Article

Not peer-reviewed version

Comparative Investigation of Traditional Machine Learning Models and Transformer Models for Phishing Email Detection

René Meléndez * , Michal Ptaszynski , Masui Fumito

Posted Date: 18 October 2024

doi: 10.20944/preprints202410.1467.v1

Keywords: Phishing detection; Phishing emails; Machine Learning; Transformer Models; Traditional 14 Models; Supervised Learning; Text Classification, Cyber threat Mitigation; Cybersecurity

 doi:10.20944/preprints202410.1467.v1

Article

Comparative Investigation of Traditional Machine Learning Models and Transformer Models for Phishing Email Detection

René Meléndez *, Michal Ptaszynski and Fumito Masui

Text Information Processing Laboratory, Kitami Institute of Technology, Kitami 090-8507, Japan
* Correspondence: rene.melendeza@gmail.com

Abstract: Phishing emails pose a significant threat to cybersecurity worldwide. There are already tools that mitigate the impact of these emails by filtering them, but these tools are only as reliable as their ability to detect new formats and techniques for creating phishing emails. In this paper we investigated how traditional models and transformer models work on the classification task of identifying if an email is phishing or not. We realized that transformer models, in particular DistilBERT, BERT, and RoBERTa had a significantly higher performance compared to traditional models like Logistic Regression, Random Forest, Support Vector Machine, and Naive Bayes. The process consisted in using a large and robust dataset of emails and applying preprocessing and optimization techniques to maximize the best result possible. roBERTa showed its outstanding capacity to identify phishing emails by achieving the maximum accuracy of 0.9943. Even though they were free successful, traditional models performed marginally worse; SVM performed the best, with an accuracy of 0.9854. The results emphasize the value of sophisticated text processing methods and the possibility of transformer models to improve email security by thwarting phishing attempts.

Keywords: phishing detection; phishing emails; machine learning; transformer models; traditional 14 models; supervised learning; text classification, cyber threat mitigation; cybersecurity

1. Introduction

The use of internet services has grown exponentially in recent years [1]. Accessibility has become widespread and affordable; for example, in the USA, the lowest cost of internet access is around $20.00 per month [2]. However, public access to free internet, although convenient, comes with security risks that should not be overlooked. As more people use email as a primary communication method between services and clients [3] The more cybersecurity risk and attempts have risen too. Among these, phishing is one of the most prevalent tactics used by malicious actors to obtain sensitive or critical information [4]. Phishing attacks exploit social engineering techniques, manipulating victims to gain their trust and ultimately extract information . In recent years, these phishing attacks have caused a $3.86 million financial loss due to incidents involving federal entities [5]. Such attacks often result in identity theft and unauthorized access to privileged accounts linked to the victim's email or other compromised personal data [6] . According to the 2023 Verizon Data Breach Investigation Report, social engineering attacks, such as phishing, now account for 36% of the most common cyberattacks [7].

Email serves as the entry point for accessing numerous online services, acting both as a communication tool and, in many cases, an authentication mechanism [8,9]. Since email access is typically password-protected, attackers can exploit weaknesses through dictionary attacks if sufficient user information is available. According to the Anti-Phishing Working Group (APWG), phishing attempts have surged in recent years [10], with over 1.2 million attacks recorded in the first quarter of 2022 alone [11]. These attempts have also evolved in sophistication, extending beyond traditional methods to target users through social media, SMS, and other platforms [12]. Phishing continues to grow rapidly as an efficient method for exploiting user trust and gaining unauthorized access to sensitive data [13].

 doi:10.20944/preprints202410.1467.v1

Phishing attacks exploit social engineering by impersonating trusted entities such as banks, businesses, online stores, and even government agencies [14]. In actuality, one of the most used methods to detect phishing in mail applications is by text processing and analyzing the content of said emails [15]. Advanced machine learning such as CNNs have proven to be an effective method to detect phishing emails and reduce false positives [16]. By implementing this solution we can drastically reduce the number of users that click the malicious links or give personal information to the phishers [17]. Normally this step prevents the possibility of being exposed to malware, or spyware that can be injected into the user's system leading to gaining access to personal information or even gaining full control of the system [18]. By encapsulating the user and the phisher apart, we hope to ensure that users do not run unnecessary risks, thereby enhancing their overall online security [19].

2. Objectives

- Compare the performance of Traditional models and Transformer models in the email classification task for phishing and non-phishing labels.Evaluating its efficacy via cuantitative metrics such as precision, recall, F1-score and accuracy.
- Explore the enhancements brought by the implementation of transformer models in text classification task via an analysis of classification accuracy and their accuracy to process complex and diverse content.
- Conduct a thorough analysis of the registries of failed classification done by Traditional models and Transformer models equally. By identifying recurring patterns and root causes of errors, this objective aims to propose actionable improvements and refinements for future phishing detection methodologies, enhancing their effectiveness and reliability.

3. Related Work

Since phishing is a social engineering attack that started with the growth and adoption of the internet worldwide, there have been numerous studies approaching this problem from different perspectives: user awareness, server and client-side solutions, and deep analysis of the composition of phishing emails and their effectiveness against users. The following will briefly explain these approaches by analyzing websites, links, and email content, with the objective of specifying the differences and advantages of each popular approach.

3.1. Analysis of Phishing Websites

This approach employs methodologies like heuristic and machine learning methods with traditional models to detect phishing websites. Heuristic strategies identify if a website is phishing or not based on the textual content of the website, performing a comparative analysis with a legitimate one. The machine learning approach also examines the content and features of a website but uses a pre-trained model to classify them. In these studies, significant results were found but were limited due to the scarce amount of samples in the dataset [20].

