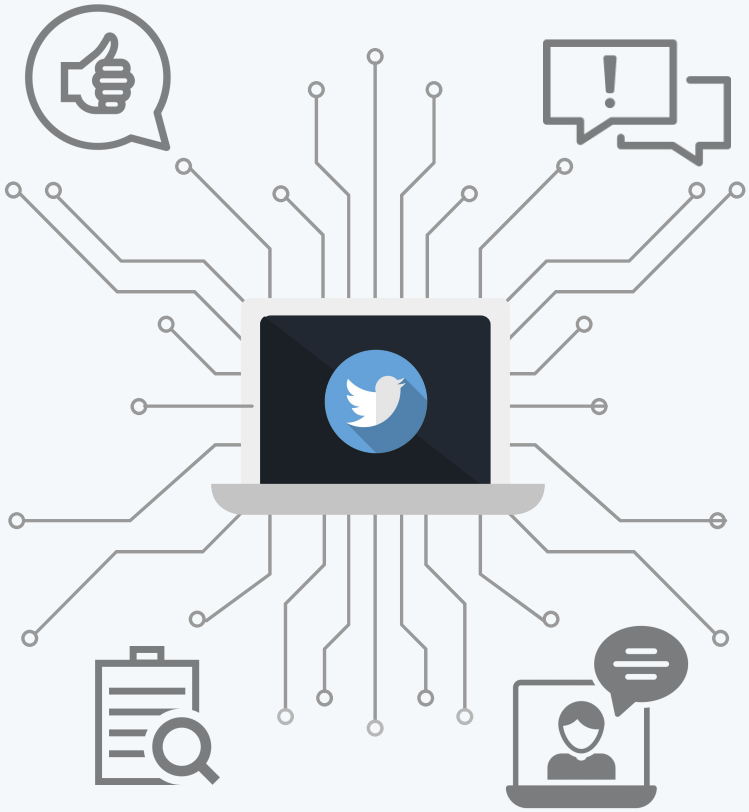
Twitter Hate Speech Detection

Can Content Moderation be Automated?

