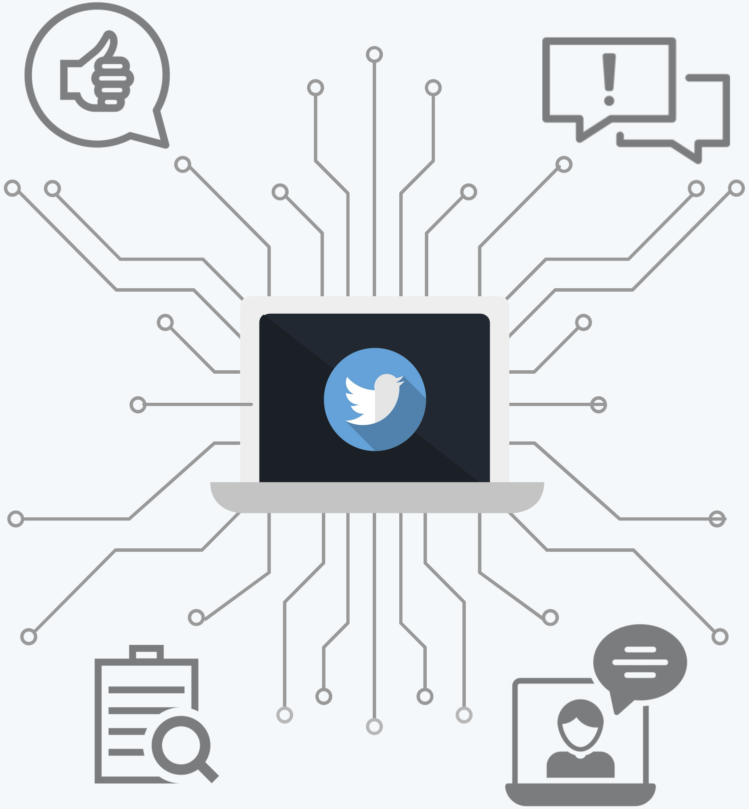
Twitter Hate Speech Detection

Can Content Moderation be Automated?