3.2. Analysis of Phishing URLs

This approach uses machine learning with traditional models to classify phishing URLs, taking into account features such as URL composition, detection of anomalies, analysis of HTML and JavaScript scripts within the URL, and domain name analysis. Each of these studies demonstrated significant effectiveness in precision. It is anticipated that using SDN and blockchain will provide a different approach to detecting these URLs and finding improved results [21].

3.3. Analysis of the Content of Phishing Emails

In this study, phishing email detection is often considered a subcategory of spam detection, using datasets or self-curated datasets to train traditional classification models, where Decision Trees and K-Nearest Neighbors (KNN) generally show the highest results, depending on the training and testing

 doi:10.20944/preprints202410.1467.v1

datasets [22,23]. Another newly added approach is using Natural Language Processing (NLP) for lexical and orthographical analysis to categorize emails. Researchers also use intentions and sentiment analysis for categorization. One of the major difficulties of this approach is the maximum number of tokens a transformer-based model can handle, which can be mitigated by splitting smaller sub-tokens, though this can affect context and accuracy [24].

Even though these advancements are helping the industry develop effective solutions for phishing detection, the problem continues to evolve over time. This study aims to contribute to the foundations of phishing detection by conducting a comparative analysis between traditional and transformer-based models, with the goal of developing a comprehensive understanding of the most effective methods, parameters, and configurations for phishing detection, and to support ongoing efforts to enhance cybersecurity globally.

3.4. Deep Learning for Phishing Detection

One of the most fascinating implementations of phishing detection is the integration of deep learning models, as demonstrated by Altwaijry et al. study. This research compares Convolutional Neural Networks (CNNs) with architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), including Bidirectional GRUs, to detect phishing emails. The 1D-CNNPD model, enhanced by adding recurrent layers, showed significant results on commonly used datasets like Phishing Corpus and SpamAssassin [25]. This approach demonstrates a significant improvement compared to traditional machine learning methods, which often face limitations when handling diverse phishing content.

3.5. Transformer Models for Phishing Detection

Phishing email detection with transformer models has gained prominence in recent years. These models excel at tasks involving the classification of text-heavy data such as emails. Transformer models like BERT and its variants are well-known for their high accuracy in identifying phishing emails. A researcher compared the performance of BERT with Recurrent Neural Networks (RNNs) for phishing email detection, achieving impressive results with a test accuracy of 96.1% for BERT. Atawneh and Aljehani explored the use of deep learning models for phishing detection and created a model combining BERT and LSTM, achieving an impressive 99.61% accuracy on a balanced dataset that included Enron emails as non-phishing emails [26].

When we research the applications of transformer models for Classification task like Phishin detection there is some models that appear with more consistency than others.For example using BERT variants, Jamal et al. (2023) worked on implementing spam and phishing email detection with transformer models like DistilBERT and RoBERTa to reduce complexity while maintaining precision and accuracy. They achieved 98.7% precision and an overall accuracy of 98.3% [27]. An important aspect of their study was the use of 8 epochs and an AdamW optimizer for each model.

Lastly, one of the most innovative approaches was introduced by Lee et al. (2020), who created a prototype model specifically for detecting phishing and spam emails. CatBERT was designed to resist adversarial attacks while maintaining its accuracy [28]. Although this model achieved 87% accuracy, its added resilience for adversarial robustness makes it suitable for real-world implementation, sacrificing a small degree of accuracy in exchange.

From this exploration, it is evident that there are various approaches to phishing detection implementations. Over the years, these methods have evolved alongside their counterparts. While heuristic methods like Decision Trees, KNN, and Support Vector Machines (SVM) have demonstrated moderate results, transformer models have gained popularity in recent years due to their ability to handle complex datasets and textual data from emails. Transformer models, particularly BERT and its variants, are the preferred choice for phishing and email detection tasks due to their high accuracy and effectiveness in handling natural language structures and text classification. Although these models require high computational resources compared to traditional methods, with some refinements

 doi:10.20944/preprints202410.1467.v1

and balanced performance, they are expected to be the mainstay of text classification and phishing detection in the future.

3.6. Datasets Used in Previous Investigations

During the preparations of this paper it was started with an exploratory investigation in which there was two objectives. First getting to know previous works and studies within this field and learning from them to being able to create a well versed methodology of work. Second of all to get to learn what datasets where used, what benign dataset where they used for said studies and how they performed in results.With this being said the number of datasets achieved was 20 which 12 had a previous article, or small study via kaggle in which they did an email classification task like, phishing detection or fraud detection or spam detection.

In these investigations we can see the prevalent use of traditional models compared to transformer models. From these we can mention Linear Regression with 94.08%, Sequential search with 99.72%, decision trees with 96.77% and Naive Bayes with 99.13%. These models showed to be effective and efficient on email classification tasks, and as said before they have the benefit of being lightweight algorithms in terms of resources needed to use them. With this being even thought the number of applications for transformer models is way lower compared to traditional models it can be seen that overall Roberta has a recognizable appearance in the results obtained. Showing that transformer models are well versed on classification task that require contextual complexity like phishing emails.

Table 1. Performance of Different Models on Various Datasets.

