

Twitter Hate Speech Detection

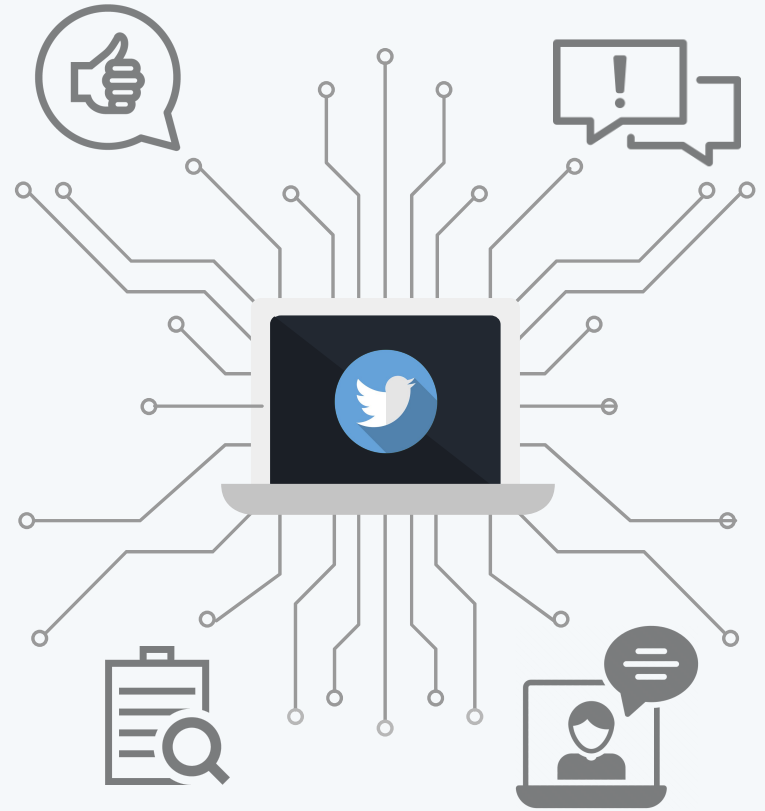
*Can Content Moderation
be Automated?*

Flatiron School Capstone
Sidney Kung

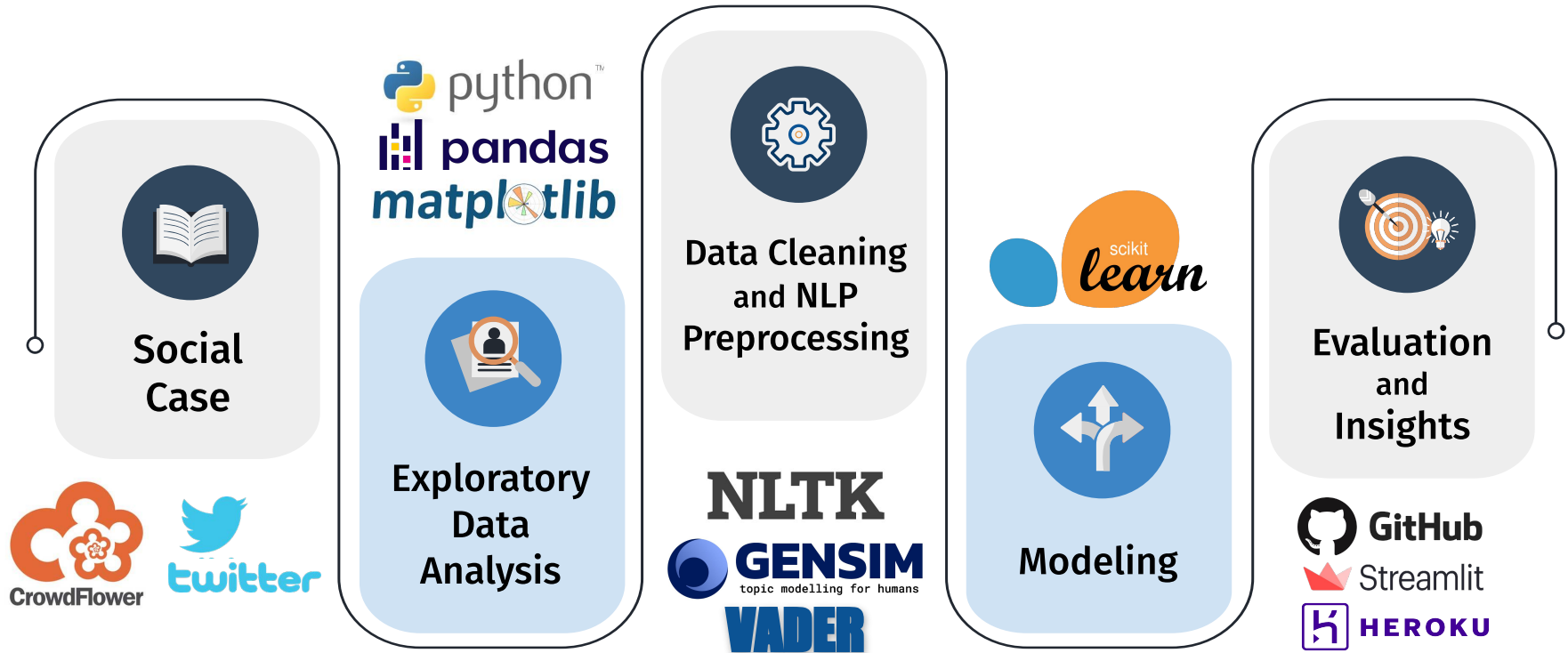


The Problem of Human Content Moderation

