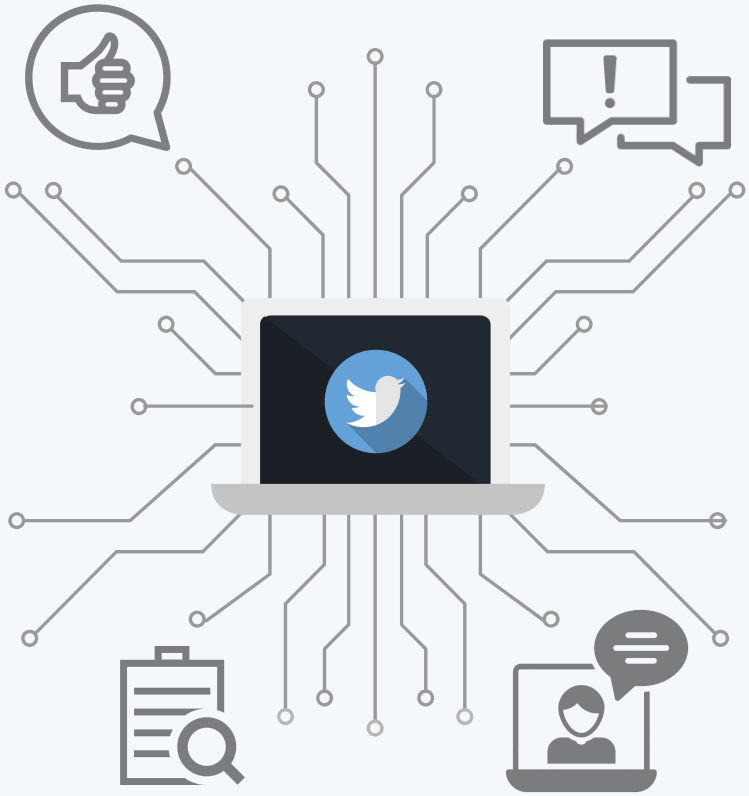
Twitter Hate Speech Detection

Can Content Moderation be Automated?

Flatiron School Capstone
Sidney Kung

Overview

- 2019 Verge article exposed Facebook's former **content moderation** contractor
- **Automating** this process could **reduce labor exploitation**
- What is **Hate Speech?**



CRISP-DM Process

Social Case

CrowdFlower

twitter

python

pandas

matplotlib

Exploratory Data Analysis

Data Cleaning and NLP Preprocessing

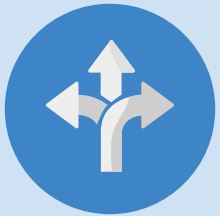
NLTK

GENSIM

topic modelling for humans

VADER

scikit learn



Modeling

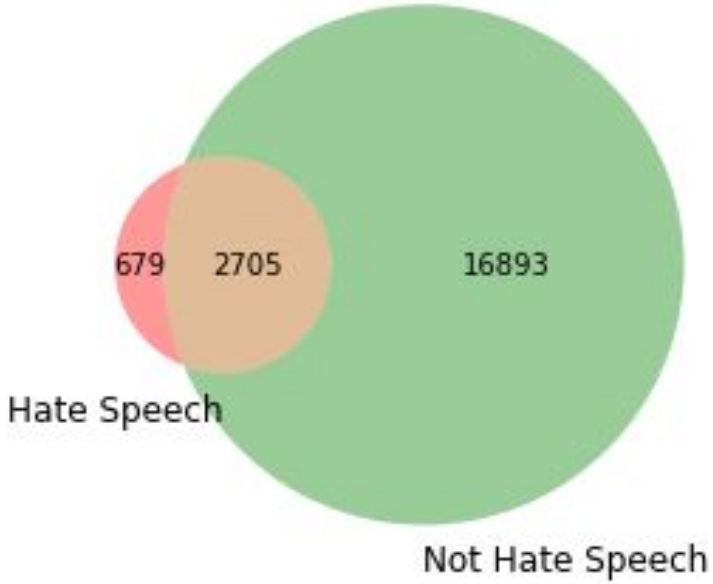
Evaluation and Insights

GitHub

Data Understanding

Sourced from 2017 Cornell University **research study**.