Year	Dataset Name	Linear Regression	Sequential	Decision Trees	Random Forest	Naïve Bayes	CNN	Roberta
2020	Email Classification [29]	94.08	0	0	0	0	0	0
2023	Email Spam Classification [30]	0	86.2	0	0	79.87	0	78.57
2001	Enron Spam Data (No Code) [31]	0	0	0	0	95	0	0
2018	Fraud Email Dataset [32]	92	0	0	0	97	0	0
2023	Phishing Email Detection [33]	0	0	93.1	0	0	97	99.36
2023	Phishing-Mail [34]	0	0	92.82	0	0	99.03	96.81
2023	Pishing Email Detection [35]	0	0	0	0	0	0	0
2018	Pishing-2018 Monkey [36]	0	0	0	0	0	0	0
2019	Pishing-2019 Monkey [37]	0	0	0	0	0	0	0
2020	Pishing-2020 Monkey [38]	0	0	0	0	0	0	0
2021	Pishing-2021 Monkey [39]	0	0	0	0	0	0	0
2022	Pishing-2022 Monkey [40]	0	0	0	0	0	0	0
2018	Private-pishing4mbox [41]	0	0	0	0	0	0	0
2023	Spam (or) Ham [42]	0	99.72	0	0	96.9	0	0
2020	Spam Classification for Basic NLP [43]	0	81	96.77	0	98.49	0	98.33
2021	Spam Email [44]	0	96.67	97.21	0	99.13	0	0
2021	Spam_assasin [45]	0	0	98.6	98.87	0	0	0
2024	Phishing Validation Emails Dataset [46]	0	0	0	0	0	0	0

4. Data and Methods

4.1. Data Collection

The data was collected manually from June 2023 to January 2024 from various online platforms, including posts about phishing on social media sites. This dataset consists of 119,148 emails, which were manually flagged as phishing or not phishing by the investigator. The origin of said data can be see in the next list of datasets with their respective source All emails are written in English as the primary language. To make this dataset more diverse, the Enron dataset was used to add a significant number of emails categorized as not phishing.

This method is common in other studies, such as Sahingoz et al. (2019), where they consolidated a dataset between Phishtank and Enron to achieve more realistic results. They created a dataset to provide a realistic assessment of phishing. It is also important to highlight that Ugochukwu et al. (2023) emphasized the importance of utilizing recent phishing emails due to the evolution of phishing tactics over the years (2016–2023). This underscores the critical need to continually refresh datasets, allowing models to be trained with the most current phishing behaviors available.

 doi:10.20944/preprints202410.1467.v1

Figure 1 shows the distribution of the dataset. It can be observed that 57.0% of these emails are flagged as phishing, which is equivalent to 67,912 emails. The distribution of emails with their sources can be observed in Figure 2.

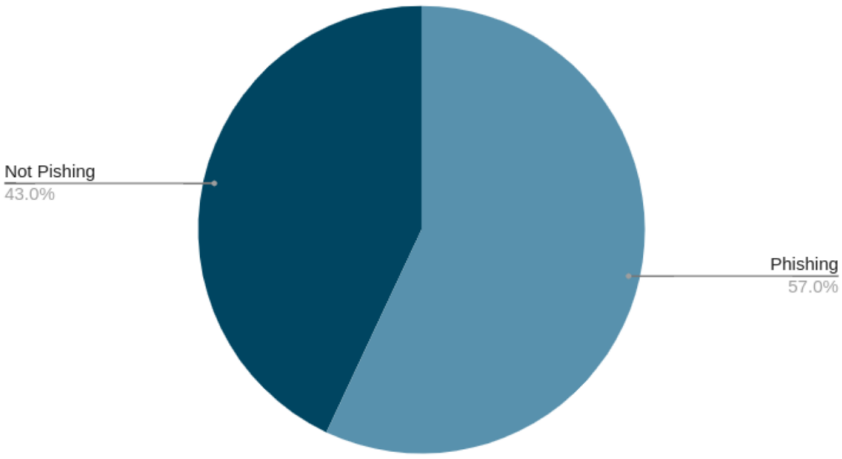

Figure 1. Distribution of the dataset. 57.8% of these emails are flagged as phishing.

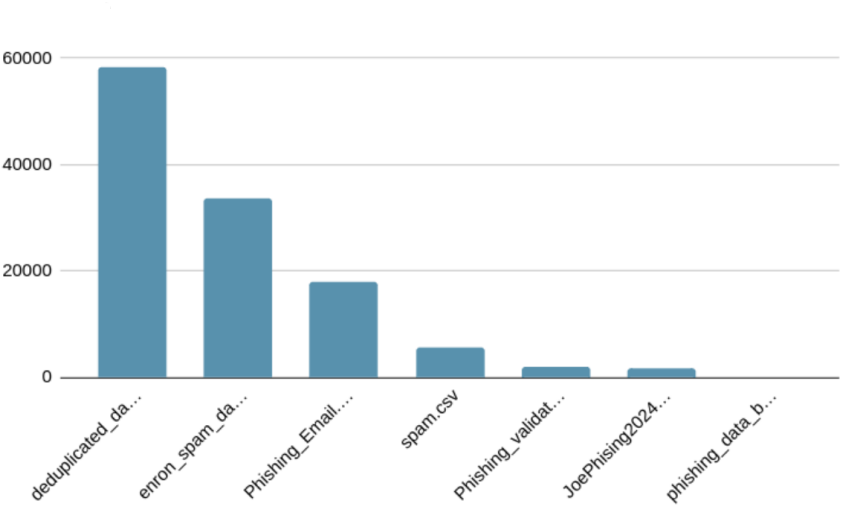

Figure 2. Distribution of emails with their sources.

