Comparative Analysis of Phishing Email Classifiers

Ashwin Venkatakrishnan
ashvenk@iu.edu

Prateek Giridhar
pgiridha@iu.edu

Prinston Rebello
prebello@iu.edu

project-ashvenk-pgiridha-prebello

Abstract

In an increasingly interconnected world, the proliferation of spam and phishing emails has presented a pressing challenge for individuals and organizations. The need for efficient email classification has led to numerous attempts to combat this issue using machine learning algorithms. This project endeavors to contribute to this ongoing effort by implementing novel methods acquired during the current semester's coursework. Our primary objective is to develop a robust email classification system capable of categorizing incoming emails as one of three types: "spam," "genuine", or "phishing." Leveraging the techniques and knowledge gained in our recent coursework, the project primarily focuses on textual analysis. By harnessing the power of data mining, we aim to construct a classification model that can effectively discern the nature of incoming emails. This project is poised to make a meaningful contribution to email security in the modern digital landscape. The ability to automatically identify and separate spam, non-spam, and phishing emails can improve user experiences and enhance cybersecurity. We envision that our data mining approach, informed by the methods learned in the semester, will not only deepen our understanding of email classification but also serve as a valuable tool in the ongoing battle against email-based threats.

Keywords

email classification, phishing, spam, exploratory data analysis, principal component analysis, decision tree, natural language processing

1 Introduction

In an age where digital communication is the lifeblood of modern society, the battle against email-based threats such as spam, phishing, and their ilk rages on. While email providers like Outlook and Gmail employ algorithms to filter unwanted messages, false positives remain an enduring challenge. Consider, for instance, legitimate emails from educational institutions with essential links, unjustly relegated to the spam folder. The core objective of our project is to advance the state of email classification by meticulously categorizing incoming emails into three distinctive groups: spam, genuine, and phishing. Our approach aims to minimize false positives by extracting textual information and insightful inferences. Our possible future work is to predict the probability of the phishing link being clicked on. Our present approach includes Exploratory Data Analysis (EDA) for pre-processing, Principal Component Analysis (PCA) if necessary, text extraction using MinHash, text analysis using Natural Language Processing

(NLP) techniques, and the generation of relationships and rules from extracted text using Apriori algorithm. Using these techniques, we will transform the original dataset to align with our objective - showcase the accuracy of algorithms such as K-Nearest Neighbors (KNN), Decision Trees, and Naive Bayes in classifying emails with minimal false positives.

Previous work

For a comprehensive understanding, consider consulting journal papers [6], [2], [5], [7], [4], [1], and the findings in the conference paper by [3], [8].

2 Methods

Our project will rely on the dataset available on Kaggle, specifically the Phishing Emails Dataset. This dataset provides essential information, including the email body and labels classifying emails as safe or unsafe. Our primary focus will be on the subset of data marked as unsafe.

The project commences with Exploratory Data Analysis (EDA) of the unsafe email dataset. The primary objective of this analysis is to identify any correlations or similarities between unsafe and safe emails. This process plays a pivotal role in our pursuit of reducing false positives in email classification.

In the course of EDA, we will employ techniques to scrutinize the data for patterns and relationships. Notably, we will leverage MinHash to gauge similarities between emails and flag any anomalies. Furthermore, we will meticulously inspect the dataset for attribute violations that might have led to misclassifications.

Following EDA, our project will transition to text extraction and analysis using NLP techniques. The aim is to extract meaningful insights from the email content. Additionally, we will employ the Apriori algorithm to generate rules based on the extracted text. These rules will serve as a foundation for our classification efforts.

We will explore various classification models to categorize emails accurately. Specifically, we will employ the Decision Tree algorithm as the cornerstone of our classification approach. A subset of the original dataset will be utilized to implement K-Nearest Neighbors (KNN) and Naive Bayes classifiers. Comparing the results from these three approaches will allow us to assess their respective accuracy. The project's outcome will be presented in a contingency table, showcasing the accuracy of each classification model. This comparative analysis will shed light on the effectiveness of our methodology.

In summary, our project integrates data mining, NLP, and machine learning techniques to enhance email classification precision. By carefully analyzing the dataset, generating rules, and implementing classification models, we aim to minimize false positives and contribute to the ongoing battle against email-based threats.

References

[1] Paul Graham. A plan for spam. `http://www.paulgraham.com/spam.html`, 2002.

[2] Asif Karim, Sami Azam, Bharanidharan Shanmugam, Krishnan Kannoorpatti, and Mamoun Alazab. A comprehensive survey for intelligent spam email detection. *IEEE Access*, 7:168261–168295, 2019.

[3] Sohail Ahmed Khan, Wasiq Khan, and Abir Hussain. Phishing attacks and websites classification using machine learning and multiple datasets (a comparative analysis). In *Intelligent Computing Methodologies: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part III 16*, pages 301–313. Springer, 2020.

[4] Sabina Kleitman, Marvin KH Law, and Judy Kay. It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling. *PloS one*, 13(10):e0205089, 2018.

[5] Gilchan Park and Julia M Taylor. Using syntactic features for phishing detection. *arXiv preprint arXiv:1506.00037*, 2015.

[6] Said Salloum, Tarek Gaber, Sunil Vadera, and Khaled Shaalan. A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access*, 10:65703–65727, 2022.

[7] Priyanka Verma, Anjali Goyal, and Yogita Gigras. Email phishing: Text classification using natural language processing. *Computer Science and Information Technologies*, 1(1):1–12, 2020.

[8] John Yearwood, Musa Mammadov, and Arunava Banerjee. Profiling phishing emails based on hyperlink information. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 120–127. IEEE, 2010.