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



CRISP-DM Process

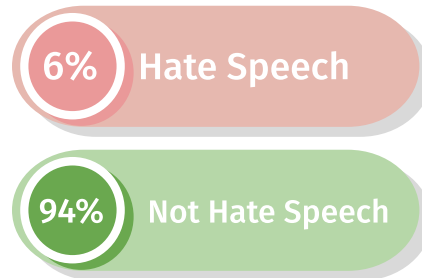
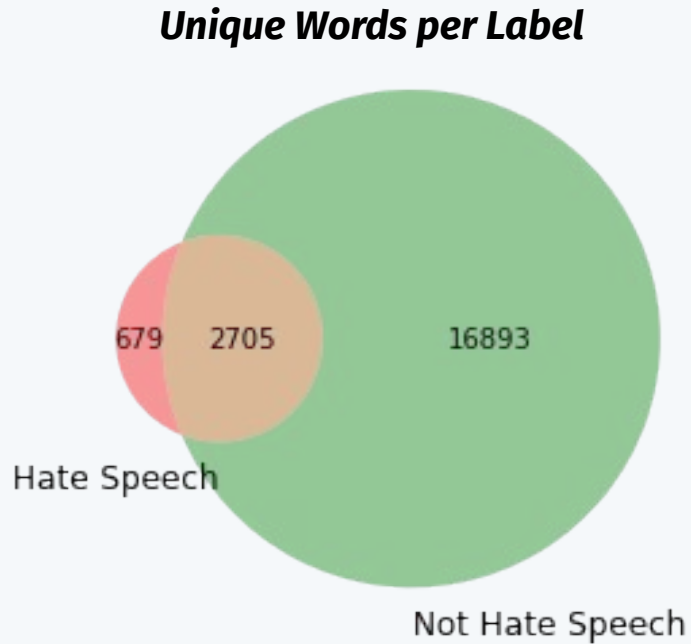


Data Understanding

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?

