A Work Project, presented as part of the requirements for the Award of a Master's degree in Management from the Nova School of Business and Economics.

Guarding Corporate Frontiers: Empirical Evidence for Recognizing Phishing Behavior Patterns in the Financial Services and Healthcare Industries

Benedikt Wiederrecht

Work project carried out under the supervision of:

Rongjiao Ji

14/06/2024

Abstract

Phishing remains a significant cyber threat, particularly for the financial services and healthcare industries, due to their high exposure. This study explored phishing susceptibility and characteristics at the industry level using feature engineering, machine learning models, and topic models. The results revealed that the healthcare industry is more susceptible to phishing than the financial services industry. The most crucial phishing indicator was the number of total characters in an email, with shorter emails indicating phishing. The number of call-to-action words in phishing emails is higher in the financial sector and fewer in the healthcare sector compared to safe emails.

Keywords

Phishing, Big Data, Feature Extraction, Logistic Regression, Machine Learning

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Contents

1 Introduction

Phishing emails are the number one most effective form of cyberattacks. In phishing attacks, cybercriminals, also known as hackers, trick individuals into revealing sensitive information or enabling malicious software by posing as trustworthy entities (Rizzoni et al. 2022). According to a study conducted by Deloitte, 91 % of cyber-attacks involve emails as the first touching point with the victim (Ho 2020). Human emotions are manipulated to gain access to sensitive data, and the spammers' strategies continue to evolve to exploit new vulnerabilities (Jáñez-Martino et al. 2023). This dynamic environment makes it challenging to develop a long-term strategy to counter phishing attacks. Consequently, there is a need for ongoing adaptation and anti-phishing training for employees (Jampen et al. 2020).

Phishing is a particular problem for businesses, as it has led to the loss of sensitive information, financial losses, and damaged reputations (Tan, Chiew, and Wong 2016). The current literature examined various phishing characteristics in emails and the susceptibility of different industries. Phishing characteristics are specific traits that identify potential phishing attempts in emails, including character lengths, URL elements, and call-to-action words. However, whether the characteristics used to detect phishing differ from industry to industry is still unclear. By looking at industry-specific characteristics, it is possible to identify patterns that are specific to that industry. Countermeasures should not be a one-size-fits-all approach but should be tailored to the organization's industry. This thesis aims to provide employees with a concise set of characteristics that can be used to identify phishing emails, depending on the industry in which they work.

Two industries, in particular, stand out: the financial services industry, which is a popular target for attacks due to a large number of transactions and highly sensitive information, and the healthcare industry, which is highly susceptible to phishing due to high click-rates (Gordon et al. 2019). This thesis, therefore, focuses on evidence from these two industries. Investigating phishing characteristics in the financial services industry and healthcare industry raises a number of questions that are addressed in the following:

RQ1. How vulnerable are the financial services and healthcare industries to phishing? The study measures the number of phishing emails compared to non-phishing emails in each

industry. Additionally, it seeks to examine the susceptibility of financial services and healthcare workers to phishing emails. This enables the identification of industries that require action and potentially highlights measures that have already had an effect. Examples of such measures include regulation and training.

RQ2. What characteristics of emails can employees working in the financial services and healthcare industries use to detect phishing? The objective is to create an understanding for employees of how phishing characteristics, such as the number of call-to-action words or the length of the email, affect the likelihood that an email is phishing. Many call-to-action words may indicate phishing in the financial services industry. However, this may not be the case in the healthcare industry.

RQ3. Which characteristics are most critical for phishing detection? This thesis aims to enable employees in the financial services and healthcare industries to assess the importance of individual phishing characteristics in emails effectively. For example, analyzing the frequency of call-to-action words may offer more valuable insights than considering the average word length in the email, greatly aiding employees in identifying phishing attempts.

RQ4. How do email topics predict phishing likelihood in the financial and healthcare sectors? Email topics, unlike the immediately evident characteristics in emails, require contextual understanding. These subjects frequently offer subtle hints regarding the legitimacy of the email. A highly sophisticated email on academic research in the healthcare industry may indicate a legitimate email, while emails containing more personal healthcare decisions can indicate phishing. The goal is to develop industry-specific awareness of email topics in both phishing and non-phishing communications.

The paper is structured as follows: Section 2 presents the current state-of-the-art research on the characteristics of phishing and non-phishing emails. Section 3 explains the methodology, including the dataset, extracted features, evaluation models, and a topic model. Section 4 presents the results of industry-specific characteristics in phishing and non-phishing emails and discusses which characteristics are most critical for determining phishing. Further, the email topics in phishing and non-phishing are discussed. Section 5 addresses limitations, and Section 6 concludes.

2 Literature Review

Besides this work, others also examined phishing tactics on an industry level. This section presents an analysis of the current state of research on phishing, focusing on papers that have addressed characteristics and methodologies similar to this paper. A comprehensive review of research on the following topics is conducted: (i) phishing in the financial services and healthcare industry, (ii) length-based characteristics of emails, (iii) behavioral cues such as call-to-action and urgency words in phishing, (iv) phishing detection techniques and (v) topic models.

Industry Analysis. Exploring industry-specific vulnerabilities, recent studies have highlighted significant differences in how the financial services and healthcare sectors respond to phishing emails. The study by Lagazio et al. (2014) analyzed cybersecurity and phishing in the financial sector and found that weak policing and weak international frameworks can increase the cost of cybercrime. Further expanding on industry differences, Tian, Jensen, and Durcikova (2023) compared phishing tactics in the financial services industry to non-financial industries. The authors conducted mock phishing attacks on 30 financial and 15 non-financial organizations and show that vulnerability varies by industry; for example, authority-based tactics are more successful for financial services organizations, and liking-based techniques are more advantageous for non-financial services organizations.

In addition to the financial services industry, a review of the literature on phishing in the healthcare industry showed how susceptible this industry is to phishing. Priestman et al. (2019) investigated phishing in healthcare organizations and highlighted that organizations are increasingly implementing digital systems, but healthcare workers lack digital literacy and awareness to detect phishing. Thus, Jalali et al. (2020), who examined why hospital employees click on phishing links so often, come to similar conclusions. According to the authors, tailored training programs should involve organizational factors, such as trust-building activities, to reduce the impact of phishing attacks in the healthcare industry.

However, industry-specific phishing research is limited because phishing email datasets are often not labeled by industry. Rule-based classification systems, which use predefined logical rules to categorize data, can help to assign email to an industry. This approach provides a high level of control and transparency. Prominent examples are Microsoft Outlook's email

rules feature and Apache SpamAssassin for spam filtering (ChatableApps 2024). In addition, Muhammad Zubair Asghar et al. (2017) also used a rule-based classification method for sentiment analysis because traditional methods were less efficient in handling emoticons, modifiers, and domain-specific words. Their lexicon-enhanced approach significantly improved sentiment analysis performance (Asghar et al. 2017).

Length-Based Characteristics. Fang et al. (2019) employed a Recurrent Convolutional Neural Network (RCNN) to construct a phishing email detection model, THEMIS, which integrated character-level and word-level representations of email headers and bodies. THEMIS reached an accuracy of 99.848 %, which further highlights the effectiveness of including length-based features in phishing detection models. Doshi et al. (2023), who employed a dual-layer architecture using character-level and word-level features, came to similar conclusions. Both character-level features and word-level features improved the model's ability to detect malicious emails, and their model achieves an accuracy of 99.51 %.

Behavioral Cues. Previous research has been conducted on urgency and call-to-action words in phishing attempts. Stojnic, Vatsalan, and Arachchilage (2021) discussed the strategic use of urgency and call-to-action words such as "please," "help," and "reward" in phishing attacks. They found that conveying urgency and call-to-action can lead to significant phishing success. Similarly, Alkhalil et al. (2021) examined specific wording that conveys a sense of urgency, such as "important" and found that urgency terms are an effective method to encourage victims to respond to attacks.

