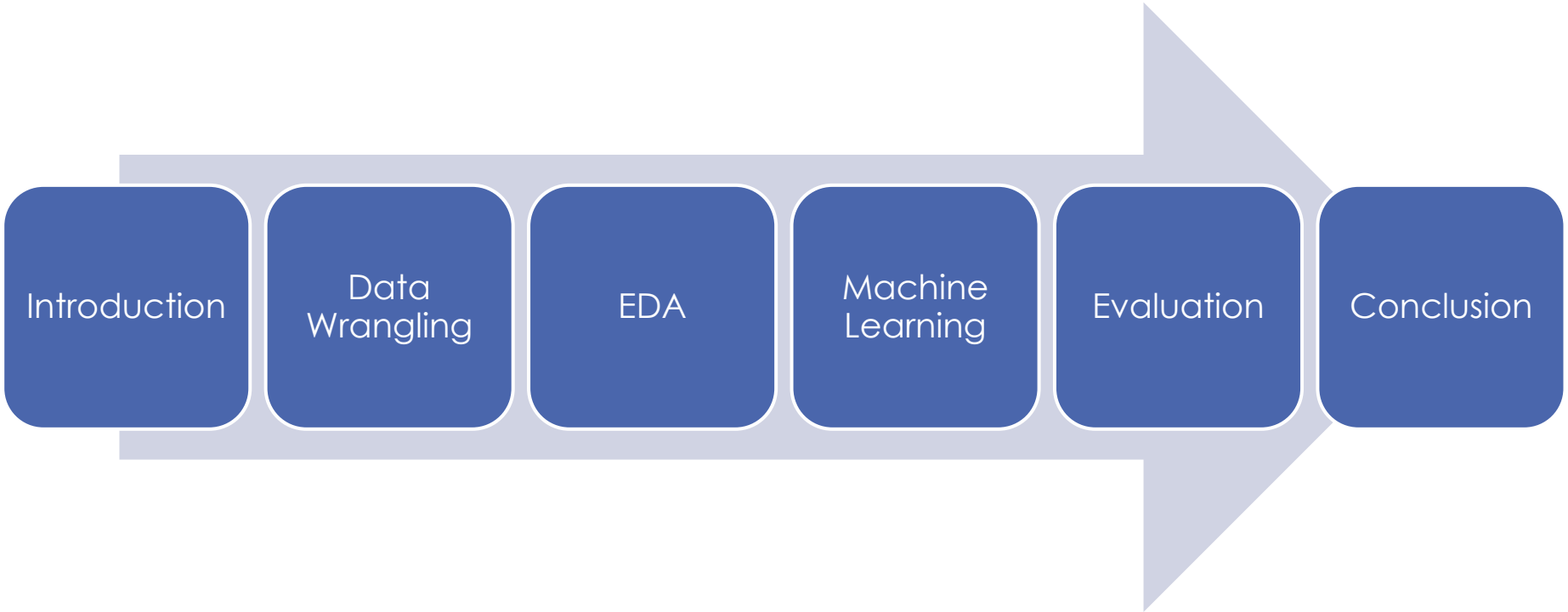
Bigotry Detection Model

FILTERING NEGATIVE COMMENTS WITH MACHINE LEARNING

Overview


Introduction
Data Wrangling
EDA
Machine Learning
Evaluation
Conclusion

Introduction

- Communication Platforms:
- Discussion
- Social Media
- Networking
- Gaming
- Other

