Summary for Part A

Git Repository : https://github.com/snirdav/project_group_9/tree/main

In Part A of this project, we focused on preparing, analyzing, and visualizing key features in a dataset related to email metadata and content. The goal was to explore and preprocess the data to ensure it was ready for model training and to gain insights into important patterns and features.

1. File Upload and Merging

• We began by uploading multiple files containing email metadata and content. These files were concatenated into a single dataset consisting of 82K rows and multiple columns, including body, sender, domain, subject, and others.

2. Data Merging and Feature Engineering

a. **Concatenation**: The files were merged to create a comprehensive dataset that combined both metadata and the body of the emails. We didn't want to analyze only the phishing-email.csv with the 82K rows of 'text-combined' because we looked for extra information from the metadata.

b. **New Features:** After initial analysis, additional columns were added to capture important features such as the domain, word count, week_day, hour, presence of capital letters, and more. These new features provided deeper insights into the email characteristics that contribute to spam detection.

4. Analysis and Feature Exploration

Key Features Identified:

- **Domain:** The domain of the email emerged as a significant feature, particularly in identifying patterns related to spam emails.
- **Word Count:** The length of the text (longer texts were less spam then shorter ones) and the frequency of certain words were important indicators of the email's nature, contributing to the analysis of spam versus non-spam.
- **Capital Letters:** The frequency of capital letters was analyzed as a potential marker for certain types of emails, especially those classified as spam.
- **Weekday & Hours:** Thursday and Friday for example had higher spam Rate than others

- **Attachments:** Emails with attachments were found to be safer, exhibiting lower fraud rates.
- **Analysis Process:**
- The dataset was analyzed to identify which features contributed most to distinguishing different classes (e.g., spam vs. non-spam). This analysis included exploring the distribution and significance of various text-related features like domain names, word length, sentiment analysis, source effect on the label, and the presence of capital letters.

Preprocessing steps that were taken:

- During the analytical phase we also handled nulls: Columns like receiver, date, sender, and URLs had a significant number of null values (~40%)
- Normalizing Features: Numeric features were normalized to ensure they have similar scales.
- Handling Outliers: Outliers were handled, likely through clipping or similar techniques, without altering the DataFrame's dimensions.
- Encoding Features: Categorical features were encoded, making them suitable for machine learning models.
- Converting Datetime: The 'date' column was converted to datetime format.
- Text Preprocessing: Text data was cleaned, including operations like lowercasing, removing stopwords, and stemming/lemmatizing.

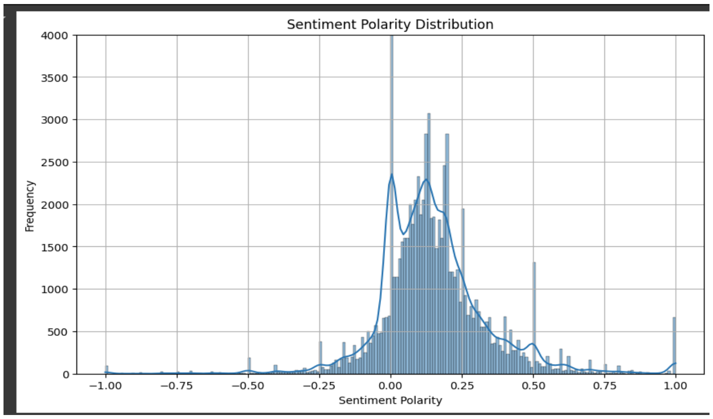
5. Visualization

Plots and Visualizations:

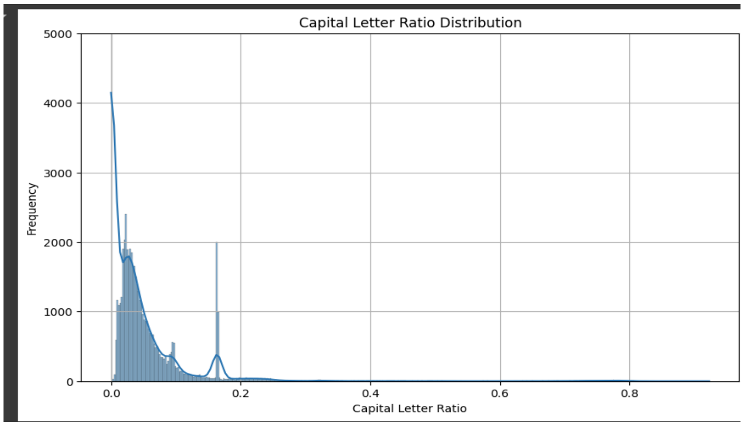
- **Word Cloud:** A word cloud was generated to visualize the most common words in the dataset from data labeled as spam and from non-spam data, it helped highlighting terms that may influence email classification.
- **Distribution Plots**: Distribution plots were used to show the distribution of text lengths before tokenization, revealing significant reductions in text length and ensuring consistent input sizes.
- **Word Frequency Analysis**: Word frequency plots were created to explore the vocabulary distribution, identifying key terms that could influence classification.
- **Feature Importance Visualization:** Bar charts and other visual tools were used to display the importance of different features, such as domain names and word length, in the dataset.
- **Correlation Matrix:** A heatmap was used to display correlations between numeric features, revealing strong correlations between word count and character count, among others.

