# University of Victoria

# **Unveiling Phishing Patterns through Machine Learning**

# MTIS & MADS Capstone Project 2024

**Prepared by:**

Paulina Gonzalez (V01045753)

Jharna kumari (V01046956)

Jheel Chirag Shah (V01033440)

Adel Agha (V01046273)

Birva Patel (V01046694)

Bilel Matmti (V00947293)

Zixin Li (V01046098)

Summer Liu (V00911652)

## **Introduction**

## Phishing emails are one of the most frequent and cunning forms of cyberattacks in today’s digital age. These deceptive messages are camouflaged as legitimate communication from trusted entities, trying to trick users into revealing sensitive information, such as login credentials, financial data, or personal details [1]. Despite advancements in cybersecurity, phishing remains a relentless threat due to its continuously evolving tactics and the human element it targets [2].

## Understanding the common patterns, characteristics, and features of phishing emails is critical in defending against these malicious phishing attempts. Users and organizations that recognize these signs and red flags can improve their resilience and reduce the likelihood of falling victim to such scams [3]. It is important for users to be alert and verify the authenticity of messages to safeguard their digital assets and privacy. In this overview, we delve into the importance of recognizing phishing email traits and how being aware can fortify cybersecurity defenses.

## **How does Phishing work? [7]**

## **Common Patterns, Characteristics, and Features**

## Phishing emails frequently have recurring themes and tactics designed to deceive users into taking action or revealing sensitive information. Understanding these patterns is crucial for recognizing and preventing such attacks. Below are some examples of common features seen in phishing emails.

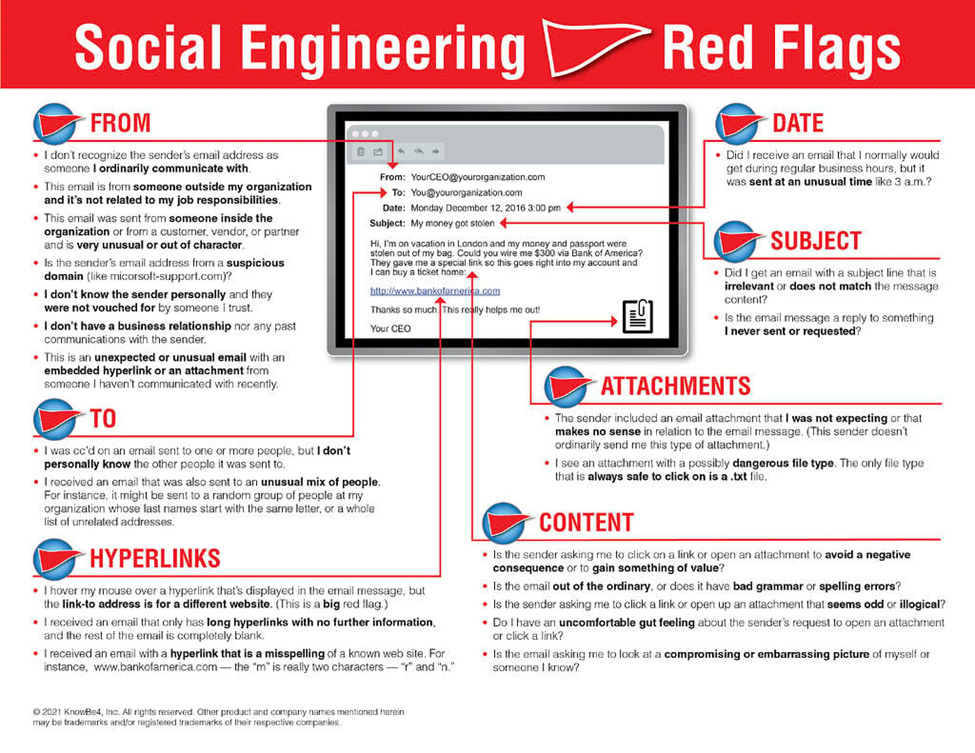
## Phishing emails often arrive unexpectedly and unsolicited [4]. Unlike legitimate emails that come from previous interactions or subscriptions, phishing messages target individuals indiscriminately or based on a demographic. Spoofed sender addresses are very common in phishing emails [5]. Malicious actors frequently impersonate trusted entities by using spoofed email addresses that appear legitimate at first glance. They tend to mimic the email domains of large companies, government services, or financial institutions to trick users into believing the message is authentic [5]. For example, there is an email claiming to be from "service@paypal.com" requesting urgent action from the user to change password, but once checking closer, the sender's address is misspelled as service@paypa1.com. Phishing emails will frequently request users to provide sensitive information, such as login credentials, account numbers, or personal details. These requests may be portrayed as an account verification, security update, or etc. to lure users into disclosing confidential data willingly [2].