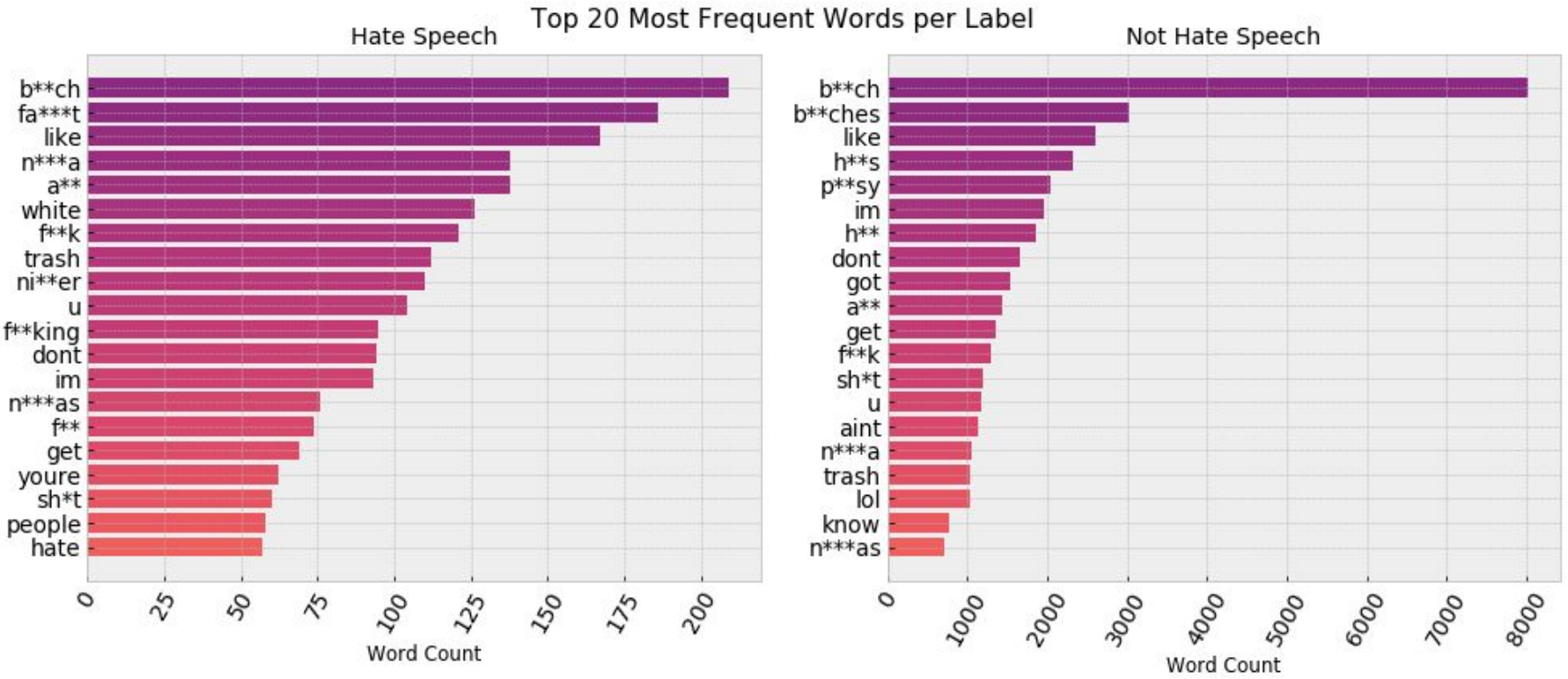
- **24,802** Tweets
- Vocabulary of **20,277 unique words**
  - 6% Hate Speech
  - 94% Not Hate Speech
- Evaluation Metric: **F1 Score**