- Toxicity: Prejudice

Introduction

Focus

- Bigotry

Business Case

- Bigotry Detection Model
- Reviewing Capacity: 10% of all comment traffic

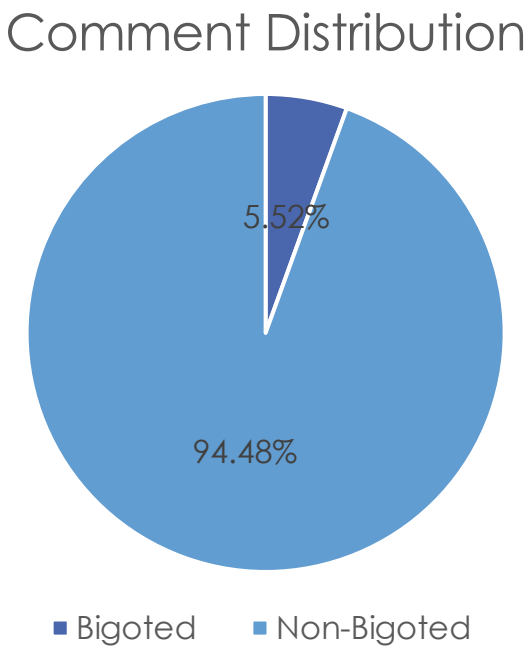
- Human in the loop (HITL)
  - 1. Model flags comments
  - 2. Team reviews flagged comments
  - 3. Team enacts penalty for offenders

Data Wrangling

- Reddit Comment Dataset
- Emotion Sensor Dataset

Data Wrangling

- Reddit Comment Dataset from Idibon: Toxic + Supportive
  - Model Columns:
    - *bigotry* - binary bigotry indicator for the comment
    - *text* – text of the Reddit comment
    - *upvotes* – number of times the comment was upvoted
  - Bigotry: How sure are we?
    - Above 2/3
    - 8814 rows with 515 bigoted comments



Data Wrangling

- Emotion Sensor Dataset from Kaggle
  - 1104 Words

- Scores:
  - Disgust
  - Surprise
  - Neutral
  - Anger
  - Sad
  - Happy
  - Fear

word	disgust	surprise	neutral	anger	sad	happy	fear
ability	0.00446429	0.04783164	0.00063800	0.02359694	0.01339286	0.01594388	0.04017858
able	0.00001730	0.00018200	0.00040900	0.00017600	0.00021900	0.00024400	0.00018600

Combining the Datasets

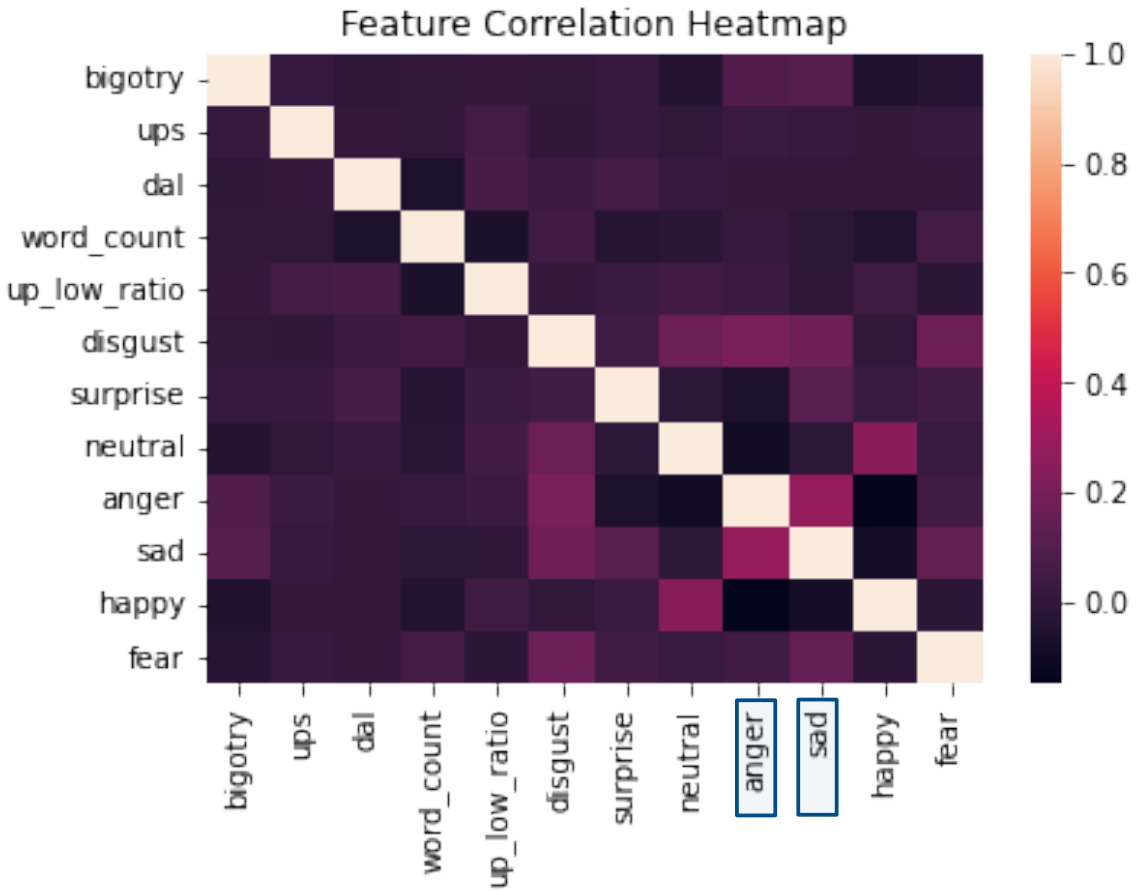
For each comment:

