Capstone Project 2025

Phishing Email Detection through Machine Learning: Pattern Analysis, Feature Extraction, and Actionable Insights

Sohpal Shaveta, shavetasohpal@uvic.ca
Zoe Zhao, yixinzhao@uvic.ca
Yashkumar Ashokbhai Sabhadiya, ysabhadiya@uvic.ca
Charina Regis, cregis@uvic.ca
Chengkai Yang, chengkaiyang@uvic.ca
Peter Ayeni, payeni@uvic.ca
Lei Wang, leiwang1012@uvic.ca
Archana Radhakrishnan Sreevidhya, archanaradhakrishnan@uvic.ca

Department of Electrical & Computer Engineering
University of Victoria, Victoria

Territory Acknowledgment

We acknowledge with respect the Lekwungen-speaking peoples on whose traditional territory the university stands and the Songhees, Esquimalt and WSÁNEý peoples whose historical relationships with the land continue to this day.

Contents

01 Data Collection & Literature Review

Provided spam email datasets were carefully selected based on their unique attributes that included subject, and body only.

02 Data Preprocessing

Tokenization and text cleaning

03 Feature Extraction

Textual data within the email transformation into a numerical format suitable for machine learning algorithm

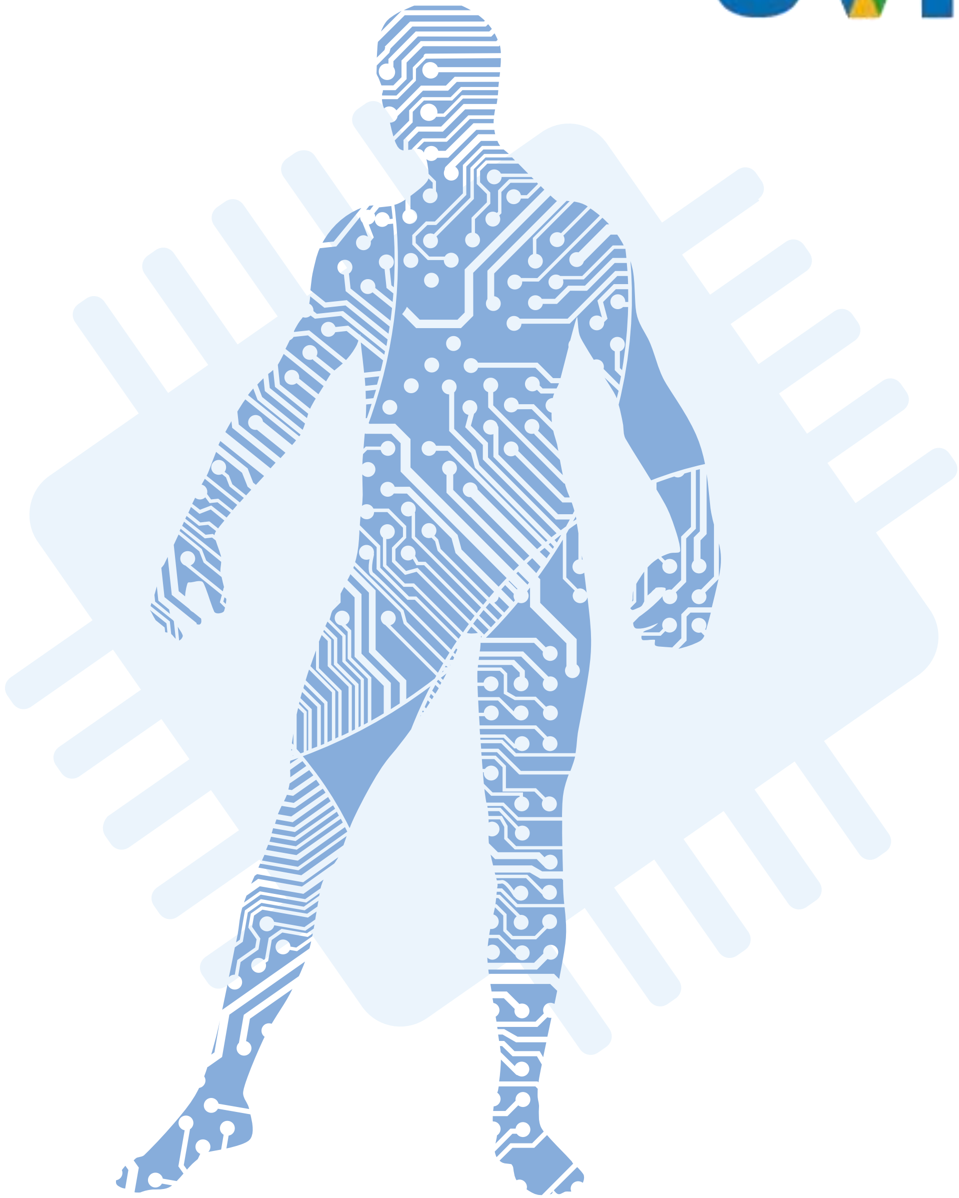
04 Data Classification

ML algorithm for classification .