Conclusion:

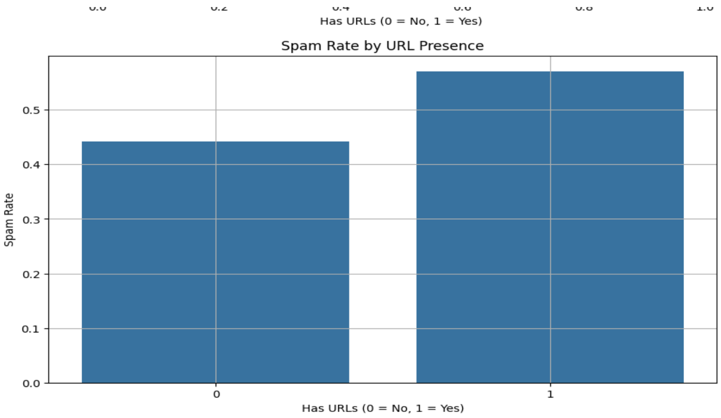
The preprocessing, analysis and visualizations confirmed that certain features, like word length and domain, play a critical role in the dataset. The consistent patterns observed in the plots supported the decision to focus on these features in subsequent analysis and model development.



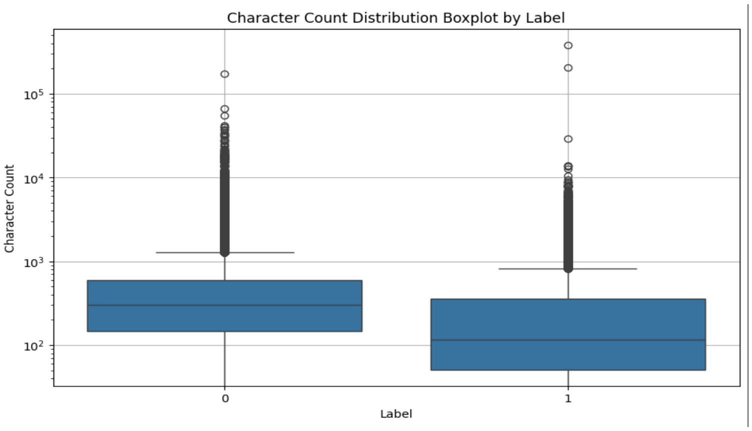
Sentiment Polarity Distribution
Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment Polarity



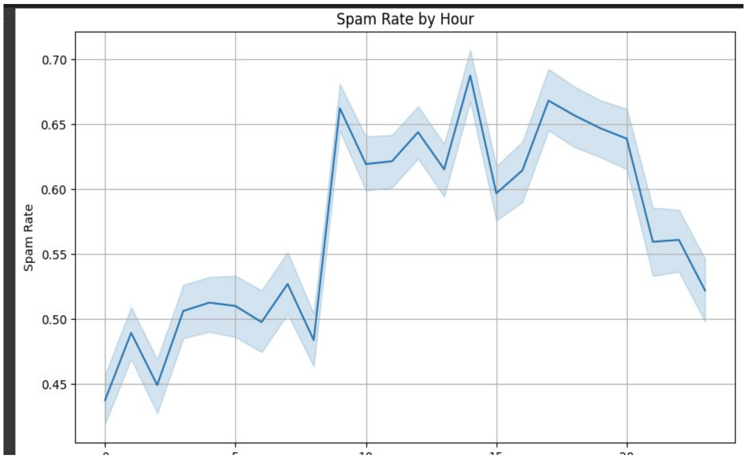
Capital Letter Ratio Distribution
Frequency
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
Capital Letter Ratio



