2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC) | 979-8-3503-2697-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/COMPSAC57700.2023.00198

The Application of the BERT Transformer Model for Phishing Email Classification

Denish Omondi Otieno
Department of Computer Science
Texas Tech University
Lubbock, TX, USA
deotieno@ttu.edu

Akbar Siami Namin
Department of Computer Science
Texas Tech University
Lubbock, TX, USA
akbar.namin@ttu.edu

Keith S. Jones
Department of Psychological Science
Texas Tech University
Lubbock, TX, USA
keith.s.jones@ttu.edu

***Abstract*—Phishing attacks pose serious challenges to users, commercial institutions and governments alike including identity theft, companies and government secrets. The weakest link in any computer security system is the people and the success of phishing attacks is substantially based on manipulating human emotions. Therefore, we cannot solely rely on humans to detect phishing attacks as phishing attacks are becoming more and more sophisticated and can bypass filters set by anti-phishing techniques. More effective and automatic phishing detection mechanisms are required and many detectors have been proposed. However, the high number of phishing emails and attacks urges additional efforts. In this study, the effectiveness of word embedding in classifying phishing emails is investigated. Pre-trained transformer model BERT (Bidirectional Encoder Representations from Transformers) is fine-tuned to execute the task of classifying emails into phishing and non-phishing. The results show that transformer technology is accurate enough to improve phishing emails detection and can complement existing classifiers.**

***Index Terms*—Social Engineering, Phishing Emails, BERT Transformer Language Model.**

I. Introduction

The digital world is rapidly expanding and evolving, and likewise, are cyber-criminals who have relied on the illegal use of digital assets especially sensitive information to inflict damage to individuals, companies, governments and alike [1]. Identity theft, which can be defined as impersonating a person's identity to steal and use their personal information is one of the most threatening crimes for all internet users [2]. Cyber-criminals have and are continuously evolving their crafty methods of stealing information and social engineering attacks remain their favorite approach [1]. Social engineering uses psychological manipulation to trick users into making security mistakes or giving away sensitive information. One of the social engineering crimes that allow cyber-criminals to perform identity theft is called phishing. Phishing functions by a form of pretexting i.e., impersonation to gain the victim's trust [3]. Phishing messages aim to nudge victims into revealing sensitive information such as usernames, passwords, credit card numbers or even click on links to malicious websites or open attachments that contain malware by impersonating legitimate entities in the cyber-space [4]. Jakobsson and Myers [3] detail the characterization of phishing attacks into three ways:

1) *The Hook*, a legitimate looking like tool (i.e., the malicious website, the email form, social media site, etc.) that the attacker employs to collect the victims sensitive information;
2) *The Lure*, the motivation, the enticement or the incentive being used to trick the victim, might be a communication encouraging the recipient to follow an included hypertext link, where the hyperlink masks a spoofed uniform resource locator (URL) of a legitimate website; and
3) *The Catch*, where the phisher uses the sensitive information about the entity to conduct illegal transactions or business.

Phishing messages can create a sense of urgency, curiosity or even fear in victims. Stojnic et al. [5] document that cybercriminals are interested in breaking into people's mindset rather than breaking into systems straightaway. In their findings, Stojnic et al. [5] advance that when crafting phishing emails the phishers consider a pattern involving:

- Devising short and catchy email subject headers that create a feeling of urgency and intrigue in victims.
- Capturing the victim's attention and gaining trust by establishing a sense of authority (i.e., claiming to be from trusted entities like financial institutions etc; and
- Giving a call to action to the victim in order to gain a response, often asking for sensitive information.

Maneriker et al. [6] report that phishing attacks continue to be a persistent problem and the number of known phishing sites are increasing. Blocking phishing attacks using a continuously growing list of known phishing sites also often fails to protect users in practice [6]. Given the significant repercussions of phishing, detection of these attacks is an active area of research and this paper investigates the application of a transformer-based model in phishing emails classification. The key contributions of this paper are as follows:

1) The paper performs sentiment analysis of phishing vs. non-phishing emails using TextBlob a Python 2 and 3 library for processing textual data [7] to find out if there exists a sentimental analysis similarity or difference between the two emails sets;
2) It also carries out topic modelling by applying Non-Negative Matrix Factorization (NMF), Latent Semantic

979-8-3503-2697-0/23/$31.00 ©2023 IEEE
DOI 10.1109/COMPSAC57700.2023.00198

Analysis (LSA) and Latent Dirichlet allocation (LDA) to point out topic trends in phishing and non-phishing emails; and ultimately,

3) It presents the transformer-based classification results of phishing and non-phishing emails based on the BERT transformer language model.

The paper is structured as follows. Section II reviews the literature. A brief background of the technical concepts utilized in this study are presented in Section III. Section IV introduces the methodology. The experimental setup is reported in Section V and Section VI details the findings while Section VII concludes the paper and sketches the future work.

