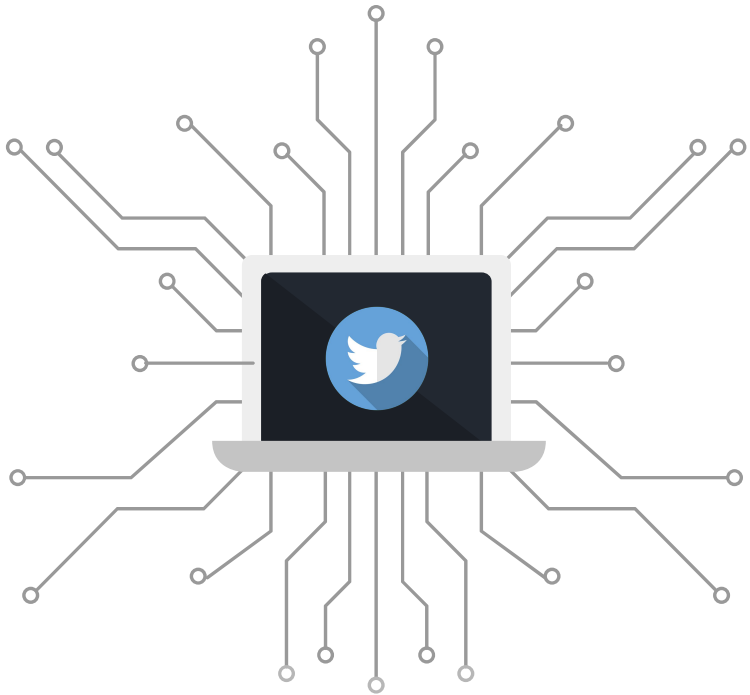
Twitter Hate Speech Detection

Can Content Moderation be Automated?

Flatiron School Capstone
Sidney Kung

Overview

- 2019 Verge article exposed Cognizant, a former Facebook **content moderation contractor**
- **Automating hate speech detection** could **reduce labor exploitation** and other human rights violations
- **Hate Speech** is defined as abusive or threatening speech that expresses prejudice against a particular group.

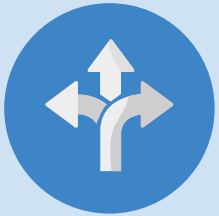


CRISP-DM Process

Social Case

Exploratory Data Analysis

Data Cleaning and NLP Preprocessing



Modeling

Evaluation and Insights

CrowdFlower

twitter

python

pandas

matplotlib

scikit learn

NLTK

GENSIM topic modelling for humans

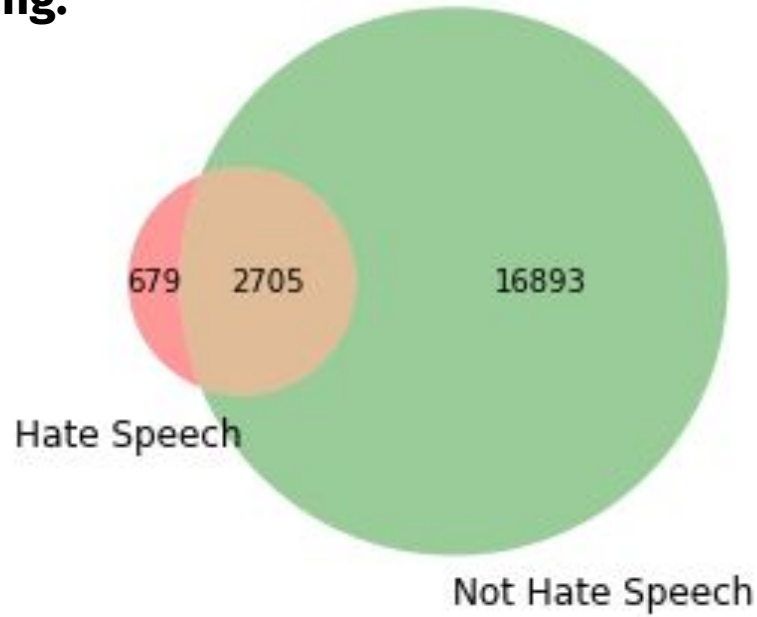
VADER

GitHub

Data Understanding

Sourced from Cornell University **research study**.
Labeled by CrowdFlower **majority-rules voting.**

