Capstone Project

Identifying Offensive Tweets

Joel Cheung DSI-28

TABLE OF CONTENTS

TABLE OF CONTENTS

1

INTRODUCTION

TWITTER

- Is a micro-blogging social media site
- 217.5 million active users globally
- 500 million tweets everyday

TWITTER

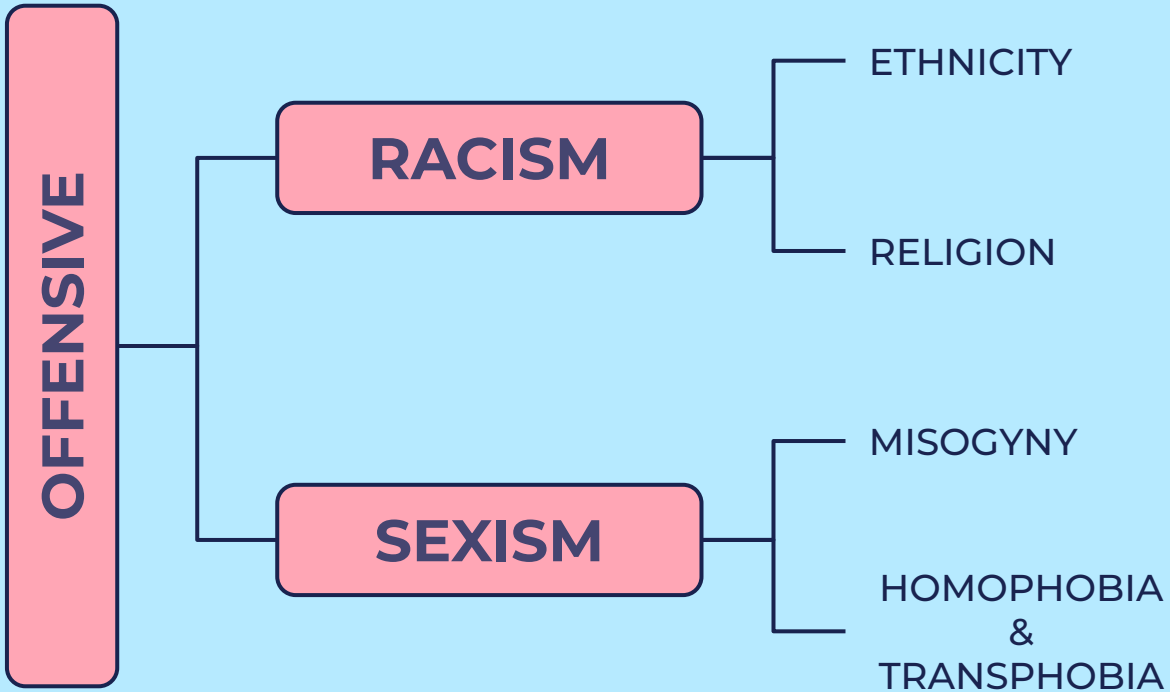
- Corporations - brand outreach
- Individuals - entertainment, news, discussions
- Freedom to discuss a wide range of topics (hashtags)

UH OH...

Conducive discussion space with potential for abuse in the forms of offensiveness*

*Areas of focus being racism and sexism

RACISM AND SEXISM



TWITTER'S POLICY

Twitter is a social broadcast network that enables people and organizations to publicly share brief messages instantly around the world. This brings a variety of people with different voices, ideas, and perspectives. People are allowed to post content, including potentially inflammatory content, as long as they're not violating the Twitter Rules. It's important to know that Twitter does not screen content or remove potentially offensive content.