Business Questions

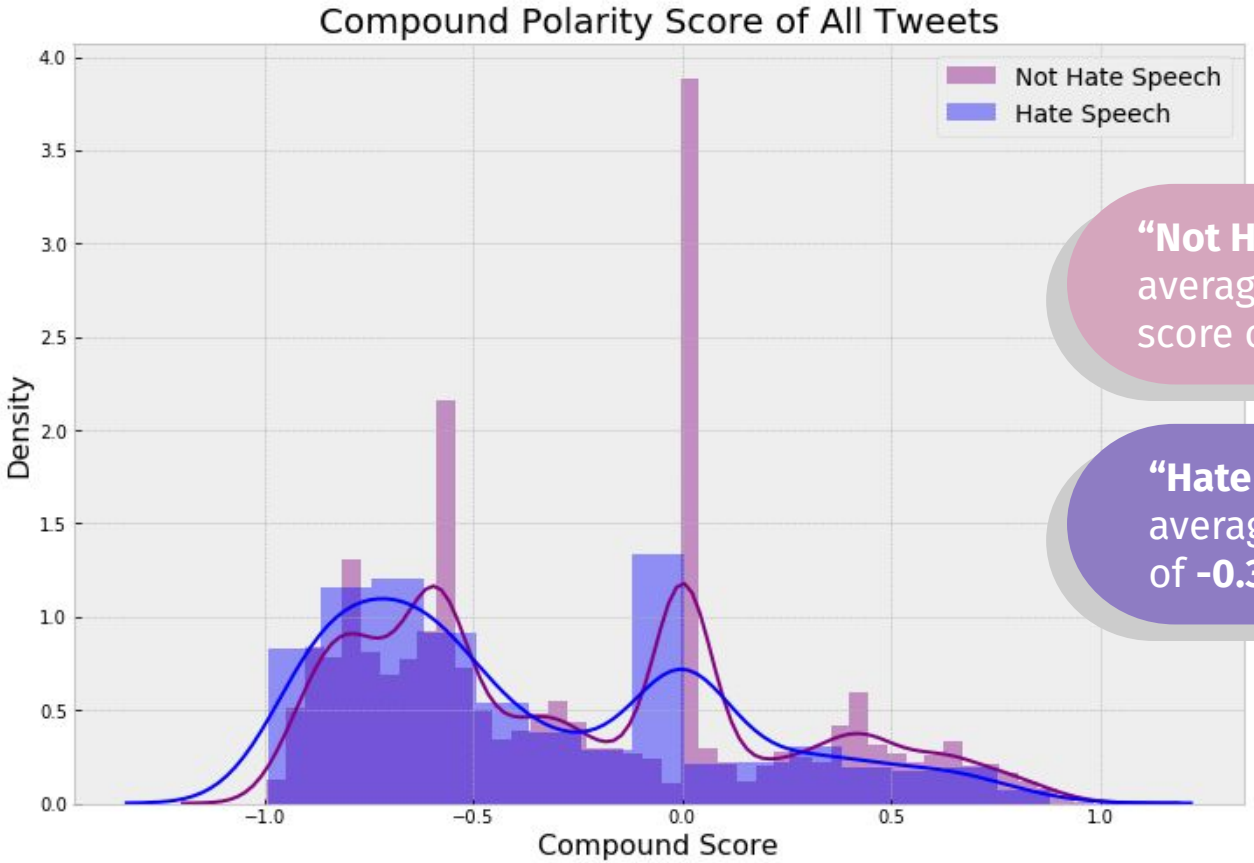
1. What are the linguistic differences between hate speech and offensive language?
2. What are the most popular hashtags of each tweet type?
3. What is the overall polarity of the tweets?