## Phishers will commonly use generic greetings like "Dear Customer" or "Greetings User" instead of addressing users by their names. This lack of personalization is a common indicator of a phishing attempt [2]. For example, an introduction to an email could be "Dear Customer, your account has been compromised. Please click the link below to secure your account." Phishing emails will often use urgency or fear to get users to do some immediate action. They may threaten account suspension, provide time limited offers, or etc. to create a sense of urgency and get around critical thinking [5]. For example, the content of an email could be something like "Your Ebay account has been compromised! Click here to verify your identity and prevent account suspension."

## It is important to keep in mind that malicious phishing emails take advantage of various psychological factors in users to manipulate them into disclosing sensitive information. Phishing emails can create a sense of **urgency** by stressing immediate threats or time sensitive actions required to avoid negative consequences, this allows users to act quickly without carefully and properly evaluating the authenticity of the message [4]. Phishing emails can trigger **fear** by threatening severe consequences, such as account compromise, financial loss, or legal repercussions, unless the user follows all the instructions. Fear creates stress and anxiety, which hinders the user’s ability to assess the situation carefully and increases the probability of compliance with the malicious actor. Moreover, phishing emails can arouse **curiosity** by promising exclusive offers, secret information, or intriguing content as an attachment [4]. Curiosity attracts users to explore further, disregarding the potential risks associated with interacting with the unknown. Phishing emails may impersonate authoritative figures, such as executives, IT administrators, or government officials, to give credibility to the message and make it easier for the user to comply. **Authority** brings trust and loyalty, making users more susceptible to manipulation [2]. There are many other psychological triggers a malicious actor could use to exploit an individual, but the ones mentioned above are some that are most used across phishing emails.

## Phishing emails will commonly contain links disguised as legitimate websites but redirect to malicious pages designed to steal login credentials or install malware. These URLs could closely look like legitimate domains but contain slight variations [1]. For example, hovering over a link in an email claiming to be from the bank reveals an address like "rbc-login.com," whereas the legitimate website is "rbc.com." Another common indicator of phishing emails is frequent grammar errors, spelling mistakes, or inconsistent formatting. These flaws show a lack of professional polish and could potentially suggest the message is not from a reputable source. For example, "We noticed unusual activity in your account. Please log in to your account to verify your identity."

**Signs of a phishing email [6]**



## **Data Preprocessing**

## **Initial Data Handling**

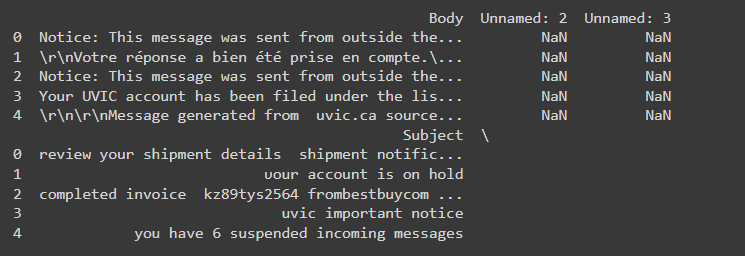
We began by loading the phishing email dataset using Pandas for data manipulation. An initial examination of the dataset revealed the structure and nature of the data. The dataset initially contained 2,576 entries with four columns. Two of these columns, 'Unnamed: 2' and 'Unnamed: 3', were entirely empty and thus provided no useful information for our analysis. Additionally, there were some missing entries in the 'Subject' and 'Body' columns which needed to be addressed.

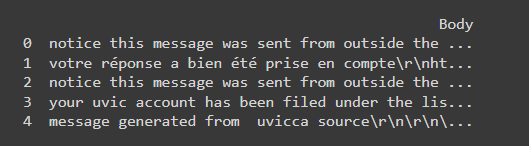
## **Data Cleaning**

Our data cleaning steps were methodical to ensure the dataset's integrity for further analysis.

* **Removing Irrelevant Columns:** The 'Unnamed: 2' and 'Unnamed: 3' columns were removed because they contained no data. This step simplified our dataset, focusing only on relevant information.
* **Handling Missing Data:** We found 109 missing entries in the 'Subject' column and 5 in the 'Body' column. Missing 'Subject' entries were filled with a single space to maintain uniformity, while rows with missing 'Body' entries were removed, considering the nature of this data for our analysis.
* **Text Normalization:** Both 'Subject' and 'Body' fields underwent normalization to convert all text to lowercase, remove special characters, and trim extra spaces. These steps standardize the format and reduce the complexity for text processing.

## **Feature Engineering**





**New features introduced into the dataset included:**

* **Body Length and URL Presence:** We calculated the length of each email's body and created a binary indicator for the presence of URLs, both of which are valuable features for phishing detection.