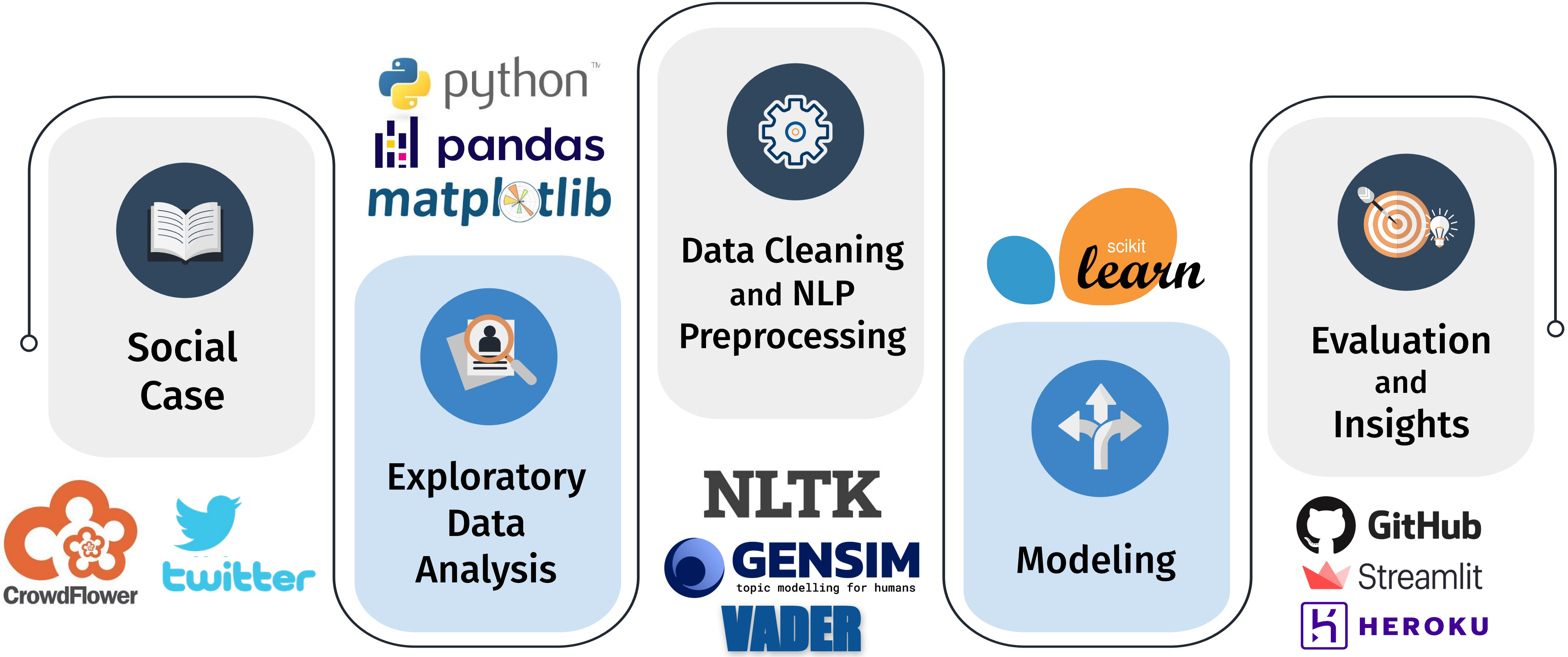
Flatiron School Capstone
Sidney Kung

The Problem of Human Content Moderation

- **Every major tech company** uses third-party contractors
- **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