II. Related Work

Two main types of technical methods for detecting phishing emails are blacklisting and machine learning [8]. Machine learning techniques can automate phishing emails detection through various methods, e.g., the use of deep learning detectors that automate phishing emails detection [8]. The blacklisting method on the other hand compares the sender's Domain Name System (DNS) address, Internet Protocol (IP) address or email address with a predefined list of phishing addresses, and if a match is flagged the email is rejected before it reaches the Simple Mail Transfer Protocol (SMTP) mail server [9]. Studies have proved the efficiency of using phishing emails detection and classification techniques in social engineering [10], [11], [12]. However, to cope with the evolution of phishing emails more approaches should be instituted that will exploit all of the emails' traits to enhance the detection capability of machine learning classifiers [10].

Bountakas et al. [10] propose a phishing email detection methodology that focuses on the detection of phishing emails by combining ensemble learning methods with hybrid features. In the study, Bountakas et al. [10] present that hybrid features provide an accurate representation of emails by fusing their content and textual traits. Two methods namely the stacking ensemble learning method and the soft voting ensemble learning are advanced by Bountakas et al. [10] as part of their phishing email detection toolkit and numerical results of an imbalanced dataset that considers the evolution of phishing emails show that soft voting ensemble learning outperforms other prominent machine learning/deep learning algorithms yielding an F1-score equal to 0.9942 [10].

Halgas et al. [11] state that online services in our daily lives have been accompanied by a range of malicious attempts to trick individuals into performing undesired actions. The most popular medium of these attempts is phishing attacks, particularly through emails and websites. In order to defend against phishing attacks, there is an urgent need for automated mechanisms to identify the malicious contents before they reach the victim. Halgas et al. [11] propose a classifier using the Recurrent Neural Networks (RNN), consisting of an encoding layer, two recurrent layers, and a linear output layer with a Softplus activation. Furthermore, Halgas et al. [11] report that they observe an accuracy of 98.91%, false positive-rate of 1.26%, false negative-rate of 1.47%, precision of 98.74%, recall of 98.53% and an F-measure of 98.63% from their first dataset which is a combination of 6951 ham emails from the SpamAssassin public corpus [13] and 4572 phishing emails from the Nazario phishing corpus [14] collected before August 2007. Likewise, from dataset two defined as a combination of the Enron email dataset [15] with the phishing emails from the Nazario phishing corpus, they [11] report an accuracy of 96.74%, false positive-rate of 2.50%, false negative-rate of 4.02%, precision of 97.45%, recall of 95.98% and an F-measure of 96.71%. Halgas et al. [11] state that the flexibility of RNNs gives their system an edge over the expert feature selection procedure, which is vastly employed in machine learning based attempts at phishing mitigation.

A study on feature extraction and selection for text classification a case study of phishing emails detection is presented by Zareapoor and Seeja [16]. The study [16] reports that in phishing emails detection, email classification is difficult due to its high dimensional sparse features that affect the generalization performance of classifiers.

Yao et al. [17] propose to use graph convolutional networks for text classification. The idea involves to represent an entire corpus as a heterogeneous graph and learn word and document embeddings mutually using graph neural networks [17]. Yao et al. [17] build a single text graph for a corpus based on word co-occurrence and document word relations, then learn a Text Graph Convolutional Network (Text GCN) for the corpus. The experimental results on multiple benchmark datasets demonstrate that a vanilla Text GCN without any external word embeddings or knowledge outperforms some state-of-the-art methods for text classification with a reported test accuracy mean $\pm$ standard deviation of 0.8634 ± 0.0009 for dataset one, 0.9707 ± 0.0010 for dataset two, 0.9356 ± 0.0018 for dataset three, 0.6836 ± 0.0056 for dataset four and 0.7674 ± 0.0020 for dataset five.

Barraclough et al. [18] advance a real-time hybrid neuro-fuzzy scheme for detecting phishing websites and protecting consumers engaging in online transactions. Barraclough et al. [18] state that despite existing approaches that utilize URL blacklists to combat phishing attacks, the approaches cannot generalize well with the ever-evolving phishing attacks strategies partly due to human weakness in verifying blacklists. Barraclough et al. [18] divulge that existing feature-based methods suffer high false positive rates and insufficient phishing features and as a result, their study proposes to introduce new inputs into a single protection platform with the objective being to utilize a Neuro-Fuzzy Scheme to detect phishing sites. Barraclough et al [18], report an overall average test accuracy of 98.5% from their study.

A study by Arya and Chamotra [19] utilizes a multi-layer detection framework for spear phishing attacks detection. The study explores sentiment analysis, context-based behavior analysis along with deception technologies to understand and profile common types of email phishing techniques such as indication of urgency in the email, package deliveries, bills and requests, law enforcement, and scanned documents. Arya and Chamotra [19] report that the latest targeted attacks which em-

ploy compromised email accounts for sending spear-phishing emails are almost impossible to detect using conventional security solutions. In their study [19] each of the detectors in the framework deployed contributes towards a suspiciousness score of the email, and when this score crosses a predefined threshold, the email is analyzed in a sandbox environment.