CBS NEWS NEWS SHOWS LIVE LOCAL Login

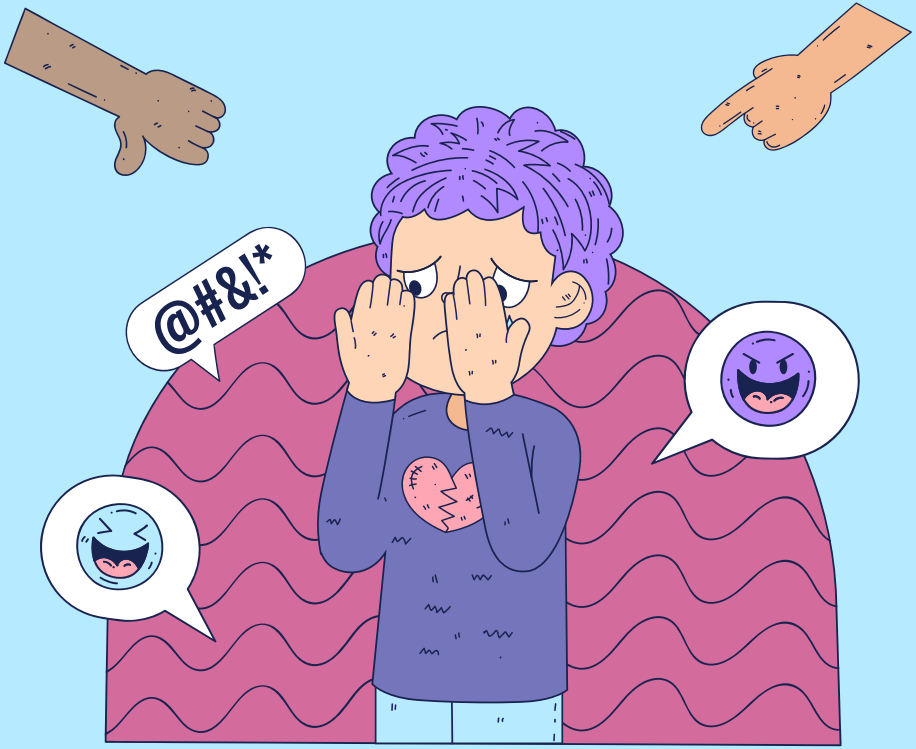
POLITICS ›

Twitter slow to remove racist and sexist tweets targeting Vice President Kamala Harris, report finds

BY MUSADIQ BIDAR, DAN PATTERSON
JUNE 9, 2022 / 7:45 PM / CBS NEWS

2

PROBLEM STATEMENT



PROBLEM STATEMENT

BALANCING ACT

Maintaining a growing user base to attract corporations, while keeping a safe space for users

IDENTIFY OFFENSIVE TWEETS

A model to flag offensive tweets to point users to resources before they publish the tweet

3

DATA SOURCES

DATA SOURCES

TRAINING DATASET

kaggle:
Classified Tweets

TEST DATASET

kaggle:
Cyberbullying
Classification challenge

4

DATA CLEANING

DATA CLEANING

- Remove emojis 😀
- Remove mentions (@username) and URLs
- Make texts lowercase
- Removing punctuations
- Removing stopwords
- Lemmatizing with Part-of-Speech (POS) tagging