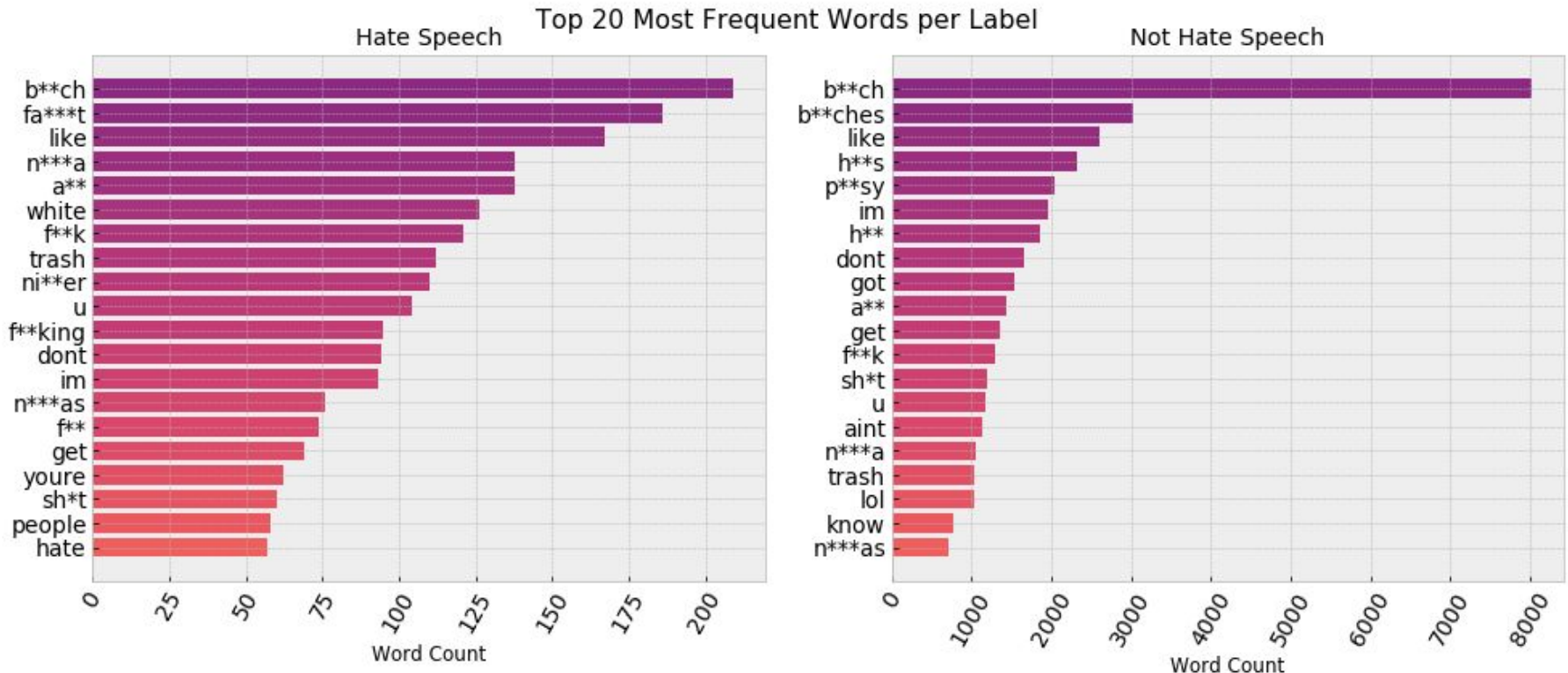
2

What are the most popular hashtags of each tweet type?

3

What is the overall polarity of the tweets?

