Module IV: Enron Email Classification

Isiah Cruz

October 2020

TABLE OF CONTENTS

PROBLEM STATEMENT

Endless possibilities.™

Email classification: Classifying emails based on their content to automate manual processes and to increase the accuracy with which emails are labeled

Tools:
Machine Learning
Natural Language Processing
Word Vectorizing
Label Encoding
TF-IDF Vectorization

BUSINESS VALUE

1

AUTO-LABELING

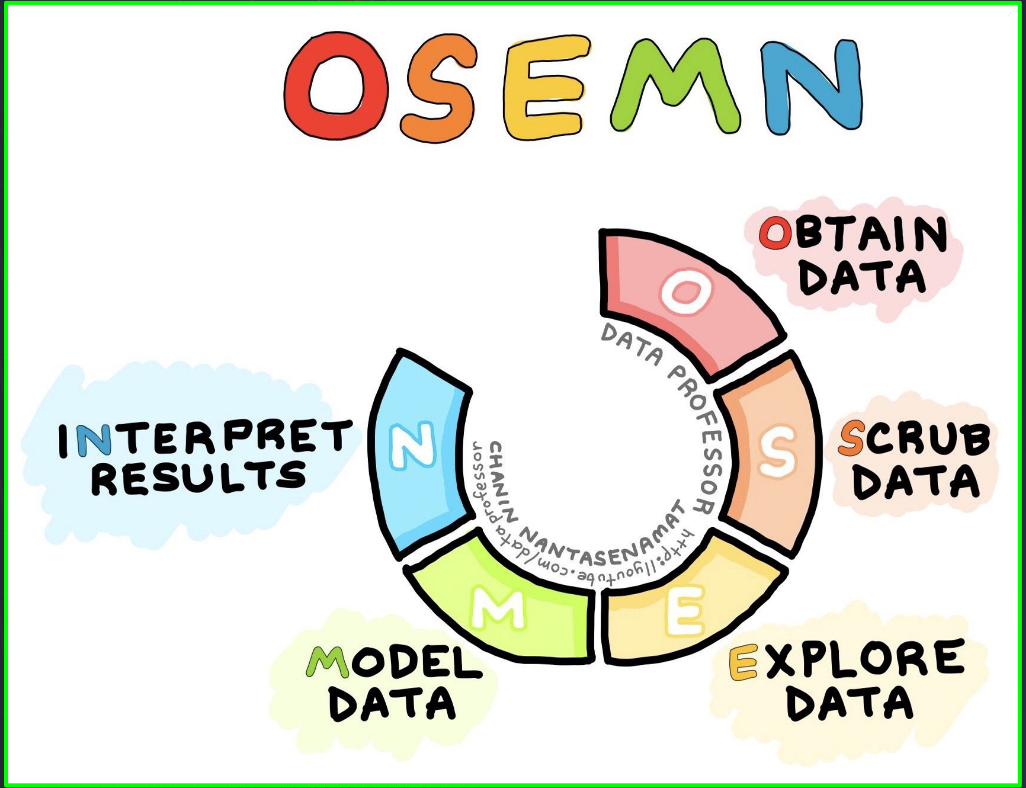
2

PREDICTION

3

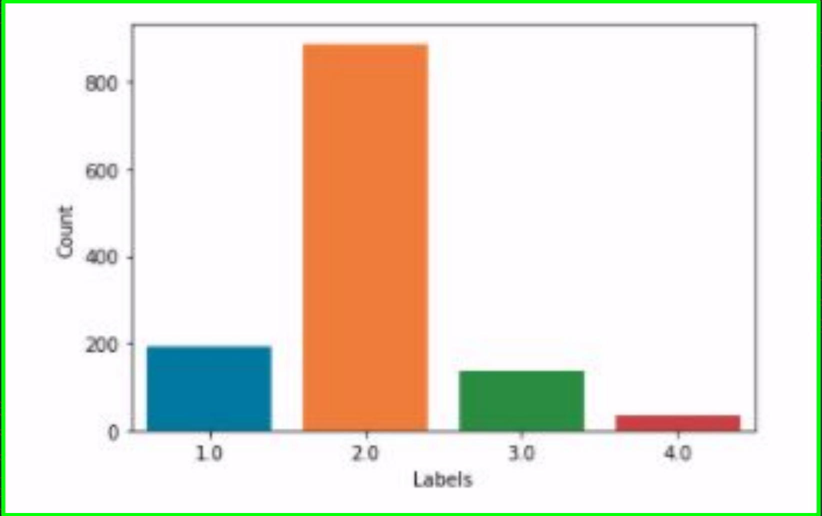
POSSIBILITIES

METHODOLOGY



OSEMN Framework

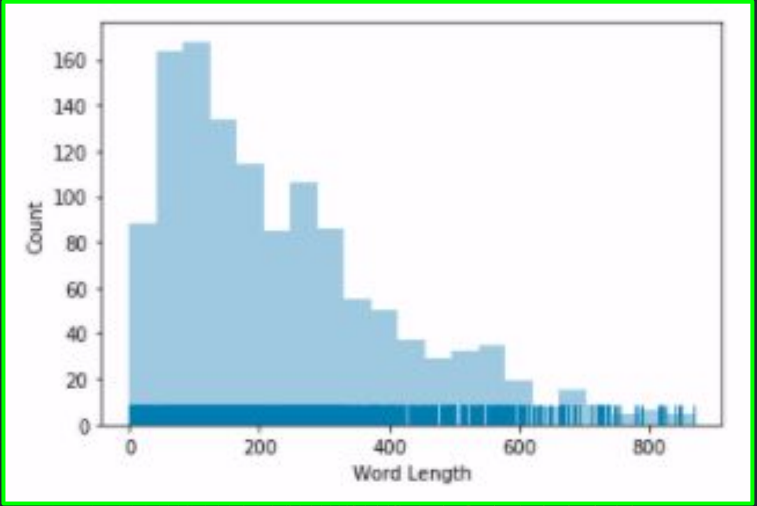
FINDINGS I

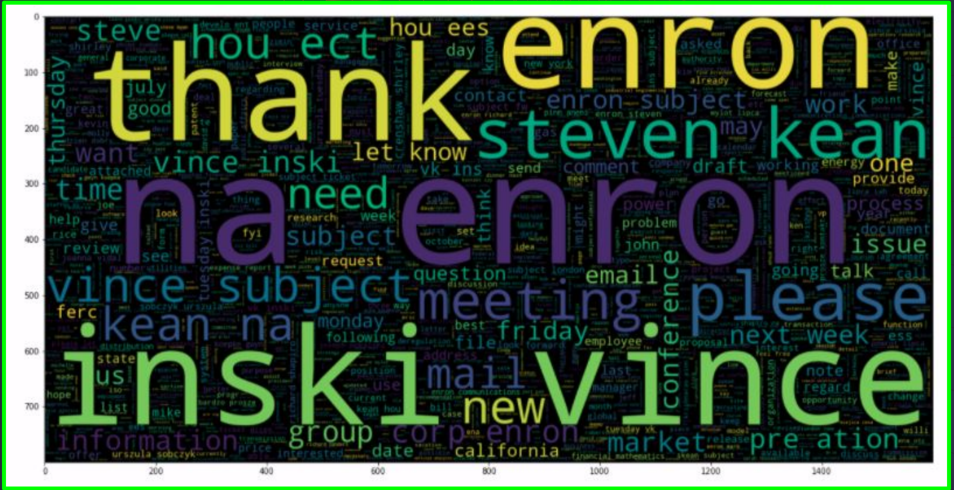


Labels

1.0: Coarse genre (company strategy, logistic arrangements, etc) — 74%