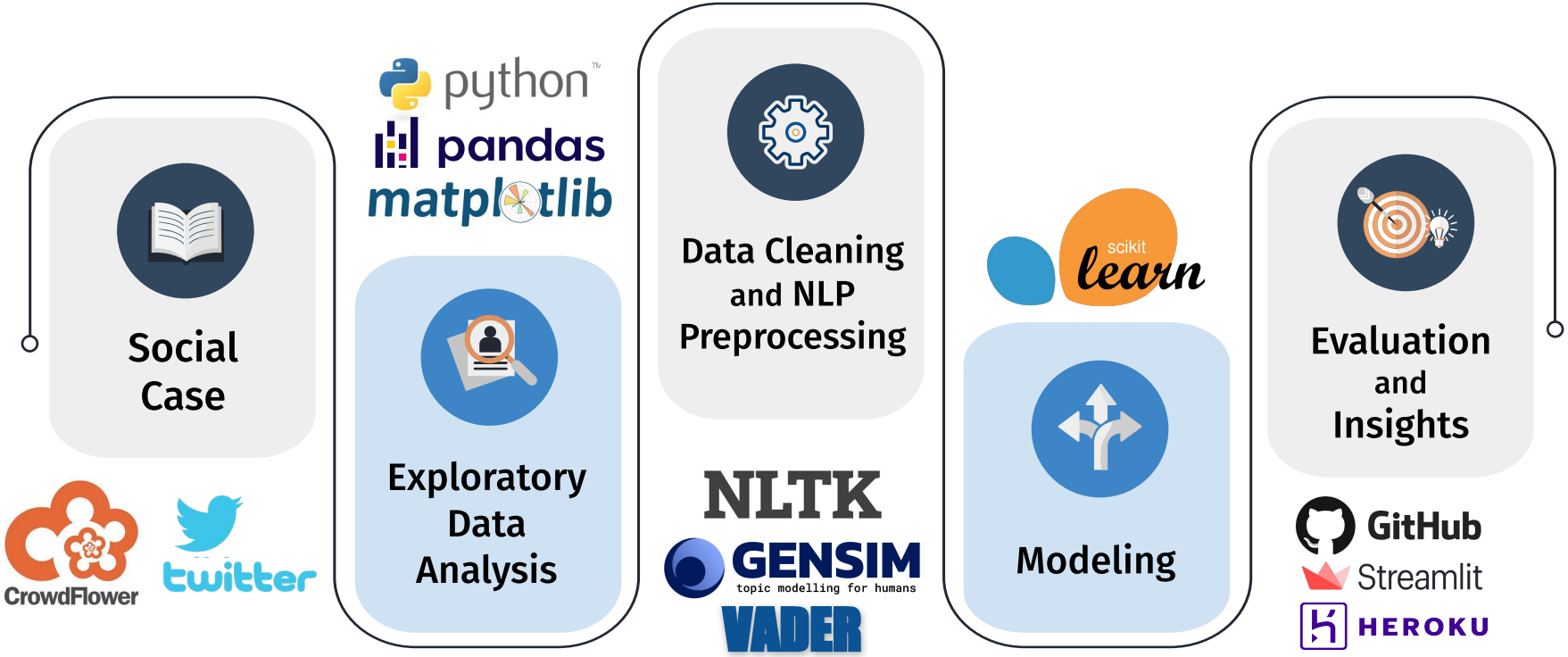
Flatiron School Capstone
Sidney Kung

The Problem of Human Content Moderation

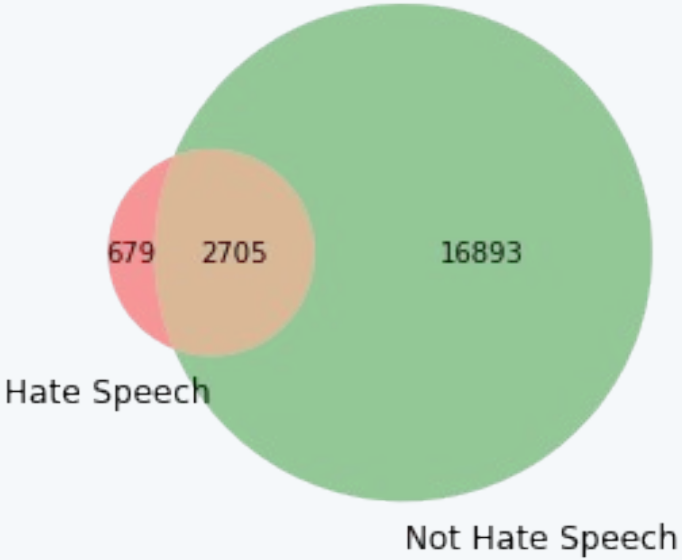
- **Every major tech company** uses third-party contractors
- **Automating** this process could **reduce labor exploitation**
- What is **Hate Speech?**