05 Visualization

Performance metrics

06 Future Recommendations



Worklog of Team-Work

Sohpal Shaveta
Zoe Zhao

01

- Literature Review
- Conclusions
- Presentation
- Presenter

Time Spent-30 hr

Yash Kumar
Charina Regis
Chengkai Yang

02-03

- Data Analyze
- Statistics
- --------

Time Spent-15 hr

Archana Sreevidhya
Peter Ayeni
Lei Wang

04-05

- Introduction
- Visualization
- ---------

Time Spent-15 hr

Flow of Activities

Basis of Literature Review-Objectives

Data Summary and Visualization , and Literature Review

Data Summary & Literature Review

Completed

Cleaning text
Removing stop words
Tokenization
Label encoding
TF-IDF / Bag of Words

Data Preprocessing & feature Extraction

Cleaning text

Logistic Regression
Random Forest
SVM / Naive Bayes
Model Training & Testing

Data Classific-ation and Visualiz-ation

Flow of Activities

Basis of Literature Review-Objectives

Accuracy
Precision
Recall
F1-score

Performances Metrics

Pattern & Feature Analysis
Identify common phishing indicators
Recommendations

Recommend-ation for Phishing Detection

Time Activity Chart

Content-01
Completed

Content-02
In Progress
Complete-22nd April

Content-03
Complete-30th April

Content-04
Complete-25th May

Content-05
Complete-30th June

Content-06
Complete-30th July

Objectives

3.Generate insights and visualizations to interpret the findings and provide actionable recommendations for phishing detection and prevention.

2.Identify common patterns, features, and characteristics indicative of phishing emails.

1.Utilize machine learning algorithms to analyze the text of email bodies and subjects in a phishing dataset

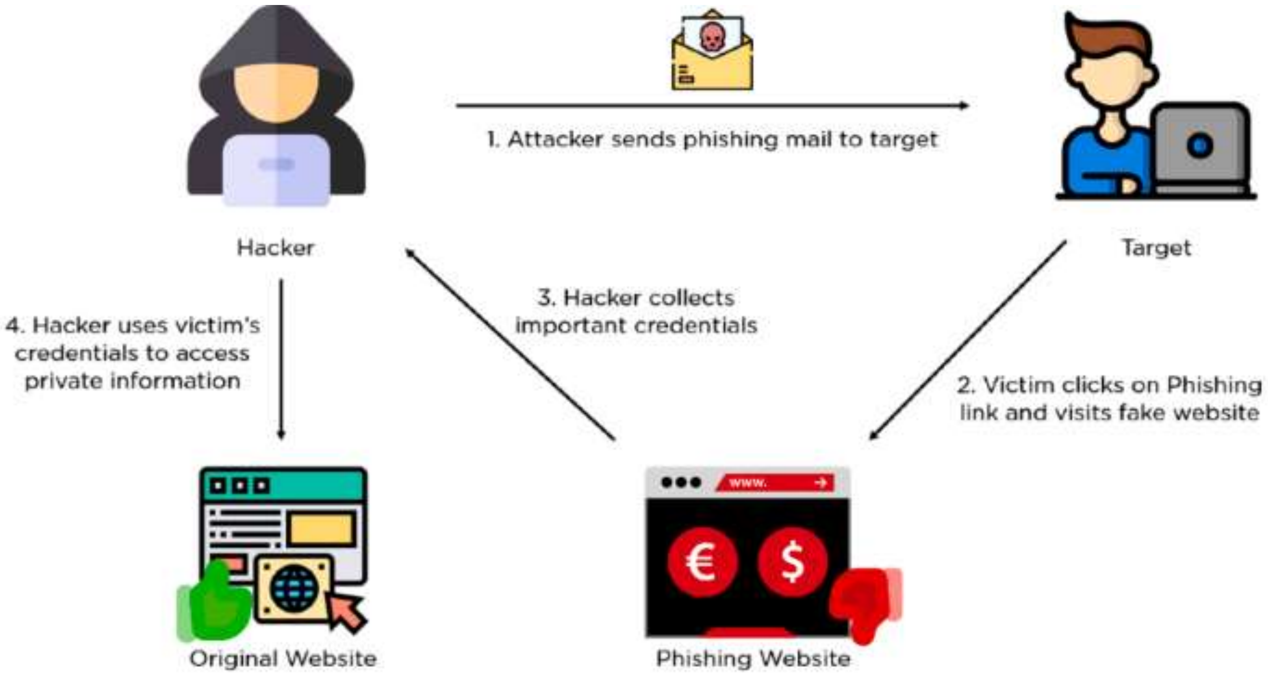
Title of Project

Phishing Email Detection through Machine Learning: Pattern Analysis, Feature Extraction, and Actionable Insights

Introduction

- In today's digital landscape, phishing remains a widespread and damaging cyber threat. These deceptive emails aim to extract sensitive information or compromise system security.