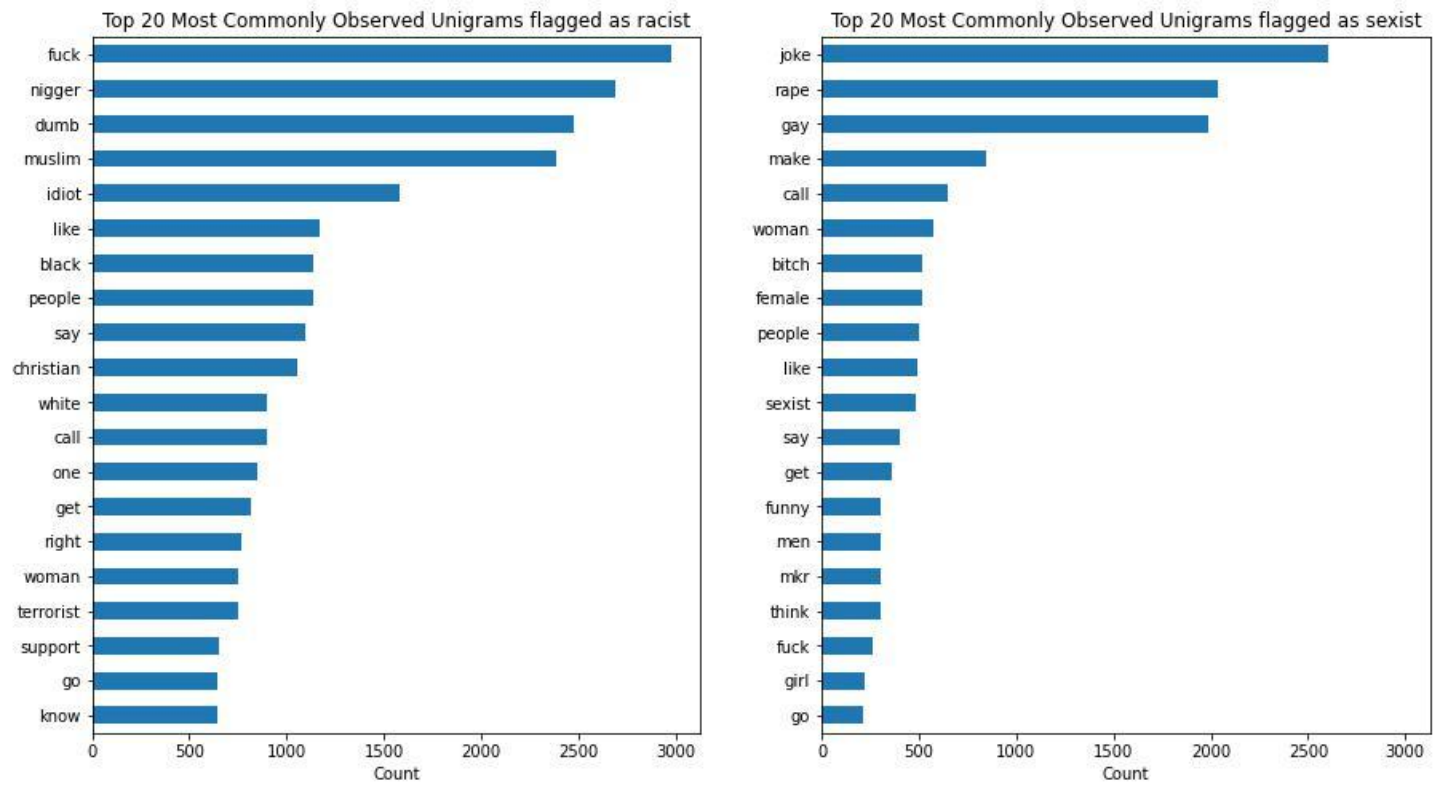
5

EXPLORATORY DATA ANALYSIS