Phishing Detection Techniques. There is a growing trend of analyzing phishing attacks using machine learning models. Deshpande et al. (2021) investigated the extent to which machine learning techniques are suitable for detecting phishing websites. The authors analyzed characteristics such as URL, keywords and characters and found that machine learning algorithms are very well suited to detect phishing. Their Random Forest model detected phishing websites with an accuracy of 97.31 %. Furthermore, they emphasized the importance of feature selection, which is also of great importance in this thesis. Rao et al. (2023) investigated a hybrid ensemble framework to detect social spam on imbalanced datasets by applying different word embedding techniques such as deep learning and machine learning techniques. They found deep learning

models to be faster and achieve better results compared to machine learning approaches. Another application of machine learning in spam detection systems was investigated by Jáñez-Martino et al. (2023). Their study dealt with evolving strategies of spammers and encourages to standardize the criteria for evaluating anti-spam filters using machine learning.

Topic Models. Stojnic, Vatsalan, and Arachchilage (2021), previously discussed for call-to-action and urgency words, successfully used natural language processing (NLP), topic modeling, and clustering techniques to analyze fraudulent emails. The models provided a structured way to examine commonly used terms and identify commonly used strategies. Further research on phishing utilizing topic models and natural language processing models in phishing, however, is limited. An et al. (2023) employed topic modeling techniques to derive marketing insights from online product reviews. They utilized clustering-based algorithms, such as the BERTopic language model, to analyze product advantages and disadvantages. The researchers find BERTopic to outperform traditional algorithms.

In summary, the current state of research on phishing attempt detection shows the following: The features industry, length characteristics, and behavioral cues appear to be highly relevant in classifying phishing and non-phishing emails. Furthermore, machine learning models have proven to be effective methods for detecting malicious emails, and language models have contributed to a better understanding of strategies in phishing emails.

3 Methodology

3.1 Dataset and Data Preprocessing

To study phishing tactics for the financial services and healthcare industries, I used the phishing dataset available on Kaggle.com and curated by SubhaJournal from 2023 (Chakraborty 2023). This dataset consists of 18650 emails, of which 11322 (60.7 %) are classified as safe and 7328 (39.3 %) as phishing. The dataset is updated annually to ensure it is current and relevant. It contains a large number of detailed emails, including sender details, keywords, and URLs, making it an excellent primary database for my phishing research.

Data preprocessing is necessary to ensure unbiased and correct analysis results. I performed

the following four steps. First, non-textual emails have no analytical value and can lead to incorrect predictive models due to their lack of content. Thus, I removed all rows that have a blank email body, which is 16. I also removed all rows containing the string ”empty”, leaving a total of 18101 emails. Second, duplicate rows lead to incorrect weighting of some of the emails, which can bias the models. Hence, I made sure that each email only appeared once in the dataset. The dataset did not contain any duplicates, ensuring the correct weight of each email. Third, both machine learning and logistic regression models require numeric input data. I converted the label ”email type” to numeric, where phishing is labeled 1 and safe emails are labeled 0. Finally, standardising text data helped to reduce model complexity and improve the robustness of natural language processing tasks. I cleaned up the email body by removing anything that negatively affects the natural language processing models. Specifically, I converted all characters to lower case so that words that are capitalised are treated the same as words that are not. Numbers were removed, which prevents models from drawing incorrect conclusions from numerical data instead of focusing on the actual text of the email. I also replaced multiple spaces with single spaces, which was crucial for accurate tokenisation. Last, I removed punctuation, reducing the number of non-alphanumeric characters, to steer the models towards analysing substantive text elements.

3.2 Feature Engineering

The feature engineering section consists of three parts: (i) the industry feature, which was derived through vocabulary analysis; (ii) length-based features; and (iii) behavioral cues features. These features were combined with the original phishing dataset to form the *phishing industry dataframe*, which serves as the baseline for the investigations conducted throughout this research. For a more comprehensive understanding of the dataset, please refer to Figure 1. The following section provides a detailed description of each feature.

Industry Feature: The objective was to identify differences across various industries rather than focusing solely on a specific sector. As the dataset lacks email industry labels, the goal was to categorize each email by its relevant industry. For instance, an email targeting a bank account manager would be identified as financial services phishing rather than generically labeled as

phishing. This approach enabled more targeted and industry-precise recommendations for countermeasures. This work used a rule-based classification method. First, a list of industry vocabulary was created for both the financial services industry and the healthcare industry. For financial services terms, words were extracted from the CFA Institute's programme glossary (CFA Institute 2024), as this is one of the most trusted resources for financial services-specific terminology. Healthcare terms were taken from the Harvard Medical Dictionary (Harvard Health Publishing 2024), known for its comprehensive and complete medical definitions. For a higher success rate, words from compound terms were counted separately.

chartered financial analysts
3085 words

Secondly, industry terms were counted for each email and recorded in a separate column. The rule for assigning an industry was set as follows: If an email contains five or more terms from either the financial industry terms list or the healthcare terms list, the email is tagged according to the industry. If the email contains less than five terms from either list, it is labeled "None." The threshold of five was determined through cross-validation by manually validating a proportion of email industry types and then conducting multiple rounds with different thresholds to identify the industry most accurately. The cross-validation process yielded a threshold of five as the optimal value. Nevertheless, the possibility of encountering comparable word counts in emails within the healthcare and financial services industries persists. Consequently, if the discrepancy between the number of words from the financial terms list and the healthcare terms list was less than three, the email was designated as "Undecided."

Table 1: Absolute Number of Emails Classified by Industry

Industry	Non-Phishing	Phishing	**Total**
Financial Services	499	750	**1249**
Healthcare	804	476	**1280**
Undecided	148	102	**250**
None	9673	5649	**15322**
Total	**11124**	**6977**	**18101**

Table 1 shows the number of emails identified as belonging to either the Financial Services or the Healthcare industry by the rules-based classification system. It also shows the number of emails identified as "Undecided" and "None". Unsurprisingly, the "None" category is by far the largest, with 15322 emails, as for most of the emails no industry could be identified or the email

belongs to a different industry, such as the technology industry, which is not investigated in this thesis. 63 % of emails in the ”None” category are non-phishing and 36.7 % are phishing, which, compared to the overall 61 % of phishing emails in the whole dataset, indicates a decrease in non-phishing emails. There are very different proportions of phishing and non-phishing emails in the two industries analysed. Almost 1300 emails could be assigned to each industry. However, 50 % more phishing emails than non-phishing emails were observed in the financial services industry. This is the opposite for the healthcare industry, where almost twice as many emails are classified as safe as phishing.

Length-Based Features. In the second step, I examined length-based features to make statements about how the semantic characteristics of emails affect the classification of phishing and non-phishing emails. Specifically, I examined each email's total characters, total words, and average word length.

Table 2: Median Number of Characters in Non-Phishing and Phishing Emails

Category	Non-Phishing Median Chars	Phishing Median Chars
Financial Services	3290.0	3105.0
Healthcare	4967.0	1570.0
Undecided	5183.5	5179.5
None	746.0	552.0

Table 2 shows the statistics. Safe emails from the financial services industry (3290 characters) were slightly longer on average than phishing emails (3105 characters). There was a much larger difference for emails from the healthcare industry, where the median character length of phishing emails (4967 characters) was more than three times that of non-phishing emails (1570 characters). The median number of characters in phishing emails was also higher for emails that could not be attributed to any industry. It is important to note that the median number of characters in the ”None” category could not be compared to the industries due to a bias. Longer emails were more likely to be assigned to an industry by the rule-based classification system than shorter emails.

For the median of the average word length per industry, Table 3 shows the results. The median presented itself around five characters regardless of the industry. The median was slightly lower for phishing emails in all categories except healthcare, but there were no significant differences.

Table 3: Median Average Word Length in Non-Phishing and Phishing Emails

Category	Non-Phishing Median Length	Phishing Median Length
Financial Services	4.89	5.00
Healthcare	5.23	5.10
Undecided	5.11	4.64
None	4.93	4.76

Behavioral Cues Features. The third and final feature deals with the behavioral language of the sender. Specifically, the occurrence of call-to-action words such as ”immediately” or ”emergency” were examined, in order to infer some sort of urgency and need for action. The aim was to determine whether pressure from the sender through call-to-action language is a widespread method in phishing and non-phishing emails. Similar to the feature engineering part, I created a list of call-to-action words and implemented a counter for each email. The call-to-action words were extracted from Metacompliance (Cubicle Ninjas 2024), a reputable source known for its comprehensive cybersecurity resources. This ensured the list was relevant and reliable for detecting potential phishing attempts. I added the call-to-action words in each email and the number of words to the phishing dataset, a new column for each. One would expect to see more call-to-action words in phishing emails. I also looked at how the use of call-to-action words differs between the financial services and healthcare industries.