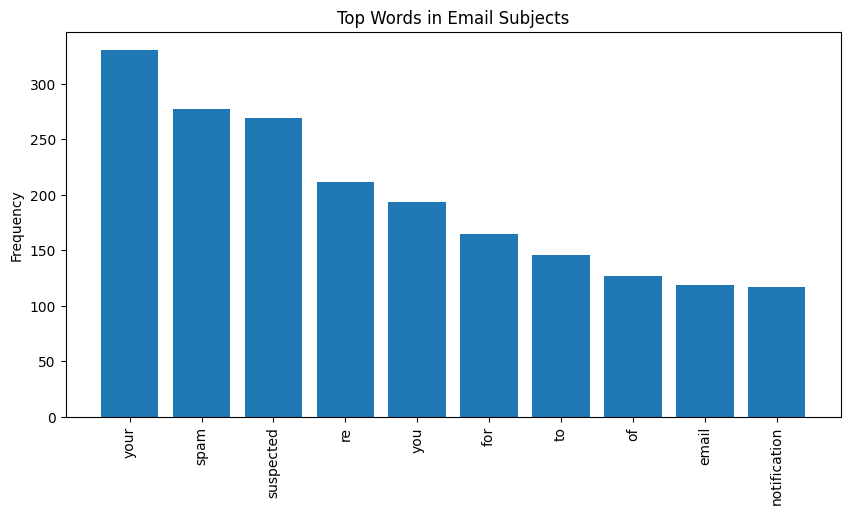
## **Visualizations**

* **Text Analysis and Visualization**

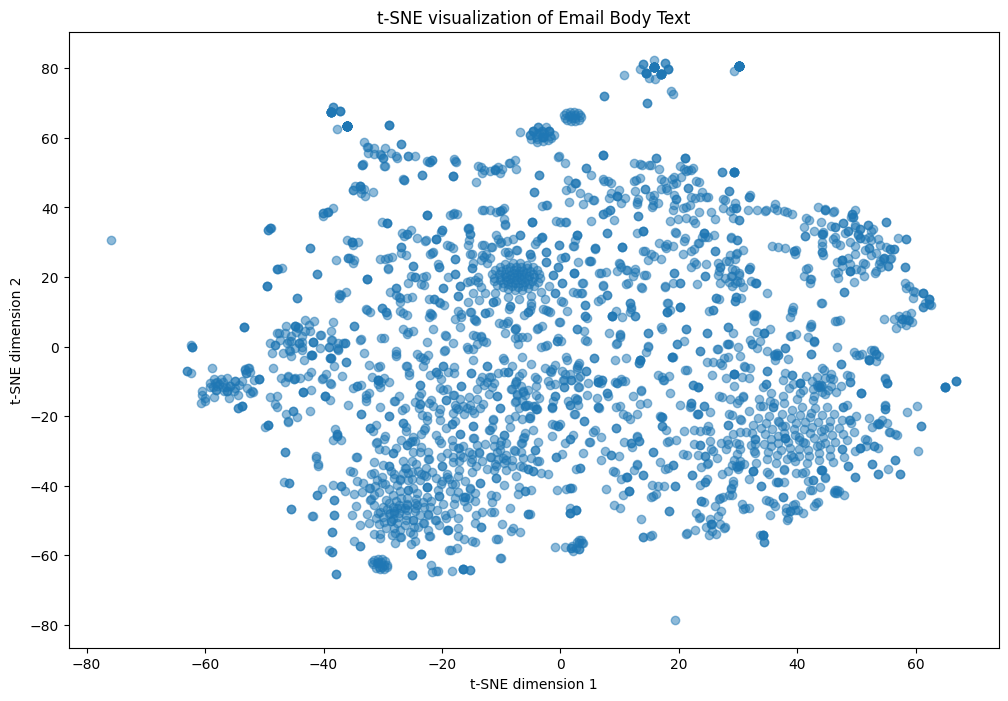
To gain deeper insights into the content of the emails, we analyzed the frequency of words after excluding common stopwords.

* **Bar Chart of Top Words in Email Subjects**

A bar chart was used to illustrate the most frequently occurring words in email subjects. This visualization helps in quickly identifying common patterns or keywords potentially indicative of phishing.