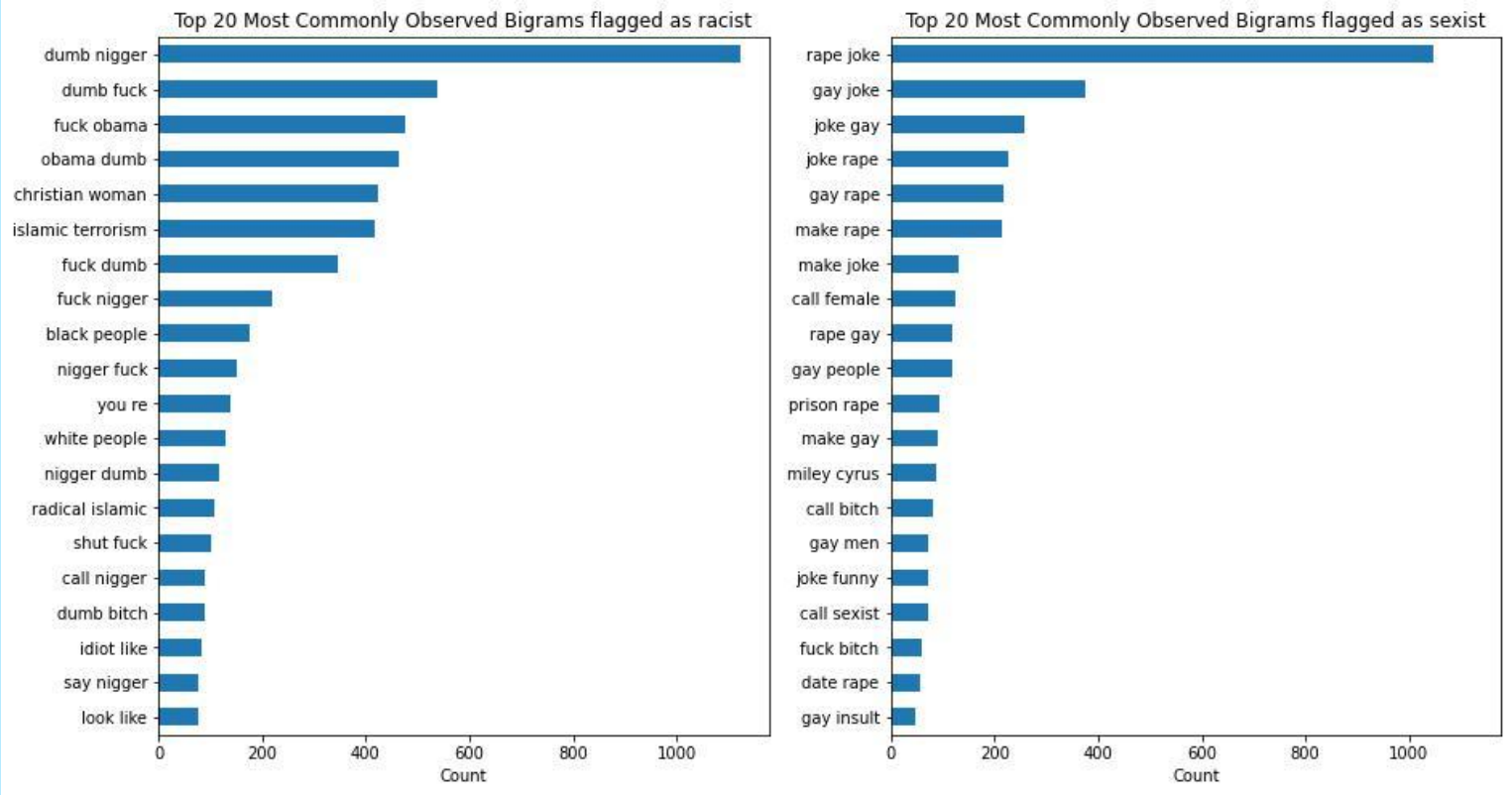
CONTENT WARNING!!