- This project leverages machine learning to analyze phishing email datasets, uncover textual patterns in subjects and bodies, and generate insights that support effective detection and enhance security awareness.



Source: Simplilearn and B. Kumar, "How Does a Phishing Attack Work?" Mar. 2023. [https://www.simplilearn.com/ice9/free_resources_article_thumb/phishing_working_2-What_Is_Phishing.PNG.

Literature Review

Author	Dataset	Method	Result (in %)
Jamal et al. 2024 [1]	Unspecified • 936 spam • 4825 ham	IPSDM BERT-based models (DistilBERT, RoBERTA)	Accuracy: 98.99 No other metrics available
Somesha et al. 2024 [2]	Nazario and SpamAssassin	Transformer based model	Accuracy: 99.51 No other metrics available
Atawneh et al. 2023 [3]	Enron, SpamAssassin, UCI	BERT, LSTM	Accuracy: 99.61 Precision: 99.87 Recall: 99.23 F1-score: 99.55 No other metrics available
Gholampour et al. 2023 [4]	Generated by GPT 2	K-Nearest Neighbor	Accuracy: 94.00 No other metrics available
Alhogail et al. 2021 [5]	CLAIR collection of fraud email [13] • 3685 spam • 4894 ham	Convolution Network (GCN) and NLP techniques (tokenization, stop word removal)	False positive rate: 0.015 Graph Accuracy: 98.2 No other metrics available
Abdul Nabi et al. 2021 [6]	Spam Base, and Spam Filter Data (Kaggle), • 3000 spa • 2000 ham	BERT transformer	Fine tune Accuracy: 98.67 F1-score: 98.66 No other metrics available

Literature Review

Author	Dataset	Method	Result (in %)
Lee et al. 2021 [7]	EES 2020 Dataset (Private)	BERT, CNN + LSTM	AUPRC: 98.51 Recall: 76.48 No other metrics available
Gangavarapu et al. 2020 [8]	SpamAssassin, Nazario • 3051 (2 class) • 3344 (2 class) • 3844 (3 class)	Random forest with fi-based feature selection	Accuracy: 98.40 No other metrics available
Gibson et al. 2020 [9]	Ling, Enron, PUA, SpamAssassin (separately) • 20,170 spam • 16,545 ham	Genetic Algorithm with SGD (GA-SGD)	Accuracy: 99.21 Precision: 98.68 Recall: 99.54 No other metrics available
Fang et al. 2019 [10]	Unspecified	RCNN using multilevel vectors and attention mechanisms with Word2Vec	Accuracy: 99.00 No other metrics available
Arif et al. 2018 [11]	Smart home dataset For sentiment analysis • SMS spam (5575 samples) • tweets (2034 samples)	XGBoost, bagged model, and generalized linear model with stepwise feature selection	Accuracy: 91.80 No other metrics available
Hijawi et al. 2017 [12]	SpamAssassin [14] • 1000 spam • 5051 ham	(MLP) Naive Bayes, random forest, and decision tree	Accuracy: 99.30 No other metrics available