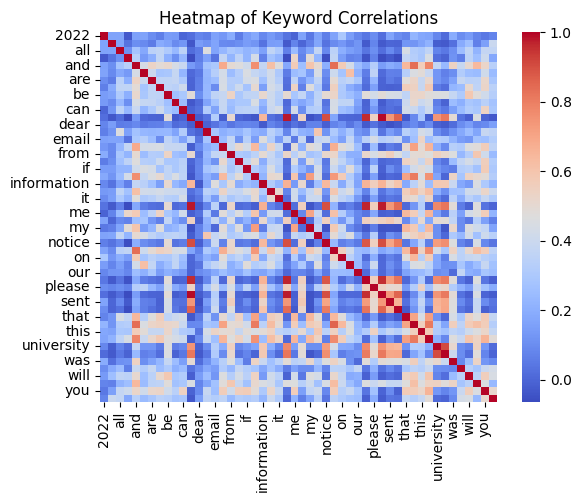
* **Interactive and Descriptive Visualizations**

****

Using t-SNE, we reduced the dimensionality of the TF-IDF vectorized text for visual inspection, which helped us visually assess clustering tendencies of the data points. An interactive Bokeh plot further enhanced this visualization by allowing for exploration of individual points with a hover tool

* **Keyword Correlation Heatmap**

The correlation heatmap provided insights into how certain keywords are related, which is crucial for identifying common phrases used in phishing

****

## **Implications and Recommendations**

## It is super important for people to know how to spot phishing emails so they can navigate the digital space securely. For companies, understanding how these phishing emails work helps them protect themselves better. If they do not address these insights, they could end up losing money, having their data stolen, and having their reputation damaged. Recommendations include implementing regular phishing awareness training for employees to improve resilience against phishing attacks [1]. Use strong email filtering systems and advanced threat detection technologies to identify and block malicious emails. Organizations should create defense strategies, incident response protocols and prevention tactics [5]. Emphasize security across all departments within an organization to remain vigilant and build culture.

## **Conclusion**

## In conclusion, our analysis of common patterns, characteristics, and features of phishing emails has provided essential insights into the deceptive tactics used by malicious actors. From spoofed sender addresses to exploiting psychological triggers, recognizing these red flags is paramount in securing and fortifying cybersecurity defenses. By educating users about signs and encouraging skepticism, organizations can empower individuals to identify and prevent phishing attempts effectively. In today's digital landscape, by staying informed and vigilant, individuals and organizations can navigate threats with confidence, safeguarding assets from the threat of phishing.

**ML techniques and recent DNN techniques**

1. **Deep Autoencoders**

**Introduction**

Autoencoders, a class of neural network models, are unsupervised models that are widely used in anomaly detection applications such as fraud, fault, intrusion, etc [8]. The technique is aimed at learning about compressed, distributed representation of a dataset for the purpose of feature extraction and dimensionality reduction of underlying patterns. The technique consists of two main parts: an encoder, a series of neural network layers that learn efficiently how to compress and encode the input data into latent space (lower dimension), and a decoder, a series of neural network layers that attempt to reconstruct the original input data representation. The idea is to recognize the most relevant aspects of the data in compressed form while ignoring noise (irrelevant information), and focusing on the features that are most relevant for distinguishing between spam and ham(non-spam) emails [8]. The reconstruction error, which measures the difference between the input and the output, is a fundamental concept in identifying anomalies. The network is designed to minimize the difference in the reconstructed error between the input and its reconstructed output. When the training phase commences, the autoencoder can learn a representation of what typical normal emails look like. When it encounters an email that deviates significantly from this norm, it will likely have a higher reconstruction error, signifying that a spam email has been detected. The deviation comes by setting a certain threshold for reconstruction error: if emails that are reconstructed with errors are greater than the threshold, then they can be flagged as spam [9].

**Variational Autoencoders**

Autoencoders are powerful, but they come with some limitations. They are prone to overfitting, where the model is coded to simply memorize the training data rather than learning to generalize to new data. Moreover, autoencoders can be limited by the size of the compressed representation; the model needs to have a balance between preserving the most relevant information in the input and minimizing the reconstruction error [9].

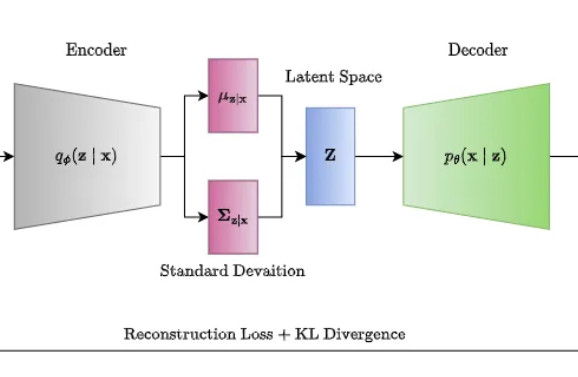


Image Source: Rosario et al. (2020) [11]