From this slide on, there will be mentions of sensitive topics such as religion, misogyny and other topics that may be offensive to many

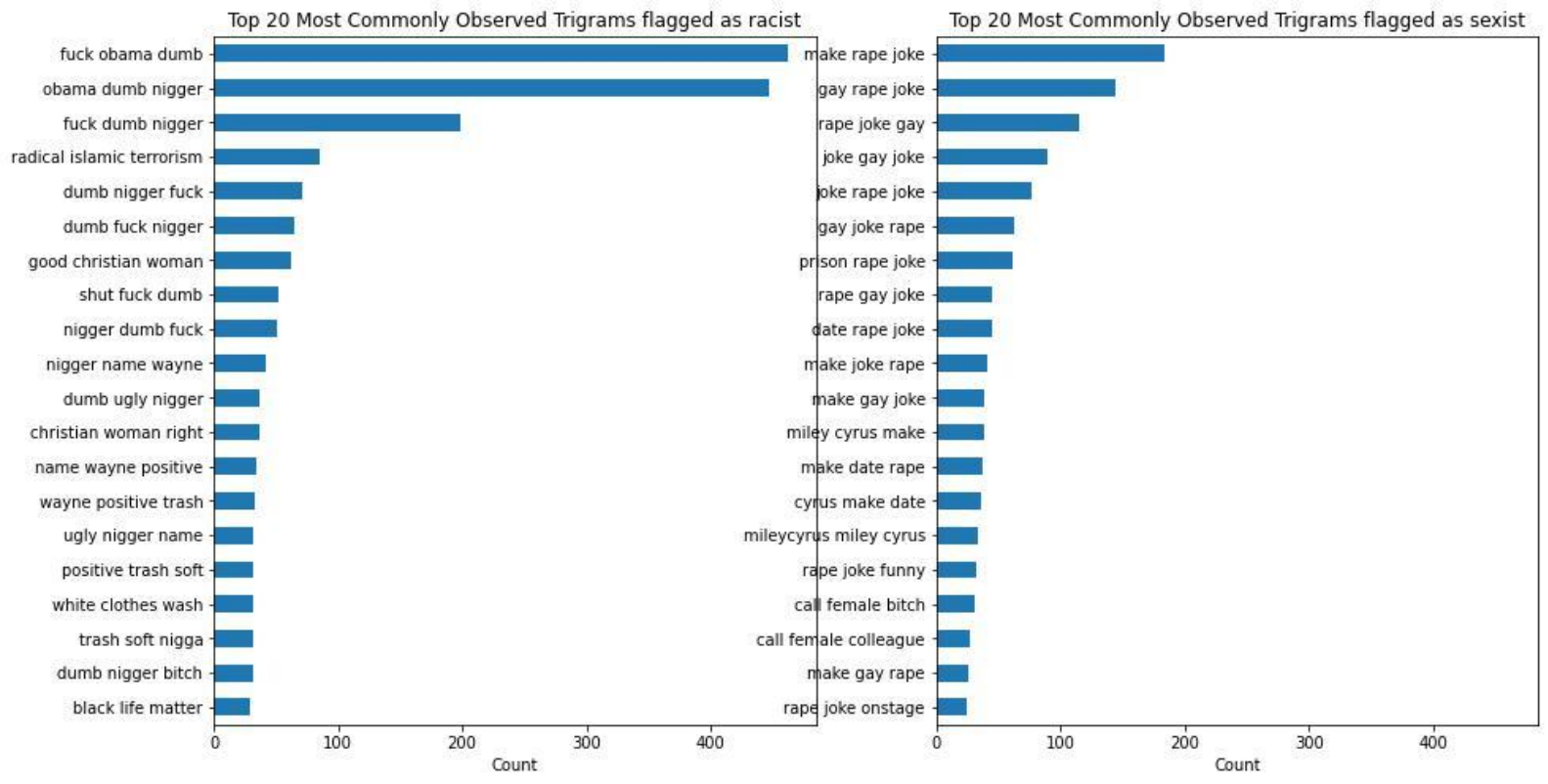
UNIGRAM


Top 20 Most Commonly Observed Unigrams flagged as racist
fuck
nigger
dumb
muslim
idiot
like
black
people
say
christian
white
call
one
get
right
woman
terrorist
support
go
know
0
500
1000
1500
2000
2500
3000
Count
Top 20 Most Commonly Observed Unigrams flagged as sexist
joke
rape
gay
make
call
woman
bitch
female
people
like
sexist
say
get
funny
men
mkr
think
fuck
girl
go
0
500
1000
1500
2000
2500
3000
Count

BIGRAM


Top 20 Most Commonly Observed Bigrams flagged as racist
dumb nigger
dumb fuck
fuck obama
obama dumb
christian woman
islamic terrorism
fuck dumb
fuck nigger
black people
nigger fuck
you re
white people
nigger dumb
radical islamic
shut fuck
call nigger
dumb bitch
idiot like
say nigger
look like
0
200
400
600
800
1000
Count
Top 20 Most Commonly Observed Bigrams flagged as sexist
rape joke
gay joke
joke gay
joke rape
gay rape
make rape
make joke
call female
rape gay
gay people
prison rape
make gay
miley cyrus
call bitch
gay men
joke funny
call sexist
fuck bitch
date rape
gay insult
0
200
400
600
800
1000
Count

TRIGRAM


Top 20 Most Commonly Observed Trigrams flagged as racist
fuck obama dumb
obama dumb nigger
fuck dumb nigger
radical islamic terrorism
dumb nigger fuck
dumb fuck nigger
good christian woman
shut fuck dumb
nigger dumb fuck
nigger name wayne
dumb ugly nigger
christian woman right
name wayne positive
wayne positive trash
ugly nigger name
positive trash soft
white clothes wash
trash soft nigga
dumb nigger bitch
black life matter
0
100
200
300
400
Count
Top 20 Most Commonly Observed Trigrams flagged as sexist
make rape joke
gay rape joke
rape joke gay
joke gay joke
joke rape joke
gay joke rape
prison rape joke
rape gay joke
date rape joke
make joke rape
make gay joke
miley cyrus make
make date rape
cyrus make date
mileycyrus miley cyrus
rape joke funny
call female bitch
call female colleague
make gay rape
rape joke onstage
0
100
200
300
400
Count

6

MODELLING

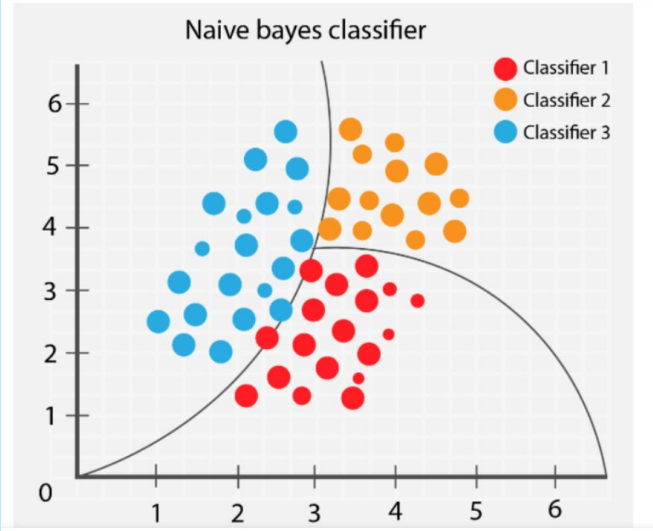
PERFORMANCE METRIC: F1-SCORE

$$F_1\text{–score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision: Of all positive predictions, how many are really positive?