The collected dataset is an important pillar for this investigation. In order to understand how this dataset was created and what data it contains, here is some statistical information for clarity (Table 2).

Table 2. Statistics of the dataset.

Specifications	Statistics
Number of Emails	119,148
Number of Words	30,478,934
Average Words per Email	255.81
Average Words per Sentence	15.53
Highest Word Count in an Email	15,828
Lowest Word Count in an Email	10
Number of Phishing Emails	68,912
Number of Non-Phishing Emails	50,236

 doi:10.20944/preprints202410.1467.v1

4.2. Applied Models

Machine learning is a pivotal step in the advancement of artificial intelligence, enabling systems to learn, identify patterns, and make decisions with minimal human intervention. The creation of models is a fundamental aspect of this process, as it automates analytical model building, simplifying the use of these models to process complex datasets. Given that the main objective of this investigation is to compare traditional models and transformer models, it is essential to distinguish between them and understand their differences.

Traditional machine learning models, such as decision trees, support vector machines (SVMs), and neural networks, have been widely used for decades. These models were introduced at the foundation of Artificial Intelligence. Using statistical algorithms and data mining, the goal of these methods is to automatically detect patterns in data. This information can be used to make predictions about future data or classification. These models rely on well-established algorithms like linear regression, k-nearest neighbors (KNN), support vector machines (SVM), and ensemble methods like random forest and gradient boosting [47,48]. All these models have specific tasks they specialize in, even though they can be adapted and fine-tuned as needed.

For example, decision trees specialize in making decisions by considering previous data, using branches to separate decision pathways within the algorithm. SVM and linear regression are commonly used for classification tasks. The problem with these models is that they lack the ability to handle more complex problems, as they use numerical or statistical approaches and are less suited for tasks like natural language processing (NLP) and image processing [49]. This makes transformer-based models especially useful for handling the evolving nature of phishing emails, which often adopt new tactics to bypass detection.

The traditional models used in this investigation are the following:

- **Logistic Regression:** A linear model utilized for binary classification. It uses probability to determine if the given data can be classified into a particular label.
- **Random Forest:** A model that uses decision trees for training and learning. After creating these decision trees, it utilizes predictions to improve the precision of the responses.
- **Support Vector Machine (SVM):** A model that classifies data by determining which hyperplane best separates the classes in the feature space.
- **Naive Bayes:** A statistical model that uses Bayes' theorem with the assumption of independence between features. It works well with large datasets.

Transformer models, introduced in 2017, represent a significant advancement in the field of machine learning. They are a specific architecture of encoder-decoder models that utilize a unique attention mechanism to derive dependencies between input and output. Initially designed for tasks like language translation, transformer models have demonstrated remarkable versatility. The key innovation of transformers is their ability to use attention as the sole mechanism for understanding and generating sequences, which has proven particularly powerful for natural language processing (NLP) tasks.

One of the primary reasons for the rapid adoption and success of transformer models in various NLP tasks is their capability for transfer learning. Pretrained transformer models can quickly and efficiently adapt to new tasks with minimal additional training, often requiring only fine-tuning with a smaller dataset. This adaptability has allowed transformers to dominate numerous NLP leaderboards and extend their applicability beyond language tasks to areas such as computer vision, audio processing, and even complex games like chess and mathematical problem-solving.

The impact of transformer models on the field of machine learning has been profound, facilitated by the integration of these models into major AI frameworks like PyTorch and TensorFlow. Furthermore, the development and commercialization of libraries such as those provided by Hugging Face have made transformers accessible to a broad audience of researchers and practitioners [50].

The transformer models used in this paper are the following:

 doi:10.20944/preprints202410.1467.v1

- **BERT (bert-base-uncased):** The BERT (Bidirectional Encoder Representations from Transformers) model is a pretrained model that utilizes bidirectional transformer logic, allowing it to analyze provided text data in both directions.
- **DistilBERT (distilbert-base-uncased):** A compact version of BERT that maintains about 97% of its accuracy while consuming fewer resources.
- **XLNet (xlnet-base-cased):** A model that generalizes BERT using permutation-based prediction, capturing dependencies without the constraint of conditional independence.
- **RoBERTa (roberta-base):** RoBERTa (Robustly optimized BERT approach) is a variant of BERT that improves training logic, including more data and steps to enhance the robustness and precision of the model.
- **ALBERT (A Lite BERT):** A lightweight version of BERT that reduces the model size through parameter sharing and embedding matrix factorization, maintaining high performance with fewer parameters.

4.3. Evaluation Metrics

- **Precision:** Precision is the proportion of true positives among all the positive predictions. It measures the accuracy of the positive predictions made by the model.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** Recall is the proportion of true positives among all the actual positive data. It measures the model's ability to capture all the positive samples.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** The F1-Score is obtained using both recall and precision. It provides a balanced measure considering both values, offering a single metric that reflects how well the model handled imbalanced data.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Accuracy:** Accuracy is the proportion of correct predictions among all the predictions made by the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **True Positive Rate (TPR):** TPR is another term for recall. It represents the proportion of actual positives correctly identified by the model.

$$\text{TPR} = \frac{TP}{TP + FN}$$

5. Evaluation Experiments

5.1. Experiment Setup

In this study, we approach phishing detection by analyzing the content of emails and training traditional models such as Logistic Regression, Random Forest, Support Vector Machine, and Naive Bayes, alongside transformer models such as DistilBERT, BERT, XLNet, RoBERTa, and ALBERT. The objective is to compare these results and determine which model (traditional or transformer) is most effective for this phishing classification task.