For this reason, Variational Autoencoders (VAE) have been introduced to overcome these limitations. They incorporate a probabilistic framework for generating the compressed representation of the input data. The encoder would work similarly as in normal encoders, except that instead of generating one point representing the latent space, it generates a probability distribution over the latent space [10]. This means that it outputs the parameters of a probability distribution, such as the mean and variance. This serves as a huge advantage because this probabilistic approach allows VAEs to learn a more structured and continuous latent space representation, which is useful for generative modeling. In addition, unlike traditional autoencoders, VAE introduces Kullback-Leibler (KL) divergence in the loss function. This term quantifies how far away the learned probability distribution is from the original, prior distribution (standard normal distribution) over the latent space. As a result, it ensures that the learned distribution is close to the prior distribution, helping to regulate the model for a meaningful structure. Another huge advantage of VAEs is that the model allows for generative modeling, which is when the model can generate new data points from the learned latent space distribution. Also, VAEs are less susceptible to overfitting than traditional autoencoders since the probabilistic nature of the encoding forces the model to learn a more robust representation of the data [9].

**Literature Review**

The use of an on demand technique like autoencoder for its powerful feature learning has shown some promising results in recent spam detection application research. A 2019 research work by Rosaria Silipo titled “Fraud Detection using a Neural Autoencoder” demonstrates a study on the autodetector algorithm being used to detect fraudulent activities. The study uses a dataset of legitimate credit card transactions, and utilizes a machine learning algorithm to reproduce the feature vector of each transaction. One successful technique that was proved to be suitable in the research is the autodetector, which is defined as a feed-forward multilayer neural network that reproduces the input data on the output layer. In the study, the technique was trained using the backpropagation algorithm against the mean squared error (MSE), a popular loss function. The study then further talks about the steps and process of implementing the technique, training it, and enabling it to detect new test cases. Python libraries such as Keras and TensorFlow were used for integration for the buildup of the encoder, as well as the Adam optimizer library for the backpropagation or learning phase of the algorithm [11].

Another study titled “Apply Stacked Auto-Encoder to Spam Detection” by Guyue Mi was explored to learn more about the capabilities of the autoencoder technique. It dives into the use case of the technique for spam detection applications and proves its accuracy and performance over the other unsupervised methods. 5 experiments were conducted, and the f score and accuracy measure metrics were used as the main criteria to determine the best algorithm. The study establishes the superiority of the autoencoder approach over other traditional methods such as Naive Bayes, Support Vector Machines, Decision Trees, Boosting, and Random Forest. This affirms the power of deep neural networks in identifying spam, promising an increase in their adoption of advanced email filtering solutions in the near future [12]. A blog from Medium by Abel G. Gebresilassie with the title “Neural Networks for Anomaly (Outliers) Detection” was explored. It outlines the steps of the whole process of the autodetector technique, going first through data preprocessing, then model building, and finally model evaluation and interpretation. The tutorial gives a good insight into the building blocks of the technique, and sets the initial groundwork for developers to expand and explore new ways [13].

1. **SVMs/One-Class SVM**

**Introduction**

*Theoretical Background:*Support Vector Machines (SVMs) are a set of supervised learning methods used for classification, regression, and outlier detection. SVMs are particularly known for their ability to create optimal hyperplanes in a multidimensional space that distinctly classifies data points into separate categories. Traditional SVMs rely on labeled training data to perform this classification. However, in many real-world applications, such as phishing detection, obtaining a comprehensive set of labeled data can be impractical or impossible due to the constantly evolving nature of phishing attacks and the scarcity of identified phishing examples.

*One-Class SVM for Unsupervised Anomaly Detection:* To address scenarios with predominantly unlabeled data, the One-Class SVM variant emerges as a useful tool. This method is tailored for unsupervised learning and is particularly adept at anomaly detection, identifying data points that deviate from the defined "normal" behavior. In the context of phishing detection, One-Class SVM is employed to model the characteristics of legitimate emails and, subsequently, to spot potential phishing emails as anomalies that do not conform to the model of normalcy.