- For words in the Emotion Sensor Dataset:
- For each emotion column:
  - Calculate average emotion score of words

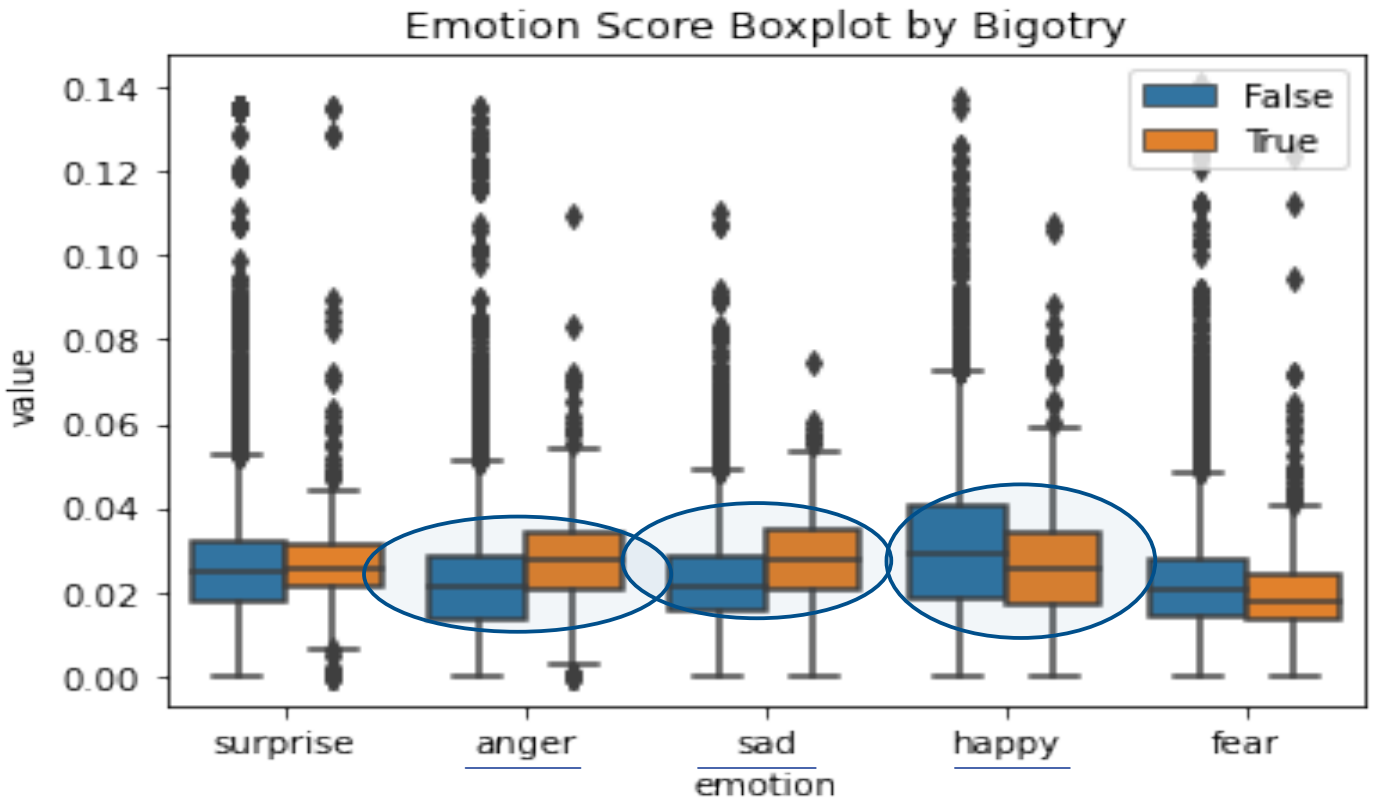
- Result:
  - Seven columns with average scores for each comment