1. What are the linguistic differences between hate speech and offensive language?



2. What are the most popular hashtags of each tweet type?

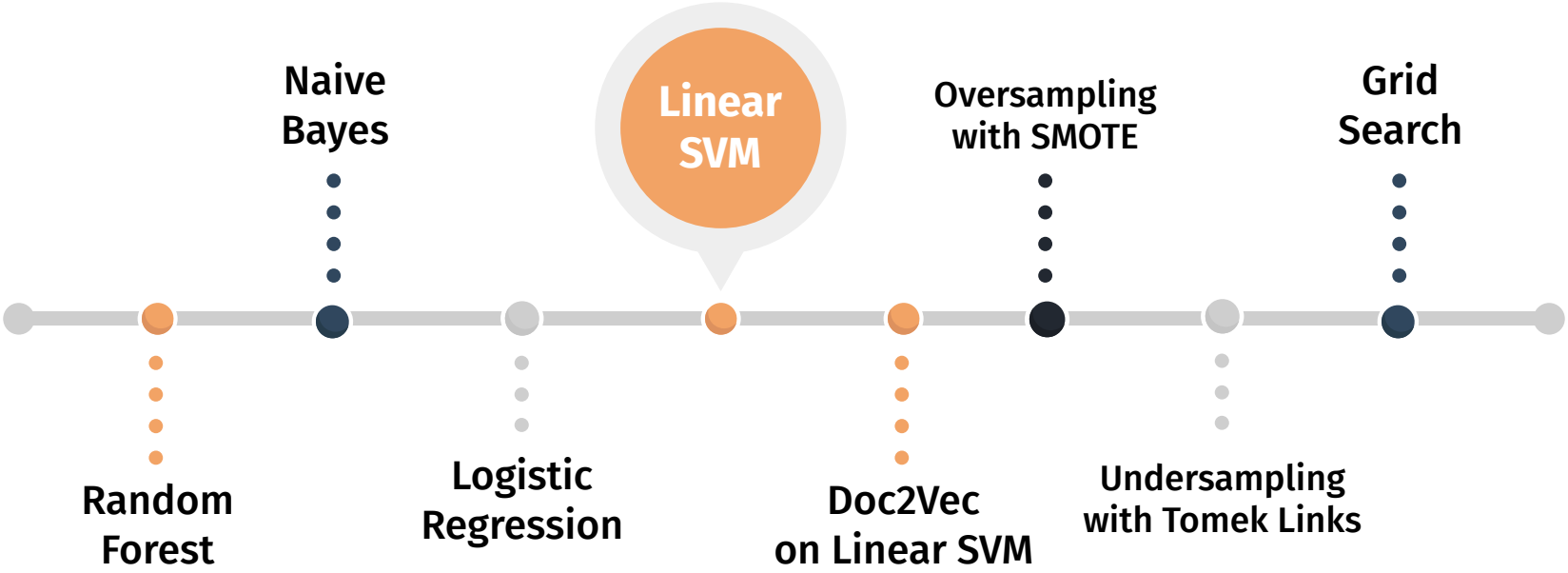
3. What is the overall polarity of the tweets?