 doi:10.20944/preprints202410.1467.v1

5.1.1. Traditional Machine Learning Model Parameters

- **Logistic Regression**
  - Maximum Iterations (max_iter): 1000
  - Random State: 42
- **Random Forest**
  - Number of Estimators (n_estimators): 100
  - Random State: 42
- **Support Vector Machine (SVM)**
  - Kernel: Linear (kernel='linear')
  - Regularization Parameter (C): 1
  - Random State: 42
- **Naive Bayes**
  - Alpha (Smoothing Parameter): 1.0
  - Random State: 42

5.1.2. Transformer Model Parameters

- **Model Names:**
  - distilbert-base-uncased
  - bert-base-uncased
  - xlnet-base-cased
  - roberta-base
  - albert-base-v2
- **Training Settings:**
  - Tokenizer: AutoTokenizer from Hugging Face's transformers library.
  - Dataset: EmailDataset class defined with:
    - Texts and labels from the dataset.
    - Tokenizer for encoding texts with special tokens, padding, and truncation.
  - Optimizer: AdamW optimizer with a learning rate of 2e-5.
  - Device: Utilizes CUDA if available, otherwise CPU.
  - Epochs: 3
- **Model Evaluation:**
  - Batch Size: 16 for both training and testing DataLoader.
  - Loss Function: Cross-entropy loss.
  - Metrics: Classification report with precision, recall, F1-score, and support.

5.2. Results and Discussion

5.2.1. Traditional Machine Learning for Phishing Email Detection

In this work, we evaluated the dataset using a range of traditional machine learning and transformer models for phishing detection, utilizing metrics such as precision, recall, F1-score, and accuracy. Support Vector Machine (SVM) stood out from the other traditional models with the highest accuracy of 0.9836 by exhibiting balanced precision, recall, and F1-scores in both phishing and non-phishing scenarios.Additionally, Random Forest and Logistic Regression performed quite well, exhibiting accuracy levels of 0.9799 and 0.9808, respectively, as well as somewhat lower but still excellent recall and precision scores. Naive Bayes was efficient, but its accuracy was lower at 0.9647 because of its lower recall for non-phishing scenarios.

 doi:10.20944/preprints202410.1467.v1

Table 3. Results for Traditional Models.

Model	Class	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Phishing	0.9863	0.9801	0.9832	0.9808
	Not Phishing	0.9736	0.9818	0.9777	
Random Forest	Phishing	0.9843	0.9804	0.9824	0.9799
	Not Phishing	0.9740	0.9791	0.9766	
Support Vector Machine	Phishing	0.9898	0.9846	0.9872	0.9854
	Not Phishing	0.9795	0.9864	0.9830	
Naive Bayes	Phishing	0.9502	0.9877	0.9686	0.9633
	Not Phishing	0.9826	0.9308	0.9560	

As seen in Figure 3, the performance of traditional models shows clear differences in their ability to distinguish between phishing and non-phishing emails. Logistic Regression, Random Forest, and Support Vector Machine exhibit similar behavior, maintaining a strong true positive rate (TPR) while keeping false positive rates (FPR) relatively low across various threshold values. Among these, Support Vector Machine slightly edges out others in terms of maintaining a higher TPR with lower FPR, showcasing better sensitivity. However, Naive Bayes underperforms in comparison to the other models, showing higher variability in its FPR, despite maintaining competitive precision. This could be attributed to Naive Bayes' strong assumptions about feature independence, which may not align as well with the dataset characteristics.

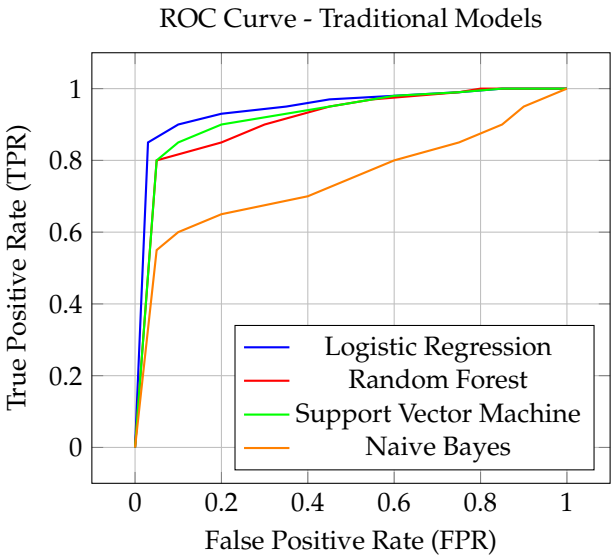

Figure 3. ROC Curve for Traditional Models.

With these evaluations, we can infer that traditional models are capable of delivering high-performance results with their ability to generalize and simplify complex data structures. However, due to the cost-efficiency and easy implementation of these traditional models, the results are highly valuable in implementations where there may be a constraint of computational resources.

5.2.2. Transformer-Based Machine Learning for Phishing Email Detection

For transformer models, roberta-base demonstrated exceptional performance, achieving the highest accuracy of 0.9943, with high precision and recall scores for both classes. Bert-base-uncased is next with an accuracy value of 0.9911. Distilbert-base-uncased also performed significantly well, achieving an accuracy of 0.9899, while albert-base-v2 had an accuracy of 0.9881. XLNet-base-cased followed closely, with an accuracy of 0.9884.

 doi:10.20944/preprints202410.1467.v1

Table 4. Results for Transformer Models.