Recall: Of all real positive cases, how many are predicted positive?

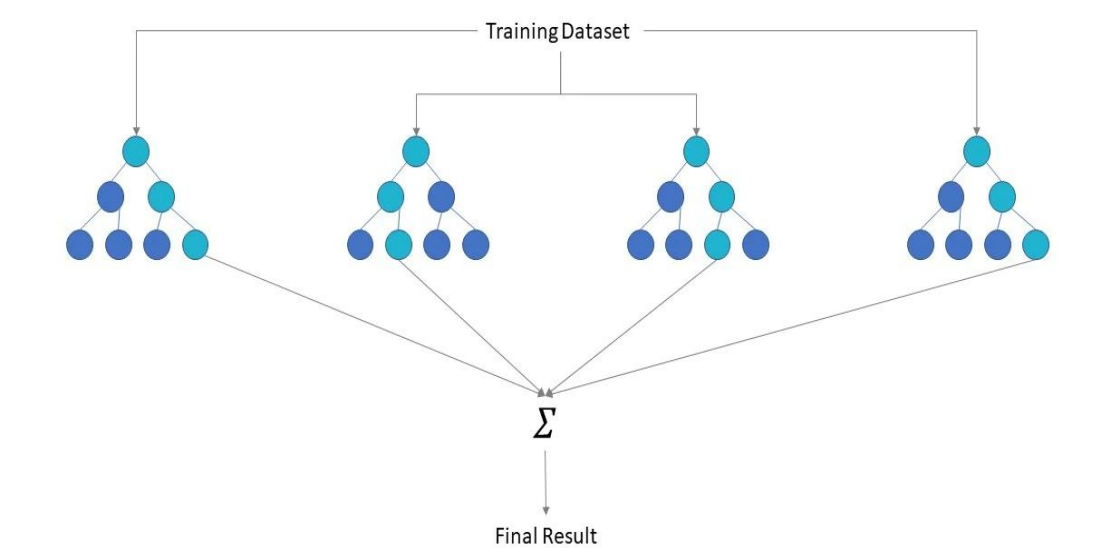
MULTINOMIAL NAIVE BAYES



- Baseline Model
- Simple, easy to train model

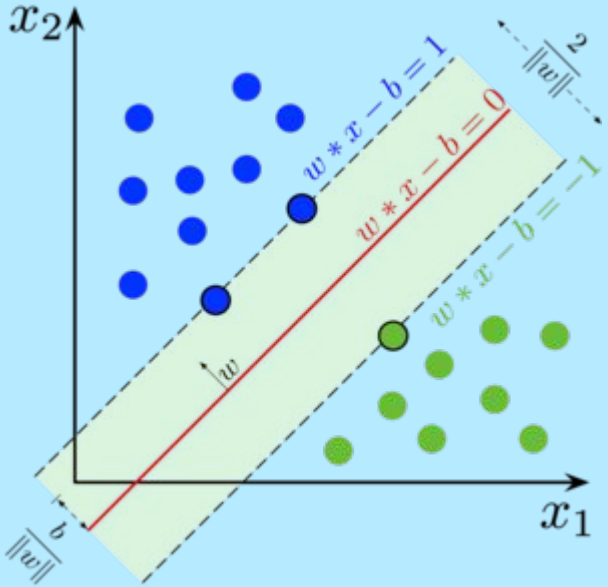
$$\Pr(Racist|text) = \frac{\Pr(text|Racist) * \Pr(Racist)}{\Pr(text)}$$

RANDOM FOREST



Algorithm trains many decision trees and returns the classification of the majority vote

SUPPORT VECTOR MACHINE



Using some representative points (support vectors), the algorithm aims to maximize the space between each category (margin)