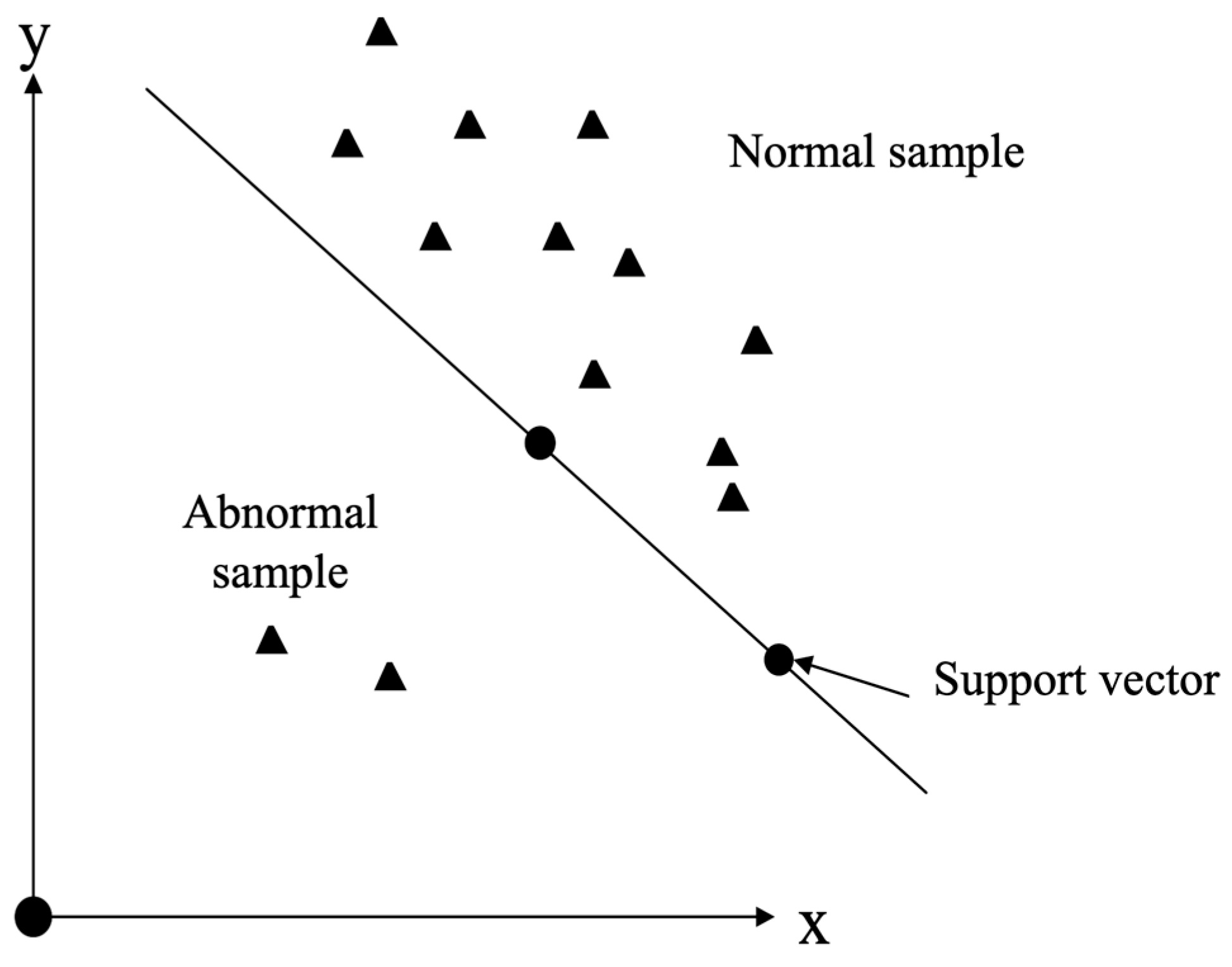


Image Source: Huang et al. (2023) [14]

**Literature Review**

*Relevance in Phishing Detection:* Phishing detection presents a unique challenge, the data is largely unstructured, imbalanced, and often unlabeled. One-Class SVM is relevant in this domain because it can learn from what is considered "normal" (non-phishing emails) without the need for a counterpart set of phishing examples. This is crucial in phishing detection, where new and unknown phishing strategies are constantly developed, and labeled examples of such attacks may not be available during training. " By leveraging One-Class SVM, cybersecurity systems can be equipped to detect phishing attempts by identifying emails that diverge from the typical patterns of legitimate communication. This capability is particularly valuable as it provides a line of defense against new and emerging phishing techniques that have not yet been labeled or cataloged " [15].

*Mathematical Foundation:*One-Class SVM is grounded in the principles of statistical learning theory, operating on the foundation of maximizing the margin between different classes of data. Unlike its counterparts, the One-Class SVM is designed to identify a single class, delineating the smallest hypersphere in a feature space that encloses the majority of the data points considered 'normal'. This approach is known as the maximum margin estimator for the dataset's support. Central to the efficacy of One-Class SVM in handling complex and high-dimensional data is the use of kernel functions. These functions implicitly map the input data into a higher-dimensional space without explicitly performing the transformation, thereby allowing the One-Class SVM to construct nonlinear boundaries. Through kernels like the Radial Basis Function (RBF), the algorithm can recognize and adapt to the intricate and abstract data distributions that are typical in textual data like emails, which is why it proves valuable in phishing detection where the delineation between normal and anomalous is often subtle and not linearly separable.