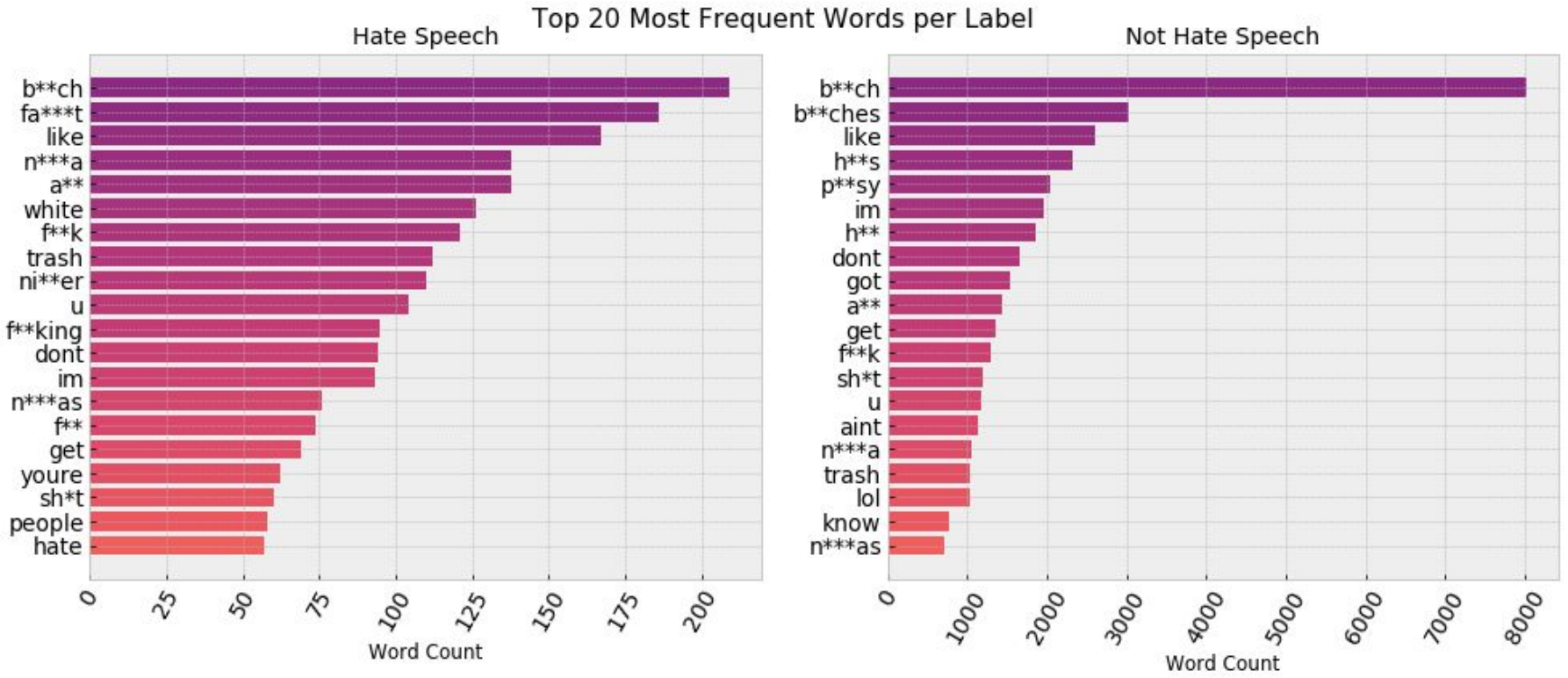
- **24,802** Tweets
- Vocabulary of **20,277 unique words**
- Binary Classification
  - 6% Hate Speech
  - 94% Not Hate Speech
- Evaluation Metric: **F1 Score**