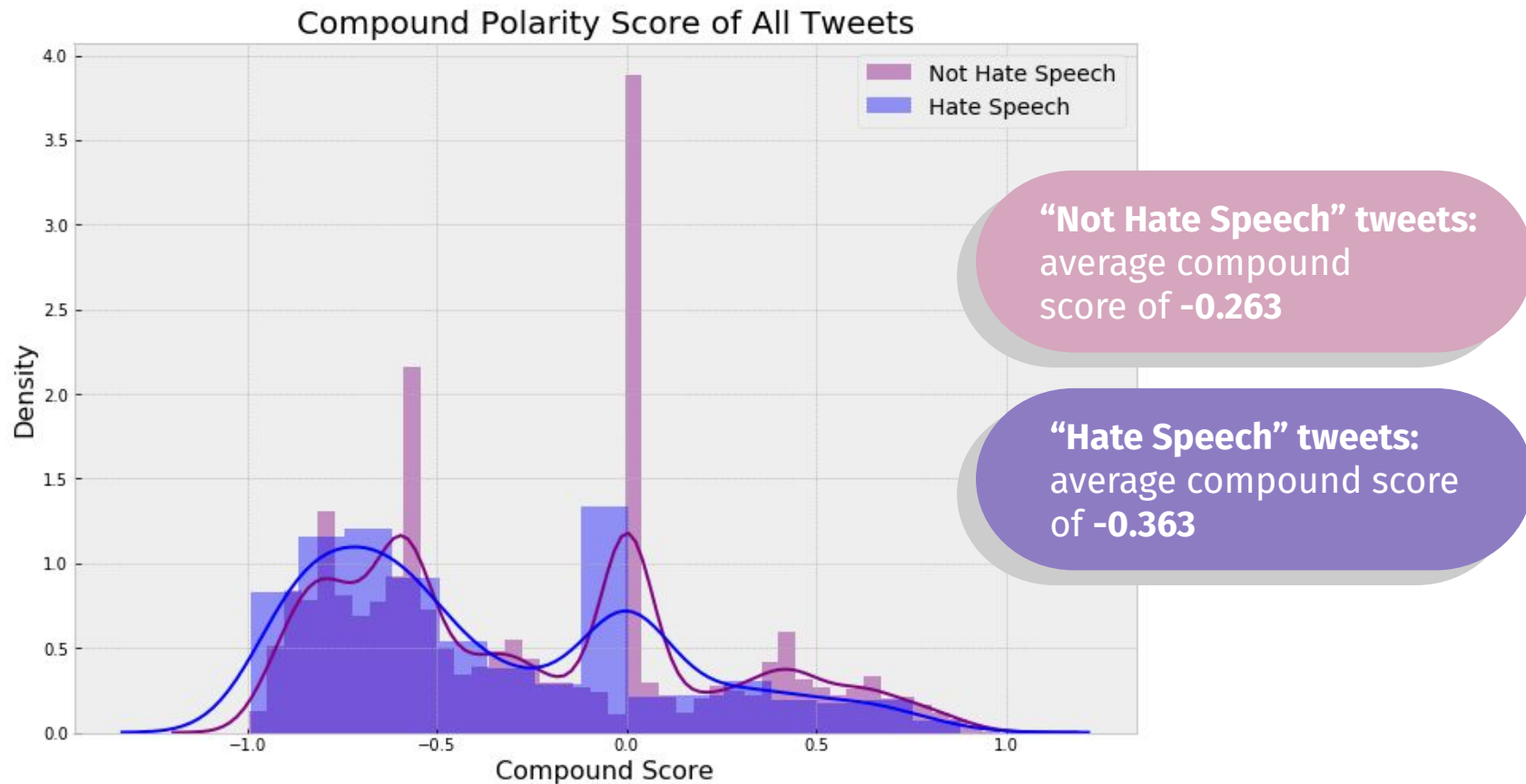
1. What are the linguistic differences between hate speech and offensive language?



