



# Module IV: Enron Email Classification

*Isiah Cruz*

**December 2020**

# TABLE OF CONTENTS

- 01 **PROBLEM STATEMENT**
- 02 **BUSINESS VALUE**
- 03 **METHODOLOGY**
- 04 **FINDINGS**

# PROBLEM STATEMENT



1. Build a model that can classify labeled Enron emails effectively
2. Take learnings from this classification and use it to classify email responses for NewtonX
3. Ultimately build an auto-response system that sends a preset message based on the label the email receives

# BUSINESS VALUE



1

**AUTO-LABELING**



2

**PREDICTION**

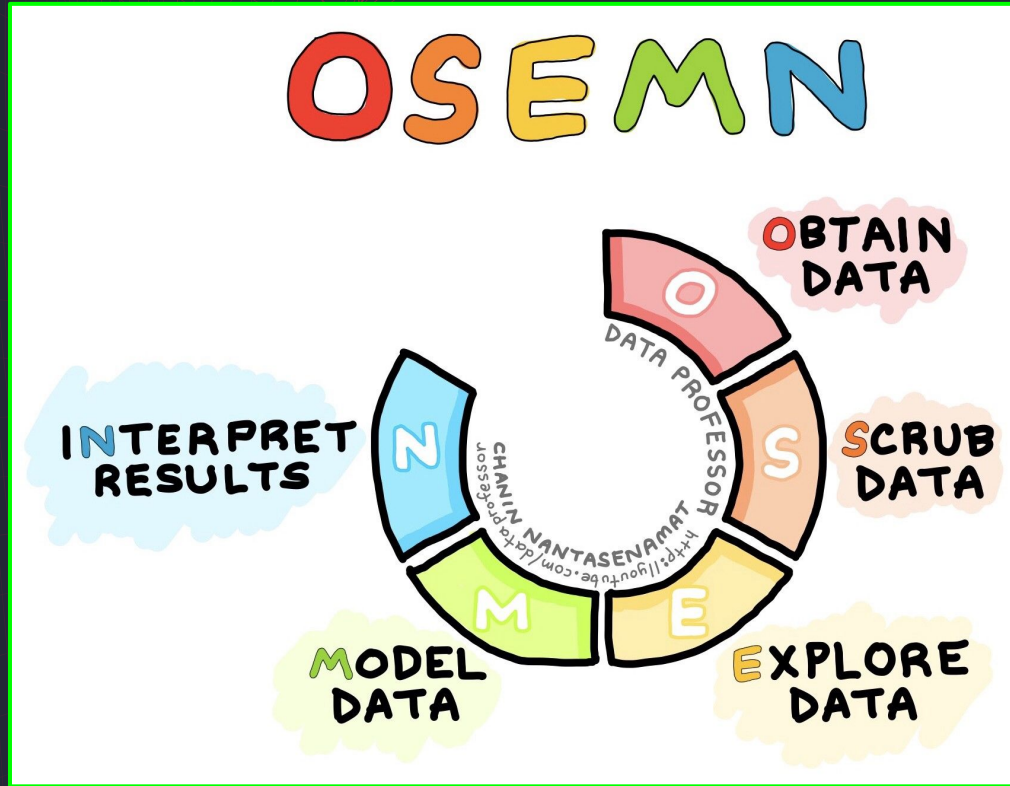


3

**POSSIBILITIES**

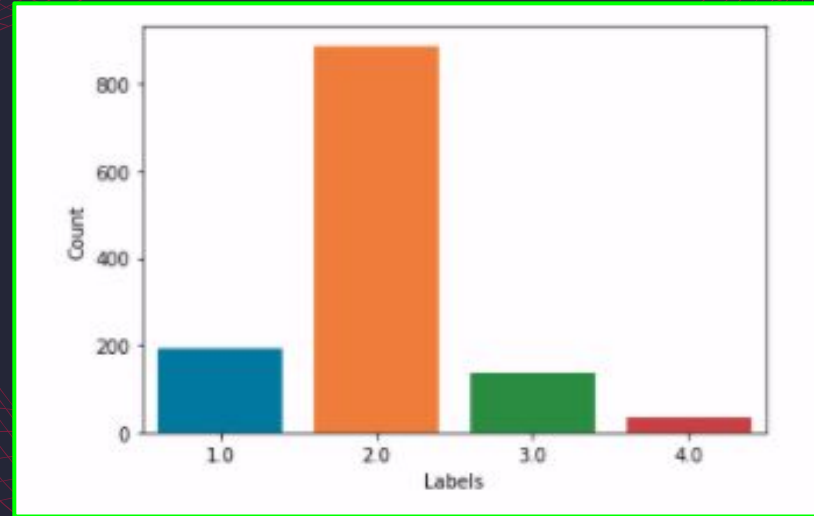


# METHODOLOGY



**OSEMN Framework**

# FINDINGS I



## Labels

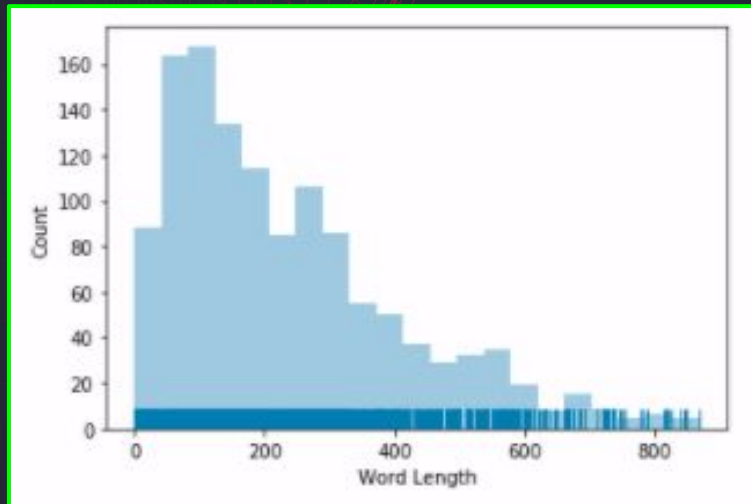