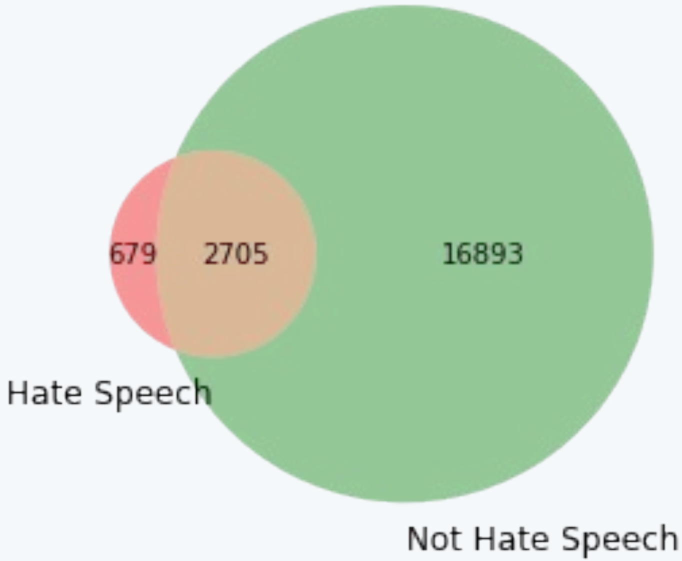
Evaluation and Insights

GitHub

Streamlit

HEROKU

Unique Words per Label



Data Understanding

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

24,802 Tweets

20,277 Word Vocabulary

6% Hate Speech

94% Not Hate Speech

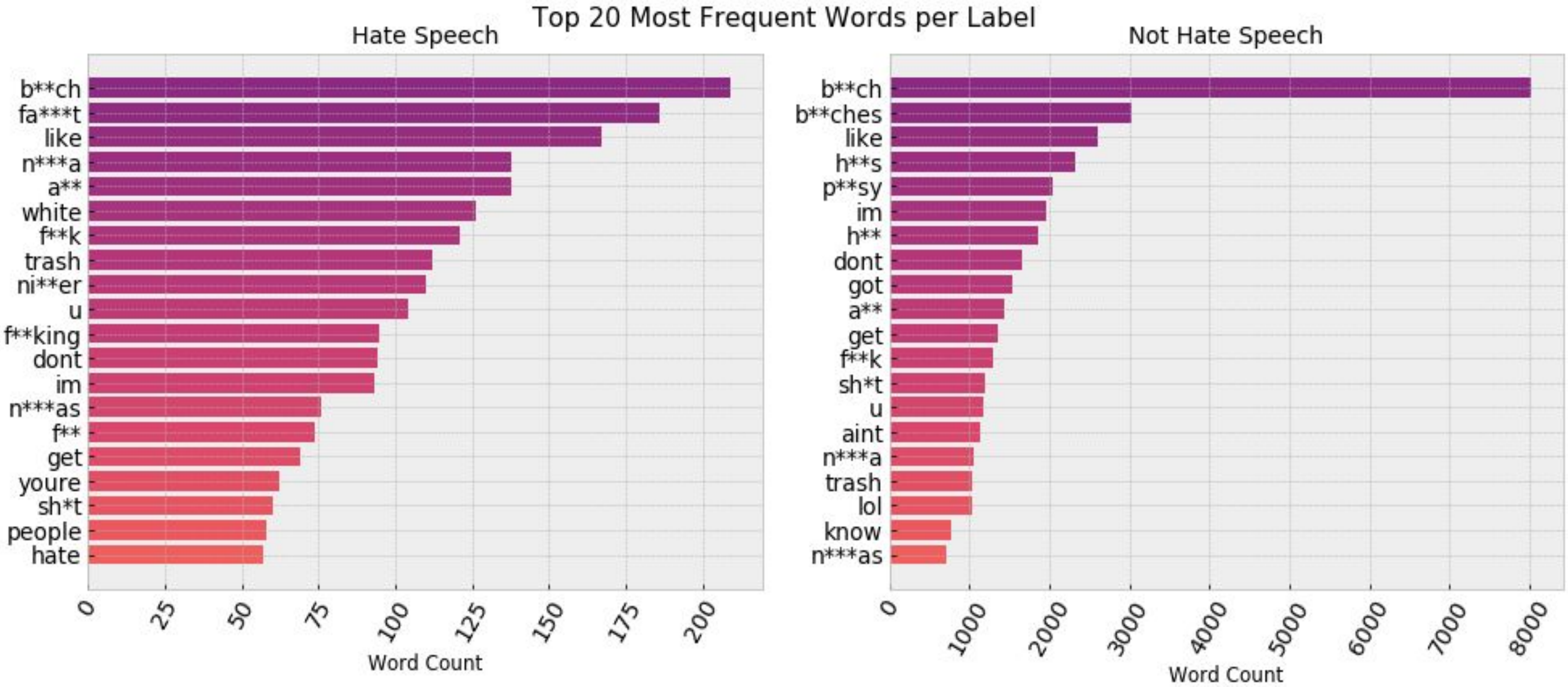
Data Analysis

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

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

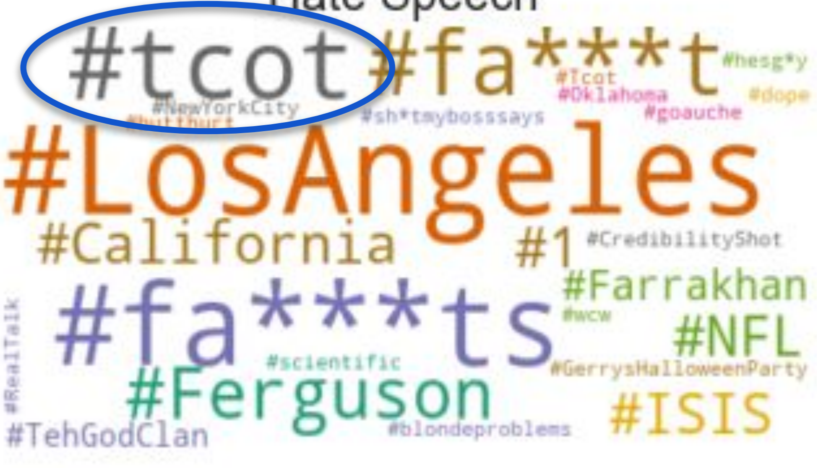
3. What is the **overall polarity** of the tweets?

