Promoting empathy among Twitter Users, in order to reduce offensive content that harms the wellness of users.

The Problem

- An increasing trend in personal attacks on Twitter has created an environment on social media where others are targeted for their race, sexual orientation, disabilities, and much more.
    - Platforms Like Twitter try to identify potentially offensive content, but struggle to distinguish between the types of personal attacks, experimented with features

Example Tweets

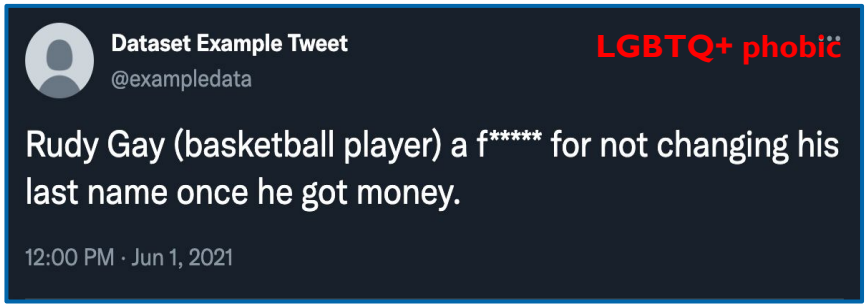
Offensive Content Warning

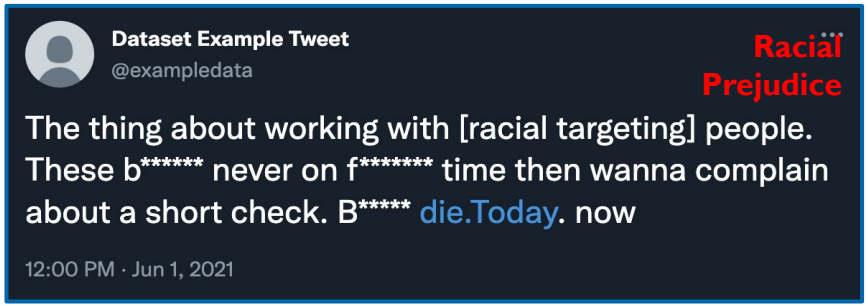
The Impact of this Problem

Eric is an impulsive tweeter, who often uses derogatory sexist language casually. Even though his intention may be joking, he often replies to friends using this language.

Emily is an online friend of Eric, who is impacted by Eric's hurtful tweets. Eric often replies to Emily, and this has impacted her mental health negatively.

Examples Tweets From Dataset





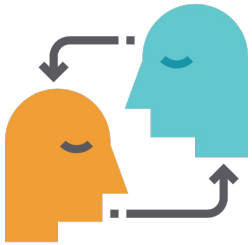
The Growing Trend of Targeting Tweets: July-December 2020

3.8 Million Tweets removed for offensive content policy

Twitter "actioned" 1,126,990 different accounts for infringing hateful conduct policy

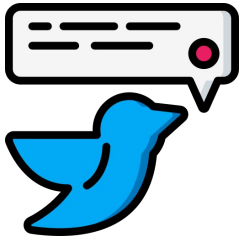
Time Magazine

The Benefits of Solving This Problem



Promote more empathy among twitter users by providing detailed feedback of their tweets

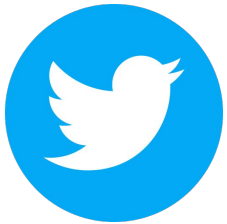
Reduce targeting online, which is tied to issues such as suicide and declining mental health



Expand Twitter's current stance on offensive content to be more comprehensive

Twitter's Current Stance on Offensive Content

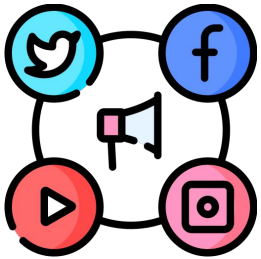
"Healthy conversation is a shared responsibility. If your Tweet reply is identified as using potentially harmful or offensive language, we may ask you, via a prompt, if you want to review it before sending."



Our Product

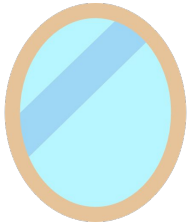
- Create a platform that identifies targeting tweets in the following categories as a proof of concept:
  - Neutral
  - General Criticism
  - Disability discrimination
  - Racial Prejudice
  - Sexism
  - LGBTQ+ phobic

Current and Future Impact



Market size is not limited to just Twitter, but **other social media outlets** where this problem is prevalent.