Chatterjee and Namin [12] introduce an approach to model the identification of phishing websites through Reinforcement Learning (RL) where an agent learns the value function from the given input URL in order to perform a classification task. The study maps the sequential decision-making process for classification using a deep neural network-based implementation of RL. Chatterjee and Namin [12] report a precision of 0.867, recall of 0.88, accuracy of 0.901 and an F-measure of 0.873 from their study.

The review of literature points towards a constant need by the research community to address the issue of phishing attacks and in this study, we seek to advance the line of study by investigating the effectiveness of word embedding in classifying phishing emails. The Pre-trained transformer model BERT is fine-tuned to build an effective phishing email detection classifier.

III. Technical Background

This section briefly describes some of the key techniques adapted in this research.

A. BERT Transformer Model

The Bidirectional Encoder Representations from Transformers (BERT) [20] uses encoders in a transformer as a sub-structure to pre-training models for Natural Language Processing (NLP) tasks. BERT, a transformer based deep learning model, is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers [20]. The bi-directionality capability of BERT allows it to read input from either right to left or left to right, simultaneously. Another advantage of BERT is that the model does not need to be retrained, it can just be fine-tuned with one additional output layer to create state-of-the-art models for a wide range of tasks such as text summarization, question answering, etc. without substantial tasks-specific architecture modifications [20]. In this study the paper aims to use BERT to pre-process a training data to generate input-mask, input-type-ids and input-word-ids for phishing and non-phishing emails classification.

B. Sentiment Analysis using TextBlob

Sentiment analysis can help us decipher the mood and emotions regarding a context. Sentiment analysis can be achieved using TextBlob. TextBlob is a Python 2 and 3 library for processing textual data. The study adopts TextBlob [7] to perform sentiment analysis on the phishing and non-phishing emails. When a sentence is passed into TextBlob it gives two outputs, which are *polarity* and *subjectivity*. Polarity lies between $[-1, 1]$ where 1 identifies a positive sentiment and -1 refers to a negative sentiment while Subjectivity lies within $[0, 1]$ and refers to personal opinions and judgments. A higher subjectivity implies the set contains personal opinion rather than factual information.

C. Non-Negative Matrix Factorization (NMF)

A matrix decomposition algorithm, Non-Negative Matrix Factorization (NMF) proposed by Lee and Seung [21], is a group of algorithms in multivariate analysis and linear algebra. It is an unsupervised learning technique that extracts sparse and meaningful features from a set of non-negative data vectors. NMF factorizes a non-negative matrix X, into the product of two lower-rank matrices W and H, such that WH approximates an optimal solution of X. NMF reduces the dimensionality of data into lower-dimensional spaces. The algorithm iteratively changes the values of W and H such that their product approaches X. This method keeps the structure of the original data intact and makes sure that both the basis and weights are non-negative. NMF terminates when the approximation error converges, or the specified number of iterations is reached.

D. Latent Semantic Analysis (LSA)

Text can be characterized by the semantic content it carries. Over time statistical computational models have been developed to create semantic representations for words encountered in text. One such model is Latent Semantic Analysis (LSA) [22]. LSA is a model for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of documents [23]. It is an unsupervised learning technique that rests on two pillars:

1) The distributional hypothesis, which states that words with similar meanings tend to appear in similar contexts; and
2) Singular Value Decomposition (SVD) a mathematical technique that performs decomposition directly on the dataset as it is.

The LSA algorithm generates a semantic space from a statistical analysis of the frequencies with which words co-occur. The process by which LSA builds a semantic space from a document collection can be termed as "training" and after training, the semantic space comprises a set of vectors containing the semantic features i.e., components of the concept associated with the lexical items for each word encountered in the document collection.

E. Latent Dirichlet Allocation (LDA)

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus [24]. It is an unsupervised machine learning technique used to recognize the latent topic structure of textual documents. The underlying assumption of LDA is that a text document will consist of multiple themes. LDA is a three-level hierarchical Bayesian model where each item of a collection of text is modeled as a finite mixture over an underlying set of topics [24]. Each topic is in turn modeled as an infinite mixture over an underlying set of topic probabilities [24]. For text modelling, the topic probabilities provide an

explicit representation of a document [25]. Additionally, a topic model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. The basic idea in LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [24].

IV. Methodology

This section describes the methodology used in this study. Figure 1 conveys the six steps employed by this paper and they are: Data Pre-processing I, Feature Engineering, Data Pre-processing II, Sentiment Analysis, Topic Modelling and Emails Classification.

