#AnalyzingHate SpeechOnTwitter

@ChristopherDeLaCruz (not my actual twitter handle)

#WhyHateSpeech?

- Hate crimes & speech have been on a steady rise since 2014
- Understanding the patterns behind recent hate speech trends will help us better understand which classes are at highest risk
- We want the model's predictions to be correct and we also want it to be able to capture most hate speech

#WhatIsHateSpeech?

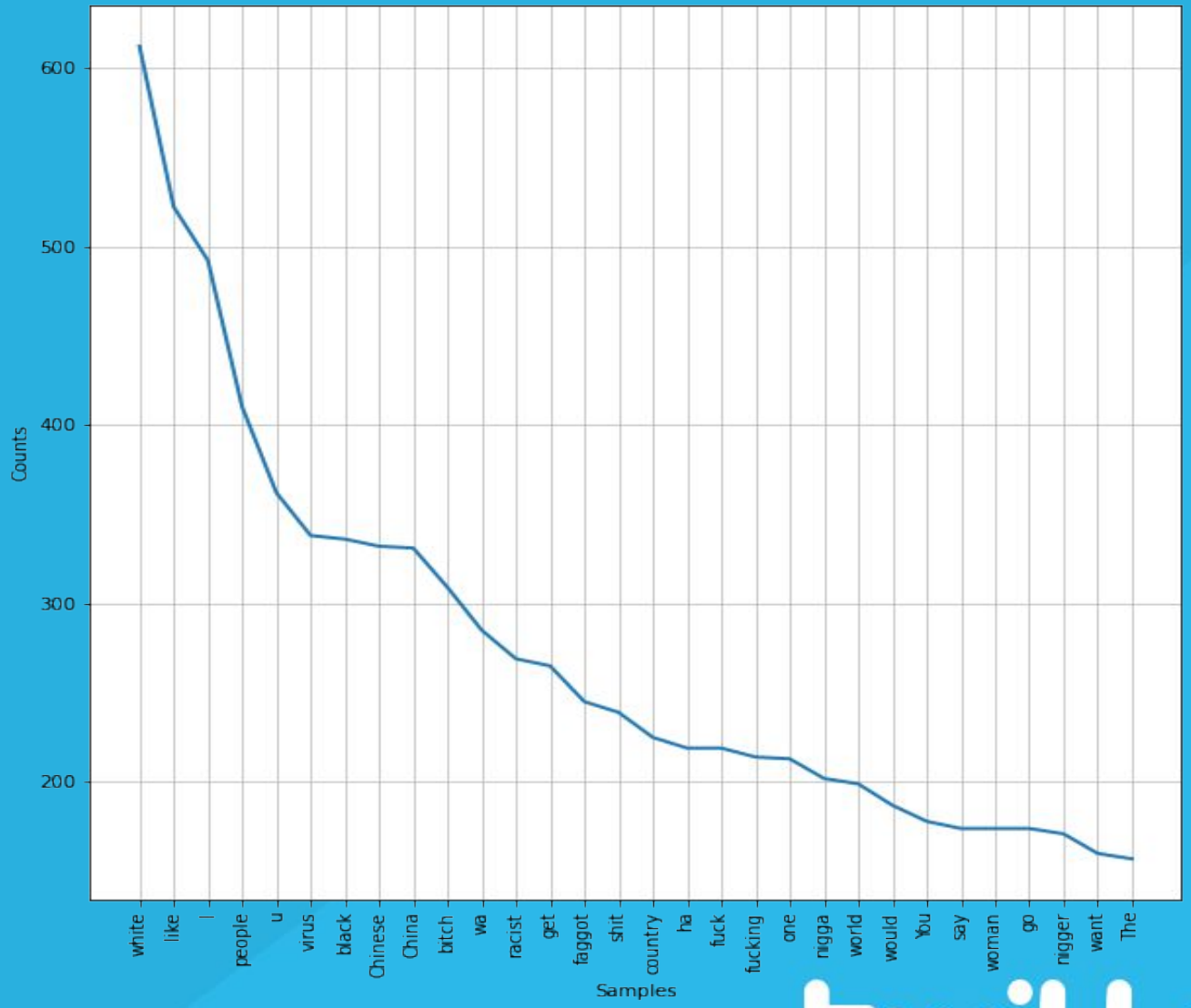
In simplest terms, hate speech is derogatory language aimed at a person based on an identification class (race, gender, class, religion, sexual orientation, etc)

#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:

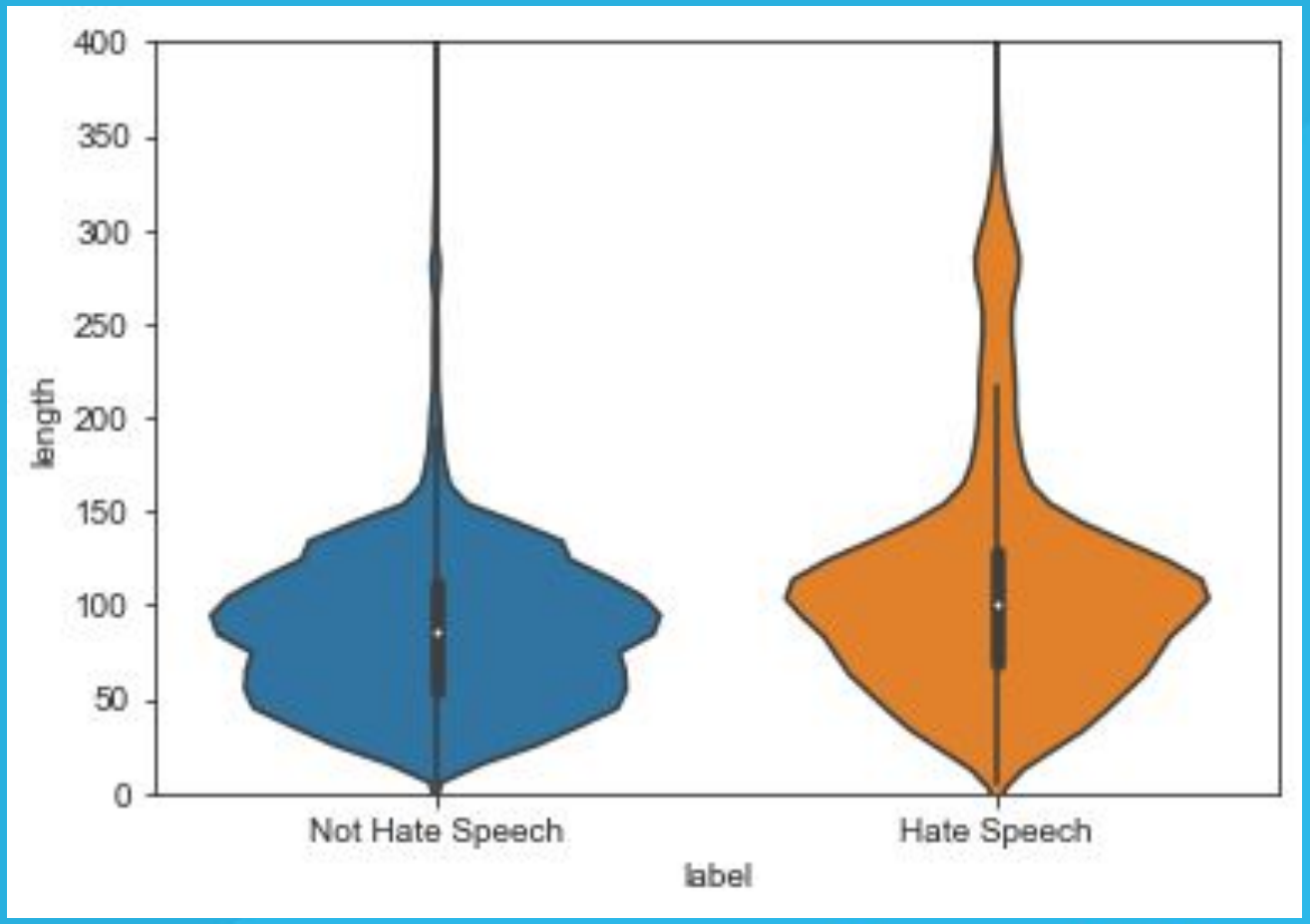
- Race: Asian
- Race: Black
- Race: White