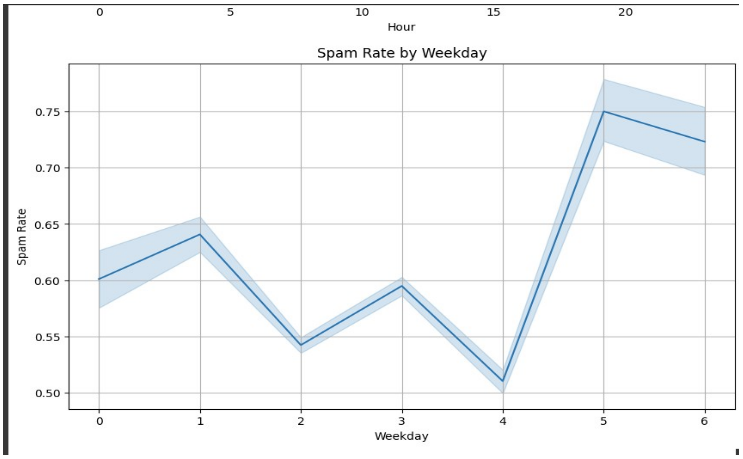
0.0
0.2
0.4
0.6
0.8
1.0
Has URLs (0 = No, 1 = Yes)
Spam Rate by URL Presence
Spam Rate
0.5
0.4
0.3
0.2
0.1
0.0
0
1
Has URLs (0 = No, 1 = Yes)



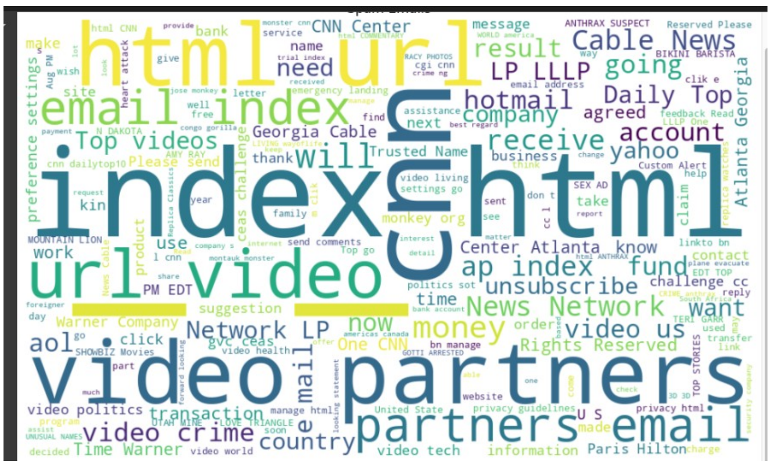
Character Count Distribution Boxplot by Label
Character Count
10^5
10^4
10^3
10^2
0
1
Label



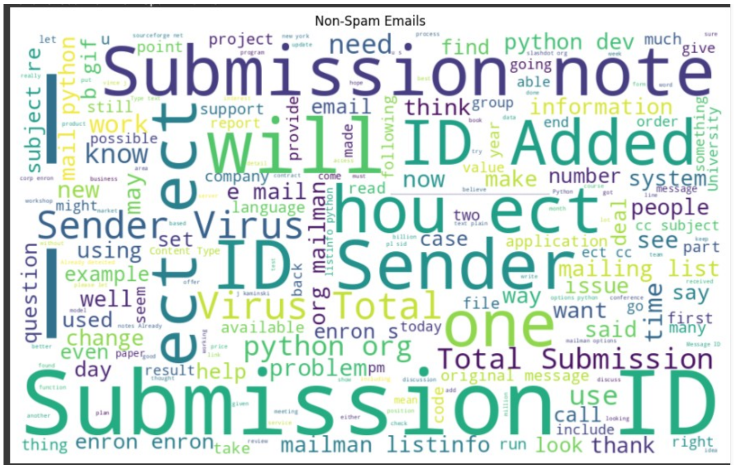
Spam Rate by Hour
Spam Rate
0.70
0.65
0.60
0.55
0.50
0.45
0
5
10
15
20



0
5
10
15
20
Hour
Spam Rate by Weekday
Spam Rate
0.75
0.70
0.65
0.60
0.55
0.50
0
1
2
3
4
5
6
Weekday



Spam Emails
html
url
index
html
cnn
video
partners
email



Non-Spam Emails
Submission
note
will
ID
Added
hou
ect
Sender
Virus
Total
one
Submission
ID