text	disgust	surprise	. . .
comment 1	d score 1	s score 1	
comment 2	d score 2	s score 2	

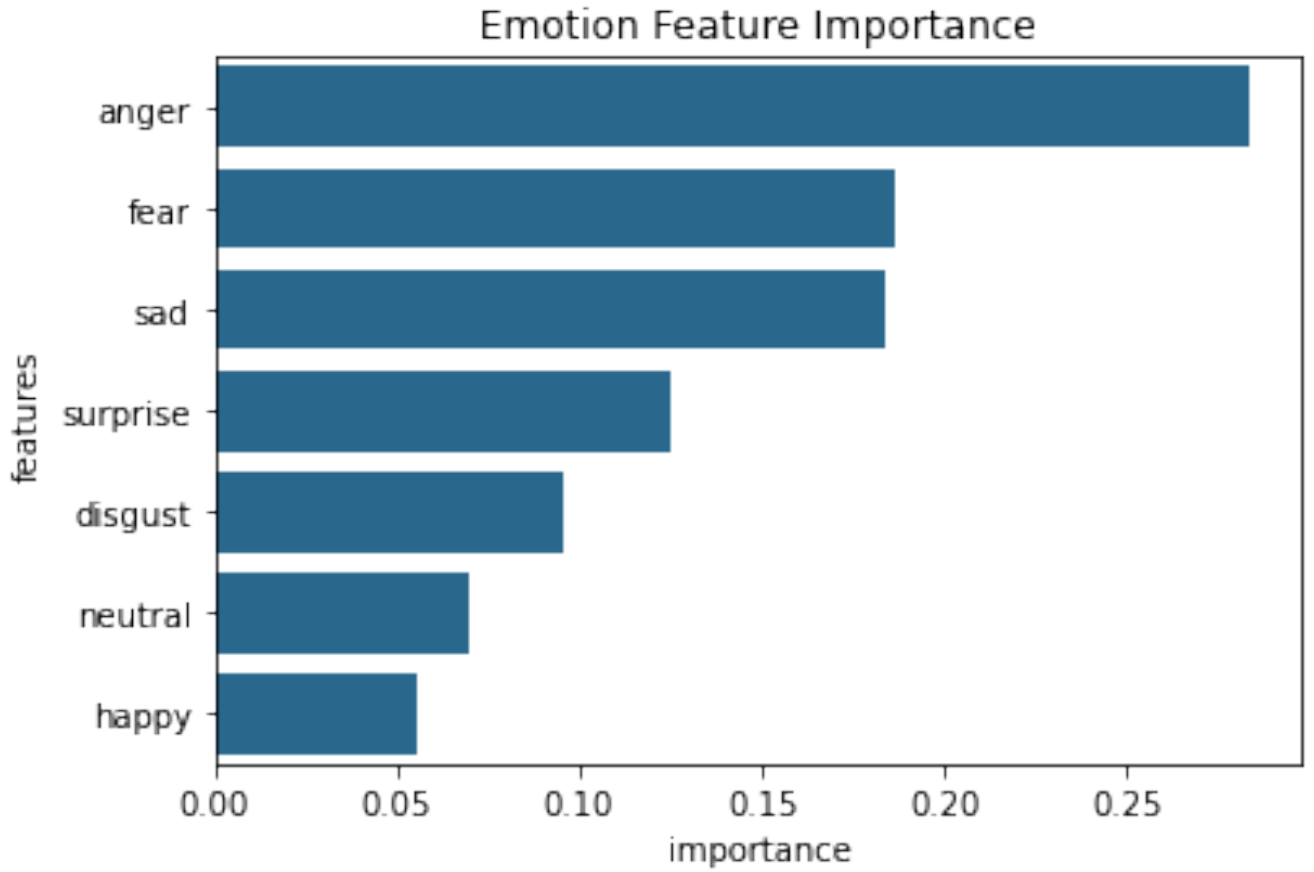
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Most Predictive Words