Dataset

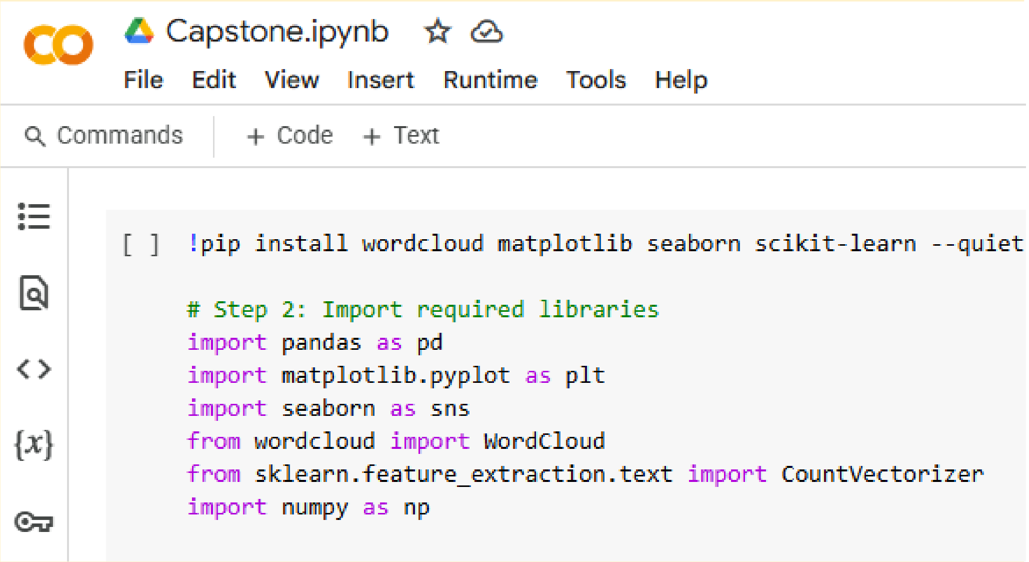
Dataset	Dataset Overview	Remarks
Dataset (Private)	**Total entries:** 2,576 **Columns:** Subject: Email subject line (2,467) Body: Email body (2,571) Unnamed: 2, Unnamed: 3: Empty columns (to be removed)	Provided email datasets were carefully selected based on their unique attributes that included subject, and body only. This dataset underwent a merging process to create a unified dataset for analysis.

```
Result
                                        Subject  \
0  ®Review your shipment details / Shipment Notif...
1                          Your account is on hold
2  Completed: Invoice # KZ89TYS2564 from-Bestbuy....
3                             UVic IMPORTANT NOTICE!
4          You have (6) Suspended incoming messages

                                              Body
0  Notice: This message was sent from outside the...
1  \r\nVotre réponse a bien été prise en compte.\...
2  Notice: This message was sent from outside the...
3  Your UVIC account has been filed under the lis...
4  \r\n\r\nMessage generated from  uvic.ca source...
```

Dataset Analysis

<table>
<tr><th>Dataset</th><th>Execution</th><th>Programme</th></tr>
<tr>
<td>Dataset (Private)</td>
<td>We utilized Google Colab, a cloud-based platform that enables the execution of Python code within a web browser, to carry out our initial dataset analysis. As part of the process, we performed the following steps:

❖ Uploaded the dataset directly into the Colab environment
❖ Loaded the data into a Pandas DataFrame for easy manipulation and analysis
❖ Examined the structure of the dataset by displaying the number of rows and columns
❖ Reviewed the column names along with their corresponding data types
❖ Checked for any missing or null values across the dataset</td>
<td>
</td>
</tr>
</table>

Dataset Analysis

Result-I (Statistics)

Capstone.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text

```
Dataset Dimensions:
Rows: 2576, Columns: 4

Data Types and Non-Null Counts:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2576 entries, 0 to 2575
Data columns (total 4 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Subject         2467 non-null   object
 1   Body            2571 non-null   object
 2   Subject_Length  2576 non-null   int64
 3   Body_Length     2576 non-null   int64
dtypes: int64(2), object(2)
memory usage: 80.6+ KB
None

Missing Values per Column:
Subject          109
Body               5
Subject_Length     0
Body_Length        0
dtype: int64
```

Result-II (Statistics)