Table 4: Average Number of Urgent Words in Non-Phishing and Phishing Emails

Email Type / Industry	Financial Services	Healthcare
Both Email Types	14.12	12.26
Non-Phishing	13.60	13.37
Phishing	14.46	10.39

The breakdown of the average number of urgent Words can be found in Table 4. For emails from the financial services industry, the number of urgent words in phishing emails (14.46) was higher than in non-phishing emails (13.6). This differs from the results for the healthcare industry, where the average number of urgent words in non-phishing emails (13.37) was higher than in phishing emails (10.39). The results section tests the differences for significance and their meaning is discussed.

The extracted features, in combination with the email text and email type, result in the

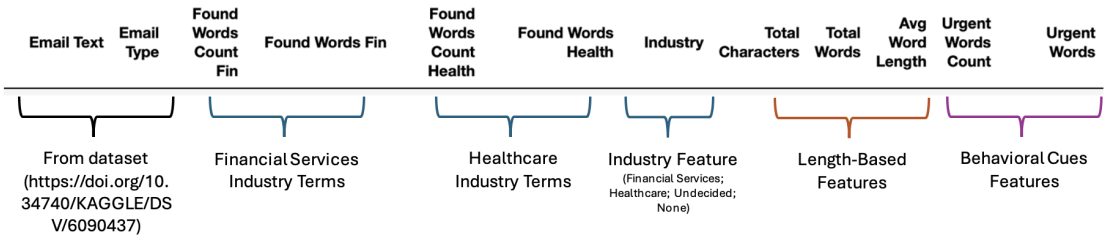

Figure 1: All Columns of the Phishing Industry Dataframe

phishing-industry data frame, depicted in Figure 1, which serves as the foundation for the analyses presented in this paper.

3.3 Evaluation Models

Following the phishing feature extraction methodology, it is important to select analytical models that evaluate the importance of each phishing feature. In this way, emphasis can be placed on the most important features that significantly impact phishing. This section examines two different models, Logistic Regression and Random Forest, and allows for comparative analysis to determine which features are consistently significant across different modeling approaches.

Model Selection. The performance metrics summary, shown in Table 12 (see Appendix 7.3), provides a comparative overview of different machine learning models tested on the phishing industry dataset. It includes four crucial indicators, such as accuracy, precision, recall, and F1 score, highlighting the models' ability to detect phishing emails effectively. Logistic Regression and Random Forest were chosen when selecting models for analysis due to their transparency and interpretability. Given the challenges in conducting training courses on phishing in the healthcare industry, where backlash from staff and unions can occur (Rizzoni et al. 2022), these characteristics are crucial since they improve comprehension and acceptance of model outcomes. Thus, Random Forest and Logistic Regression are particularly suitable choices, even though Logistic Regression only performed with a medium accuracy of 0.64. On the other hand, Random Forest performed with an accuracy of 0.73. Although Gradient Boosting had the same accuracy as Random Forest, this model would require more careful tuning, which can result in models that are less interpretable (Manning 2017).

Logistic Regression Analysis. Logistic regression is a well-established statistical method

with a history of successful application in a variety of classification problems across different fields (Hosmer Jr, Lemeshow, and Sturdivant 2013). This study makes use of these properties when classifying emails as phishing or non-phishing and investigates the significance of potential explanatory variables. The effect of each feature (industry, length characteristics, and behavioral cues) on the probability of phishing or non-phishing is analyzed.

Prior to establishing the equation, it is necessary to consider the issue of multicollinearity, as multicollinearity threatens the bias and interpretability of coefficients in logistic regression models (Allison 2012). The variance inflation factor (VIF) for all feature variables measures the extent to which the variance of a regression coefficient increases due to multicollinearity in the model (O'brien 2007). Table 14 in the Appendix shows the VIF results for all features. It can be observed that the variables *Total Words* and *total characters* have high VIF scores, indicating that they are strongly correlated. Consequently, *total words* was removed from the model as it is less informative than *total characters*. Furthermore, *found words count fin* and *found words count health* also show a moderate to high VIF value, possibly due to their strongly association with the industry variables. Both variables were excluded from the model. The logistic regression model to test the significance of the features, which are presented in the Results section, was set up as follows:

$$
\begin{aligned}
(1)\ EmailType^{1} = {} & \beta_0 + \beta_1 IndustryFinancialServices + \beta_2 IndustryHealthcare \\
& + \beta_3 IndustryUndecided + \beta_4 TotalCharacters \\
& + \beta_5 AvgWordLength + \beta_6 UrgentWordsCount + e
\end{aligned}
$$

Random Forest Classifier. Random Forest is a machine learning model suitable for classification problems (Akinyelu and Adewumi 2014). Ziegler & König (2013) further emphasized the versatile application of Random Forests in real-world scenarios due to its speed, flexibility, and robustness. The following steps were implemented to obtain a robust and informative Random Forest model. First, continuous numerical features, such as total characters, average word length, and urgent word count, were converted into four categorical bins: low, medium, high,

1. 1=Phishing, 0=Non-Phishing

and very high. In contrast to logistic regression, which allows for straightforward interpretation of continuous variables due to the linear correlations between explanatory and dependent variables, Random Forest benefits from converting continuous variables into categorical bins. This is due to the nature of Random Forest operations, where continuous variables possess numerous potential split points, complicating the analysis of feature importance across different types of variables. Secondly, the dataset was split into two distinct sets: features and the target variable *email type*. These two sets were then divided into training and testing sets. Standardization was employed to ensure that the features were comparable, and thus, the Random Forest model was trained on the scaled training data.

Finally, model performance was evaluated using accuracy and a detailed classification report, including precision, recall, and F1-score. A feature importance analysis was conducted to identify the most influential features in the classification process, providing insights into key indicators of phishing emails and possibly comparing them with the coefficients of the logistic regression analysis. This approach to data preprocessing, model training, and evaluation ensured that the Random Forest classifier effectively enhances phishing detection, yielding robust and interpretable results.

3.4 Email Topic Analysis

Phishing characteristics such as the length of the email or call-to-action words are easy to spot and can be recognized and classified fairly quickly when first skimming the email. However, analyzing the topics covered requires a more thorough understanding, yet they may serve as similarly good indicators of phishing. Email topics refer to the main subjects or themes discussed within the email content, such as education, pharmaceuticals, or financial transactions. The aim is to sensitize employees in the financial services and healthcare industry to email topics that indicate phishing and non-phishing.

To explore the topics discussed in emails, I used BERTopic to analyze the email text. BERTopic uses state-of-the-art language models, such as BERT, to capture contextual information more efficiently than traditional models (Grootendorst 2022). When applying BERTopic, emails were converted into high-dimensional embeddings, representing text forms as vectors. The

dimensions were then reduced to allow clustering. The clustering method was based on the density of the documents' points in the embedded space, with each cluster representing a possible topic while retaining important words in topic descriptions (Samsir et al. 2023).

In this study, I used BERTopic to identify different topics automatically. In order to optimize the topic model, I experimented with a limit of 5 to 10 topics and evaluated them based on the distinctiveness and meaningfulness of the topics. For topic models to be optimized, distinctiveness and meaningfulness are essential because they guarantee that the topics that the identified topics are relevant and clear and offer insights that can be put to use. For a sample of 1000 emails from the entire dataset, limiting to a maximum of 8 topics resulted in the most meaningful and distinctive topics. I applied BERTopic to sub-sets of the phishing data, such as the dataset only containing financial services emails or the dataset only containing healthcare emails. For the topics generated by BERTopic, I generated labels. Table 15 in the Appendix shows the detected topics in the full dataset. The topic names were based on the top words per category, as shown in Figure 4.

