#AnalyzingHateSpeechOnTwitter

@ChristopherDeLaCruz (not my actual twitter handle)

#WhyHateSpeech?

- Over the past few years, multiple organizations (including the FBI) have reported increases in hate speech
- Hate crimes have also increased and generally reflect similar patterns shown in hate speech trends
- In order to do a better job of protecting these classes, we must map a better understanding of current trends and patterns in hate speech

#WhatIsHateSpeech?

Is calling someone dumb hate speech?

Depends.

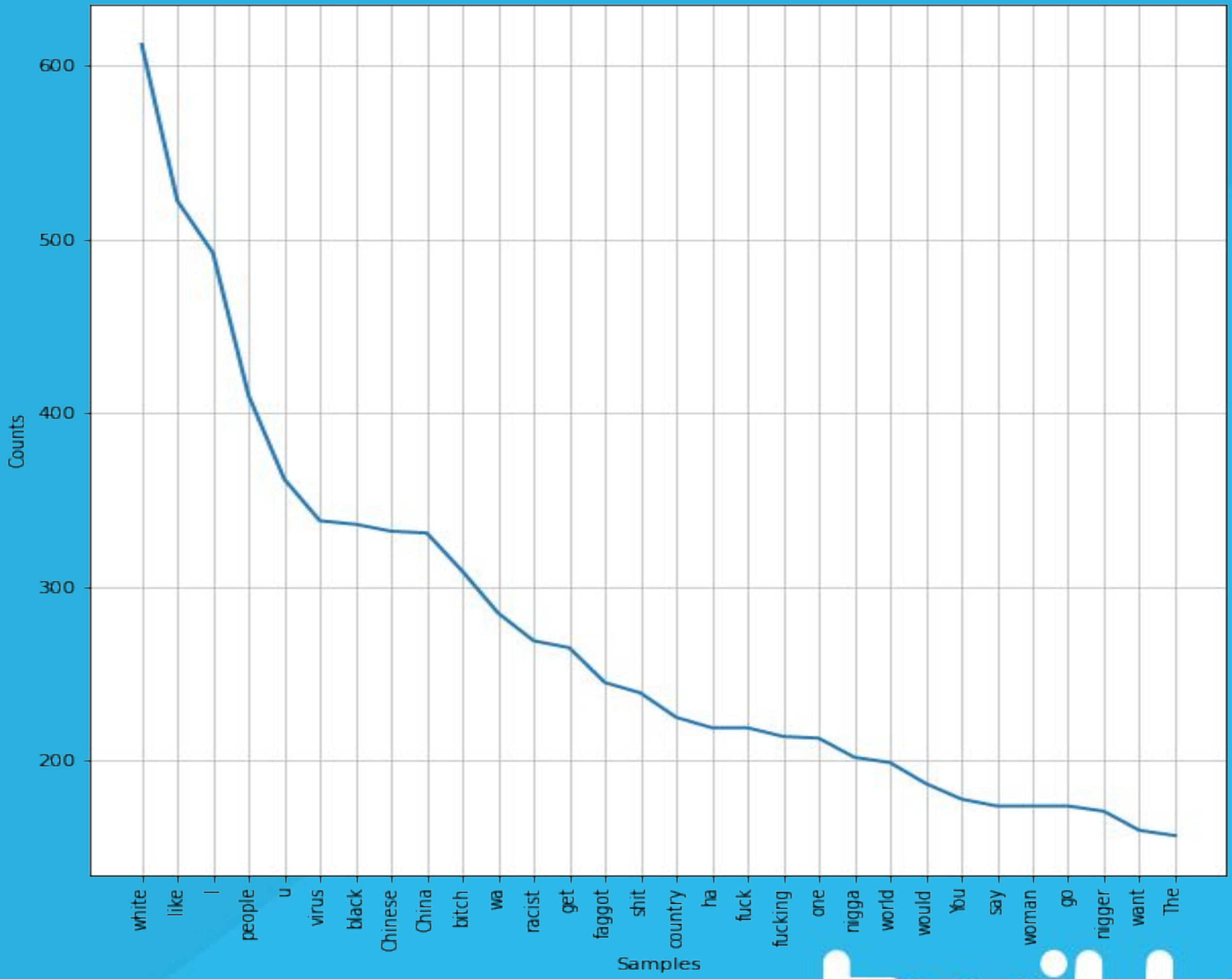
Hate speech is derogatory language aimed at a person based on an identification class (race, gender, class, religion, sexual orientation, etc)