**1.0: Coarse genre (company strategy, logistic arrangements, etc) – 74%**

**2.0: Included/forwarded information (forwarded emails, press releases, etc) – 14%**

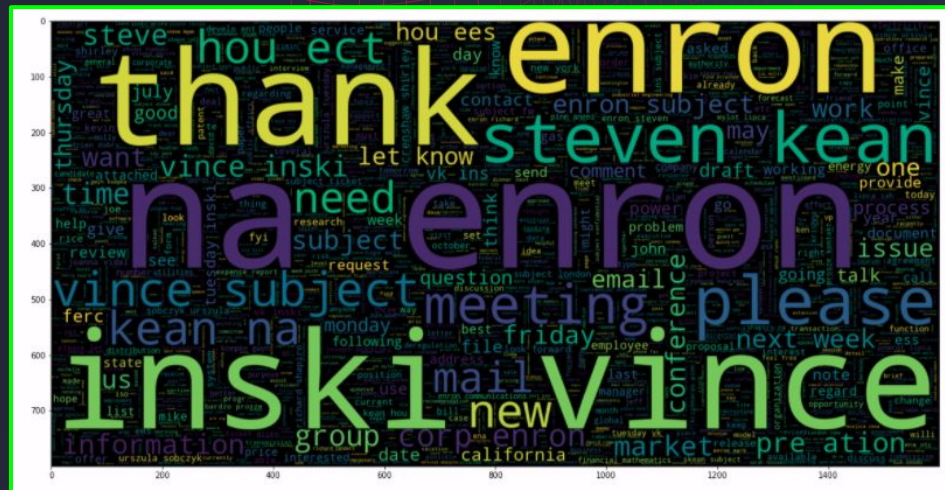
**3.0: Primary topics (meeting minutes, regulations, etc) – 10%**

**4.0: Emotional tone (jubilation, sarcasm, etc) – 2%**

## FINDINGS I



## Word length of emails is sizeable

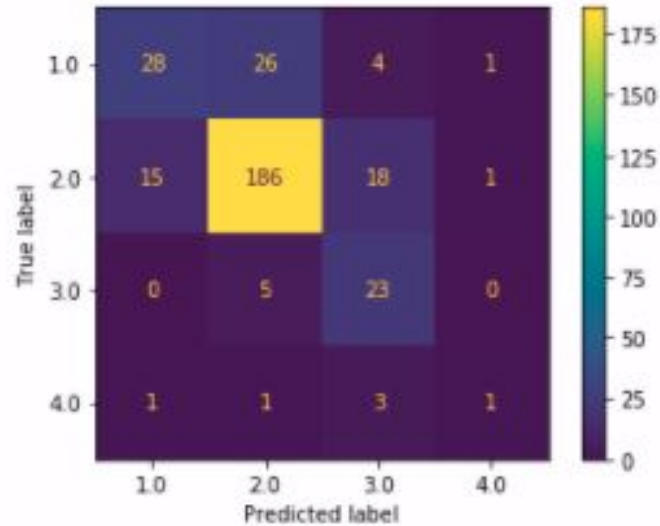


## Internal whistleblowers as topic of discussion for Category 1.0

## 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**

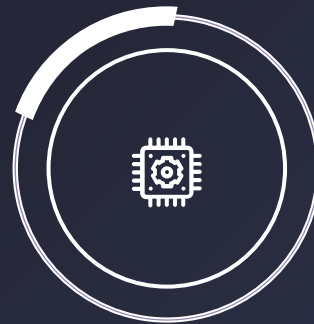


# FUTURE WORK



## 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