Capstone.ipynb

File Edit View Insert Runtime Tools Help

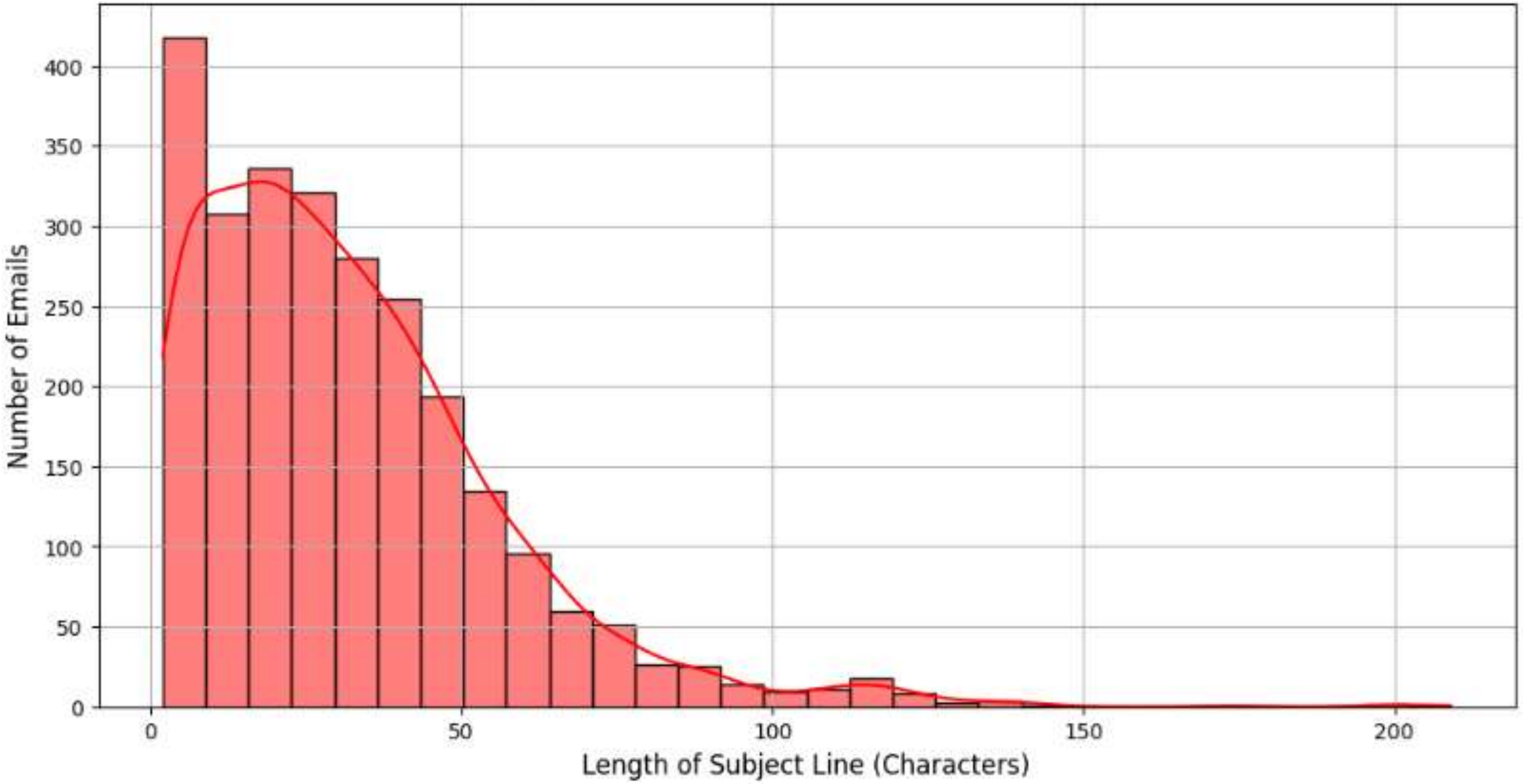
Commands + Code + Text

```
        Subject_Length   Body_Length
count      2576.000000   2576.000000
unique             NaN           NaN
top                NaN           NaN
freq               NaN           NaN
mean         31.916537    804.414596
std          24.698988   1096.254825
min           2.000000      2.000000
25%          14.000000    289.000000
50%          27.000000    527.500000
75%          44.000000    934.000000
max         209.000000  16423.000000

 Unique Values in Each Column:
Subject: 1939 unique values
Body: 2283 unique values
Subject_Length: 127 unique values
Body_Length: 1242 unique values
```