Gender: Women

- Sexual Orientation: Gay

twitter

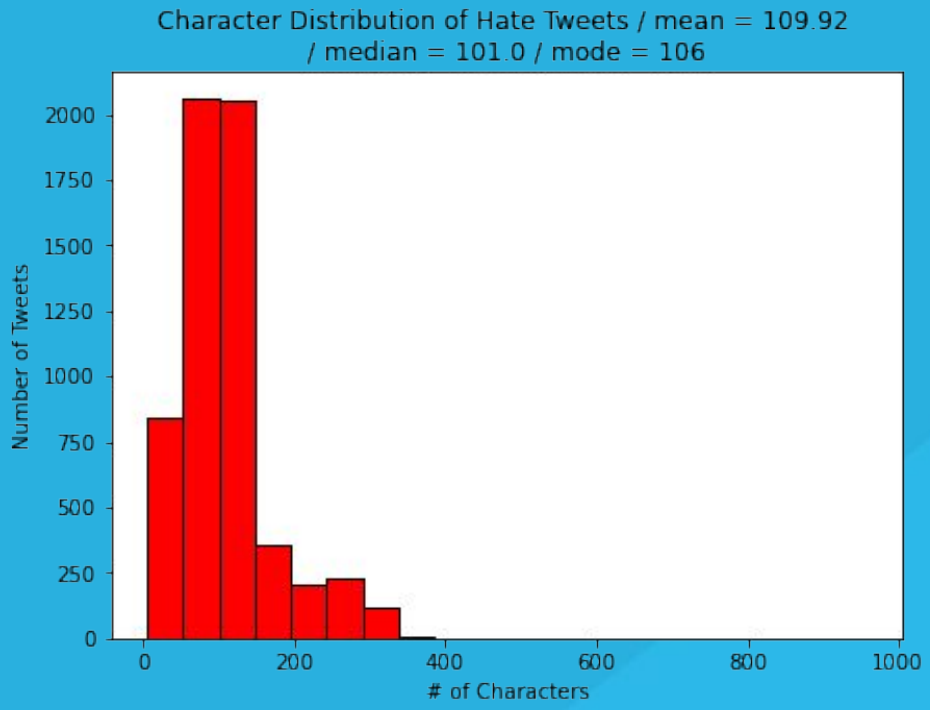
#HateSpeechResults

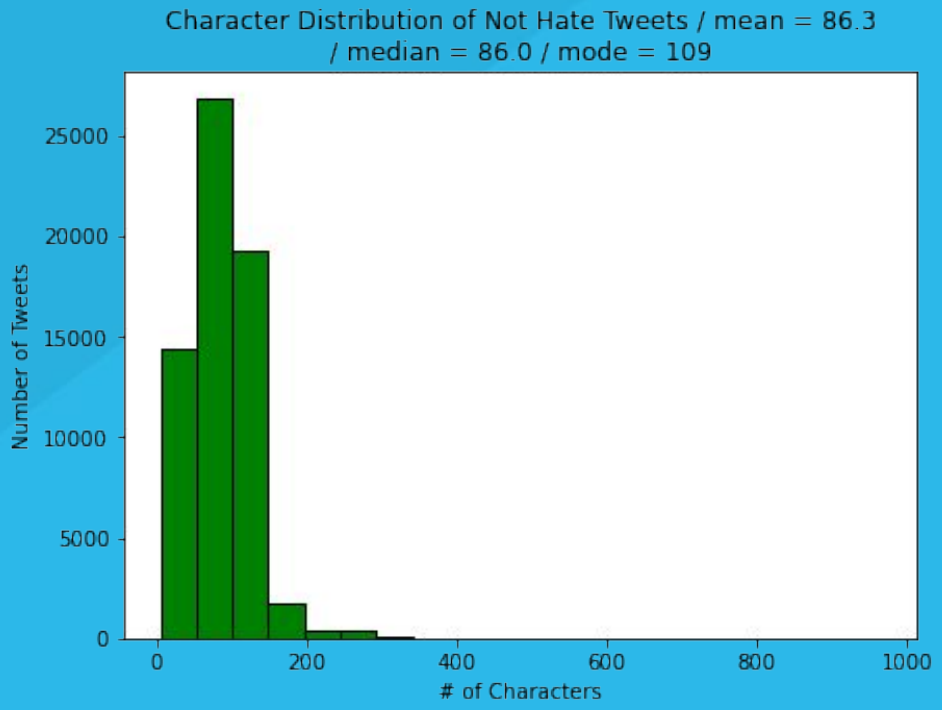


Hate tweets look longer and the math agrees but how much longer?

twitter

#HateSpeechResults





Hate tweets are about 24 characters longer on average

twitter

#HateSpeechResults

Baseline Model - 93% accurate if it simply predicts “Not Hate Speech” everytime

Ultimately, most important metric is how accurate the model is about predicting Hate Speech only

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

Ultimately the best model was...

#DRUMROLL

like a literal drumroll, this is not the name of the model

#HateSpeechResults

Pipeline: TF-IDF -> SMOTE -> Support Vector Classifier!

- Out of the 440 tweets that the model predicted as hate speech, 376 tweets were actually hate speech (85% precision)
- Out of 1,175 hate tweets, the model correctly identified only 376 of these tweets (32% recall)

twitter

#HateSpeechResults

The model excels at identifying hate speech revolving around current events (pandemic, politics, etc).

<u>Trending Hashtags on Hate Tweets Model Got Correct</u>

- #allahsoil (regarding a book titled ‘Allah’s Oil’)
- #politics
- #trump
- #libtard
- #liberal
- #chinesevirus

#NextSteps

- We need more tweets. A LOT more tweets.

- Look into ways to continue improving the model's ability to detect and classify hate speech (more data, different sampling techniques, etc)

#ThankYou!

Presentation By Christopher de la Cruz

Email: cdelacruz2013@gmail.com

Github: cdlc01

LinkedIn: https://www.linkedin.com/in/christopherdelacruzhamilton/

#Bonus

Hate Speech Predictor Application

Test Subjects

Hate Tweet: you might be a libtard if... #libtard #sjw #liberal #politics

Not Hate Tweet: diet coke trash