1) *Data Pre-processing I and II.* Data pre-processing is divided into two phases. Phase I and Phase II. In phase I an emphasize on dropping the duplicates, removing the (NaNs) i.e., emails with null values, removing the URLs, punctuations, deleting the special characters and applying lower-casing is realized. Likewise, in phase II elimination of the stop-words i.e., words which serve a syntactic purpose rather than give content to a sentence is achieved. The paper justifies the two-phase data pre-processing in that a need exists to quantify the number of stop-words in the phishing and non-phishing datasets as part of the study's feature engineering process.
2) *Feature Engineering.* Feature Engineering necessitates transforming textual data into numerical data. It is a flexible way of extracting features from documents. This study employs the "count model" to describe the occurrence of words within a corpus. The features of interest to the study include character count, word count, unique word count, mean word length, stop words count and non-stop words count.
3) *Sentiment Analysis.* The paper applies sentiment analysis using TextBlob to determine the emotional tone of the phishing and non-phishing emails. While the focus is on the polarity of the phishing and non-phishing emails i.e., positivity, negativity and neutrality, the paper progresses to put forward the distribution of the top ten positive words from the phishing and non-phishing emails.
4) *Topic-Modelling.* Topic modelling is an approach that can scan a series of documents to find words and phrases patterns within them. Topic modelling clusters word groupings and related expressions that best represent a set, a corpus, or an article. To uncover the latent topics hidden in the phishing and non-phishing emails, the paper applies NMF a factorization-based algorithm, LSA an unsupervised learning approach for extracting relationships between words and LDA where statistical and graphical concept are used to find correlations between documents in a corpus.
5) *Emails Classification.* Gmail and Outlook have email classification systems, i.e., in Gmail, emails are categorized into tabs such as Promotions, Updates, Forums, Social and Primary by extracting the email's sender, the email's contents and by learning the user behavior, the system sorts emails into the preferred folders. An email classification system is a monitoring system from within an email's inbox, designed to monitor the real-time flow of emails by automatically reviewing the contents from the emails. The system classifies emails based on the criteria set by an institution and with the advancement in social engineering attacks, emails classification systems can be critical additions to the existing social engineering attacks management policies.

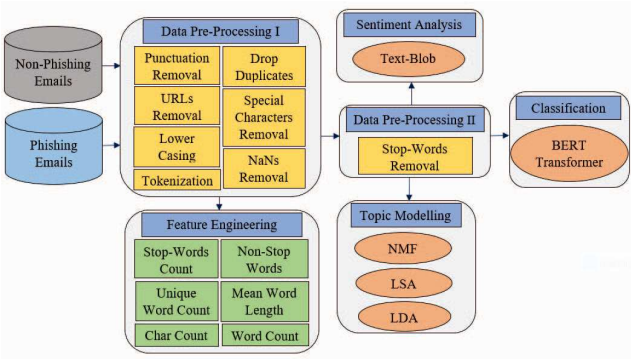


Fig. 1: The flowchart of methodology.

V. Experimental Setup

This study concatenates two datasets into one dataset for the experiments in the study i.e.,

1) The Nazario phishing corpus by Nazairo [14]; and
2) The Enron email dataset [15].

Google Colab, a Google research product is adopted as the preferred study environment.

A. Data Sets

The Nazario phishing corpus [14] consists of a good number of phishing emails collected and recorded in each year. The phishing emails are subdivided into subsets and this study zeros in on the mbox subsets concatenating the phishing0.mbox subset, the phishing1.mbox subset, the phishing2.mbox subset, the phishing3.mbox subset and the private-phishing4.mbox subset into a subset. A preliminary data preparation after concatenation drops the NaNs and all the columns except the `Body` column where the study extracts 8094 phishing emails.

The second dataset, the Enron email dataset [15], is a collection of a total of about 0.5M real-world email messages from about 150 users and the data was originally made public and posted on the web. This study applies the May 7th, 2015, version of the Enron email dataset and extracts a random sample of 8094 non-phishing emails from the `Body` column in respect to the extracted phishing emails from the Nazario phishing corpus [14].

This study is performed based on the `Body` contents of the emails.

B. Data Preparation

Data labeling is a part of the preprocessing stage when developing machine learning/deep learning models. It denotes

TABLE I: TextBlob Sentiment Analysis of Phishing vs Non-Phishing Emails.

#	Class	Positive	Negative	Neutral
1	Phishing Emails	67.0%	20.6%	12.4%
2	Non-Phishing Emails	67.9%	11.5%	20.6%

the identification of raw data and then the addition of one or more labels to the data to specify its context to the machine learning/deep learning models. The 8094 extracted phishing emails from the Nazario phishing corpus [14] are thus labeled as (1) implying phishing emails and equally the random sample of 8094 non-phishing emails from the Enron email dataset [15] are labeled as (0) to represent non-phishing emails. To better explore the selected datasets, this study develops Python scripts in which it utilizes the NLP analysis libraries such as nltk, pandas, numpy, string, random, textblob, matplotlib, gensim, itertools, counter, sklearn and their related modules. The extracted phishing emails from the Nazario phishing corpus [14] and the random sample of non-phishing emails from the Enron email dataset are likewise concatenated to create a new data frame. An analysis of the concatenated new data frame after lower casing reveals 363 duplicates from the random non-phishing emails sample of the Enron email dataset and 2459 phishing emails duplicates from the Nazario phishing corpus [14]. All the duplicate emails are discarded from this study.