Rank	Word	Probability
1	fa**ot	0.8792
2	ni**a	0.8675
3	ni**er	0.8359
4	bigoted	0.7443
5	misogynistic	0.7443
6	aboriginal	0.7443
7	penis	0.7081
8	motherf***er	0.6599
9	gay	0.6359
10	c*nt	0.6335
11	atheist	0.5927
12	slut	0.5927
13	motherf***ing	0.5927
14	homophobic	0.5927
15	butthurt	0.5927

Least Predictive words

Rank	Word	Probability
1	gem	0.002
2	glass	0.0052
3	thank	0.0056
4	le	0.008
5	favorite	0.0084
6	nope	0.0087
7	keyboard	0.009
8	minute	0.0095
9	small	0.0096
10	sound	0.0106
11	before	0.0107
12	nice	0.0108
13	since	0.0112
14	used	0.0119
15	wish	0.012

Machine Learning

Upvotes	Dale-Chall Readability	Word Count	Caps Lock Ratio	7 Emotion Columns	Vectorized Word Matrix

- Trackable
- Computable

Count-Vectorized Word Matrix

word1	word2	word3	. . .
0	1	0	1

Model Comparison

- GridSearchCV for ROC-AUC

- Best Model:
  - Logistic Regression

Model Name	ROC-AUC Score	Best Hyperparameters
Logistic regression	0.7774	C=0.16238, penalty='l1', solver='liblinear'
Random forest classifier	0.7734	criterion='entropy', max_depth=15, max_features=None, n_estimators=500
Gradient Boosting	0.7558	learning_rate=0.1, max_depth=6, n_estimators=100
Support vector classification	0.7322	C=1, gamma=0.1, kernel='rbf'