Model	Class	Precision	Recall	F1-Score	Accuracy
distilbert-base-uncased	Phishing	0.9933	0.9890	0.9911	0.9899
	Not Phishing	0.9853	0.9911	0.9882	
bert-base-uncased	Phishing	0.9947	0.9897	0.9922	0.9911
	Not Phishing	0.9863	0.9929	0.9896	
xlnet-base-cased	Phishing	0.9828	0.9971	0.9899	0.9884
	Not Phishing	0.9961	0.9768	0.9863	
roberta-base	Phishing	0.9928	0.9974	0.9951	0.9943
	Not Phishing	0.9964	0.9903	0.9934	
albert-base-v2	Phishing	0.9939	0.9853	0.9896	0.9881
	Not Phishing	0.9806	0.9919	0.9862	

One notable observation is the consistently high performance for all transformer models, with accuracy normally scoring above 0.9881 across all models. This suggests that transformer-based models, in particular BERT and its variants, are highly effective at distinguishing phishing from non-phishing emails. Although DistilBERT and ALBERT have slightly lower scores compared to the other models implemented in this study, with accuracy values of 0.9899 and 0.9881, these values cannot be diminished. DistilBERT, in particular, stands out since it is a distilled version of BERT, being about 60% faster than normal BERT while still managing to achieve a similar score. This trade-off between accuracy and speed of analysis makes it an excellent choice for implementations that need a lightweight model and can be easily deployed without requiring significant resources.

As showed in Figure 4 it can be observed a similar performance between most of Transfomer models. RoBERTa and BERT still show a strong performance by indicating a high true positive rate(TPR) and a relatively low false positive rate(FPR). In comparison XLNet and ALBERT perform slightly lower but still achieve significant rates in comparison to traditional models.

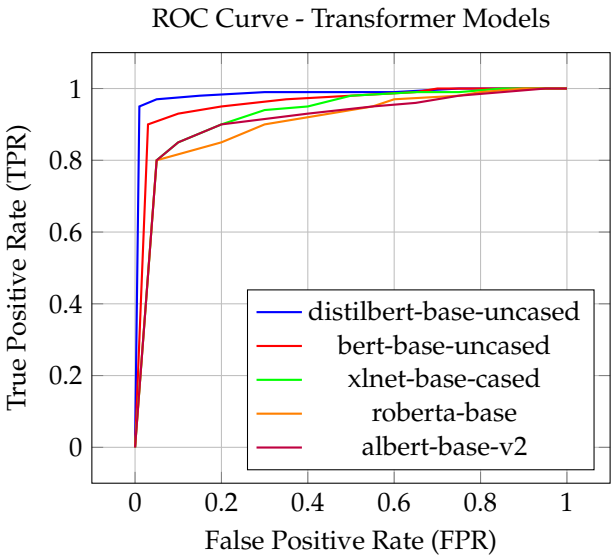

Figure 4. ROC Curve for Transformer Models.

By understanding all the results and studying the intrinsic characteristics of each model, organizations can make an informed decision when selecting a model for phishing email detection, taking into account variables like accuracy, computational cost, and scalability.

 doi:10.20944/preprints202410.1467.v1

5.3. Analysis of Failed Predictions

After taking in account the numerical values of the results obtained in these papers we also have saved all the failed predictions per model in csv format for it to be analyzed. Said failed records were saved in a different file in order to analyze them in depth.

As it can bee seen, there is a significant gap in efficiency and accuracy when comparing these results. The transformer models clearly outperforms traditional models by having 65.21% fewer misclassifications. This reductions in failed predictions highlights the superior performance for transformer models, making then a reliable choice for email classification tasks.

5.3.1. Error Analysis for Traditional Model

After reviewing the failed prediction cases in the task classification task for emails. There were some noticeable patterns that can impact its precision. First of all, grammatical errors and informal type of writing is often classified as non-phishing, this error happens due to the low weight of said words and misspells and also the lack of context understanding for traditional models. Sentiment analysis techniques could improve the accuracy of text classification tasks like phishing detection. Another Factor that affects traditional models is the handling of HTML formatted content, which is often detected as not phishing. This pattern may be due to the lack of examples that have similar content in the training dataset used, limiting the models ability to identify phishing emails that contain HTML code or malicious code.

Other important observation is that in some cases emails contained mixed-language mails where both English and other languages and the use of leatspeak (a form of writing where letters are replaced by letters of symbols) also causes misclassification errors as non-phishing this also due to the to the lack of proper training for said variations.

Traditional models struggle with grammatical errors, mixed languages, leatspeak, and HTML code in emails, highlighting the need of enhanced or fine-tuned datasets for training and parameter adjustments for these models to better detect these characteristics.

5.3.2. Error Analysis for Transformer Models

Even though transformer models have significantly more accuracy compared to traditional models, they still have a significant number of miscategorized emails. A pattern that is recurring is that the mails contained common words like “Click here” or “Password” and these emails were wrongfully categorized as phishing emails. This only reflects on how rigid a pattern-based classification approach is. The use of leatspeak and Phishing URLs also confuses transformer models as they are not trained to detect this type of writing, leading to misclassification. Also the use of HTML based content is another issue that appears to be misclassified as non-phishing. Lastly emails with formal or length format, including multiple breaks in lines, are often classified as phishing , suggesting that the model might have learn presentation or lyrical patterns that not necessarily represent fraudulent intent, It would be essential to add last points in great number of registries for the training dataset so they could better reflect the current phishing methods and behavior.