"Not Hate Speech" tweets: average compound score of **-0.263**

"Hate Speech" tweets: average compound score of **-0.363**

Modeling Process

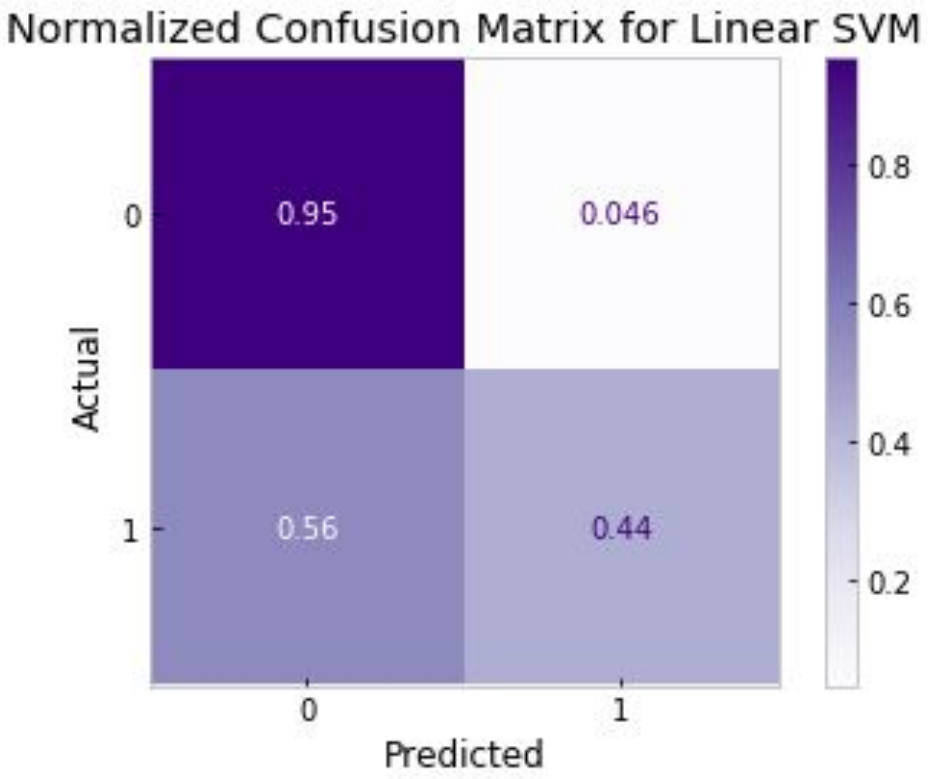


Final Model Analysis

Linear SVM Classifier

F1 Score: 0.3955
Recall: 0.437

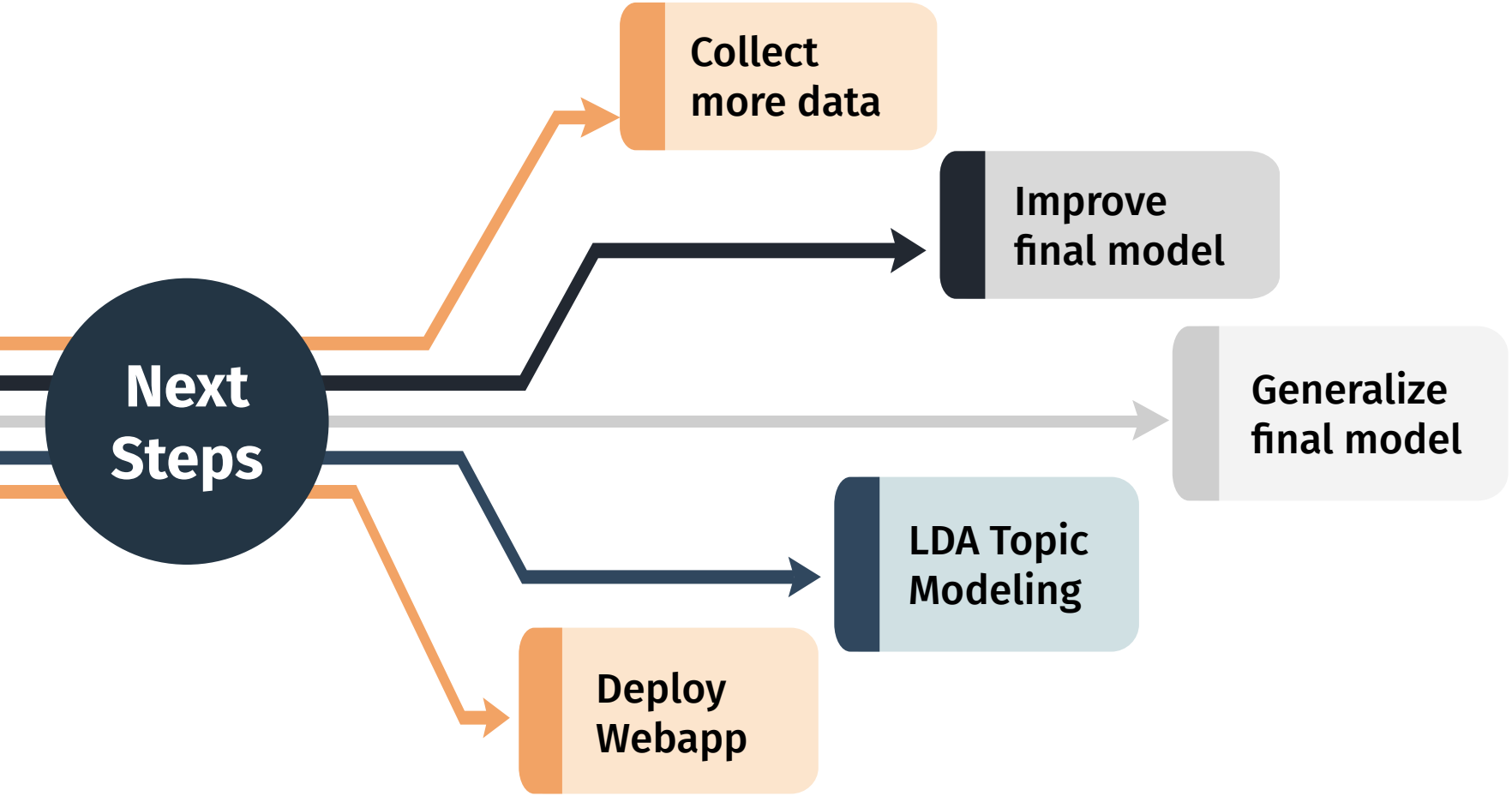
High **True Negative** Rate
Low **False Positive** Rate
Moderate **True Positive** Rate



Conclusion

Two major roadblocks:

1. Massive class imbalance
2. Model's ability to “understand” hate speech


Next Steps
Collect more data
Improve final model
Generalize final model
LDA Topic Modeling
Deploy Webapp

Thank You!

GitHub Repository
https://github.com/sidneykung/twitter_hate_speech_detection

SidneyJKung@gmail.com

https://www.linkedin.com/in/sidneykung/

@Sidney_K98

Presentation Template:
SlidesGo