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

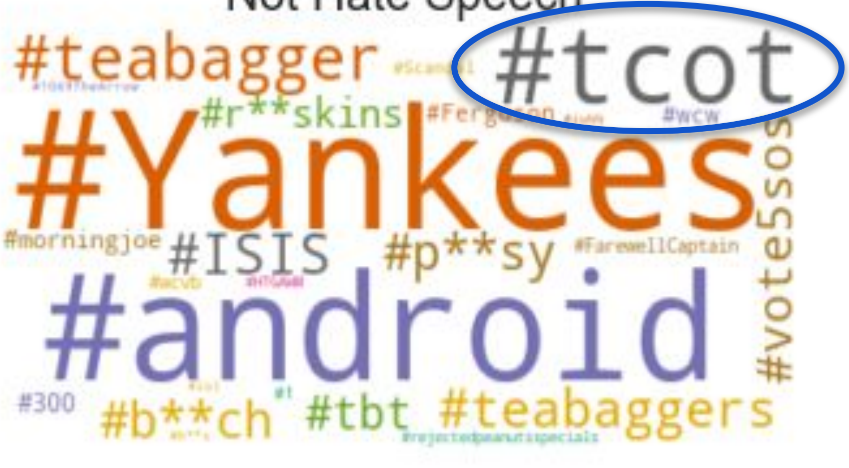


What are the **popular hashtags** of each tweet type?

