, Kirda E. Emailprofiler: Spearphishing filtering with header and stylometric features of emails. In2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC) 2016 Jun 10 (Vol. 1, pp. 408-416). IEEE.
[39] Xiujuan W, Chenxi Z, Kangfeng Z, Haoyang T, Yuanrui T. Detecting spear-phishing emails based on authentication. In2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) 2019 Feb 23 (pp. 450-456). IEEE.
[40] Ding X, Liu B, Jiang Z, Wang Q, Xin L. Spear phishing emails detection based on machine learning. In2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD) 2021 May 5 (pp. 354-359). IEEE.
[41] Mishra S, Jabin S. Anomaly detection in surveillance videos using deep autoencoder. International Journal of Information Technology. 2023 Dec 24:1-2.
[42] Rajak A, Tripathi R. DL-SkLSTM approach for cyber security threats detection in 5G enabled IIoT. International Journal of Information Technology. 2023 Dec 18:1-8.
[43] Jain G, Sharma M, Agarwal B. Optimizing semantic LSTM for spam detection. International Journal of Information Technology. 2019 Jun 4;11:239-50.
[44] Priya CS, Deepalakshmi P. Sentiment analysis from unstructured hotel reviews data in social network using deep learning techniques. International Journal of Information Technology. 2023 Oct;15(7):3563-74.
[45] Nmachi Wosah P. A framework for securing email entrances and mitigating phishing impersonation attacks. arXiv e-prints. 2023 Dec:arXiv-2312.