Business Questions

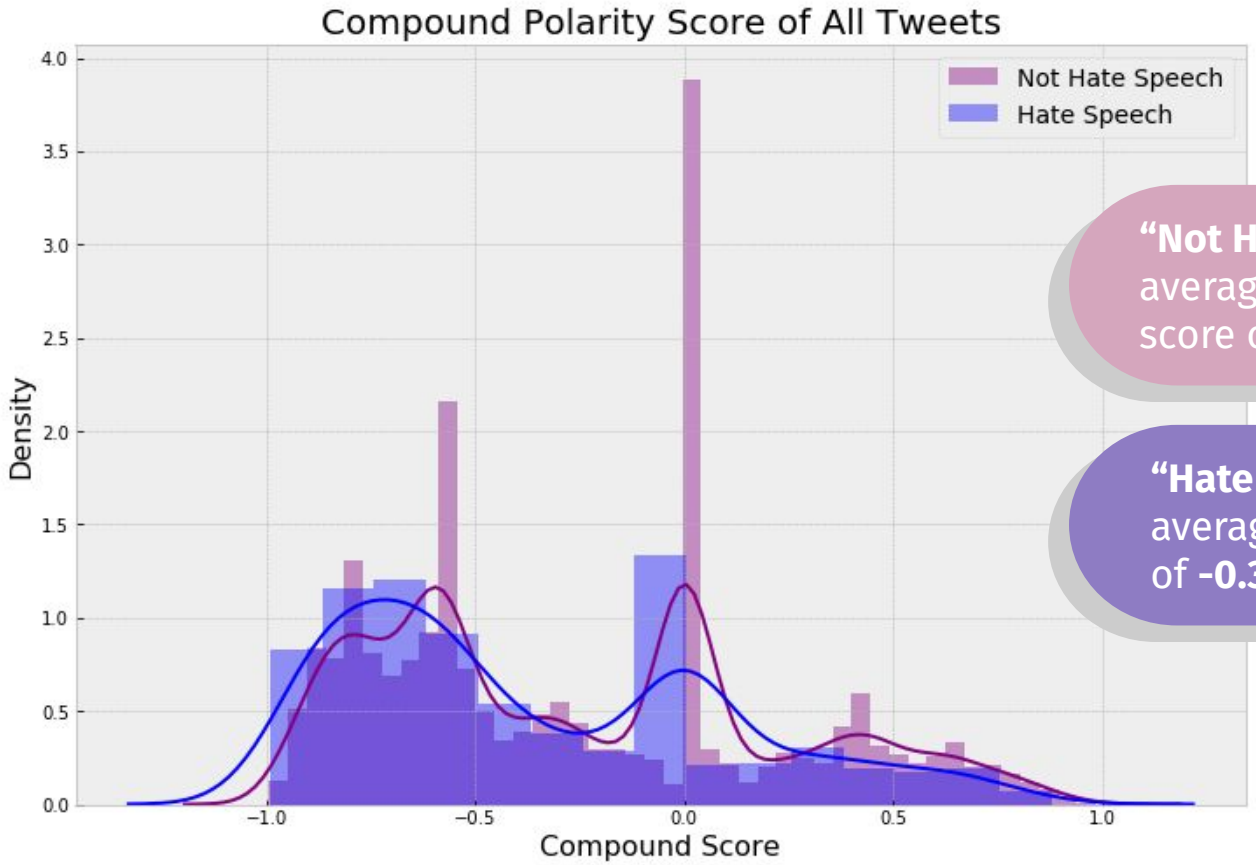
1. What are the linguistic differences between hate speech and offensive language?
2. What are the most popular hashtags of each tweet type?
3. What is the overall polarity of the tweets?