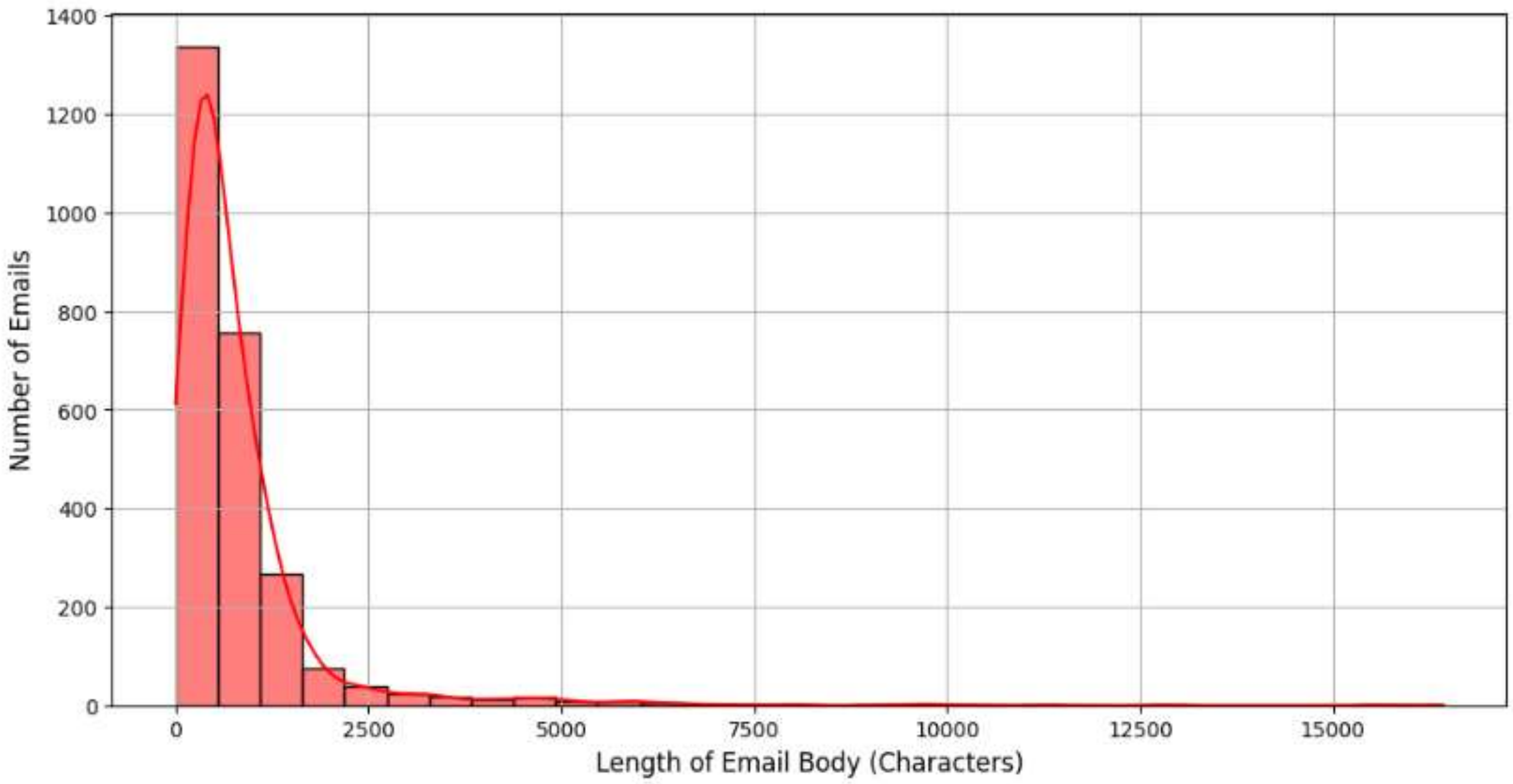
Dataset Analysis

Subject Length Distribution



Dataset Analysis

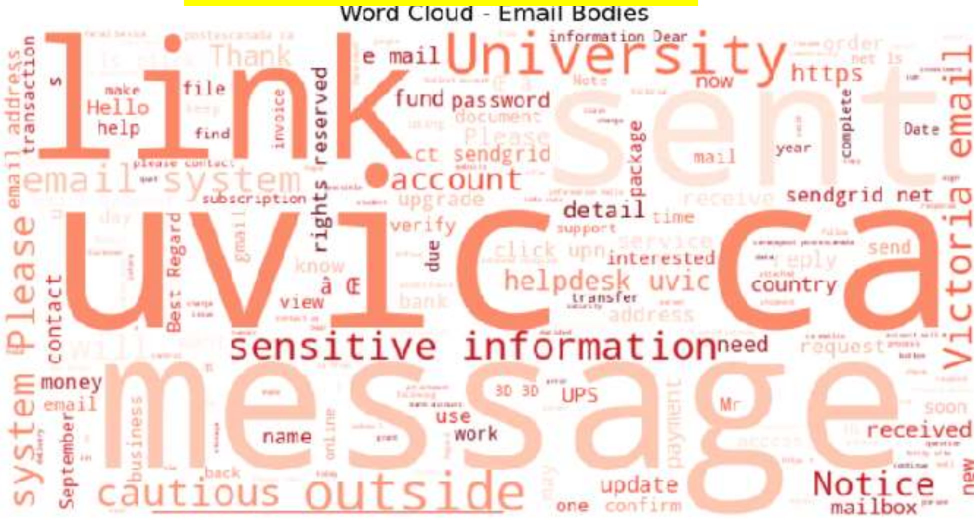
Body Length Distribution



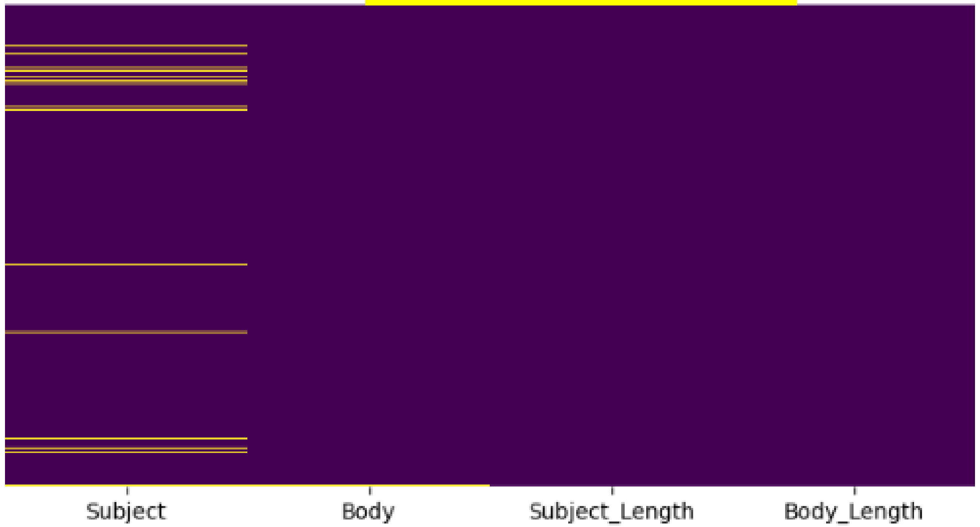
Dataset Analysis

Word-Email Subject

Word-Email Bodies

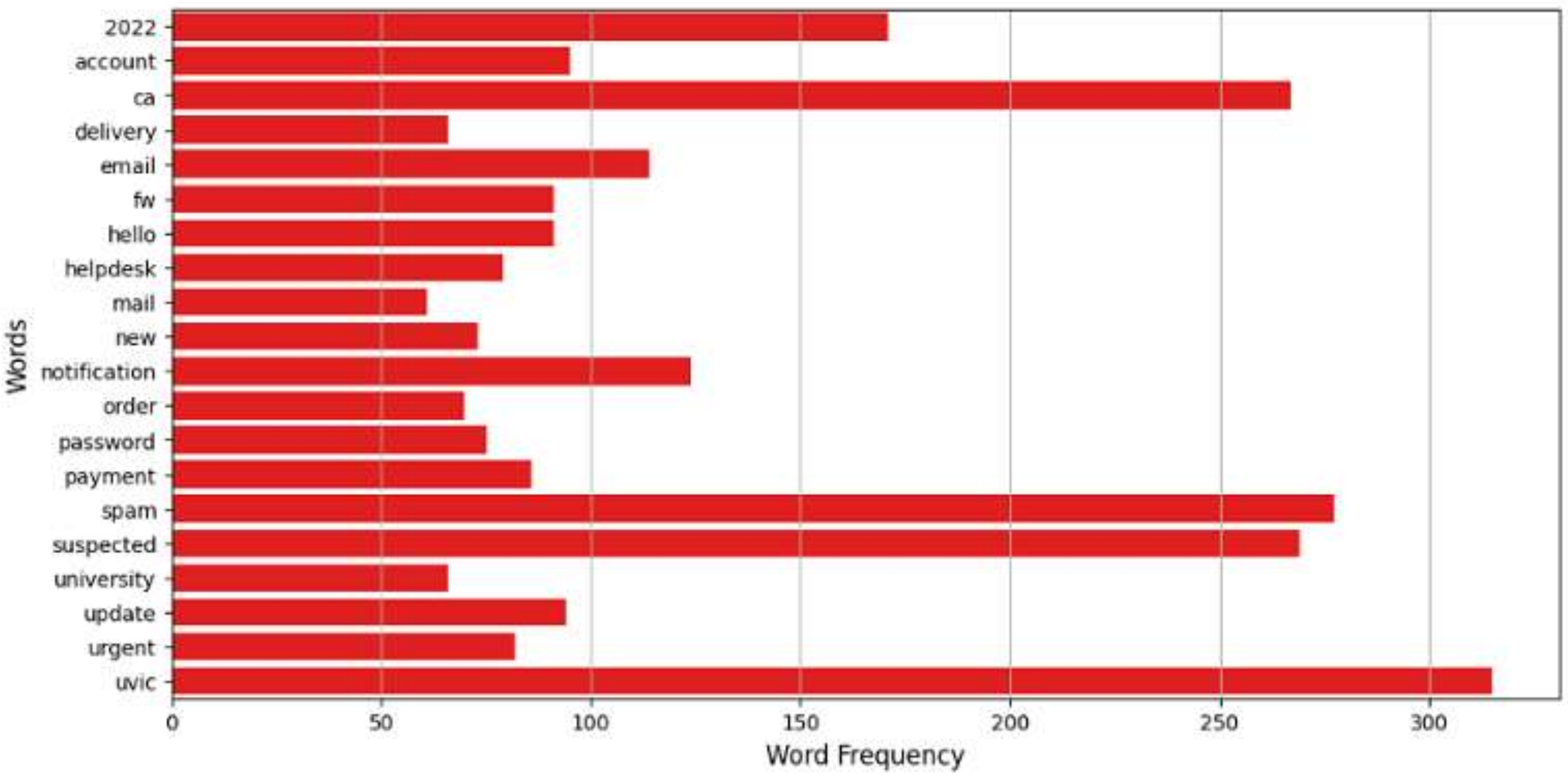


Missing Heat Map



Dataset Analysis

Top 20 Most Frequent Words in Email Subjects



Conclusions

- The dataset comprises structured email metadata with two main text fields: Subject and Body.
- These categorical textual features allow for natural language processing.

- Basic statistical analysis shows that both Subject and Body fields are populated for most entries, with low null count (<5%).
- Descriptive statistics (mean, median, and range) on the email subject and body length highlight a diverse sample distribution

Data Set

Statistical Analysis

Capstone Project

Length Distribution

Word Frequency

- The histogram of subject lengths reveals a unimodal distribution with a peak around 40–60 characters.
- The body length distribution exhibits high variance and a right tail (positive skewness), with a standard deviation likely greater than the mean.

- Word frequency across Subject and Body indicates significant lexical divergence: shorter, high-impact words dominate subjects; more varied vocabulary in bodies.
- Missing value heatmap shows isolated nulls (<2%) mostly in body fields, which could be imputed or filtered.

