SPAMALOT:
SPAM OR NOT?

Laura Bishop, Brian Kath, Jake Moen, Katy Yelle

01

Introduction

02

Meet the Data

Data Source, Data Cleaning and General Insights

Data Source

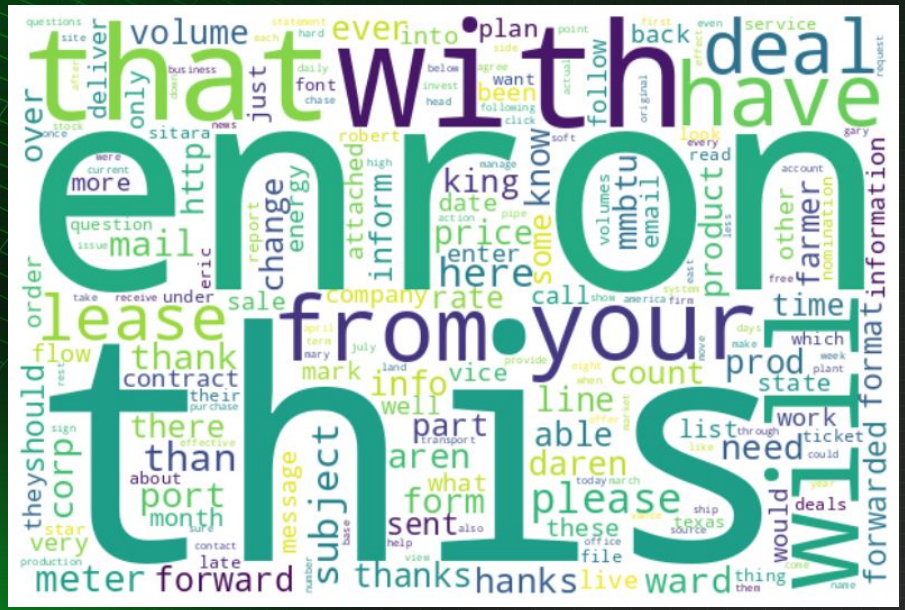
Dataset created by Balaka Biswas
Downloaded from Kaggle 11/30/2023 (approx 4yrs old)

https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv

5172 rows, 3002 columns

- Email Name
- 3000 most common words in all the emails, w/ count per email
- Spam label
    - 3672 Not Spam (Ham)
    - 1500 Spam

Data Overview

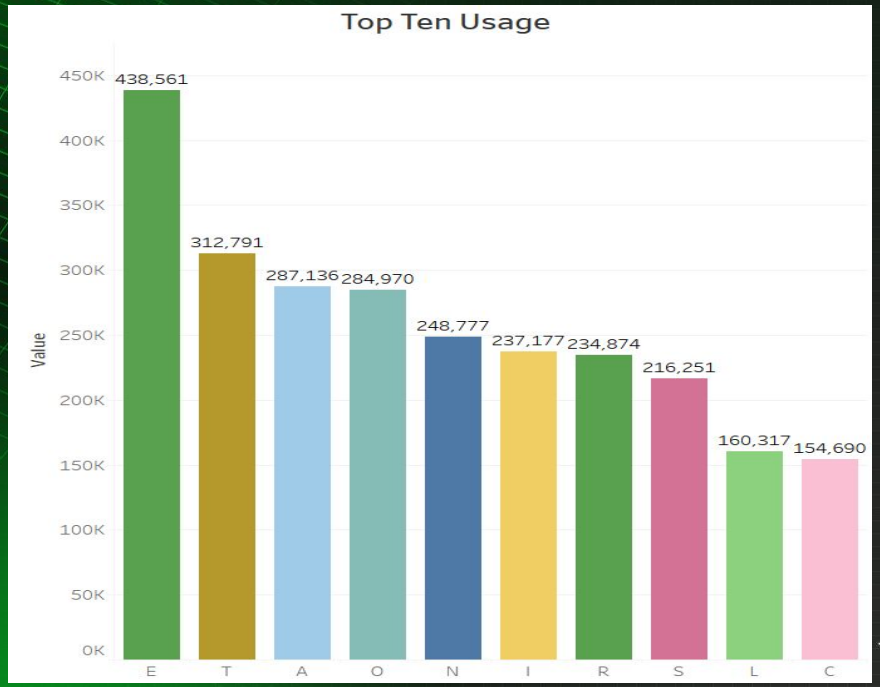


Word Analysis

- Minimum word length: 1
- Average word length : 6
- Maximum word length : 16

. . . .