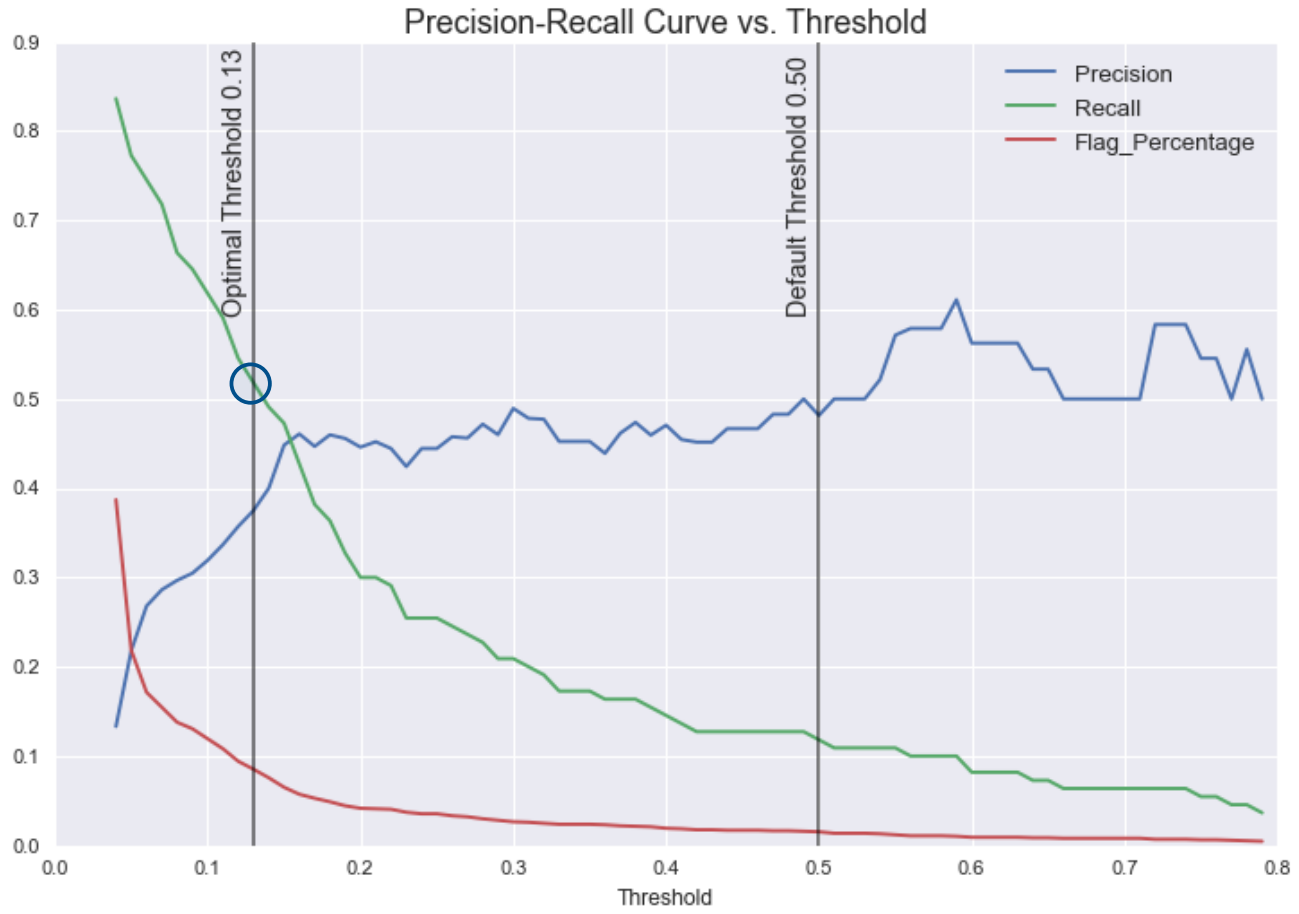
Thresholding

- Context:

10% Comment Traffic

All Flags Reviewed

Maximize Recall



Model Evaluation

- Classification Report

	precision	recall	f1-score	support
0	0.97	0.93	0.95	1675
1	0.34	0.52	0.41	110
accuracy			0.91	1785
macro avg	0.66	0.73	0.68	1785
weighted avg	0.93	0.91	0.92	1785

Model Evaluation +

- Appeals
  - Low Precision => False Flags
- Predicted Probability
  - Slow Precision Increase
- Reporting
  - New data, but missing data

	precision	recall	f1-score	support
0	0.97	0.93	0.95	1675
1	0.34	0.52	0.41	110
accuracy			0.91	1785
macro avg	0.66	0.73	0.68	1785
weighted avg	0.93	0.91	0.92	1785

Conclusion

Model Success:

- Detected over 51% of all bigoted comments

EDA Success:

- Identified Most Predictive Words

Future Work:

- More complex threshold specific model
- Use different emotion sensor score functions
- Replicate on milder data

Works Cited

- Images on Overview Slide
  - https://iteachu.uaf.edu/online-discussions/
  - https://favpng.com/png_view/network-computer-network-career-job-clip-art-png/rH4nDLnC
  - https://commons.wikimedia.org/wiki/File:Social-media-g6e011fe56_1920.png
  - https://pngtree.com/freepng/game-gaming-internet-multiplayer-online-flat-color-icon-vect_4986457.html
  - https://thenounproject.com/icon/profanity-15388/

- Dataset links
  - Reddit Comment Dataset Article:
    - *https://web.archive.org/web/20160304012220/http://idibon.com/toxicity-in-reddit-communities-a-journey-to-the-darkest-depths-of-the-interwebs/*
    - Data from anonymous source
  - Emotion Sensor Dataset:
    - https://data.world/elie707/emotions-sensor-dataset