CRISP-DM Process



Unique Words per Label

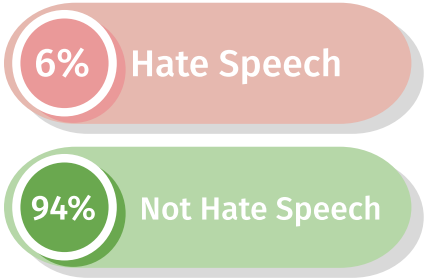


Data Understanding

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

24,802 Tweets

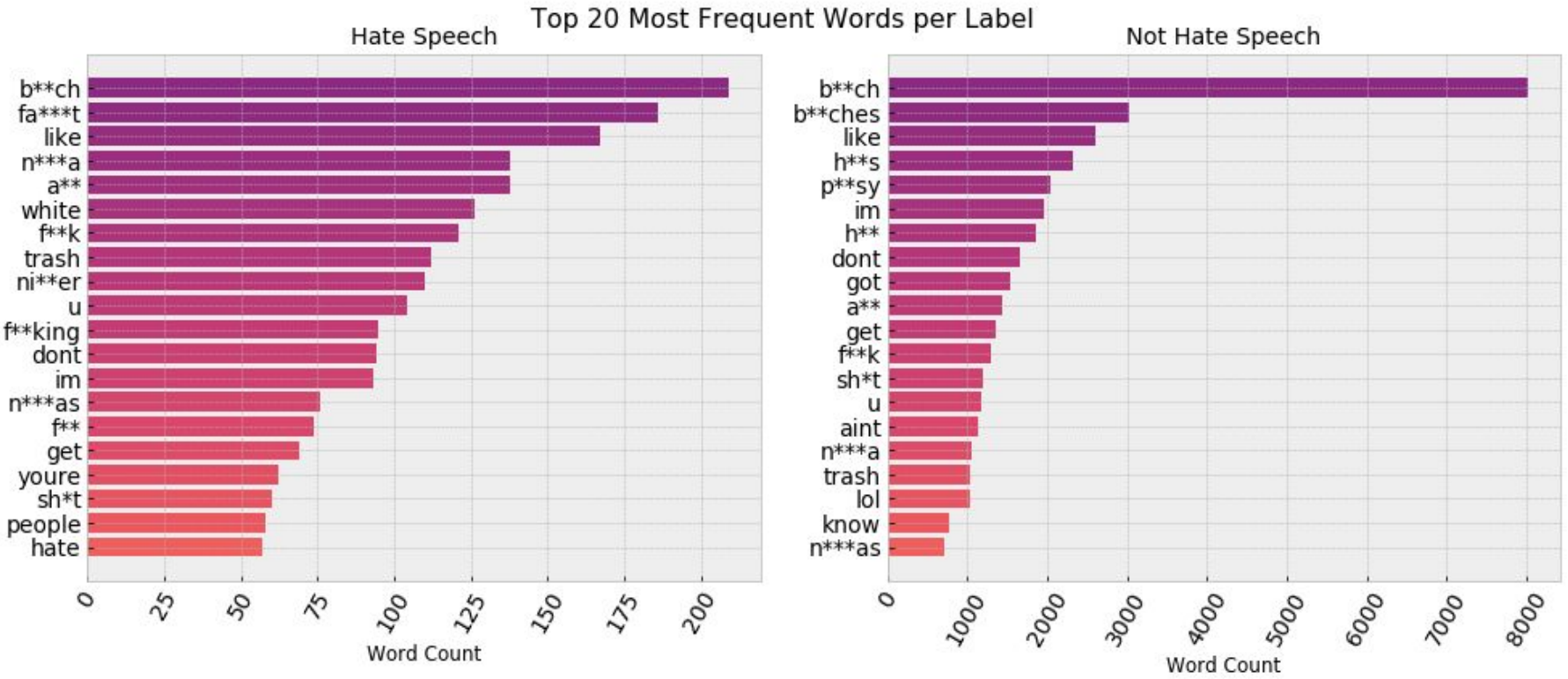
20,277 Word Vocabulary