4 Results and Discussion

4.1 Feature Analysis

Industry Feature. The results of Table 1 on the absolute number of emails classified by industry show that the financial services industry encountered more phishing emails than the healthcare industry. This prompts the investigation of independence between email type and industry type. If there is a significant discrepancy in the number of phishing emails per industry, it would underline the necessity for countermeasures to be tailored to the industry. This is due to unique vulnerabilities and attack methods within each sector. Therefore, I tested the following hypothesis based on Table 1:

(2) H_0 : Industry Type and Email Type are independent

Table 5 shows the results of the χ^2 test for independence from hypothesis (2). With a p-value close to 0, indicating significant results at a very low level, it can be concluded that the industry of the email is not independent of its type. Importantly, this demonstrates a meaningful difference

Table 5: χ^2 Test of Independence Results for Industry and Email Type

Metric	Value
χ^2 Statistic	263.514
p-value	0.000
Degrees of Freedom	3
Significant at $\alpha = 0.05$	True

in the number of phishing emails an employee is likely to encounter, depending on whether they are employed in the financial services or healthcare industry. Employees in the financial services industry are exposed to significant more phishing emails than healthcare workers.

Although employees in the financial services industry find themselves more frequently exposed to deceptive emails, it cannot be concluded that the financial sector suffers greater damage as a result. A better indicator of damage done is the susceptibility to phishing. The "2023 Phishing by Industry Benchmarking Report" looked at differences in phishing behavior between industries (KnowBe4 2023). They conducted simulated phishing attacks across multiple industries and measured employee susceptibility through the Phish-Prone Percentage (PPP), quantifying the likelihood of an employee responding to a phishing attempt. The results show that Banking, the main sector of the financial services industry, is one of the most targeted sectors but has a PPP score of only 3 %. The reason is high regulations and good employee training. In contrast, while also a frequent target, the healthcare industry experiences phishing attempts at a significantly lower rate than the financial sector. However, with a PPP score of 32.3 %, the healthcare industry is more vulnerable, indicating lower training standards or fewer preventative measures.

The findings thus lead to the conclusion that the financial services industry is better equipped to avoid falling for phishing emails despite a significantly higher number of phishing emails. It suggests that training and regulation are successful. In contrast, the healthcare industry, which receives fewer phishing emails, is highly vulnerable, so training to recognize phishing is necessary.

Length-based Features. Having shown that the healthcare industry needs to improve its ability to detect phishing emails, it is necessary to determine which phishing characteristics can be used to identify phishing. More specifically, what email characteristics can employees in the

financial services and healthcare industries spot to determine whether an email is secure?

To assess this, I first look at length-based features. The number of total characters in an email may indicate whether the email is phishing. Figure 2 shows the differences in median characters for the full dataset, financial services, and healthcare industries. As described in Section 3.2, there are differences between the median character lengths of phishing and non-phishing in each industry. To test whether these differences in median character length are significant, I performed a Mann-Whitney U test between phishing and non-phishing, shown in Section 7.1 of the Appendix. This test is crucial to determine if the observed variations in email length can reliably distinguish between phishing and legitimate emails. The results show that the median number of characters for phishing emails is significantly lower than for non-phishing emails across all datasets.

Cormack et al. (2007) shed light on why phishing emails are often shorter than safe emails. Traditional bag-of-words-based spam filters have difficulties recognizing relevant spam characteristics in shorter emails, so short emails often bypass spam filters. Aware of this vulnerability, hackers exploit it to enhance the effectiveness of their phishing attempts.

Consequently, the results point to this conclusion: short emails indicate phishing. Especially in the healthcare industry, the difference in total characters between phishing and safe emails is particularly large. When faced with a suspicious email that may be phishing, healthcare workers, and to some extent, financial services workers, should first check the length of the email and be skeptical if it is short. It could be a tactic to bypass the spam filter.[2]

Behavioral Cues Features. Besides paying attention to short emails, behavioral cues in emails, such as call-to-action or urgency words, can also indicate phishing. Figure 3 shows the average number of urgent words per email type across the full dataset, the financial services emails-only dataset, and the healthcare-only emails dataset. Mann-Whitney-U tests were performed to determine whether the difference between the number of average urgent words differed between phishing and non-phishing; see Section 7.2 in the Appendix. The tests were essential to statistically validate whether urgent words are consistent indicators of phishing across

2. Other length-based features, such as the average word length and the total number of words, are not being discussed. The average word length is not an effective indicator for identifying phishing emails; see Section 4.2. The total number of words strongly correlates with the number of total characters; see Section 7.5, which suggests that it is sufficient to discuss the total characters.

different industries. The results indicate that the average number of urgent words in the financial services dataset is significantly higher for phishing emails, while in comparison, the number is significantly lower in the healthcare dataset.

The findings are notable because, although Stojnic, Vatsalan, and Arachchilage (2021) and Alkhalil et al. (2021) discovered that call-to-action and urgency words are effective tactics for hackers to get victims to respond, these types of words were significantly less prevalent in phishing emails compared to secure emails within the healthcare industry. One possible explanation for this is that healthcare emails are often urgent. Email correspondence with the healthcare sector often concerns urgent medical issues that require immediate care and action, resulting in the high usage of call-to-action words.

The research on behavioral cues indicates that financial industry phishing emails tend to contain more urgent words than safe emails. Therefore, financial services employees should be alert to an unusually high number of urgent or call-to-action words in emails. The opposite is true for healthcare workers, where urgent words are more common in safe emails due to the frequent need for immediate attention and action. Therefore, healthcare workers should be alert if an unusually low number of urgent or call-to-action words in emails is observed.

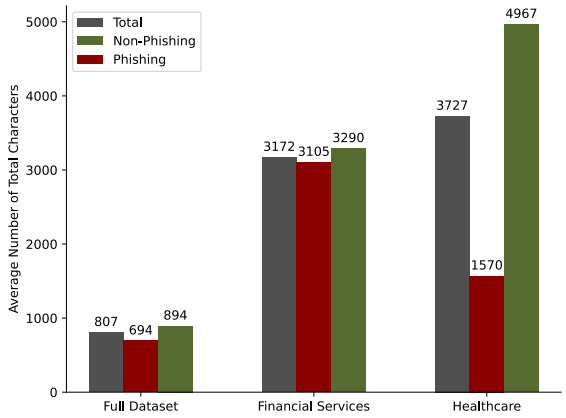

Figure 2: Breakdown of Median Characters by Dataset and Email Type

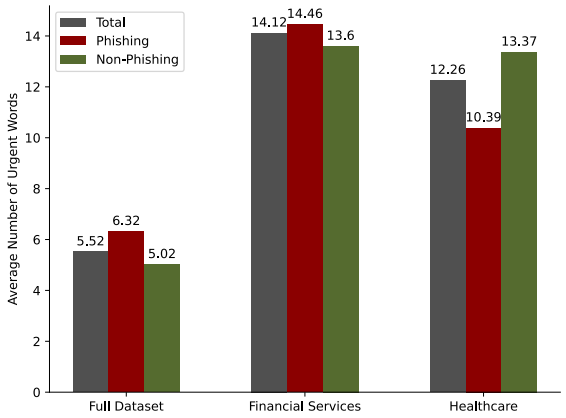

Figure 3: Breakdown of Average Urgent Words by Dataset and Email Type

4.2 Evaluation Models

This study has thus far examined various characteristics of phishing, including industry-specific differences in susceptibility to phishing and email characteristics that indicate phishing. The

question arises as to which characteristics are most crucial for phishing detection. Therefore, the characteristics are examined and ranked using Logistic Regression and Random Forest to determine the most influential features in predicting the email type.

Logistic Regression Analysis. Table 6 presents the results of the Logistic Regression model. All variables, including total characters, the number of urgent words, and industry types, except for average word length, were found to be significant at the 1 % level. This indicates that the features are effective predictors of email type. Average word length was found to be an inadequate indicator of email type, meaning it should be disregarded when scanning an email for phishing indicators. Furthermore, when examining the industry variables, those employed in the financial sector are at a highly increased risk of receiving phishing emails. The same applies to healthcare workers, the effect, however, is not as pronounced as in the financial services industry. This suggests that the two industries are particularly attractive to hackers. In addition, the number of total characters has a negative effect on the email type. Each additional character in the email decreases the likelihood that the email is a phishing email. Finally, as expected, the number of average urgent words also increases the likelihood of the email being phishing.