C. Classification Using BERT

The study loads the data into BERT to generate a contextualized embedding vector. This study further applies pre-processing using the pre-processor object and pass the pre-processed text to its model to generate the contextualized embedding vector. To maximize performance the study balanced the dataset and used a dropout layer to regularize the model in a bid to prevent over-fitting. The dependencies applied by the study include Tensorflow-hub, Tensorflow-text in addition to the NLP analysis libraries listed in the data preparation (Section V-B). An examination of the dataset reveals an imbalanced set of 5635 emails labeled as (1) implying phishing emails and 7731 emails labeled as (0) denoting non-phishing emails. The imbalance might affect the study's model and the study uses regularization i.e., down-sampling for the majority class. The study takes any random minority number of samples 5635 for the majority class 7731 and creates a balanced dataset. The investigation splits the balanced dataset into 80-20 ratio with (80%) being training and (20%) as the testing, for evenness of the data the study applies stratification "stratify" to ensure the same ratio of both categories are loaded for each case with the aim being to prevent over-fitting.

VI. Results and Analysis

This section reports the results of this study.

1) Sentiment Distribution: Sentiment Analysis analyses a text, document, or a corpus and tells whether the underlying sentiment is positive, negative, or neutral. Table I illustrates the percentage of polarity for both the phishing and non-phishing emails. While 20.6% of the phishing emails are detected to be negative, 20.6% of the non-phishing emails are observed to be neutral. There appears to be a close variation in the percentage of positivity for both sets of emails with only a 0.9% difference between the phishing and non-phishing emails. 12.4% of the phishing emails are declared to be neutral and 11.5% of the non-phishing emails are found to be negative. Furthermore, Table II shows the distribution of the top 10 positive words in the phishing and non-phishing emails, where "new" at 14.1%, "email" at 14.0% and "message" at 12.5% are the most frequent positive words in the non-phishing emails, while "click" at 5.4%, "online" at 5.5% and "access" at 5.9% are recorded at the base of the top 10 frequent positive words in the phishing emails with "account" at 26.8% being the top most positive word noticed from the phishing emails. The results from Table I and Table II show that sentiment analysis alone could not present a higher classification ability between the phishing and non-phishing emails other than semantic relations and the frequency vectors of terms.

TABLE II: Distribution of The Top 10 Positive Words in Phishing vs Non-Phishing Emails.

#	Phishing	%	#	Non-Phishing	%
1	Account	26.8%	1	New	14.1%
2	Paypal	13.1%	2	Email	14.0%
3	Email	11.1%	3	Message	12.5%
4	Ebay	10.8%	4	Sent	10.8%
5	Information	9.3%	5	Time	9.6%
6	Security	6.3%	6	Original	9.0%
7	Bank	5.9%	7	Information	9.0%
8	Access	5.9%	8	Need	7.8%
9	Online	5.5%	9	Business	6.8%
10	Click	5.4%	10	Call	6.5%

2) Spearman Features Correlation Matrixes: Correlation is a statistic that measures the degree to which two variables move in relation to each other, it shows the strength of a relationship between two variables. The two main types of correlation are positive and negative. Positive correlation occurs when two variables move in the same direction and there is a linear relationship between them, i.e., as one increases, so does the other and a negative correlation occurs when two variables move in opposite directions i.e., as one increases, the other decreases. Figure 2a demonstrates a representation of the correlation of features from the phishing emails. It shows that "word-count" and "char-count" and "non-stop-words" and "word-count" are having strong positive correlations. Likewise, Figure 2b indicates that features from the non-phishing emails i.e., "word-count" and "char-count", "non-stop-words" and "word-count" are having perfect collinear relationship. The analysis of Figure 2a and Figure 2b show that existing methods such as correlation plots and heat maps are insufficient in the classification of phishing emails as most of the correlation features from the phishing and non-phishing emails appear to be having closely related or similar patterns with no strong distinctive features.

3) Topic-Modelling: This study extracts 10 topic words modelled by NMF, LSA and LDA from both the phishing and

TABLE III: Non Negative Matrix Factorization (NMF) Emails Topic Modelling Results.

NMF Part A: Phishing Emails		
Topic #	10 Topic Words	Possible Categorization
1	'0.025*"email" + 0.011*"credit" + 0.010*"mail" + 0.009*"national" + 0.008*"card" + 0.008*"business" + 0.007*"city" + 0.007*"address" + 0.006*"corporate" + 0.006*"table"'	Personal/Sensitive Information
2	'0.061*"bank" + 0.043*"service" + 0.039*"customer" + 0.034*"online" + 0.022*"business" + 0.022*"form" + 0.014*"dear" + 0.011*"complete" + 0.011*"national" + 0.010*"client"'	Online Service Request
3	'0.107*"ebay" + 0.031*"message" + 0.030*"email" + 0.025*"item" + 0.025*"send" + 0.020*"policy" + 0.019*"member" + 0.017*"question" + 0.016*"trademark" + 0.015*"privacy"'	Communication Updates
NMF Part B: Non-Phishing Emails		
Topic #	10 Topic Words	Possible Categorization
1	'0.025*"image" + 0.009*"time" + 0.009*"stock" + 0.008*"news" + 0.007*"company" + 0.007*"share" + 0.006*"trade" + 0.006*"week" + 0.006*"right" + 0.006*"billion"'	Trade/Business
2	'0.008*"service" + 0.008*"email" + 0.006*"send" + 0.006*"information" + 0.006*"include" + 0.006*"business" + 0.006*"need" + 0.005*"work" + 0.005*"message" + 0.005*"provide"'	Notification
3	'0.020*"report" + 0.008*"account" + 0.008*"committee" + 0.008*"financial" + 0.007*"business" + 0.007*"company" + 0.006*"management" + 0.005*"executives" + 0.005*"work" + 0.005*"deal"'	Trade/Business