References

[1] Jamal S, Wimmer H, Sarker IH. An improved transformer-based model for detecting phishing, spam and ham emails: a large language model approach. SECURITY AND PRIVACY Apr. 2024. https://doi.org/10.1002/spy2.402.

[2] Somesha M, Pais AR. Phishing classification based on text content of an email body using transformers. In: Information Security, Privacy and Digital Forensics. 1075; 2024. p. 343–57. https://doi.org/10.1007/978-981-99-5091-1_25. Springer, Singapore

[3] Atawneh S, Aljehani H. Phishing email detection model using deep learning. Electronics (Basel) Oct. 2023;12(20):4261. https://doi.org/10.3390/ electronics12204261.

[4] Mehdi Gholampour P, Verma RM. Adversarial robustness of phishing email detection models. In: Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics. New York, NY, USA: ACM; Apr. 2023. p. 67–76. https://doi.org/10.1145/3579987.3586567

[5] Alhogail A, Alsabih A. Applying machine learning and natural language processing to detect phishing email. Comput Secur Nov. 2021;110:102414. https://doi. org/10.1016/j.cose.2021.102414.

[6] AbdulNabi I, Yaseen Q. Spam Email Detection Using Deep Learning Techniques. Procedia Comput Sci 2021;184:853–8. https://doi.org/10.1016/j. procs.2021.03.107

[7] Lee J, Tang F, Ye P, Abbasi F, Hay P, Divakaran DM. D-Fence: a flexible, efficient, and comprehensive phishing email detection system. In: 2021 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE; Sep. 2021. p. 578–97. https://doi.org/10.1109/EuroSP51992.2021.00045.