Random Forest Analysis. The Random Forest model's feature importance ranking provides insights into the characteristics and their explanatory value in classifying emails as phishing. The total characters of an email, categorized as *total characters very high*, *total characters high*, *total characters medium*, and *total characters low*, emerge as a crucial factor. *total characters very high* is identified as the most critical factor, with an importance score of 0.134. This indicates that phishing emails often differ significantly in character count from legitimate ones, particularly tending towards higher character counts. Furthermore, the level of urgency indicated by the call-to-action words used, captured by the variables *urgent words count very high*, *urgent words count high*, *urgent words count medium*, and *urgent words count low*, proves to be a significant factor. The model demonstrates that the frequency of urgent words by the sender significantly influences the classification of email types, especially when these words are present at very high or very low frequencies. Interestingly, variations in average word length, from *avg word length high* to *avg word length low*, and the presence of a specific industry identified, such as *industry financial services* and *industry healthcare*, offer additional, although less impactful,

cues. It can be stated that the random forest model was less successful in deriving insights from the average word length and the industries regarding email classification.

Although the accuracy of the random forest model is of secondary importance, given the objective of identifying the most influential features, the 73 % accuracy rate of the random forest model can still be considered noteworthy. The majority of the emails could be correctly identified. Table 13 in the Appendix shows the classification report for the Random Forest model. Fang et al. (2019) achieved an accuracy of 99.848 % for their phishing detection model THEMIS through an advanced feature extraction method that includes multilevel vectors to represent emails. This method also analyzes emails at the character level, word level, email header, and body simultaneously, thus enabling more accurate determination of phishing emails. In the future, incorporating their multilevel analysis techniques can enhance feature selection and model accuracy in broader phishing detection efforts.

The results of logistic regression and random forest models in phishing email detection indicate similarities and differences in their insights. Both models identified the number of total characters and the urgency of words as significant predictors of phishing. However, they differ in their emphasis on other features. While logistic regression points to the importance of industry-specific risks, highlighting increased vulnerabilities in the financial and healthcare sectors, the random forest model suggests that these industry factors are less impactful. Furthermore, the logistic regression model finds that the average word length is not a significant predictor, a finding that is underlined by the Random Forest's lesser emphasis on variations in word length. These discrepancies highlight the importance of leveraging multiple modeling techniques to understand the factors influencing phishing detection comprehensively.

The findings highlight the following points: the length of an email is the most significant indicator when scanning for suspicious content. In general, safe emails are longer, thus examining total characters of an email should be prioritized. The second most crucial factor is the number of call-to-action words, which vary depending on the industry. For further insight, refer to the results presented in Section 4.1. Last, the average word length should be disregarded when determining whether an email is phishing.

Table 6: Logistic Regression Results

	Dependent variable:
	Email.Type
IndustryFinancialServices	0.93^{***} (0.07)
IndustryHealthcare	0.19^{***} (0.07)
IndustryUndecided	0.38^{**} (0.15)
TotalCharacters	-0.0004^{***} (0.00)
AvgWordLength	-0.002 (0.01)
UrgentWordsCount	0.16^{***} (0.01)
Constant	-0.87^{***} (0.05)
Observations	18,101
Log Likelihood	$-10{,}961.38$
Akaike Inf. Crit.	21,936.75
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Random Forest Model: Feature Importance Ranking

Rank	Feature	Importance
1	Total_Characters_VeryHigh	0.134
2	Urgent_Words_Count_VeryHigh	0.124
3	Urgent_Words_Count_Low	0.088
4	Total_Characters_High	0.079
5	Urgent_Words_Count_High	0.069
6	Total_Characters_Medium	0.069
7	Urgent_Words_Count_Medium	0.060
8	Industry_None	0.058
9	Avg_Word_Length_Low	0.050
10	Total_Characters_Low	0.049
11	Avg_Word_Length_VeryHigh	0.048
12	Industry_Financial Services	0.048
13	Industry_Healthcare	0.046
14	Avg_Word_Length_Medium	0.036
15	Avg_Word_Length_High	0.033
16	Industry_Undecided	0.009

4.3 Email Topic Analysis

The previously discussed phishing characteristics help to recognize indications of phishing at the first glance at an email. Leveraging email topics to detect suspicious emails, on the other hand, requires a more detailed consideration. The aim is to provide comprehensive information on email topics that may indicate phishing. Exploring various topics across the financial services and healthcare datasets using BERTopic has produced interesting insights. Considering the differences between phishing and non-phishing topics, employees will be able to make more informed decisions regarding the safety of an email by understanding the thematic focus of phishing efforts.

Table 8: Topic Distribution for Financial Services

Category	**Topics**
Common	Corporate Affairs, Digital & Technology, Global Finance
Only Phishing	Real Estate & Loans, Financial Transactions, Energy Sector, Public Funding, Personal Planning
Only Non-Phishing	State Governance, Linguistics & Education, Media & Communication

Table 8 shows the topic distribution for financial services emails, categorizing them into common, phishing-only, and non-phishing-only topics. Topics shared across both legitimate and phishing emails include *Corporate Affairs*, *Digital & Technology*, and *Global Finance*, indicating overlap in email topics that are typically very relevant in financial contexts. This

overlap implies that hackers strategically copy frequently occurring topics in emails from the financial services industry to increase the legitimacy of their phishing attempts.

The most common topics in non-phishing emails in the financial services industry are *State Governance*, *Linguistics & Education*, and *Media & Communication*. These topics often involve content that shapes public opinion and spreads information. This suggests that non-commercial and educational emails are indicators of legitimacy. One example is an email invitation for an upcoming webinar on financial literacy. Conversely, the most common topics in phishing emails in the financial services industry, such as *Real Estate & Loans*, *Financial Transactions*, and *Personal Planning*, focus on transactions and personal financial dealings. These areas are susceptible to exploitation due to their financial nature. An email asking to verify your account for an upcoming transaction might indicate phishing. Analyzing email topics in the financial services industry has revealed that emails centered around financial transactions and personal financial planning are more likely to be phishing attempts. In contrast, those discussing non-monetary and educational topics tend to be safer.

Table 9: Topic Distribution for Healthcare

Category	**Topics**
Common	Media & Communication
Only Phishing	Pharmaceuticals, Corporate Affairs, Health & Wellness, Geopolitical Entities, Aging & Fitness, Disease Prevention, Weight Management
Only Non-Phishing	Academic Research, Linguistics & Education, Media & Communication

In the financial services industry, there is a certain degree of overlap in phishing and non-phishing, making it more challenging to identify phishing using email topics. In contrast, in the healthcare industry, the only common topic is *Media & Communication*. This indicates distinct differences in email topics between phishing and non-phishing emails.

Table 9 shows the distribution of topics across phishing and non-phishing emails within the healthcare industry. Notably, topics traditionally associated with academic and educational content, such as *Academic Research* and *Linguistics & Education*, are predominantly found in non-phishing emails. This indicates that legitimate interactions in the research-heavy healthcare sector commonly concentrate on these academic topics. An email on new health guidelines

to inform healthcare professionals could be an example of a legitimate email. In contrast, topics related to personal health issues, including *Pharmaceuticals*, *Aging & Fitness*, *Disease Prevention*, and *Weight Management*, are more frequently observed in phishing emails. One example is an email advertisement for a new pharmaceutical product designed for weight loss, potentially indicating phishing. This trend shows that phishing attempts often exploit sensitive health-related topics, potentially because they evoke urgent responses from recipients. Thus, while emails centered on more sophisticated academic topics are typically legitimate, those focusing on health-related decisions are more susceptible to phishing attempts.

Tables 15 to 19 in the Appendix provide a detailed summary of the frequency of topics, along with the five most significant words that contributed to the identification of these topics.

5 Limitations

Threats to Internal Validity. Limitations within this thesis originate primarily from potential biases in the dataset and feature extractions. The emails were not randomly selected, meaning they may not be representative. Other measurement errors cannot be ruled out; for example, the rule-based classification system may have incorrectly assigned an email to an industry. Another concern relates to the model specification in the machine learning models. Since I considered only a few features, important variables may have been omitted, which could have led to an underspecification.