*Applying One-Class SVM in Phishing Detection:*A study titled "Anomaly Detection in Emails using Machine Learning and Header Information"[16] demonstrates the effectiveness of One-Class SVM in identifying email anomalies, particularly phishing and spam. This research primarily analyzes the content of the email body and subject by instead focusing on email header information. By extracting features solely from headers, the study addresses the challenges posed by the varied linguistic styles found in email content, aiming for a language-independent anomaly detection method. The findings from this study guide us to different research directions, such as the development of hybrid models that combine One-Class SVM with other machine learning techniques to enhance predictive accuracy and reliability.

*Future Directions:*SVMs represent a robust and versatile technique for detecting phishing emails, offering powerful mechanisms for handling both labeled and unlabeled data. When working with labeled datasets, traditional SVMs provide a highly effective tool for classification, capable of distinguishing between phishing and legitimate emails with a high degree of accuracy. In scenarios where labeled data is scarce or entirely absent, One-Class SVM adapts to these challenges by focusing on anomaly detection, modeling what normal email behavior looks like, and flagging deviations from this norm as potential phishing attempts.

For situations involving unlabeled datasets, an additional option or strategy could involve using a separate labeled dataset to train the model. This allows the trained model to then be applied to the unlabeled dataset, facilitating the identification of phishing emails through supervised learning techniques. Such an approach not only leverages the strengths of SVMs under different data conditions but also enhances the model’s applicability and accuracy by providing it with foundational knowledge from labeled examples.

As we advance in our project, the choice between using a traditional SVM or a One-Class SVM will highly depend on the nature of our dataset. The availability of labeled data will guide our initial approach, but regardless of the path chosen, our next steps involve a thorough evaluation and integration of other machine learning and deep learning algorithms to enhance our detection capabilities.

1. **K-Means**

**Introduction**

The K-means clustering algorithm is a popular unsupervised learning technique used widely in data mining and machine learning to partition a dataset into K distinct clusters. K-means enhances the understanding of data distribution and structure by grouping data based on feature similarity, which would be particularly useful in phishing email detection. In a real application, K-means can identify and categorize emails into clusters based on specific features like content, sender information, usage patterns, etc. By learning from the clustering results, it becomes possible to identify potential phishing emails and reduce the risk of security breaches.

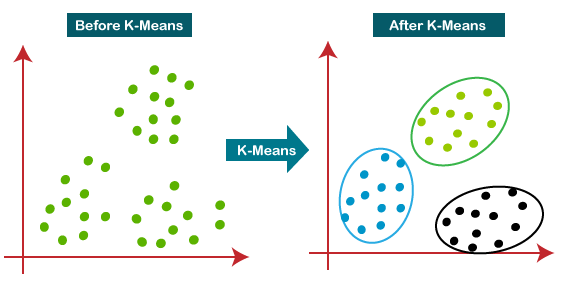
****

Image Source:K. Thareja (2023) [17]

**Literature Review**

Two related papers can help in understanding and implementing K-means better. The first research paper [18] by M. Basavaraju and R. Prabhakar “*Novel Method of Spam Mail Detection Using Text-Based Clustering Approach*” presents a text-based clustering method for effectively classifying spam mail. The K-means clustering technique is focused on in this paper due to its scalability and grouping efficiency. The core methodology used in this paper makes use of the vector space model, which treats every email as a vector in a multidimensional space, making it easier to apply K-means for efficient grouping.

Further, the paper “*Detection of Phishing Attacks: A Machine Learning Approach”* [19] written by Basnet, Mukkamala, and Sung concentrates on using machine learning techniques to detect phishing emails, which is one of the most prevalent problems nowadays. ⁤⁤This paper explores and compares the use of several machine learning algorithms, including Support Vector Machines (SVM), Neural Networks, Self-Organizing Maps (SOMs), and unsupervised K-means clustering, to improve the efficiency of the detection of such emails, Each technique is applied to the dataset, which contains both phishing and normal emails.

K-means clustering is an unsupervised learning algorithm where class labels are unknown. This method is used to explore inherent groupings within the data that could potentially distinguish phishing from normal emails. The authors describe the algorithm’s iterative process of assigning emails to clusters based on feature similarity to improve the model's ability to generalize from unlabeled data [19]. This is crucial for phishing detection because such behaviors always require adaptation to unseen characteristics.

The paper [19] also provides a quantitative evaluation of the machine learning models used. The K-means algorithm and other algorithms are assessed based on the accuracy of the classification. The results demonstrate that K-means has a good accuracy level, which emphasizes its feasibility for the detection of phishing emails and its utility in situations that lack labeled data.

In all, by using K-means clustering in these papers, the studies not only demonstrate the adaptability of unsupervised learning algorithms K-means in cybersecurity but also encourage the refinement of these techniques in further research.