6. Conclusions and Future Work

6.1. Conclusions

In this paper, we conducted an investigation centered in the detection of phishing emails via traditional models and transformer models used in machine learning. Achieve this by doing an exhaustive analysis of a self-created email dataset. We evaluated the effectiveness of various traditional models such as Logistic Regression, Random Forest, SVM, NB, and transformer models such as DistilBERT, BERT, XLNet, RoBERTa and ALBERT, using metrics like precision, recall, F1-Score and global precisión.

 doi:10.20944/preprints202410.1467.v1

Our analysis was based on a dataset consisting of 119,148 English-language email samples, which we strengthened by incorporating examples from various public sources. The results revealed that transformer models significantly outperformed traditional models. For example, RoBERTa achieved the highest accuracy at 0.9943, with a high F1-score of 0.9951 for phishing detection, demonstrating its superiority in identifying phishing emails accurately. In contrast, traditional models like Logistic Regression and Random Forest showed slightly lower performance, with Logistic Regression achieving an accuracy of 0.9808 and an F1-score of 0.9832 for phishing detection.

XLNet and BERT also performed very well among transformer models, with accuracy of 0.9884 and 0.9911, respectively. In comparison, traditional models like SVM, with an accuracy of 0.9854, were still effective but did not match the precision of transformer models. The lowest-performing traditional model, Naive Bayes, had an accuracy of 0.9633 and struggled particularly with recall, misclassifying many phishing emails.

In our error analysis, we identified several patterns that traditional models struggle with. Grammatical errors, mixed-language content, and leatspeak were often classified as non-phishing, while HTML code and phishing URLs were not accurately detected. Transformer models showed similar difficulties with phishing URLs and certain formal email formats, leading to misclassifications. Addressing these weaknesses is crucial for improving the model's performance.In summary this investigation demonstrates the effectiveness of transformer models in phishing detection and underscores the importance of detailed email content analysis to protect users from cyber threats.

6.2. Future Work

To further improve the results obtained in this study, future models should incorporate sentiment analysis techniques to detect social engineering tactics and to better understand the emotional tone and intent behind the email content. By identifying these elements, sentiment analysis could enhance the models ability to detect benign emails from phishing attempts. Improving the precision by being context-aware

Taking in account the results of this study, we propose developing a specialized NLP model for phishing detection. This model would take advantage of the transfer models and the inherited text processing techniques to accurately identify and filter phishing attempts text given information. We think that training a model that has good use of resources and has a significant preciseness can be applied and protect users from the ever-evolving phishing threat.

References

1. Laudon, K.; Traver, C. *E-commerce 2023: Business, Technology, Society*; Pearson, 2023.
2. Cellucci, N.; Moore, T.; Salaky, K. How Much Does Internet Cost Per Month? 2024. Accessed: 2024-9-15.
3. Kirchner, A.S.; Schilling, J. Email communication trends in the digital age. *Journal of Internet Services* **2022**, *10*, 123–130.
4. Blanchard, D.G. Cybersecurity challenges in the evolving landscape of online communication. *Cybersecurity Today* **2021**, *14*, 45–55.
5. of Investigation, F.B. 2022 Internet Crime Report. *FBI Annual Reports* **2022**, *12*, 5–10.
6. Smith, M. Identity theft and its economic impacts: A review. *Journal of Cybersecurity* **2021**, *15*, 18–25.
7. Verizon. 2023 Data Breach Investigations Report. *Verizon DBIR Reports* **2023**, *16*, 10–20.
8. Johnson, R. The role of email in modern application security. *Journal of Information Security* **2021**, *11*, 150–160.
9. Williams, P.; Hamilton, D. Email as a critical step in online authentication systems. *Cybersecurity Journal* **2022**, *18*, 22–30.
10. Group, A.P.W. 2022 Phishing Activity Trends Report. *APWG Annual Reports* **2022**, *7*, 2–15.
11. Anderson, D. The rise of phishing attacks in the digital era. *Cybersecurity Insights* **2022**, *19*, 30–40.
12. Liu, C. Phishing in the modern world: Beyond emails. *Journal of Digital Forensics* **2023**, *20*, 5–15.
13. Patel, N. The increasing effectiveness of phishing in social engineering attacks. *Information Security Review* **2023**, *24*, 50–60.