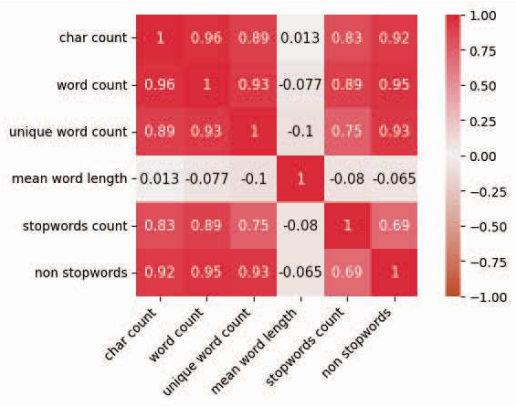
TABLE IV: Latent Semantic Analysis (LSA) Emails Topic Modelling Results.

LSA Part A: Phishing Emails		
Topic #	10 Topic Words	Possible Categorization
1	'0.625*"account" + 0.341*"paypal" + 0.315*"ebay" + 0.231*"email" + 0.193*"information" + 0.145*"access" + 0.120*"security" + 0.115*"bank" + 0.106*"send" + 0.106*"update"'	Update Information
2	'-0.721*"ebay" + 0.337*"account" + 0.251*"paypal" + -0.204*"message" + -0.173*"item" + -0.146*"email" + -0.141*"send" + -0.117*"policy" + -0.111*"question" + -0.107*"member"'	Personal/Sensitive Information
3	'-0.581*"bank" + 0.540*"paypal" + -0.254*"online" + -0.241*"service" + -0.210*"customer" + -0.123*"business" + -0.112*"form" + -0.091*"dear" + 0.082*"ebay" + -0.077*"national"'	Online Service Request
LSA Part B: Non-Phishing Emails		
Topic #	10 Topic Words	Possible Categorization
1	'0.312*"price" + 0.194*"electricity" + 0.183*"time" + 0.158*"report" + 0.143*"company" + 0.129*"plan" + 0.127*"plant" + 0.106*"million" + 0.105*"billion" + 0.100*"year"'	Financial Information
2	'-0.368*"price" + 0.293*"report" + -0.239*"electricity" + 0.159*"firm" + 0.157*"board" + 0.143*"account" + 0.134*"financial" + 0.133*"company" + 0.130*"committee" + 0.119*"million"'	Financial Information
3	'-0.668*"error" + -0.422*"database" + -0.300*"occur" + -0.299*"attempt" + -0.299*"engine" + -0.299*"initialize" + -0.076*"operation" + -0.066*"close" + -0.065*"perform" + -0.013*"message"'	System/Computer Error

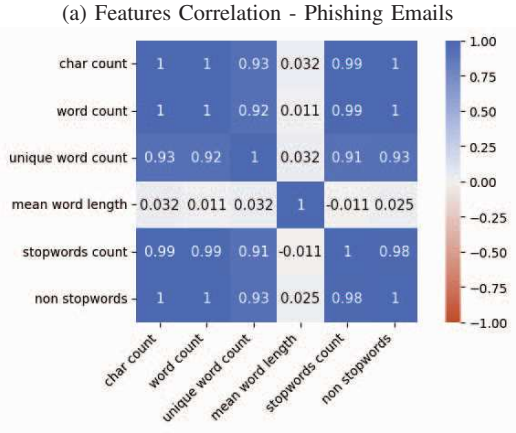
TABLE V: Latent Dirichlet Allocation (LDA) Emails Topic Modelling Results.

LDA Part A: Phishing Emails		
Topic #	10 Topic Words	Possible Categorization
1	'0.055*"ebay" + 0.041*"email" + 0.033*"message" + 0.020*"send" + 0.015*"account" + 0.013*"item" + 0.010*"user" + 0.010*"member" + 0.010*"policy" + 0.010*"question"'	Personal/Sensitive Information
2	'0.071*"account" + 0.030*"paypal" + 0.025*"email" + 0.019*"information" + 0.019*"update" + 0.015*"click" + 0.014*"security" + 0.012*"online" + 0.012*"bank" + 0.011*"access"'	Click and Update
3	'0.045*"bank" + 0.022*"service" + 0.020*"customer" + 0.017*"online" + 0.015*"form" + 0.013*"business" + 0.013*"email" + 0.011*"dear" + 0.010*"update" + 0.008*"mail"'	Update Information
LDA Part B: Non-Phishing Emails		
Topic #	10 Topic Words	Possible Categorization
1	'0.007*"group" + 0.006*"trade" + 0.005*"business" + 0.005*"report" + 0.005*"service" + 0.004*"work" + 0.004*"company" + 0.004*"meet" + 0.004*"market" + 0.004*"million"'	Trade/Business
2	'0.007*"send" + 0.006*"email" + 0.006*"message" + 0.006*"price" + 0.005*"need" + 0.005*"time" + 0.004*"information" + 0.004*"image" + 0.004*"trade" + 0.004*"think"'	Notification
3	'0.009*"price" + 0.006*"time" + 0.006*"email" + 0.005*"electricity" + 0.004*"send" + 0.004*"order" + 0.004*"plan" + 0.004*"meet" + 0.004*"work" + 0.004*"plant"'	Notification