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



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

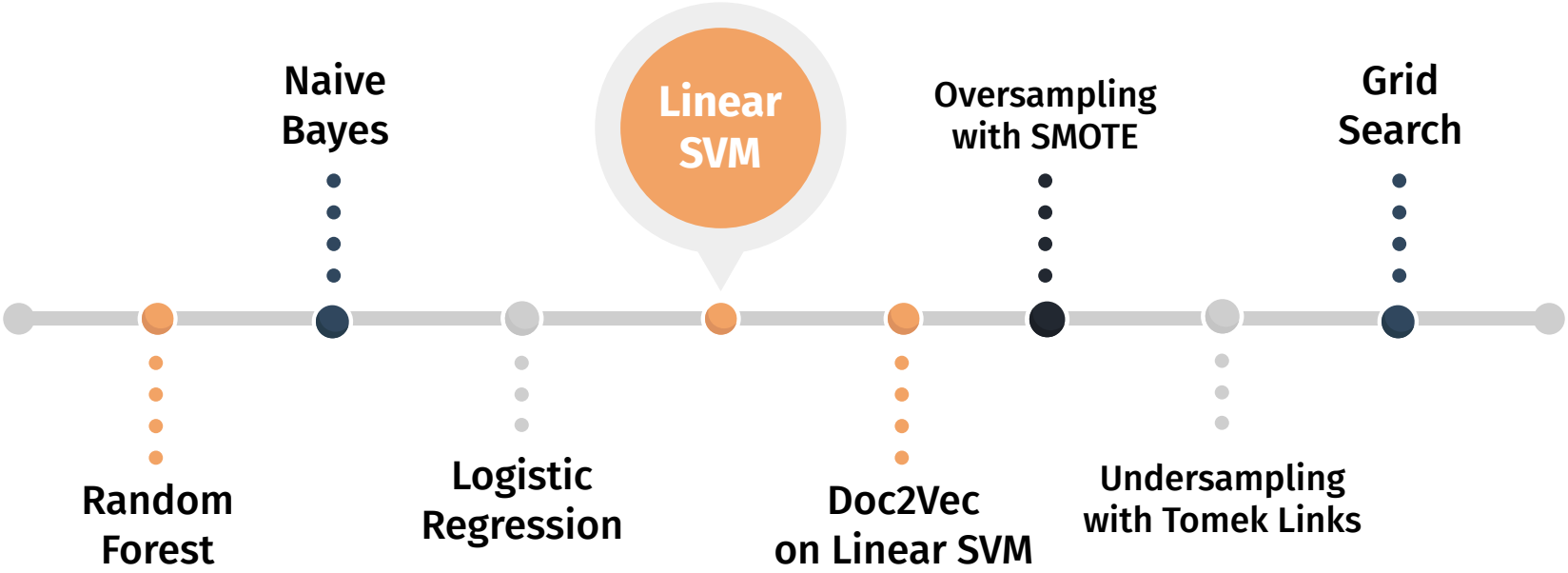
3. What is the overall polarity of the tweets?