User and Market size within Twitter would be **1.1 million+ disciplined users**, and those involved in feature beta testing



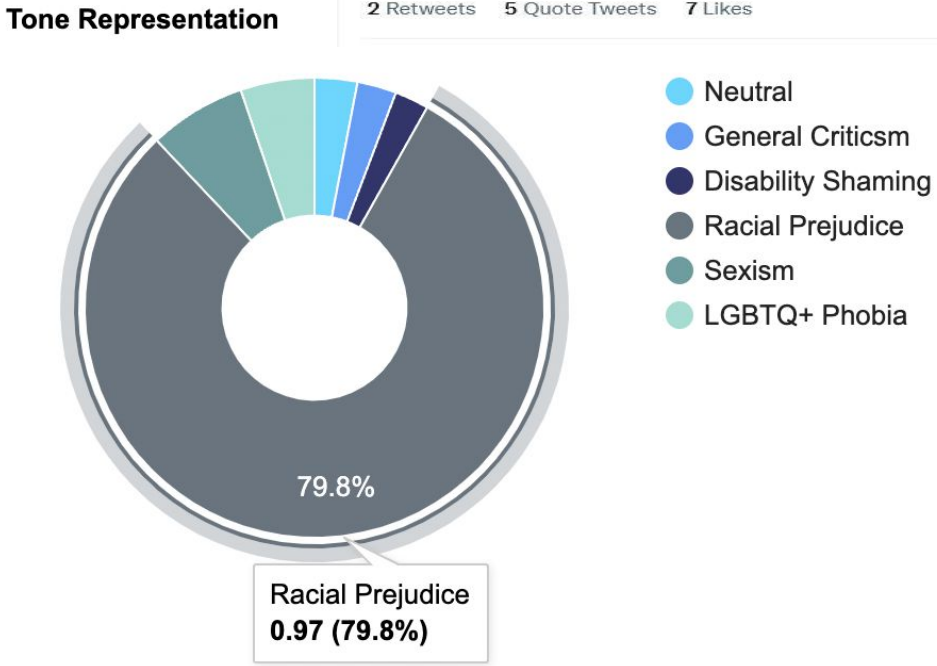
Without policing, make Twitter users aware of the **type of speech they are projecting** on the platform (ex: racism, etc)

Live Demo

Demo Tweet


Paul Reed
@Bball_paul
Why is it so many n s dressin up like girls and puttin on wigs tryna be funny. Shit gay💯😤
1:52 PM · Jun 24, 2015 from South Apopka, FL · Twitter for Android
2 Retweets
5 Quote Tweets
7 Likes

Demo Tweet Results



Data Generation

Hate Speech Dataset

(0) Neutral
(1) General Criticism
(2) Disability Shaming
(3) Racial Prejudice
(4) Sexism
(5) LGBTQ+ Phobia

I convinced hitler was a [redacted], no [redacted], no kids, all those leather uniforms, wanted to be alone with his closest men lock
Free wop
BREAKING: Charlie Crist files emergency motion for a mulligan.
I'm dat [redacted] that these [redacted] just can't stand...
I got called a [redacted] for buying girl toms so now I'm gonna [redacted] that person [redacted]
Keyair angels; female party promoters are not for me.. Most party promoters are [redacted]; them angel [redacted] are
a majority of my news feed is people arguing about what it means to be a [redacted] #vermontproblems
@vinyldanyl @BleedingLSD hes a lyin [redacted] [redacted] lmao
Don't love yo [redacted], just [redacted] yo [redacted] when I got time to do it
@KimW16 so she's [redacted]?
where da ratchet [redacted] at?
Oh I forgot we lived in the ghetto....#hoodrats
That's my young [redacted] I don't want nuthin old but her bank roll

Tweet	(0)	(1)	(2)	(3)	(4)	(5)
Person A	0	1	0	1	0	0
Person B	0	1	0	0	0	0
Person C	0	1	0	1	0	0
Person D	0	0	0	1	0	0
Person E	0	1	0	1	0	0