References

[8] Gangavarapu T, Jaidhar CD, Chanduka B. Applicability of machine learning in spam and phishing email filtering: review and approaches. Artif Intell Rev Oct. 2020;53(7):5019–81. https://doi.org/10.1007/S10462-020-09814-9/METRICS

[9] Gibson S, Issac B, Zhang L, Jacob SM. Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms. IEEE Access 2020;8: 187914–32. https://doi.org/10.1109/ACCESS.2020.3030751.

[10] Fang Y, Zhang C, Huang C, Liu L, Yang Y. Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism. IEEE Access 2019;7:56329–40. https://doi.org/10.1109/ACCESS.2019.2913705

[11] Arif MH, Li J, Iqbal M, Liu K. Sentiment analysis and spam detection in short informal text using learning classifier systems. Soft comput Nov. 2018;22(21): 7281–91. https://doi.org/10.1007/S00500-017-2729-X/METRICS.

[12] Hijawi W, Faris H, Alqatawna J, Al-Zoubi AM, Aljarah I. Improving email spam detection using content based feature engineering approach. In: 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). IEEE; Oct. 2017. p. 1–6. https://doi.org/10.1109/

[13] Dragomir Radev, " CLAIR collection of fraud email," ACL Data and Code Repository, ADCR2008T001. Jun. 2008.

[14] Spam Assassin Project. Spam assassin project. Spam Assassin Public Corpus; 2015.

Question Answer Session

Thank You