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

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

Modeling Process

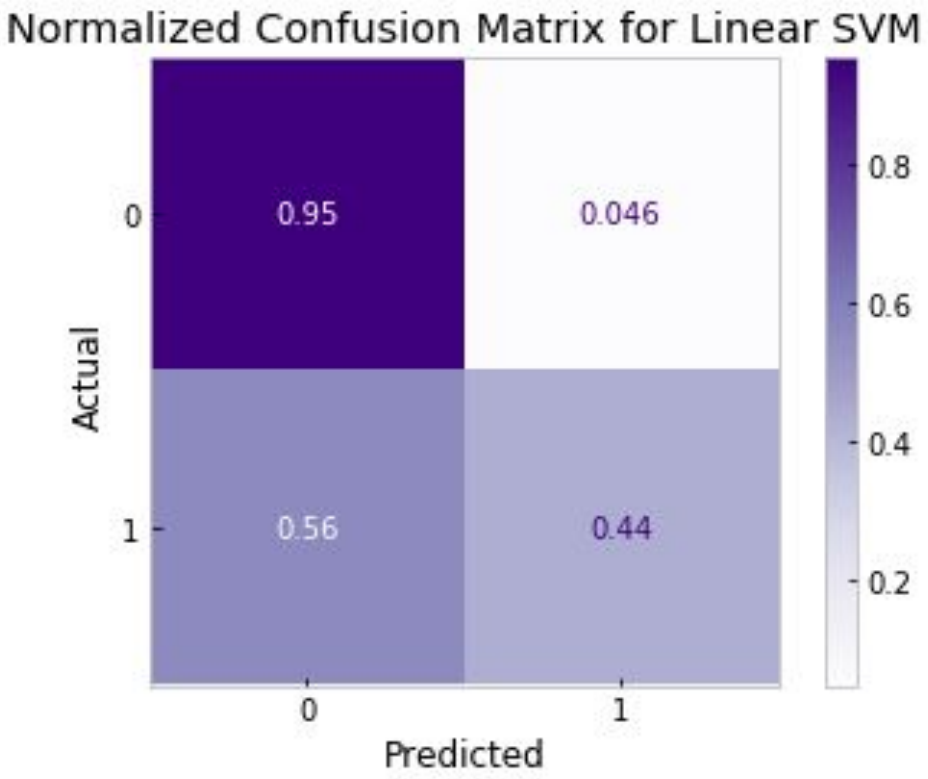


Final Model Analysis

Linear SVM Classifier

F1 Score: 0.3955
Recall: 0.437

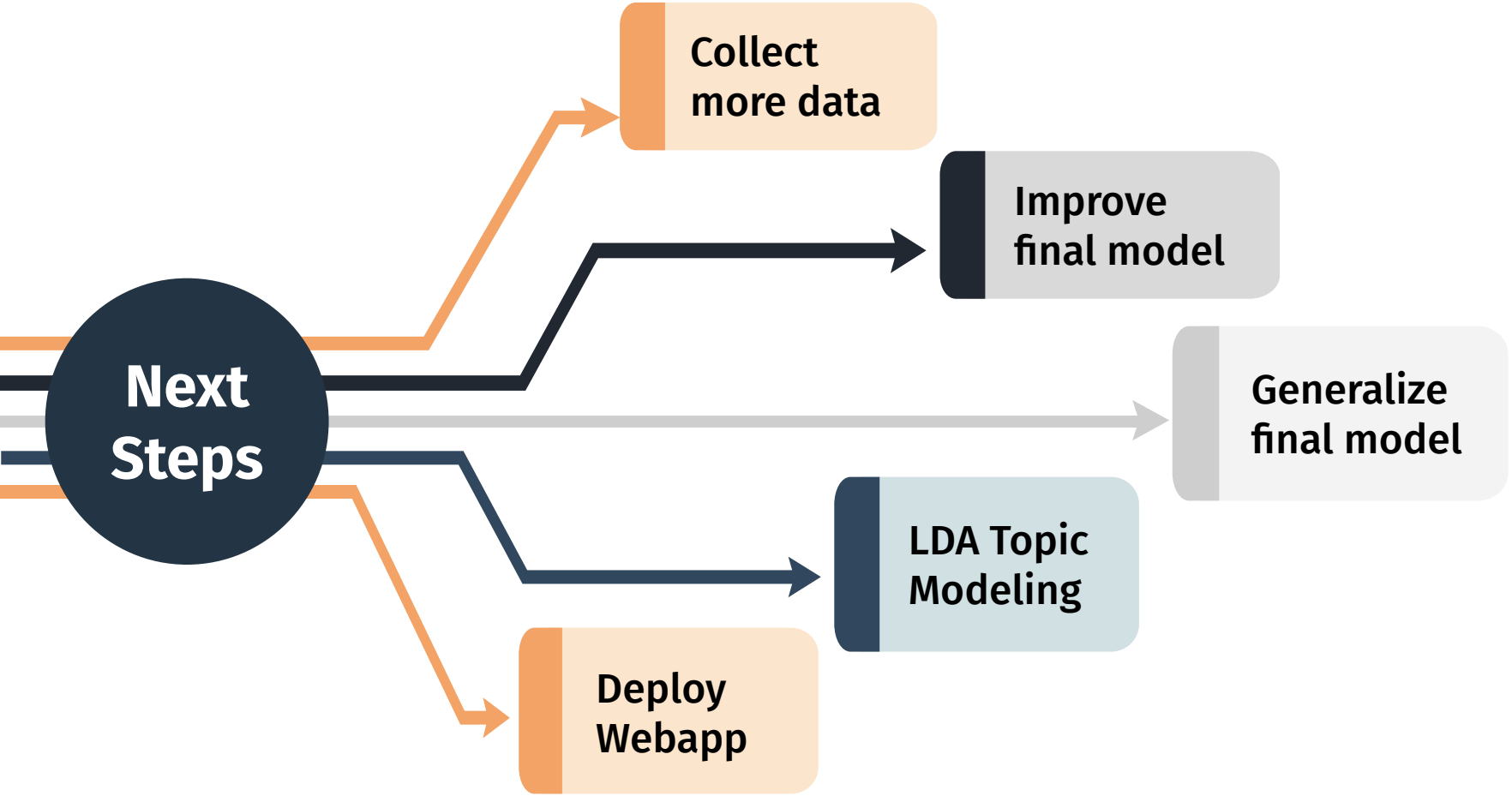
High **True Negative** Rate
Low **False Positive** Rate
Moderate **True Positive** Rate



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