Threats to External Validity. Regarding the generalizability of my findings, my study focuses solely on the healthcare and financial services industries, which may limit the applicability of the results to other industries. Additionally, the continued evolution of phishing tactics may make my findings less applicable over time. Finally, there may be variations in the transferability of my results to different countries and cultures due to regional differences in phishing susceptibility.

6 Conclusion

Phishing emails remain the most common form of cyberattacks, deceiving individuals into revealing sensitive information. This harms individuals and can have far-reaching consequences for businesses and entire industries. Despite existing research, there remains a gap in understanding how phishing characteristics differ by industry and how to recognize and effectively combat them. Due to their high exposure to phishing emails, the financial services and healthcare industries were analyzed.

This thesis examined 18,101 distinct emails and identified an industry for a subset. The data underwent feature engineering, focusing on email characteristics, including length-based features and behavioral cues. These were evaluated on an industry level. The importance of individual phishing characteristics was ranked using machine learning models. Additionally, email topics in emails were analyzed using a topic model. The aim was to investigate phishing susceptibility in the financial services and healthcare industries and to identify clear recommendations that help employees detect phishing.

Investigating phishing characteristics at industry level poses several questions to which my thesis has found the following answers:

RQ1. How vulnerable are the financial services and healthcare industries to phishing? I observed a meaningful difference in the number of emails encountered by employees in the financial services industry compared to the healthcare industry. Financial sector workers are exposed to significantly more phishing emails than healthcare workers. However, susceptibility to phishing is much higher in healthcare than in the financial services industry due to regulation and training in financial institutions. This suggests that the financial services industry is adequately equipped to combat phishing, whereas the healthcare industry may require further training and regulation.

RQ2. What characteristics of emails can employees working in the financial services and healthcare industries use to detect phishing? Shorter emails are an indicator for phishing, with the aim of bypassing the spam filter. Especially phishing emails in the healthcare industry are particularly short. Employees should therefore scan the length of a suspicious email as the very first step. Urgent words are another good indication of phishing in the financial industry.

However, this does not apply to the healthcare industry, where, on average, there are fewer urgent words in phishing emails. Consequently, employees should evaluate the signals from urgent words differently depending on their industry.

RQ3. Which characteristics are most critical for phishing detection? The length of an email is the most effective indicator when scanning for suspicious content. Furthermore, the number of urgent words is the second most effective indicator. The average word length is unreliable and should be disregarded when looking for signs of phishing.

RQ4. How do email topics predict phishing likelihood in the financial and healthcare sectors? The analysis of email topics in the financial services industry shows that emails about financial transactions and personal financial planning are more likely to be phishing attempts. In contrast, those on non-monetary topics are generally safer. In healthcare, emails on sophisticated academic topics are usually legitimate, but those concerning health-related decisions are often phishing attempts.

To gain a more accurate understanding of the future outlook, it is essential to continue refining phishing detection methods through adaptive strategies. Focusing on strategies tailored to industries is important, as there are apparent differences in phishing behaviors across different sectors. Additional industries, particularly the tech industry, should be examined because of its connection to the digital landscape. Furthermore, integrating machine learning and topic models could further enhance the identification of phishing emails.

For the healthcare industry, which is highly susceptible to phishing, further research should be conducted into the reasons for the industry's high susceptibility to phishing. Implementing unannounced assessments in which healthcare workers are observed could assist in comprehending the circumstances under which they fall for phishing emails. For the financial services industry, which is significantly less susceptible, investigations should be carried out into which training and regulations were successful and to what extent these measures are transferable to other industries.

References

Akinyelu, Andronicus A, and Aderemi O Adewumi. 2014. "Classification of phishing email using random forest machine learning technique." *Journal of Applied Mathematics* 2014.

Alkhalil, Zainab, Chaminda Hewage, Liqaa Nawaf, and Imtiaz Khan. 2021. "Phishing attacks: A recent comprehensive study and a new anatomy." *Frontiers in Computer Science* 3:563060.

Allison, Paul D. 2012. *Logistic regression using SAS: Theory and application.* SAS institute.

An, Yusung, Dongju Kim, Juyeon Lee, Hayoung Oh, Joo-Sik Lee, and Donghwa Jeong. 2023. "Topic Modeling-Based Framework for Extracting Marketing Information From E-Commerce Reviews." *IEEE Access* 11:135049–135060.

Asghar, Muhammad Zubair, Aurangzeb Khan, Shakeel Ahmad, Maria Qasim, and Imran Ali Khan. 2017. "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme." *PloS one* 12 (2): e0171649.

CFA Institute. 2024. *CFA Program Glossary.* https://www.cfainstitute.org/programs/cfa/curriculum/glossary. Accessed: 2024-04-11.

Chakraborty, Subhadeep. 2023. *Phishing Email Detection.* https://doi.org/10.34740/KAGGLE/DSV/6090437. https://www.kaggle.com/dsv/6090437.

ChatableApps. 2024. *The Ultimate Guide to the Classification of Emails: Everything You Need to Know.* https://chatableapps.com/technology/the-ultimate-guide-to-the-classification-of-emails-everything-you-need-to-know/. Accessed: 2024-04-11.

Cormack, Gordon V, José María Gómez Hidalgo, and Enrique Puertas Sánz. 2007. "Spam filtering for short messages." In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management,* 313–320.

Cubicle Ninjas. 2024. *Call to Action Phrases.* https://cubicleninjas.com/call-to-action-phrases/. Accessed: 2024-04-11.

Deshpande, Atharva, Omkar Pedamkar, Nachiket Chaudhary, and Swapna Borde. 2021. "Detection of phishing websites using Machine Learning." *International Journal of Engineering Research & Technology (IJERT)* 10 (05).

Doshi, Jay, Kunal Parmar, Raj Sanghavi, and Narendra Shekokar. 2023. "A comprehensive dual-layer architecture for phishing and spam email detection." *Computers & Security* 133:103378.

Fang, Yong, Cheng Zhang, Cheng Huang, Liang Liu, and Yue Yang. 2019. "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism." *IEEE Access* 7:56329–56340.

Gordon, William J, Adam Wright, Ranjit Aiyagari, Leslie Corbo, Robert J Glynn, Jigar Kadakia, Jack Kufahl, Christina Mazzone, James Noga, Mark Parkulo, et al. 2019. "Assessment of employee susceptibility to phishing attacks at US health care institutions." *JAMA network open* 2 (3): e190393–e190393.

Grootendorst, Maarten. 2022. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." *arXiv preprint arXiv:2203.05794.*

Harvard Health Publishing. 2024. *Medical Dictionary of Health Terms: A-C.* https://www.health.harvard.edu/a-through-c. Accessed: 2024-04-11.

Ho, Siew Kei. 2020. "91% of all cyber attacks begin with a phishing email to an unexpected victim." Deloitte Report.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression.* John Wiley & Sons.

Jalali, Mohammad S, Maike Bruckes, Daniel Westmattelmann, and Gerhard Schewe. 2020. "Why employees (still) click on phishing links: investigation in hospitals." *Journal of medical Internet research* 22 (1): e16775.

Jampen, Daniel, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. 2020. "Don't click: towards an effective anti-phishing training. A comparative literature review." *Human-centric Computing and Information Sciences* 10 (1): 33.

Jáñez-Martino, Francisco, Rocío Alaiz-Rodríguez, Víctor González-Castro, Eduardo Fidalgo, and Enrique Alegre. 2023. "A review of spam email detection: analysis of spammer strategies and the dataset shift problem." *Artificial Intelligence Review* 56 (2): 1145–1173.

KnowBe4. 2023. *2023 Phishing by Industry Benchmarking Report.* https://www.brighttalk.com/resource/core/436449/2023-phishing-by-industry-benchmarking-report-pdf_931268.pdf. Accessed: 2024-05-01.

Lagazio, Monica, Nazneen Sherif, and Mike Cushman. 2014. "A multi-level approach to understanding the impact of cyber crime on the financial sector." *Computers & Security* 45:58–74.