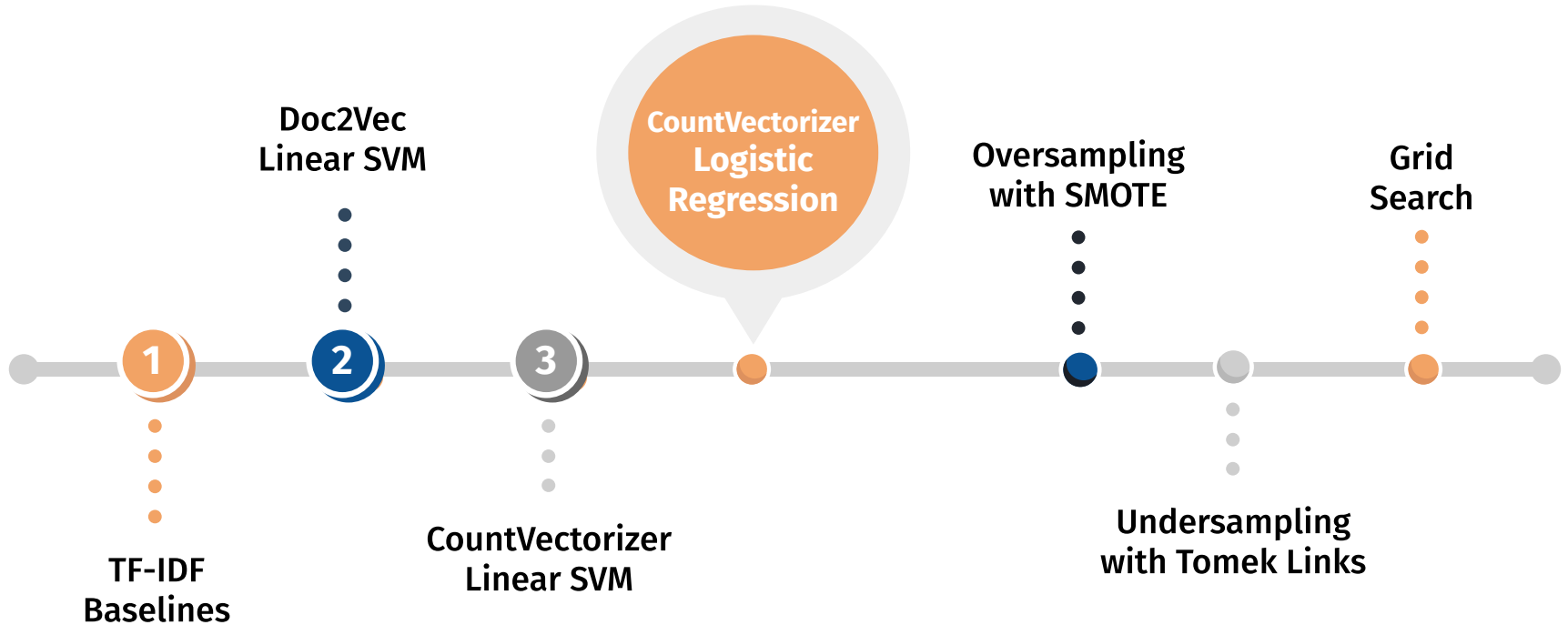
2. What are the most popular hashtags of each tweet type?