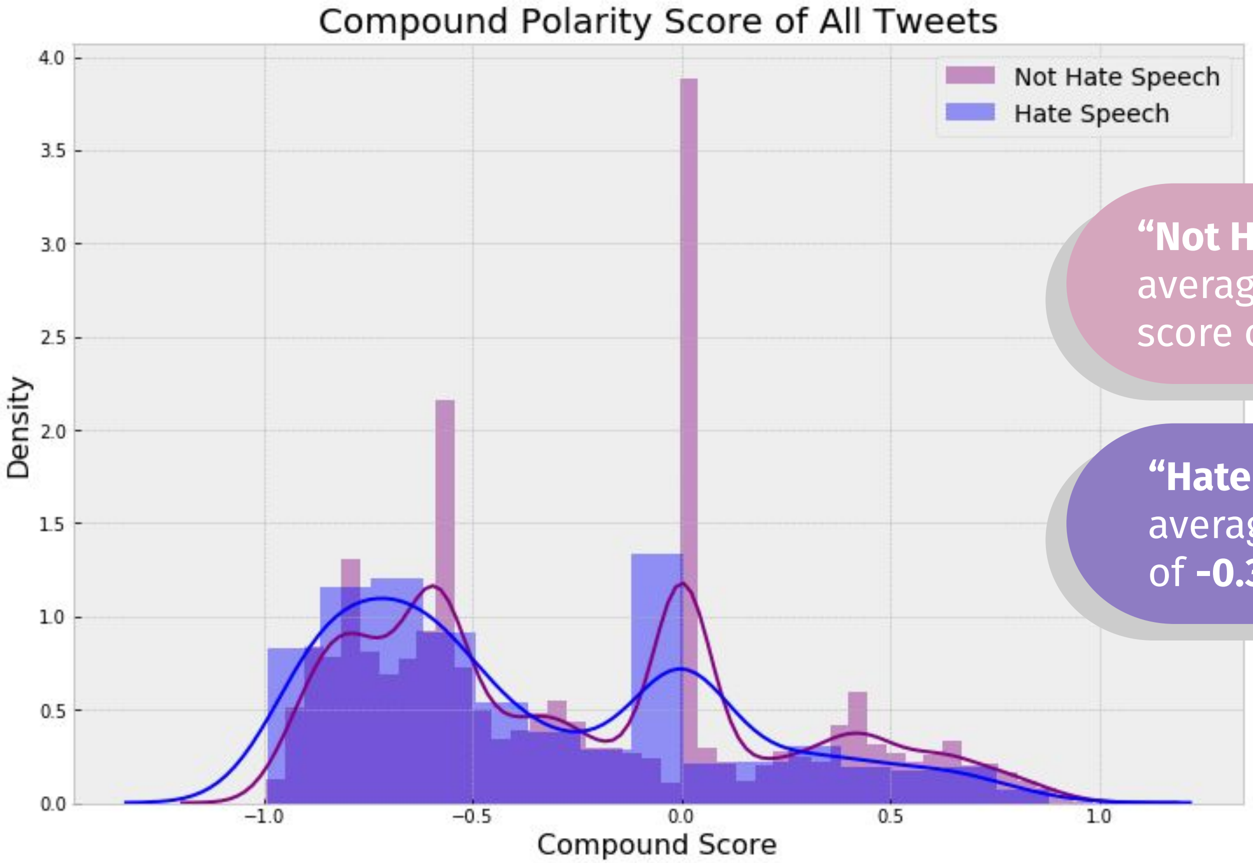
Hate Speech



Not Hate Speech



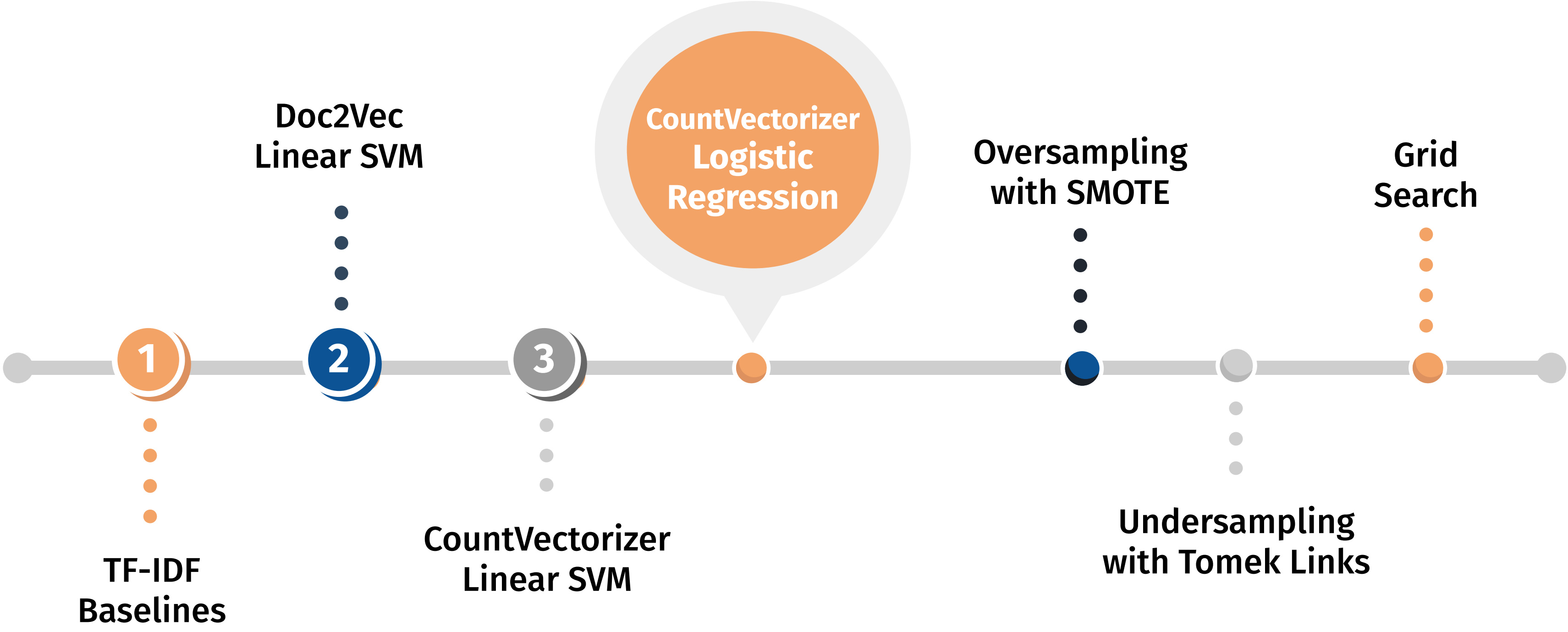
What is the **overall polarity** of the tweets?