Data Analysis

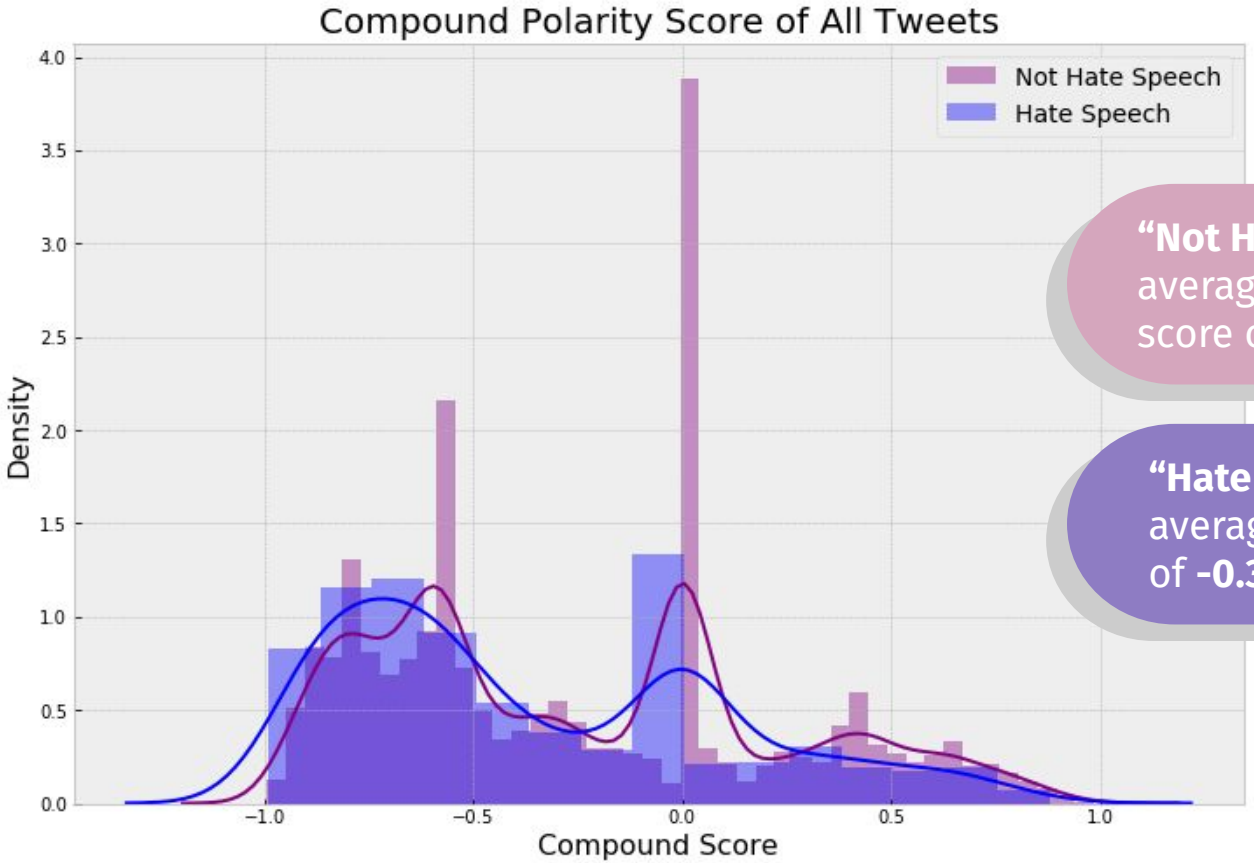
1. What are the linguistic differences between hate speech and offensive language?
2. What are the most popular hashtags of each tweet type?
3. What is the overall polarity of the tweets?