BERT

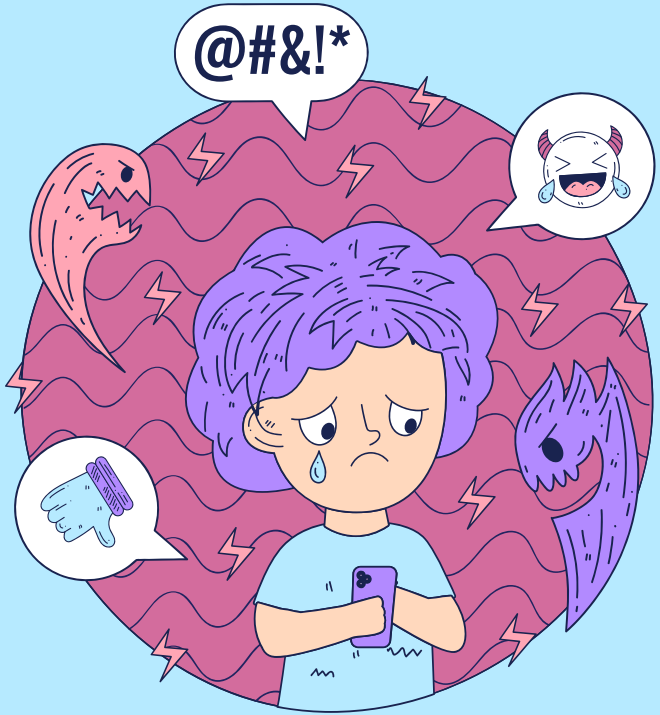
- Bidirectional Encoder Representations from Transformers (BERT)
- Pre-trained model created by Google in 2018
- Trained on the entire English Wikipedia and BookCorpus by having it fill in masked words (blanks)
    - Total of 3.3 billion words
    - Useful for contextual word embeddings
    - Different representation for different contexts (despite being same word e.g. dog's bark vs tree bark
- Takes into account words before and after a particular word
    - Truly bidirectional compared to LSTM

7

MODELLING RESULTS

MODELLING RESULTS

Model	F1-Score		
	Train	Validation	Test
Multinomial NB	0.922	0.889	0.640
Random Forest	0.999	0.931	0.883
SVM	0.963	0.927	0.882
BERT	0.999	0.927	0.856



8

LIMITATIONS & FUTURE STEPS

LIMITATIONS

Dataset issues

Training dataset might not be comprehensive enough in terms of region and timeframe; no data dictionary available

Computing Constraints

Long training time, may not achieve true optimal hyperparameters

FUTURE STEPS

More data

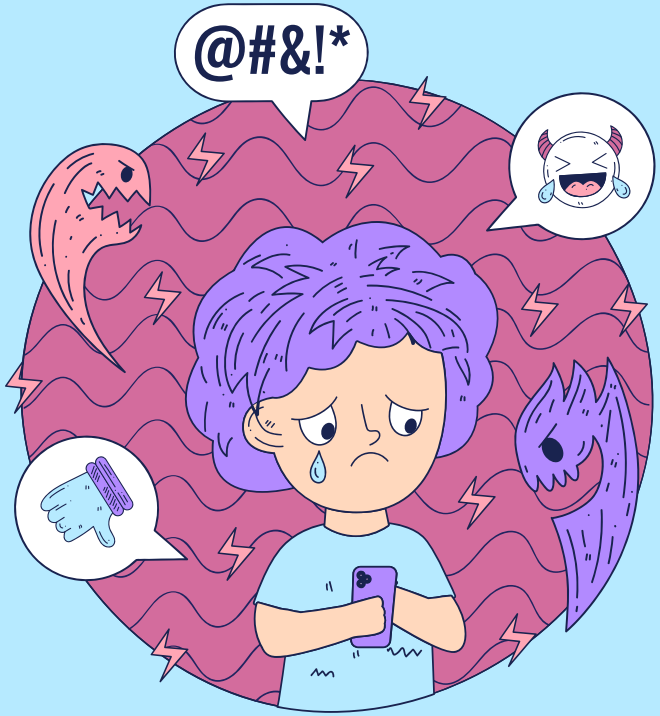
More tweets from all over the globe in order to generalize better

More types of data

User information

Acquire better computing equipment

Allows for more granular GridSearch



9

CONCLUSION

CONCLUSION

Model

Multinomial Naive Bayes for **speed**, Random Forest for **better predictions**

THANKS!