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

“Hate Speech” tweets: average compound score of **-0.363**

Modeling Process



Baseline Model

Naive Bayes
with TF-IDF Vectorization

F1 Score	0.1923
Recall	0.1254

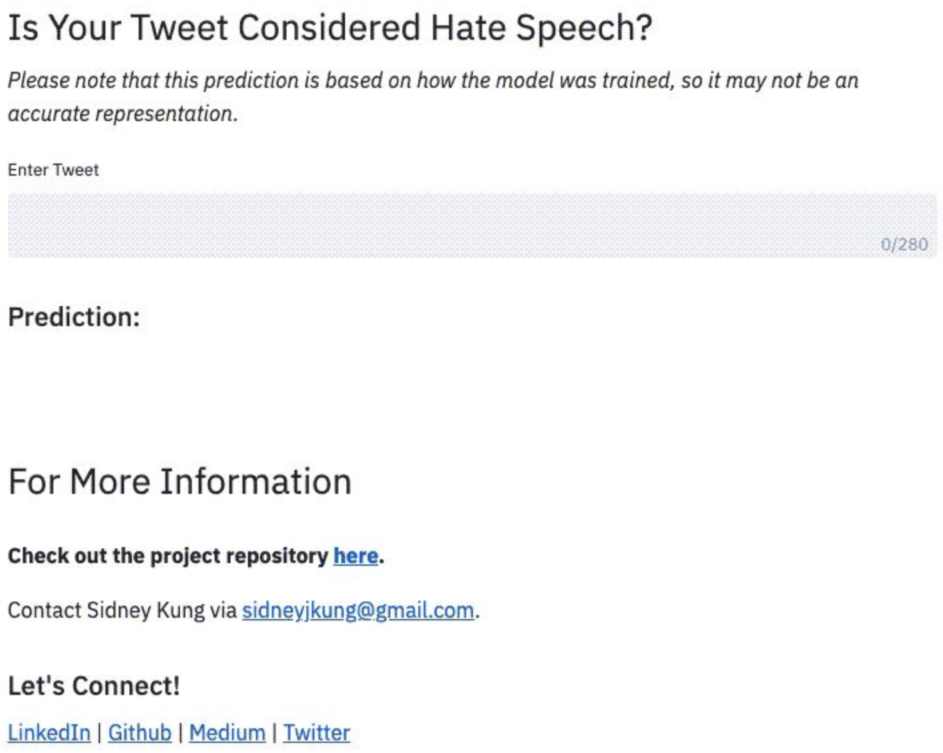
Final Model

Logistic Regression
with Count Vectorization

F1 Score	**0.3958**
Recall	**0.624**

Model Deployment

hate-speech-predictor.herokuapp.com



Is Your Tweet Considered Hate Speech?

Please note that this prediction is based on how the model was trained, so it may not be an accurate representation.