Preprocessing and train-test split

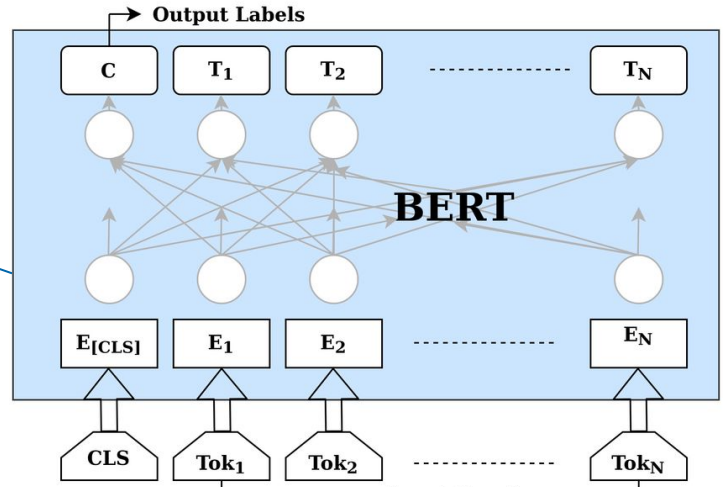
1) Removal of duplicate tweets and NA values
   a) Led to deletion of 836 tweets in total.
2) Removal of URLs, hashtags, usernames, emojis, numbers, and RT
   a) (https://~, #, @, 😉 1234.., RT)
   b) Maintained casing of letters
3) Implemented a 90-5-5 train-validation-test split of our dataset
   a) Train size: 20,372 tweets
   b) Validation size: 1,132 tweets
   c) Test size: 1,132 tweets
4) Tokenization (cased-BERT tokenizer)

['[CLS]', 'For', 'the', 'record', 'No', '##H', '##omo', 'but', 'don', '##t', 'care', 'who', 'is', 'unless', 'I', 'gotta',]

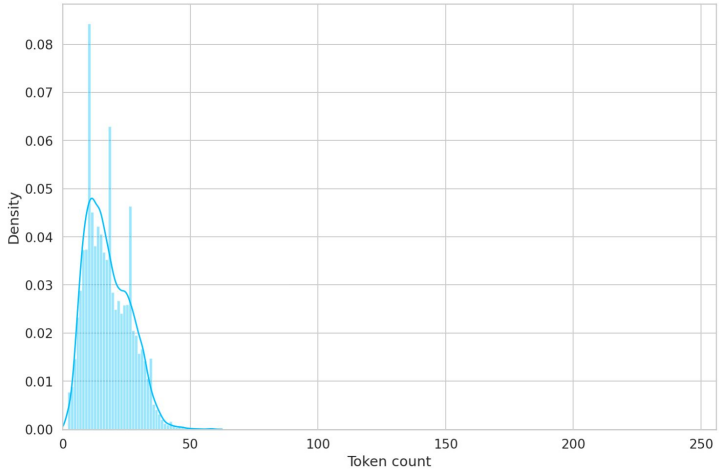
Base Framework

- Train for 4 epochs
- Batch = 32
- Learning rate = 2e-5 (0.00002)
- Maximum token length = 50

(BERT model architecture diagram)