non-phishing emails and performs a possible categorization of the topic words. Tables III, IV and V record the top 3 extracts per section i.e., the phishing and non-phishing sections. The possible categorization columns of Tables III, IV and V perform an annotation of the topic words to establish the possible themes of the topic words. The theme of "Personal/Sensitive Information" characterized by words i.e., "credit", "card" and "account" in relation to the extracted topic words is recorded in NMF part A: Phishing emails topic number 1, LSA part A: Phishing emails topic number 2 and LDA part A: Phishing emails topic number 1 of Tables III, IV and V. The themes of "Update your Personal Information" is observed in LSA part A: Phishing emails topic number 1 and LDA part A: Phishing emails topic number 3, equally a theme of "Click and Update" is observed in LDA part A: Phishing emails topic number 2. The study points out to a possible categorization of an "Online Service Request" theme, symbolized by words i.e., "online", "service", "form" in regards to the adjacent extracted topic words as recorded by NMF part A: Phishing emails topic number 2 and LSA part A: Phishing emails topic number 3



(a) Features Correlation - Phishing Emails

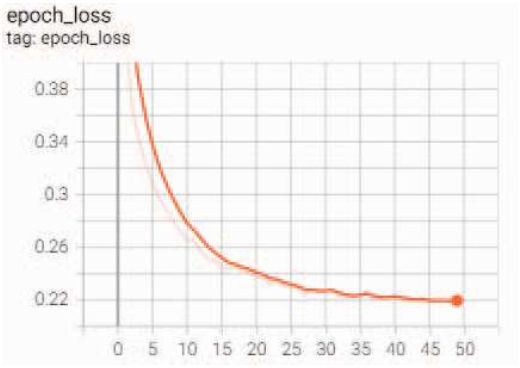


(b) Features Correlation - Non-Phishing Emails

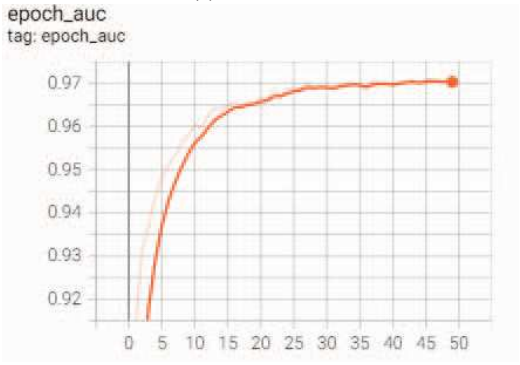
Fig. 2: Spearman Features Correlation Matrixes (Heat-maps) of Phishing vs Non-Phishing Emails.

of Tables III and IV.

The possible categorization of "Trade/Business" and the categorization theme of "Financial Information" which might be considered as a subset of "Trade/Business" appear to be dominant in the non-phishing emails topic words. This is evidenced by Table III, NMF part B: Non-phishing emails, topic number 1 and 3 and Table V, LDA part B: Non-phishing emails, topic number 1. Table IV, LSA part B: Non-phishing emails, topics number 1 and 2 are observed to record the possible categorized theme of "Financial Information". The Enron email dataset comprises a collection of real-word company emails and a theme of "Notification" is observed to be possible in Table III, NMF part B: Non-phishing emails, topic number 2 and Table V, LDA part B: Non-phishing emails, topics number 2 and 3. The non-phishing emails topic words also present a possible categorization of "System/Computer error" as recorded in Table IV, LSA part B: Non-phishing emails, topic number 3. This study finds that topic modelling appears to be giving some insights into distinguishing between phishing and non-phishing emails. However, topic modelling alone might not be sufficient for differentiating between phishing and non-phishing emails due to the ever-evolving nature of phishing attacks.



(a) Loss Curve



(b) Area Under the Curve

Fig. 3: Model Learning Curves.

TABLE VI: The BERT Transformer Model Classification Report.

	Precision	Recall	F1-Score
Phishing Emails	0.92	0.94	0.93
Non-Phishing Emails	0.94	0.92	0.93
Accuracy			0.93

4) BERT Model and Evaluation: The evaluation metrics of precision, recall, accuracy and the F1-score are used as the assessment metrics of the study's BERT model. The study trains its model for 50 epochs and the loss curve details the training process and the direction in which the model learned. Figure 3a, the loss curve, with its X-axis, denoted by values 0 to 50 implying the 50 epochs and the Y-axis recording values 0.22 to 0.38, shows that the loss of the study's trained model is almost always lower on the training than the validation and continued training could have led to over-fitting. The evaluation metrics of precision, recall, and F1-score are recorded in Table VI, with the phishing emails reporting a precision of 92%, recall of 94% and an F1-Score of 93%, while the non-phishing emails record an F1-Score of 93%, recall of 92% and a precision of 94%. The area under the curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Figure 3b, the area under the curve, records values 0.92 to 0.97 in its Y-axis in relation to the 50 epochs accounted for on its X-axis. The evaluation of the area under the curve for this study's model

implies the study's classifier can to some degree correctly distinguish between phishing and non-phishing emails with Table VI recording an accuracy of 93%.