Enter Tweet

0/280

Prediction:

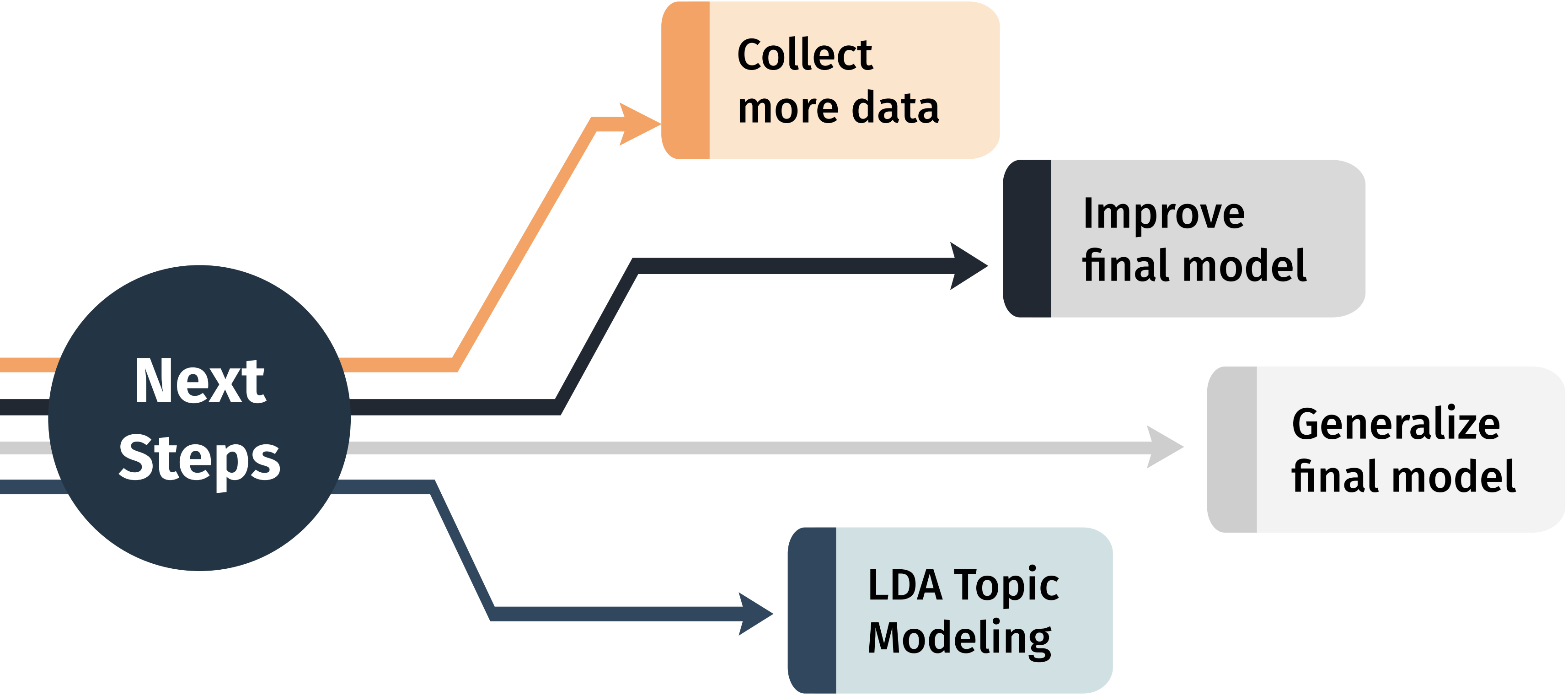
For More Information

Check out the project repository here.

Contact Sidney Kung via sidneyjkung@gmail.com.

Let's Connect!

LinkedIn | Github | Medium | Twitter



Next Steps

Collect more data

Improve final model

Generalize final model

LDA Topic Modeling

Thank You!

GitHub Repository
github.com/sidneykung/twitter_hate_speech_detection

Web App on Heroku
hate-speech-predictor.herokuapp.com

SidneyKung.com

SidneyJKung@gmail.com

linkedin.com/in/sidneykung

@Sidney_K98

Presentation Template: SlidesGo