 doi:10.20944/preprints202410.1467.v1

14. Thompson, B. Social engineering in phishing: A review of strategies. *Journal of Information Warfare* **2021**, *17*, 35–45.
15. Patel, A.J. Text analysis as a defense against phishing emails. *Journal of Cybersecurity Technology* **2022**, *9*, 65–75.
16. Zhang, M.; Brown, D. CNN-based phishing email detection. *Journal of Machine Learning Security* **2021**, *12*, 100–110.
17. Wilson, J. Reducing phishing attacks through machine learning techniques. *Cybersecurity Advances* **2023**, *8*, 25–35.
18. Gonzalez, L. Malware and spyware in phishing attacks: A threat to personal data. *Journal of Information Security* **2022**, *21*, 12–20.
19. Yamamoto, S. Enhancing online security by mitigating phishing threats. *Cybersecurity Innovations* **2022**, *6*, 15–25.
20. Tang, L.; Mahmoud, Q.H. A survey of machine learning-based solutions for phishing website detection. *Machine Learning and Knowledge Extraction* **2021**, *3*, 672–694. doi:10.3390/make3030034.
21. Samad, A.S.; Balasubaramanian, S.; Al-Kaabi, A.S.; Sharma, B.; Chowdhury, S.; Mehbodniya, A.; Webber, J.L.; Bostani, A. Analysis of the performance impact of fine-tuned machine learning model for phishing URL detection. *Electronics* **2023**, *12*. doi:10.3390/electronics12071642.
22. Agrawal, G.; Kaur, A.; Myneni, S. A review of generative models in generating synthetic attack data for cybersecurity. *Electronics* **2024**, *13*. doi:10.3390/electronics13020322.
23. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Next-generation spam filtering: Comparative fine-tuning of LLMs, NLPs, and CNN models for email spam classification. *Electronics* **2024**, *13*. doi:10.3390/electronics13112034.
24. Salloum, S.; Gaber, T.; Vadera, S.; Shaalan, K. A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access* **2022**, *10*, 65703–65727. doi:10.1109/ACCESS.2022.3183083.
25. Altwaijry, N.; Al-Turaiki, I.; Alotaibi, A. Deep learning for phishing email detection. *Journal of Information Security and Applications* **2020**, *53*. doi:10.1016/j.jisa.2020.102526.
26. Newaz, I.; Jamal, M.K.; Hasan Juhas, F.; Patwary, M.J.A. A Hybrid Classification Technique using Belief Rule Based Semi-Supervised Learning. 2022 25th International Conference on Computer and Information Technology (ICCIT), 2022, pp. 466–471. doi:10.1109/ICCIT57492.2022.10055390.
27. Jamal, K.; Hossain, M.A.; Mamun, N.A. Improving Phishing and Spam Detection with DistilBERT and RoBERTa. *arXiv preprint* **2023**, [2311.04913]. doi:10.48550/arXiv.2311.04913.
28. Lee, Y.; Saxe, J.; Harang, R.E. CATBERT: Context-Aware Tiny BERT for Detecting Social Engineering Emails. *ArXiv* **2020**, *abs/2010.03484*.
29. Awe, T. Email Classification. https://www.kaggle.com/datasets/taiwoawe/email-classification, 2020. Accessed: 2024-02-01.
30. Tapakah68. Email Spam Classification. https://www.kaggle.com/datasets/tapakah68/email-spam-classification, 2023. Accessed: 2024-02-01.
31. Wiechmann, M. Enron Spam Data (No Code). https://www.kaggle.com/datasets/marcelwiechmann/enron-spam-data, 2001. Accessed: 2024-02-01.
32. LL, L. Fraud Email Dataset (has more than 4 codes for Naïve Bayes). https://www.kaggle.com/datasets/llabhishekll/fraud-email-dataset, 2018. Accessed: 2024-02-01.
33. Journal, S. Phishing Email Detection. https://www.kaggle.com/datasets/subhajournal/phishingemails/data, 2023. Accessed: 2024-02-01.
34. Mourya, S. Phishing-Mail. https://www.kaggle.com/datasets/somumourya/fishing-mail, 2023. Accessed: 2024-02-01.
35. Journal, S. Pishing Email Detection. https://www.kaggle.com/datasets/subhajournal/phishingemails, 2023. Accessed: 2024-02-01.
36. Jose. Pishing-2018 Monkey. https://monkey.org/~jose/phishing/, 2018. Accessed: 2024-02-01.
37. Jose. Pishing-2019 Monkey. https://monkey.org/~jose/phishing/, 2019. Accessed: 2024-02-01.
38. Jose. Pishing-2020 Monkey. https://monkey.org/~jose/phishing/, 2020. Accessed: 2024-02-01.
39. Jose. Pishing-2021 Monkey. https://monkey.org/~jose/phishing/, 2021. Accessed: 2024-02-01.
40. Jose. Pishing-2022 Monkey. https://monkey.org/~jose/phishing/, 2022. Accessed: 2024-02-01.

 doi:10.20944/preprints202410.1467.v1

41. Jose. Private-pishing4mbox. https://monkey.org/~jose/phishing/, 2018. Accessed: 2024-02-01.
42. Sivapragasam, A. Spam (or) Ham (Has a lot of cases for Naïve and Decision trees). https://www.kaggle.com/datasets/arunasivapragasam/spam-or-ham?select=spam+%28or%29+ham.csv, 2023. Accessed: 2024-02-01.
43. Naidu, C. Spam Classification for Basic NLP. https://www.kaggle.com/datasets/chandramoulinaidu/spam-classification-for-basic-nlp, 2020. Accessed: 2024-02-01.
44. Rhitazajana. Spam Email. https://www.kaggle.com/datasets/rhitazajana/spam-email, 2021. Accessed: 2024-09-15.
45. Olalekan, G. Spam_assasin. https://www.kaggle.com/datasets/ganiyuolalekan/spam-assassin-email-classification-dataset, 2021. Accessed: 2024-02-01.
46. Miltchev, R.; Rangelov, D.; Evgeni, G. Phishing validation emails dataset, 2024. doi:10.5281/zenodo.13474745.
47. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006.
48. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 ed.; Springer: New York, 2009.
49. Batutin, A. Choose Your AI Weapon: Deep Learning or Traditional Machine Learning? https://shelf.io/blog/choose-your-ai-weapon-deep-learning-or-traditional-machine-learning, 2023. Accessed: 2024-09-19.
50. Amatriain, X. Transformer models: an introduction and catalog. *ResearchGate* **2023**. doi:10.48550/arXiv.2302.07730.