Manning, Benjamin. 2017. "Extreme gradient boosting and behavioral biometrics." In *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 31. 1.

O'brien, Robert M. 2007. "A caution regarding rules of thumb for variance inflation factors." *Quality & quantity* 41:673–690.

Priestman, Ward, Tony Anstis, Isabel G Sebire, Shankar Sridharan, and Neil J Sebire. 2019. "Phishing in healthcare organisations: Threats, mitigation and approaches." *BMJ health & care informatics* 26 (1).

Rao, Sanjeev, Anil Kumar Verma, and Tarunpreet Bhatia. 2023. "Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data." *Expert Systems with Applications* 217:119594.

Rizzoni, Fabio, Sabina Magalini, Alessandra Casaroli, Pasquale Mari, Matt Dixon, and Lynne Coventry. 2022. "Phishing simulation exercise in a large hospital: A case study." *Digital Health* 8:20552076221081716.

Samsir, Samsir, Reagan Surbakti Saragih, Selamat Subagio, Rahmad Aditiya, and Ronal Watrianthos. 2023. "Using BERTopic Model for Abstracts Classification." *JURNAL MEDIA INFORMATIKA BUDIDARMA* 7 (3).

Stojnic, Tatyana, Dinusha Vatsalan, and Nalin AG Arachchilage. 2021. “Phishing email strategies: understanding cybercriminals’ strategies of crafting phishing emails.” *Security and privacy* 4 (5): e165.

Tan, Choon Lin, Kang Leng Chiew, and KokSheik Wong. 2016. “PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder.” *Decision Support Systems* 88:18–27.

Tian, Chuan Annie, Matthew L Jensen, and Alexandra Durcikova. 2023. “Phishing susceptibility across industries: The differential impact of influence techniques.” *Computers & Security* 135:103487.

Ziegler, Andreas, and Inke R König. 2014. “Mining data with random forests: current options for real-world applications.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4 (1): 55–63.

List of Figures

List of Tables

7 Appendix

7.1 Length-based Characteristics

Mann-Whitney U Test on character length differences between Phishing and Non-Phishing Emails Across Industries. The following hypotheses are tested:

(3) H_0 : *The median number of characters in phishing emails is* ***greater*** *than or equal to the median number of characters in non-phishing emails.*

(4) H_0 : *The median number of characters in phishing emails is* ***greater*** *than or equal to the median number of characters in non-phishing emails for the financial services industry.*

(5) H_0 : *The median number of characters in phishing emails is* ***greater*** *than or equal to the median number of characters in non-phishing emails for the healthcare industry.*

Table 10: Mann-Whitney U Test Results for Character Length

Metric/Industry	(1) Full Dataset	(2) Financial Services	(3) Healthcare
U-statistic	35177195.5	176548.5	81003.0
p-value	0.000	0.045	0.000
Significant at $\alpha = 0.05$	True	True	True

Summary of Mann-Whitney *U* Test for Character Length:
The results of the character length analysis of phishing and non-phishing emails indicate that, across the full dataset and specific industries, phishing emails tend to be shorter. For the financial services industry, the difference is marginally significant, while it is highly significant for the healthcare industry.

7.2 Behavioral Cues

Mann-Whitney *U* test on average urgent words differences between phishing and non-phishing emails across industries. The following hypotheses are tested:

(6) H_0 : *The median number of characters in phishing emails is* ***greater*** *than or equal to the median number of characters in non-phishing emails.*

(7) H_0 : *The median number of characters in phishing emails is* ***greater*** *than or equal to the median number of characters in non-phishing emails for the financial services industry.*

(8) H_0 : *The median number of characters in phishing emails is* ***smaller*** *than or equal to the median number of characters in non-phishing emails for the healthcare industry.*

Table 11: Mann-Whitney *U* Test Results for Behavioral Cues

Metric/Industry	(1) Full Dataset	(2) Financial Services	(3) Healthcare
U-statistic	44435458.0	214369.0	139105.0
p-value	0.000	0.000	0.000
Significant at $\alpha = 0.05$	True	True	True

Summary of Mann-Whitney *U* Test for Average Urgent Words:
The results of the urgent words analysis of phishing and non-phishing emails indicate that, in the full dataset and the financial services dataset, the average number of urgent words is significantly higher for phishing emails, while the number is significantly lower in the healthcare dataset.

7.3 Model Selection

Table 12: Performance Metrics of Various Machine Learning Models for Classification Tasks Tested on the Phishing Industry Dataset

kursiv Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.64	0.63	0.64	0.57
K-Nearest Neighbors	0.66	0.66	0.66	0.66
Support Vector Classifier	0.63	0.62	0.63	0.55
Decision Tree	0.68	0.69	0.68	0.68
Random Forest	0.73	0.73	0.73	0.73
Gradient Boosting	0.73	0.73	0.73	0.72
MLP Classifier	0.72	0.71	0.72	0.70

Summary of Table 12:

The table shows that Random Forest and Gradient Boosting models outperform other tested machine learning models on the Phishing Industry Dataset, both achieving the highest metrics across accuracy, precision, recall, and F1-score at 0.73. In contrast, Support Vector Classifier exhibits the lowest performance, with accuracy scores of 0.63 respectively.

7.4 Random Forest

Table 13: Classification Report for Random Forest Model on Phishing Industry Dataset

Class	Precision	Recall	F1-Score	Support
0	0.78	0.80	0.79	3384
1	0.65	0.62	0.64	2047
accuracy			0.73	5431
macro avg	0.72	0.71	0.71	5431
weighted avg	0.73	0.73	0.73	5431

Summary of Table 13:

Table 13 presents a classification report for the Random Forest model applied to the Phishing Industry Dataset, showing Non-phishing (Class 0) achieving higher precision, recall, and F1-score than phishing (Class 1), with values of 0.78, 0.80, and 0.79 respectively. The model's overall accuracy stands at 0.73, with a weighted average for precision, recall, and F1-score also at 0.73 across a total support of 5431 instances.

7.5 Variance Inflation Factor

Before a logistic regression equation can be set up, it must be ensured that there is no multicollinearity. To test this, I calculated the variance inflation factor of all variables, which measures the extent to which the variance of a regression coefficient increases due to multicollinearity in the model (O'brien 2007)

Table 14: Variance Inflation Factor for Explanatory Variables

Feature	VIF
Found Words Count Fin	3.521398
Found Words Count Health	4.450374
Total Characters	15933.257291
Total Words	15904.762571
Avg Word Length	1.040123
Urgent Words Count	2.711624
Industry_None	NaN[1]
Industry_Financial Services	1.948934
Industry_Healthcare	2.117389
Industry_Undecided	1.128829

[1] Industry_None is the baseline category

Summary of Table 14:
High VIF Scores: The variables *Total Characters* and *Total Words* show extremely high VIF scores, indicating multicollinearity. **Moderate VIF Scores:** *Found Words Count Fin* and *Found Words Count Health* show a moderate VIF, suggesting moderate multicollinearity.

Variable	**Explanation**
Found Words Count Fin	Finance-related words in email.
Found Words Count Health	Healthcare-related words in email.
Total Characters	Total number of characters in email.
Total Words	Total number of words in email.
Avg Word Length	Average length of words in email.
Urgent Words Count	Number of urgency-indicative words.
Industry_None	Baseline, no industry specified.
Industry_Financial Services	Email pertains to financial services.
Industry_Healthcare	Email pertains to healthcare industry.
Industry_Undecided	Not clear between Financial Services and Healthcare.