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



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

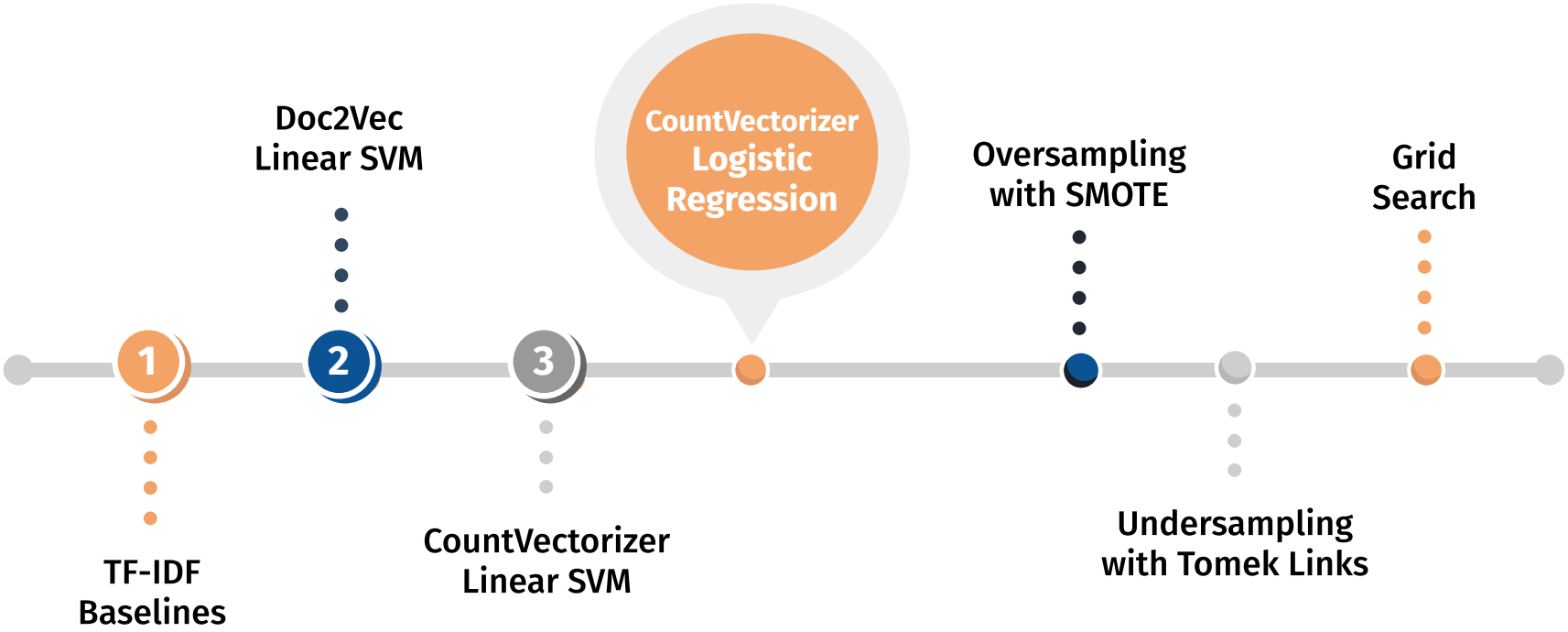
3. What is the overall polarity of the tweets?