2.0: Included/forwarded information (forwarded emails, press releases, etc) — 14%
3.0: Primary topics (meeting minutes, regulations, etc) — 10%
4.0: Emotional tone (jubilation, sarcasm, etc) — 2%

FINDINGS I





Word length of emails is sizeable

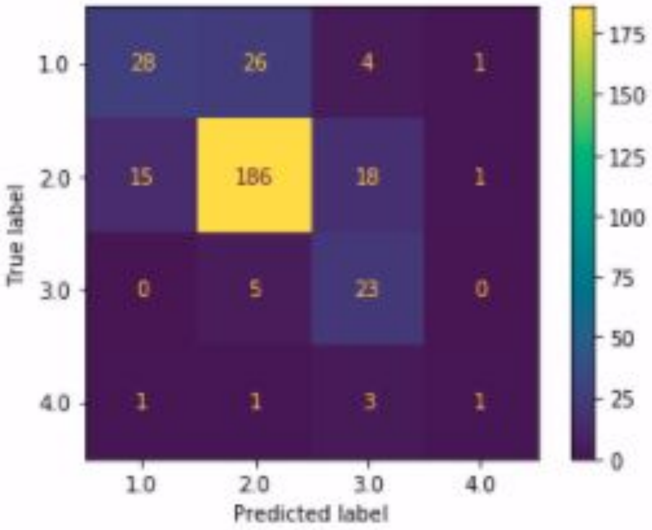
Internal whistleblowers as topic of discussion for Category 1.0

7

FINDINGS 3

```
Train Accuracy: 0.9872754491017964
Train Accuracy: 0.7603833865814696


Accuracy Score for model: 76.04%
Precision Score for model: 76.89%
Recall Score for model: 76.04%
F1 Score for model: 75.79%
```



MODEL #1: GRADIENT BOOSTING

FUTURE WORK

LABELING

Use this model to label the 98K or so emails in our original dataset that are unlabeled

AUTO-RESPONSES

Create an auto-response tool that responds to emails according the what label they receive