7.6 BERTopic

7.6.1 Full Dataset

Table 15: Topic Distribution in **Full Dataset**

Topic	Name	Count
0	Corporate Affairs	201
1	Pharmaceuticals	336
2	Linguistics & Education	134
3	Email Management	112
4	Digital & Technology	100
5	Software Development	59
6	Online Marketing	20
7	Data Management	14

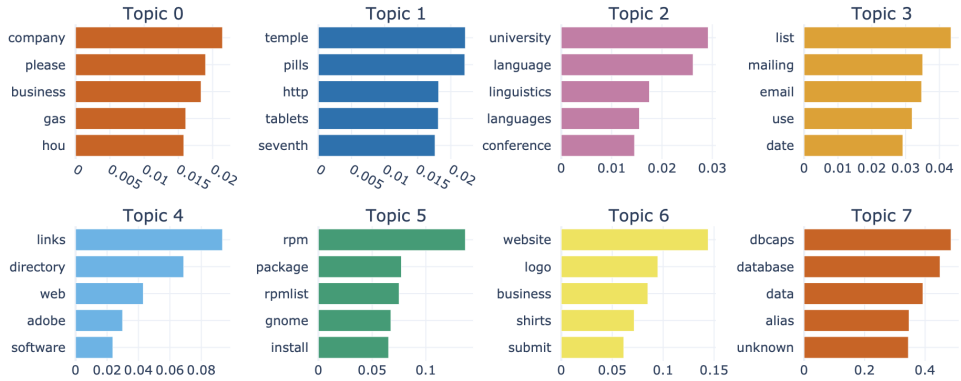

Figure 4: Breakdown of Topic Distribution in **Full Dataset** Showing Top 5 Words for Every Topic. X-axis Presents the Relevance Scores of Words.

Summary of Table 15:

Table 15 displays the topic distribution across the Full Dataset, with ”Pharmaceuticals” as the most prevalent topic at 336 instances, followed by ”Corporate Affairs” at 201. Figure 4 visualizes the top five words for each topic, showing key terms such as ”pills” for Pharmaceuticals and ”company” for Corporate Affairs, illustrating their significant thematic roles within the dataset.

7.6.2 Financial Services Dataset

Table 16: Topic Distribution in **Only Financial Services Phishing Dataset**

Topic	Name	Count
0	Corporate Affairs	254
1	Real Estate & Loans	164
2	Financial Transactions	127
3	Energy Sector	67
4	Digital & Technology	38
5	Public Funding	17
6	Global Finance	17
7	Personal Planning	13

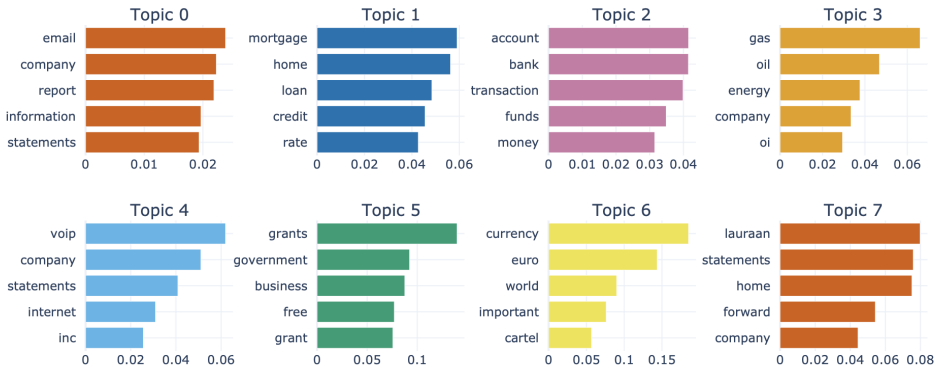

Figure 5: Breakdown of Topic Distribution in **Only Financial Services Phishing Dataset** Showing Top 5 Words for Every Topic. X-axis Presents the Relevance Scores of Words.

Summary of Table 16:

Table 16 details the topic distribution within a financial services phishing dataset, showing ”Corporate Affairs” as the most common topic with 254 occurrences, followed by ”Real Estate & Loans” and ”Financial Transactions” with 164 and 127 instances, respectively. Figure 5 visualizes the relevance scores of the top five words for each topic, highlighting key terms such as ”email” for Corporate Affairs and ”mortgage” for Real Estate & Loans, indicating the thematic focus in phishing content.

Table 17: Topic Distribution in **Only Financial Services Non-Phishing Dataset**

Topic	Name	Count
0	State Governance	29
1	Linguistics & Education	13
2	Corporate Affairs	345
3	Digital & Technology	28
4	Media & Communication	14
5	Global Finance	67

Figure 6: Breakdown of Topic Distribution in **Only Financial Services Non-Phishing Dataset** Showing Top 5 Words for Every Topic. X-axis Presents the Relevance Scores of Words.

Summary of Table 17:

Table 17 displays the topic distribution within the Only Financial Services Non-Phishing Dataset, highlighting ”Corporate Affairs” as the dominant topic with 345 instances, followed by ”Global Finance” with 67. Figure 6 visually represents the top five words associated with each topic, showcasing the relevance of terms such as ”risk” in Corporate Affairs and ”company” in Global Finance, which help identify thematic elements typical of non-phishing financial communications.

7.6.3 Healthcare Dataset.

Table 18: Topic Distribution in **Only Healthcare Phishing Dataset**

Topic	Name	Count
0	Pharmaceuticals	189
1	Corporate Affairs	95
2	Media & Communication	59
3	Health & Wellness	34
4	Geopolitical Entities	28
5	Aging & Fitness	21
6	Disease Prevention	16
7	Weight Management	12

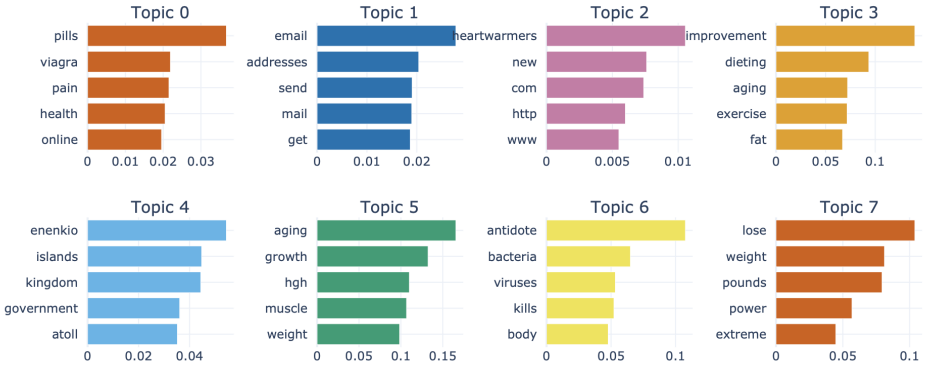

Figure 7: Breakdown of Topic Distribution in **Only Healthcare Phishing Dataset** Showing Top 5 Words for Every Topic. X-axis Presents the Relevance Scores of Words.

Summary of Table 18:

Table 18 presents the topic distribution in the Only Healthcare Phishing Dataset, showing ”Pharmaceuticals” as the leading topic with 189 occurrences, followed by ”Corporate Affairs” with 95. Figure 7 provides a breakdown of the top five words for each topic, illustrating how specific terms like ”pills” in Pharmaceuticals and ”email” in Corporate Affairs dominate the lexical landscape, suggesting thematic priorities in healthcare-related phishing attempts.

Table 19: Topic Distribution in **Only Healthcare Non-Phishing Dataset**

Topic	Name	Count
0	Academic Research	482
1	Digital & Technology	31
2	Linguistics & Education	68
3	Media & Communication	221

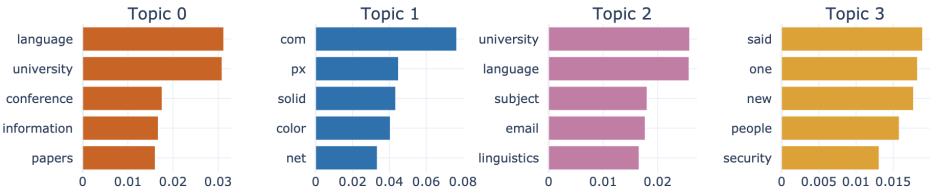

Figure 8: Breakdown of Topic Distribution in **Only Healthcare Non-Phishing Dataset** Showing Top 5 Words for Every Topic. X-axis Presents the Relevance Scores of Words.

Summary of Table 19:

Table 19 outlines the topic distribution in the Only Healthcare Non-Phishing Dataset, with ”Academic Research” dominating at 482 instances, followed by ”Media & Communication” with 221. Figure 8 visually details the top five words for each topic, such as ”language” and ”university” for Academic Research, reflecting key terms that characterize non-phishing communications in healthcare settings.