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

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

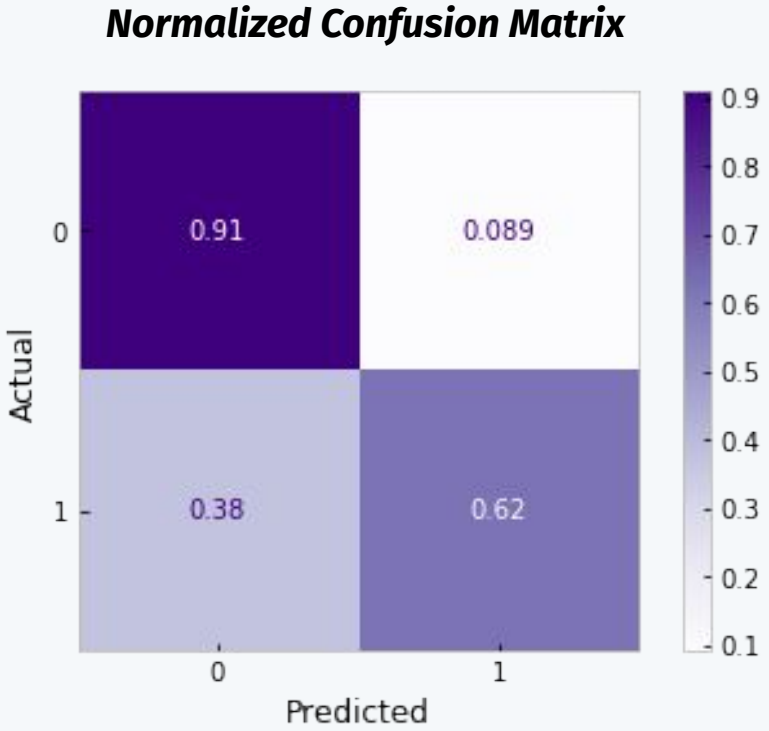
Modeling Process



Final Model Evaluation

Logistic Regression Classifier
with CountVectorizer

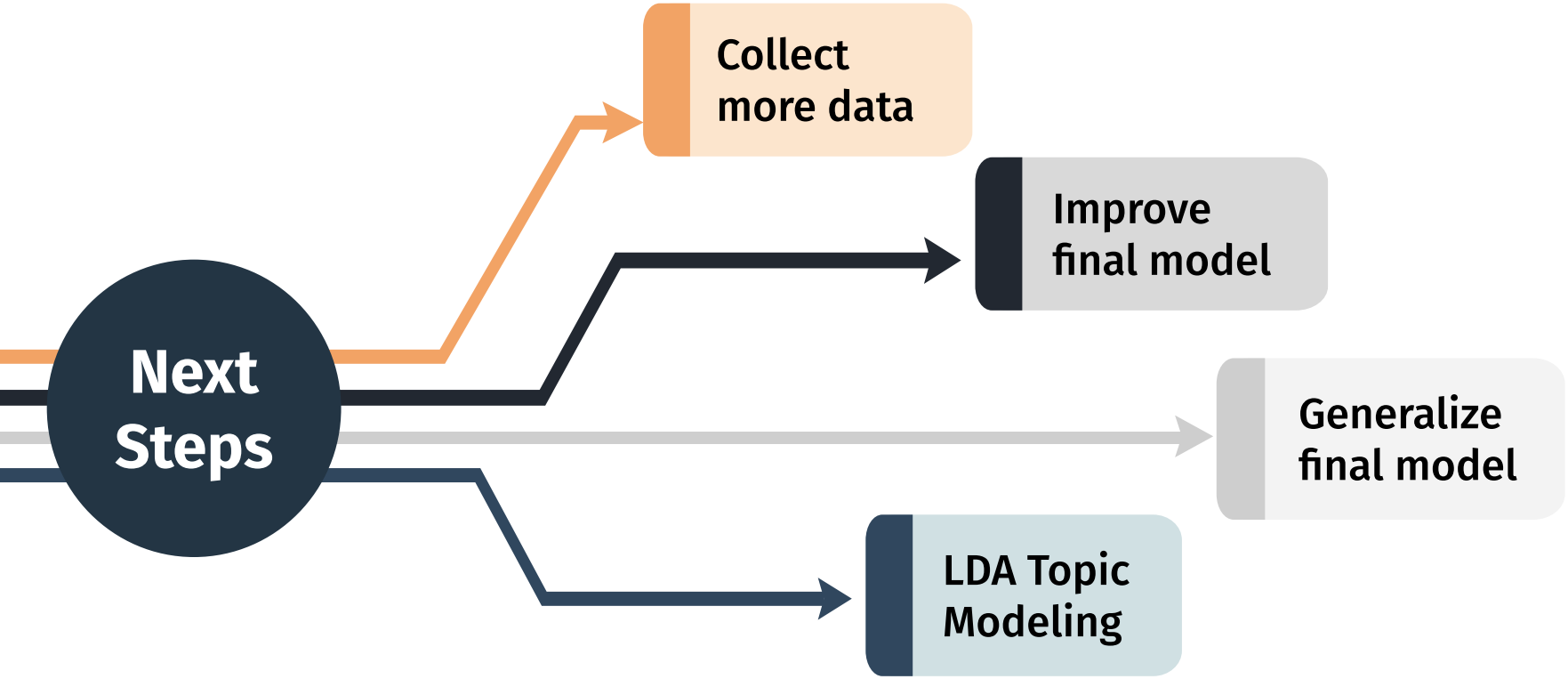
F1 Score	**0.3958**
Recall	0.624



Conclusion

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

Thank You!

GitHub Repository
github.com/sidneykung/twitter_hate_speech_detection

SidneyJKung@gmail.com

linkedin.com/in/sidneykung/

@Sidney_K98

Presentation Template:
SlidesGo