3. What is the overall polarity of the tweets?



Modeling Process



Final Model Evaluation

Logistic Regression Classifier
with CountVectorizer

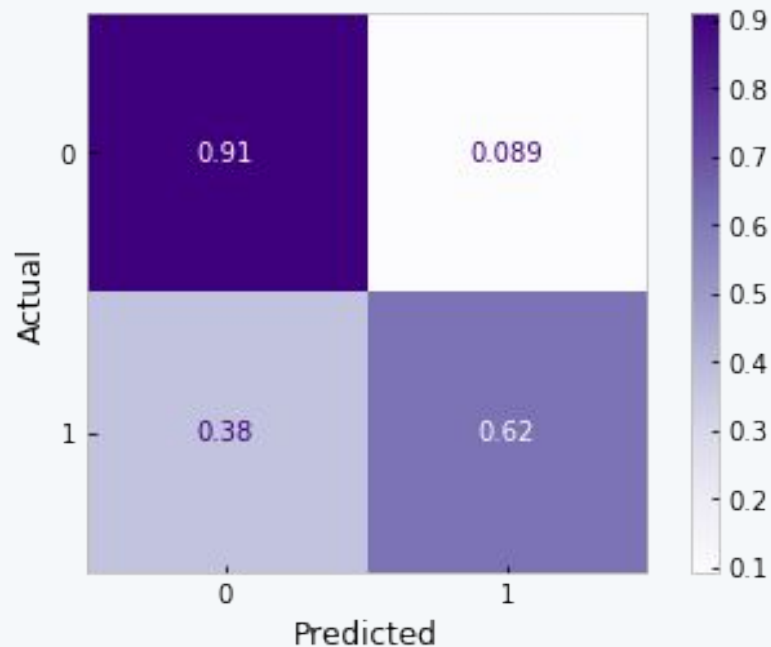
F1 Score

0.3958

Recall

0.624

Normalized Confusion Matrix

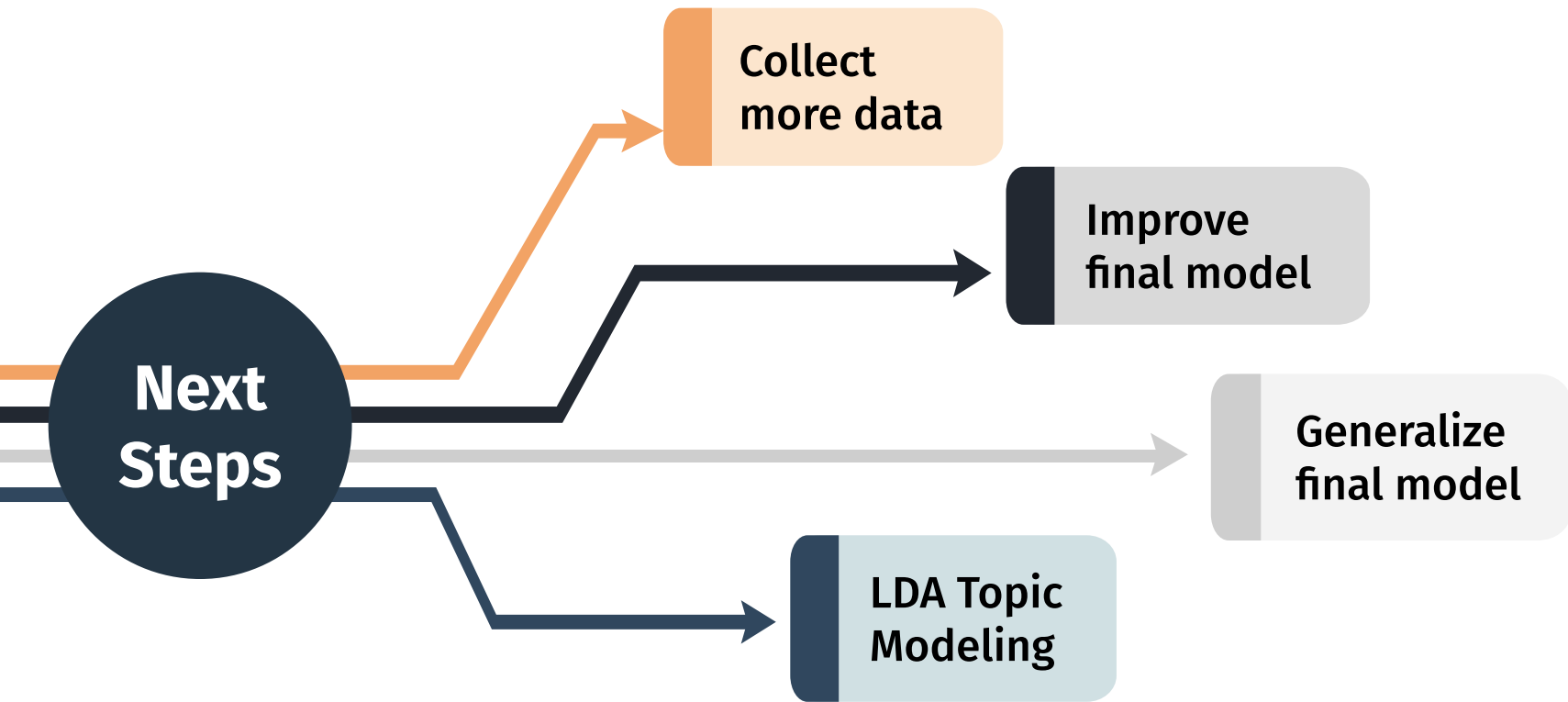


Conclusion

Major Roadblocks

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





Thank You!



GitHub Repository

[github.com/sidneykung/
twitter_hate_speech_detection](https://github.com/sidneykung/twitter_hate_speech_detection)



Web App on Heroku

[hate-speech-predictor.
herokuapp.com/](https://hate-speech-predictor.herokuapp.com/)



SidneyKung.com



SidneyJKung@gmail.com



linkedin.com/in/sidneykung/



[@Sidney_K98](https://twitter.com/Sidney_K98)

Presentation Template:
SlidesGo