#HateSpeechData

Approximately 70,000 labeled tweets (Hate tweets and Not Hate Tweets) were gathered from four different Kaggle datasets and one dataset from Georgia Tech’s CLAWS

Labels are:
1 - Hate Speech
0 - Not Hate Speech

#HateSpeechResults



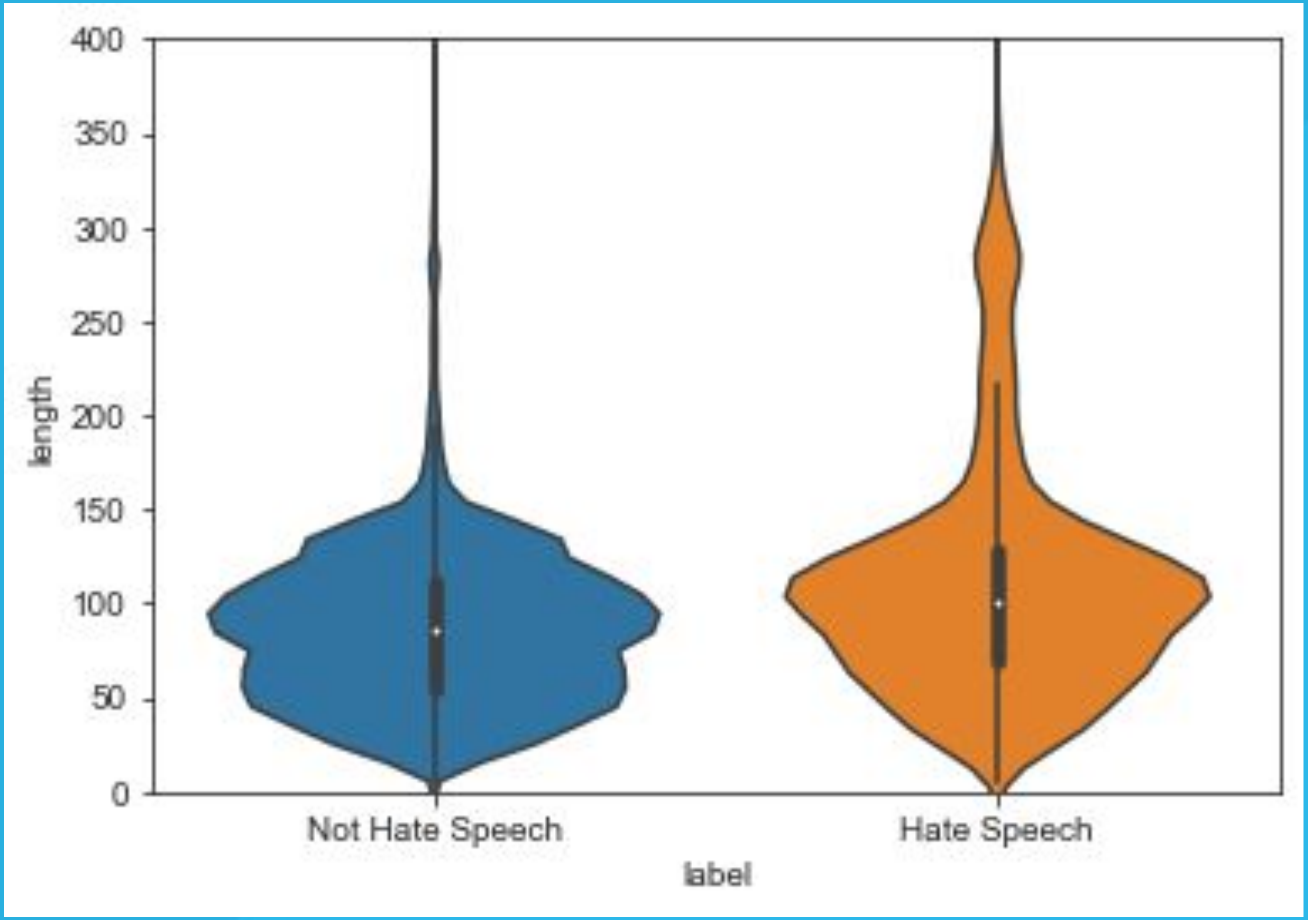
Identity Classes Mentioned Most Often:

- Race: Asian
- Race: Black
- Race: White

Gender: Women

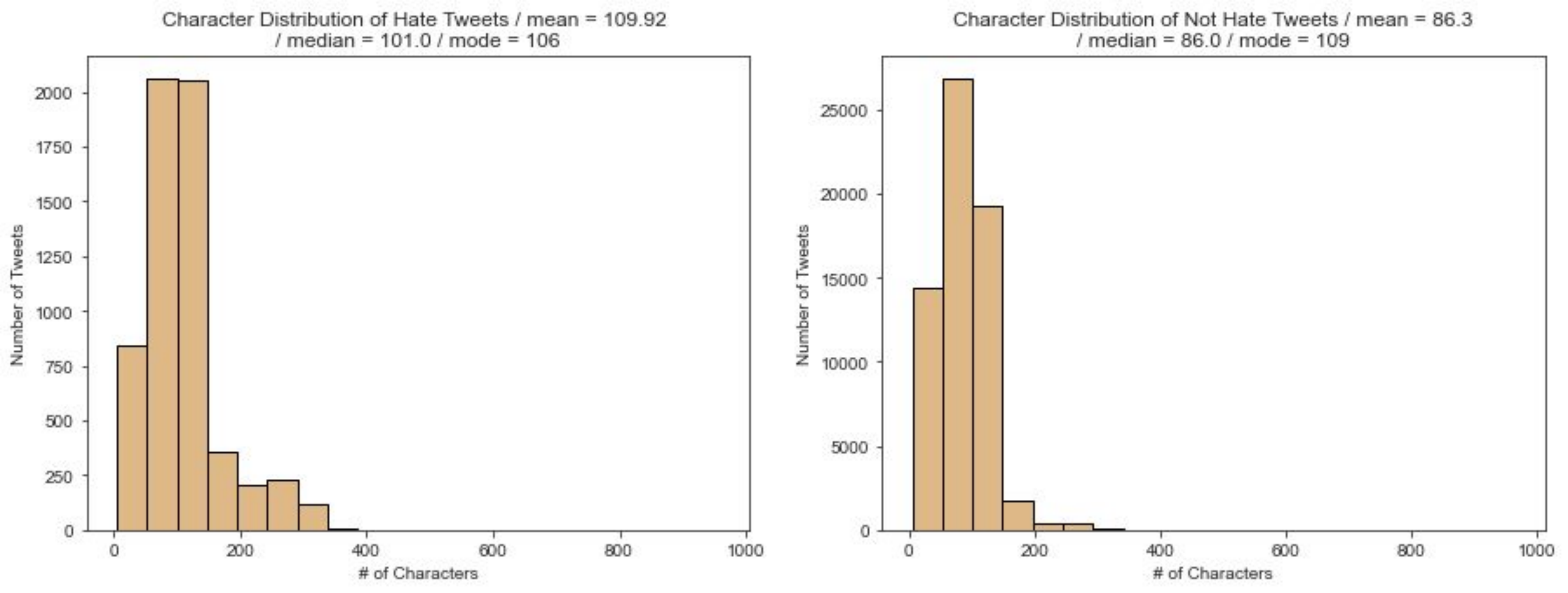
- Sexual Orientation: Gay

#HateSpeechResults



Hate tweets look longer and the math agrees but how much longer?

#HateSpeechResults



Hate tweets are about 24 characters longer on average

#HateSpeechResults

Over 7,000 models were run but there could only be ONE winner.

First priority - high precision in detecting hate speech

Second priority - high recall in detecting hate speech

Ultimately the best model was...

#DRUMROLL

#HateSpeechResults

Support Vector Classifier!

- Able to predict hate speech with 85% precision!
- Recall is 33% and cannot be brought up without sacrificing precision

#NextSteps

- Looking into ways to continue improving the model’s ability to detect hate speech (more data, different sampling techniques, etc)

- Deploying the model on freshly scraped tweets to gather additional information about hate speech and establishing deeper patterns