VII. Conclusion and Future Work

This study investigates the classification of phishing emails with the objective being to develop a classification model that can classify phishing emails from non-phishing emails. The study examined the ideas of sentiment analysis, features correlations analysis as well as topic modelling and found out that they cannot conclusively be relied upon to classify phishing emails from non-phishing emails. However, the study noted that sentiment analysis, features correlations analysis and topic modelling offer valuable insights of the phishing and non-phishing emails. The study encounters a good number of machine learning-based classification models that can be adapted for classifying phishing emails and adds to the field of knowledge by presenting the possibility of developing effective and robust classifiers using machine learning-based transformers i.e., BERT an emerging language modeling technique.

Acknowledgement

This research work is supported by National Science Foundation (NSF) under Grant No: 1723765.

References

[1] L. N. Zainab Alkhali, Chaminda Hewage and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, 563060, 2021.

[2] V. Ramanathan and H. Wechsler, "phishgillnet—phishing detection methodology using probabilistic latent semantic analysis, adaboost, and co-training," *EURASIP Journal on Information Security*, vol. 2012, no. 1, pp. 1–22, 2012.

[3] M. Jakobsson and S. Myers, *Phishing and countermeasures: understanding the increasing problem of electronic identity theft*. John Wiley & Sons, 2006.

[4] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Computers & Security*, vol. 68, pp. 160–196, 2017.

[5] T. Stojnic, D. Vatsalan, and N. A. Arachchilage, "Phishing email strategies: understanding cybercriminals' strategies of crafting phishing emails," *Security and Privacy*, vol. 4, no. 5, p. e165, 2021.

[6] P. Maneriker, J. W. Stokes, E. G. Lazo, D. Carutasu, F. Tajaddodianfar, and A. Gururajan, "Urltran: Improving phishing url detection using transformers," in *MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM)*. IEEE, 2021, pp. 197–204.

[7] S. Loria, "textblob documentation," *Release 0.15*, vol. 2, p. 269, 2018.

[8] A. Alhogail and A. Alsabih, "Applying machine learning and natural language processing to detect phishing email," *Computers & Security*, vol. 110, p. 102414, 2021.

[9] B. B. Gupta, N. A. Arachchilage, and K. E. Psannis, "Defending against phishing attacks: taxonomy of methods, current issues and future directions," *Telecommunication Systems*, vol. 67, pp. 247–267, 2018.

[10] P. Bountakas and C. Xenakis, "Helphed: Hybrid ensemble learning phishing email detection," *Journal of Network and Computer Applications*, vol. 210, p. 103545, 2023.

[11] L. Halgaš, I. Agrafiotis, and J. R. Nurse, "Catching the phish: Detecting phishing attacks using recurrent neural networks (rnns)," in *Information Security Applications: 20th International Conference, WISA 2019, Jeju Island, South Korea, August 21–24, 2019, Revised Selected Papers 20*. Springer, 2020, pp. 219–233.

[12] M. Chatterjee and A.-S. Namin, "Detecting phishing websites through deep reinforcement learning," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2019, pp. 227–232.

[13] SpamAssassin, https://spamassassin.apache.org/old/publiccorpus/.

[14] J. Nazairo, "Phishing corpus," Accessed January 2022. [Online]. Available: https://monkey.org/~jose/phishing/

[15] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15*. Springer, 2004, pp. 217–226.

[16] M. Zareapoor and K. Seeja, "Feature extraction or feature selection for text classification: A case study on phishing email detection," *International Journal of Information Engineering and Electronic Business*, vol. 7, no. 2, p. 60, 2015.

[17] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7370–7377.

[18] P. A. Barraclough, M. A. Hossain, M. Tahir, G. Sexton, and N. Aslam, "Intelligent phishing detection and protection scheme for online transactions," *Expert Systems with Applications*, vol. 40, no. 11, pp. 4697–4706, 2013.

[19] S. Arya and S. Chamotra, "Multi layer detection framework for spear-phishing attacks," in *Information Systems Security: 17th International Conference, ICISS 2021, Patna, India, December 16–20, 2021, Proceedings 17*. Springer, 2021, pp. 38–56.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[21] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 788-791, 1999.

[22] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[23] S. Bergamaschi and L. Po, "Comparing lda and lsa topic models for content-based movie recommendation systems," in *Web Information Systems and Technologies*, V. Monfort and K.-H. Krempels, Eds. Cham: Springer International Publishing, 2015, pp. 247–263.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[25] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.