Those papers are particularly relevant for researchers focusing on the application of clustering algorithms like K-means in the field of cybersecurity. They offer foundational approaches to using K-means for spam detection, providing a valuable resource for enhancing email security protocols and developing more efficient anti-spam systems.

## References

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

[2] R. Dhamija, J. D. Tygar and M. Hearst, "Why phishing works," in CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montreal, 2006.

[3] IBM, "What is phishing?," IBM, [Online]. Available: <https://www.ibm.com/topics/phishing>. [Accessed 7 April 2024].

[4] Federal Trade Commision, "How to Recognize and Avoid Phishing Scams," Federal Trade Commision, September 2022. [Online]. Available: <https://consumer.ftc.gov/articles/how-recognize-and-avoid-phishing-scams>. [Accessed 7 April 2024].

[5] Phishing.org, "What Is Phishing?," KnowBe4, [Online]. Available: <https://www.phishing.org/what-is-phishing>. [Accessed 7 April 2024].

[6] Z. Alkhalil, C. Hewage, L. Nawaf and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," Frontiers in Computer Science, vol. 3, 2021.

[7] B. K. Jena, "What is Phishing Attack in Cyber Security - Complete Guide," simplilearn, 15 September 2023. [Online]. Available: <https://www.simplilearn.com/tutorials/cryptography-tutorial/what-is-phishing-attack>. [Accessed 7 April 2024].

[8] U. Michelucci, “An introduction to autoencoders,” arXiv (Cornell University), Jan. 2022, doi: 10.48550/arxiv.2201.03898. [Online]. Available: <https://www.jeremyjordan.me/autoencoders/>

[9] J. Rocca, “Understanding Variational Autoencoders (VAEs)”, Towards Data Science, [Online].Available:<https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73>

[10] A. J, “Variational autoencoder based anomaly detection using reconstruction probability,” SNU Data Mining Center, vol. 2, p. 1, Jan. 2015, [Online]. Available: <https://cir.nii.ac.jp/ja/crid/1370848655242575246>

[11] Rosaria Silipo, Kathrin Melcher, and Maarit Widmann, “Fraud Detection Using a Neural Autoencoder,” *arXiv (Dataversity)*, April. 2019, [Online]. Available: [Fraud Detection Using a Neural Autoencoder - DATAVERSITY](https://www.dataversity.net/fraud-detection-using-a-neural-autoencoder/)]

[12] G. Mi, Y. Gao, and Y. Tan, “Apply Stacked Auto-Encoder to spam detection,” in Lecture notes in computer science, 2018, pp. 3–15. doi: 10.1007/978-3-319-20472-7\_1.]

[13] [Abel G. Gebresilassie, “Neural Networks for Anomaly (Outliers) Detection,” *Good Audience(Medium)*, June. 2018, [Online]. Available: [Neural Networks for Anomaly (Outliers) Detection | by Abel G. Gebresilassie | Good Audience](https://blog.goodaudience.com/neural-networks-for-anomaly-outliers-detection-a454e3fdaae8) ]

[14] Huang, G., Chen, J., & Liu, L. (2023). One-Class SVM Model-Based Tunnel Personnel Safety Detection Technology. Applied Sciences, 13(3), 1734. Image retrieved from <https://www.mdpi.com/applsci/applsci-13-01734/article_deploy/html/images/applsci-13-01734-g001.png>

[15] Altaher, A. (2017). Phishing Websites Classification using Hybrid SVM and KNN Approach. International Journal of Advanced Computer Science and Applications, 8(6). Retrieved from <http://thesai.org/Publications/ViewIssue?volume=8&issue=6&code=IJACSA>.

[16] C. P. Beaman and H. Isah, "Anomaly Detection in Emails using Machine Learning and Header Information," ArXiv, abs/2203.10408, 2022. [Online]. Available: <https://arxiv.org/abs/2203.10408>

[17] K. Thareja, “K-means clustering & it’s real use-case in the security domain.,” Medium, https://khushithareja.medium.com/k-means-clustering-its-real-use-case-in-the-security-domain-a9ac265167dc (accessed Apr. 17, 2024). Image retrieved from <https://khushithareja.medium.com/k-means-clustering-its-real-use-case-in-the-security-domain-a9ac265167dc>]

[18] M. Basavaraju and Dr. R. Prabhakar, “A novel method of spam mail detection using text based clustering approach,” *International Journal of Computer Applications*, vol. 5, no. 4, pp. 15–25, Aug. 2010. doi:10.5120/906-1283

[19] R. Basnet, S. Mukkamala, and A. H. Sung, “Detection of phishing attacks: A machine learning approach,” Soft Computing Applications in Industry, pp. 373–383, Feb. 2008. doi:10.1007/978-3-540-77465-5\_19