Data Overview



<u>Word Usage Analysis</u>

- Minimum word use: 21 (explosion, returns, flw, dorcheus, offsystem, greatest, allowing)
- Average word use: 34,345
- Maximum word use : 438,561

03

Data Models

Logistic Regressions

Model 1

Long words (>6 letters) Only
-Reduced to 1198 columns
-92% Accuracy

Model 2

Common Words Only
-Reduced to 307 columns
-95% Accuracy

Model 3

All Columns
-97% Accuracy

SVM & Random Forest

Model 4

Support Vector Machine (SVM)
-All columns
-95% Accuracy

Model 5

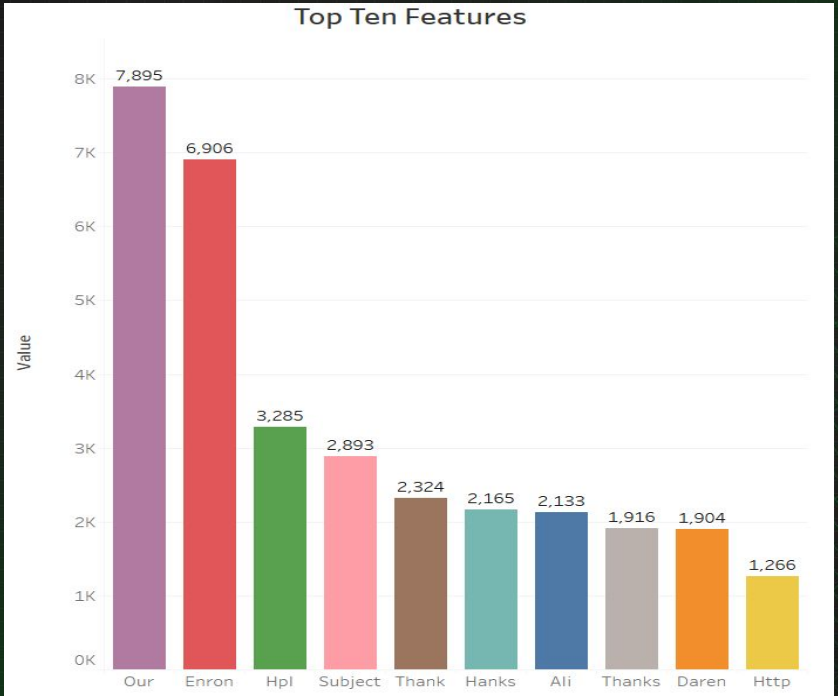
Random Forest
-All columns
-98% Accuracy

Feature Importances

Feature	Score
‘enron’	.0215
‘http’	.0149
‘hpl’	.0144
‘thanks’	.0126
‘hanks’	.0123

Feature	Score
‘ali’	.0113
‘thank’	.0096
‘daren’	.0095
‘our’	.0093
‘subject’	.0089

Feature Importances



Model Results Summary

Model	Accuracy
Logistic Regression Long Words Only	92%
Logistic Regression Common Words Only	95%
Logistic Regression All Columns	97%
Support Vector Machine	95%
Random Forest	98%

04

Conclusions

Conclusions

- Focusing exclusively on email content (Words vs Date/Time) was a key factor for the success of all our models

- Three Models Tested:
  - Logistic Regression
  - Support Vector Machine
  - Random Forest – 98%!

- Accurate spam detection is important! Not just for the convenience, but to deter Bad Faith Actors
  - Phishing attempts
  - Trojan programs
  - Malware

05

Next Steps

What We Would Do if Given Unlimited Time and Resources

Next Steps

- Optimize the random forest model utilizing information gained from the feature importances

- Create a program that would generate the data similar to our dataset utilizing more recent / current data
    - How does that change the dataset?
    - Does it impact model performance?
    - Does it support our hypothesis that using words is ideal?

06

Project Website

THANKS !

To Hunter, Sam, and Randy!
Thank you so much for these last 24 weeks, giving us this opportunity to grow and develop a whole new skill set built for success!