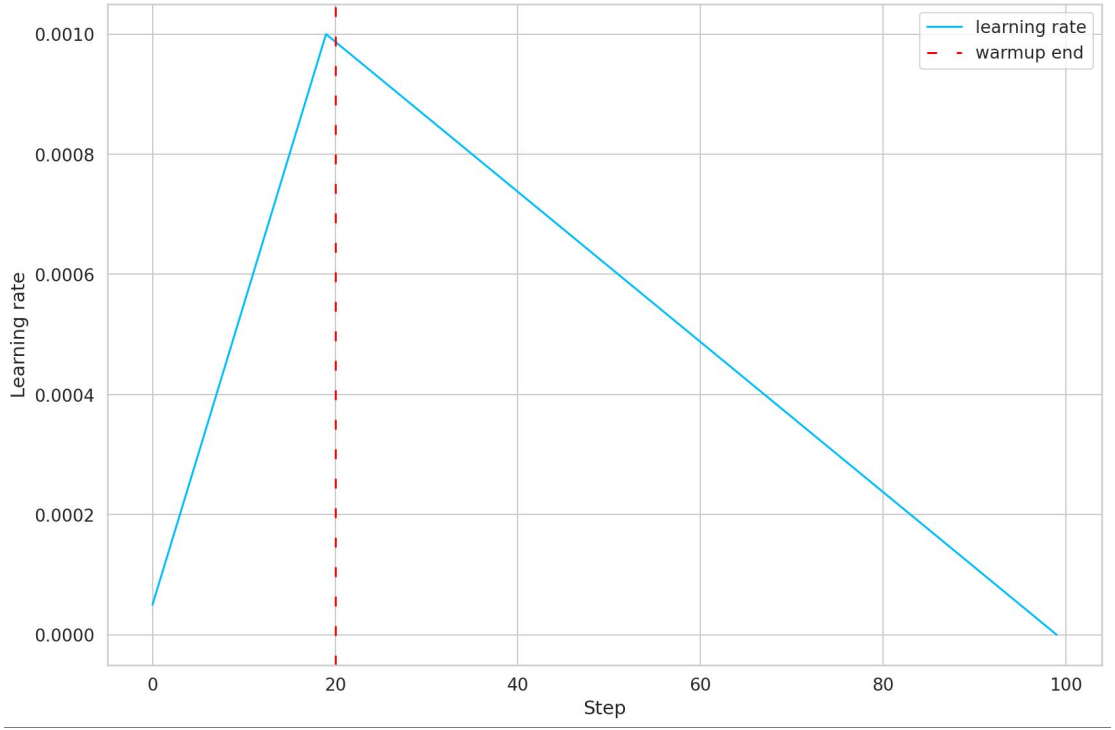
(Token length distribution)



Dynamism of our Model

PyTorch Lightning:

- ModelCheckpoint
- Optimal number of 2 epochs
-

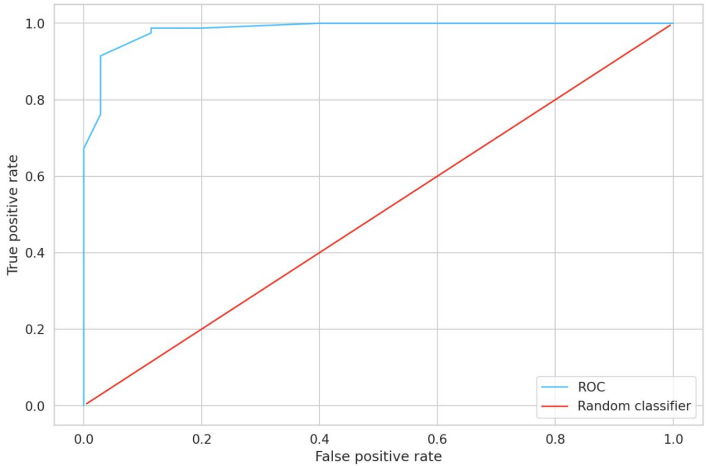


Plot of learning rate vs. step per epoch

Model Evaluation

Sigmoid Function → Binary Cross Entropy Loss

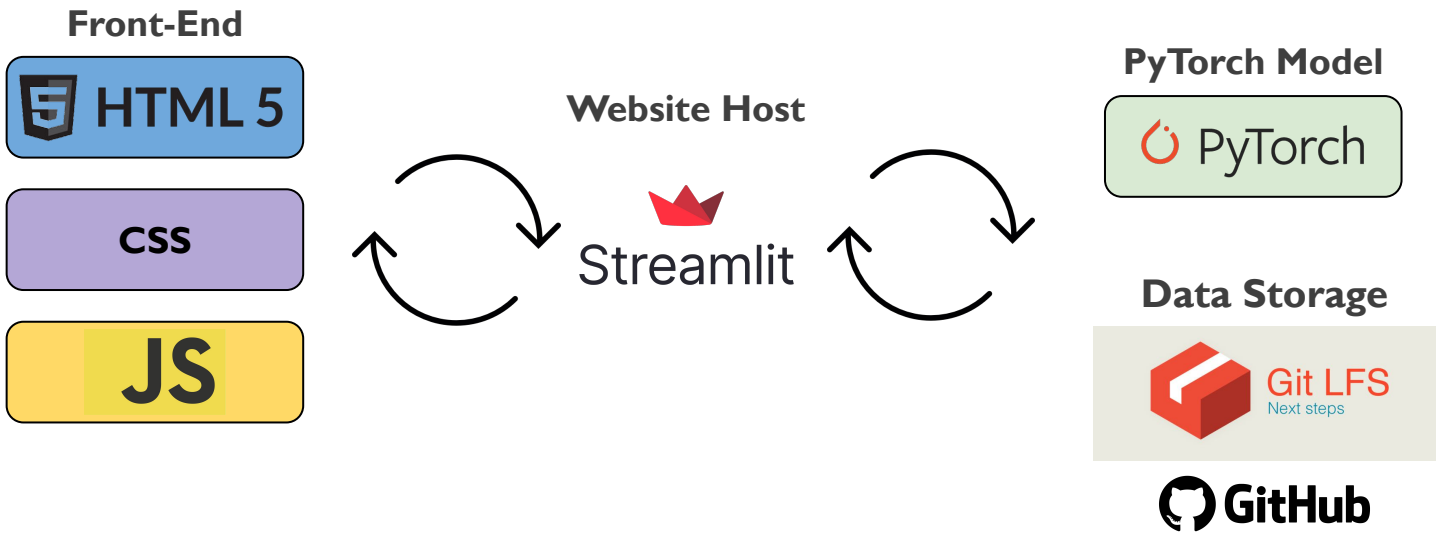
98.4% Accuracy



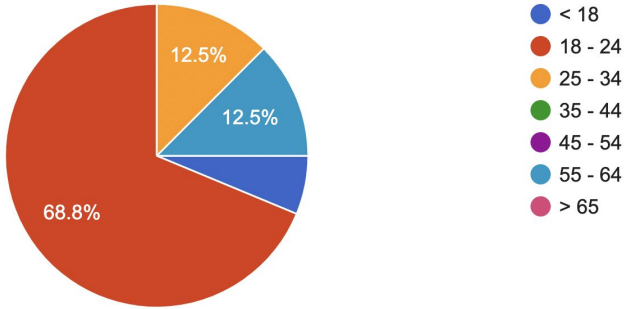
	precision	recall	f1-score	support
disability shaming	1.00	0.84	0.91	19
racial prejudice	0.99	0.99	0.99	97
sexism	1.00	1.00	1.00	747
lgbtq+ phobia	1.00	1.00	1.00	66
micro avg	1.00	0.99	1.00	929
macro avg	1.00	0.96	0.98	929
weighted avg	1.00	0.99	1.00	929
samples avg	0.72	0.71	0.72	929

Return to demo

Data Engineering Pipeline

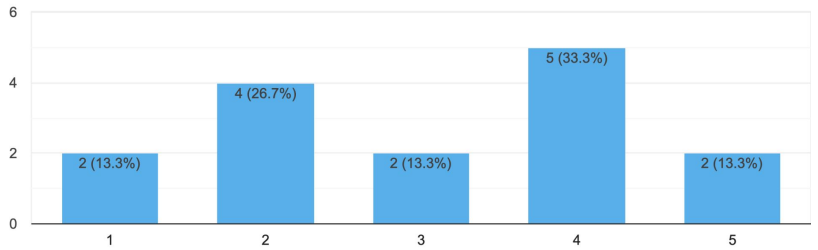


Usability Testing



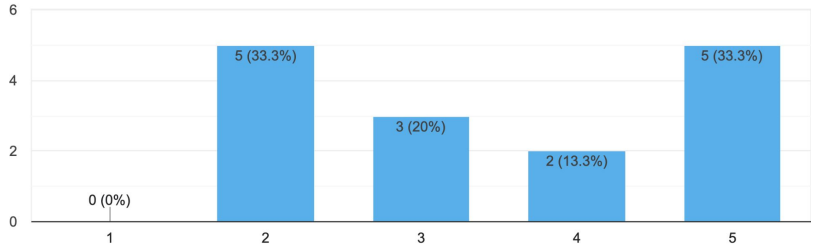
On a scale 1-5, how helpful do you believe a donut chart representation (See Image Below) that breaks down the types of hate speech present in y...ording and feel more confident tweeting content?

15 responses



ow confident do you feel tweeting content online?

responses



Key Takeaways

List 5 key takeaways summary:

1. Sdfs
2. Sdf
3. Sdf
4. Sdfs
5. sdf

Next Steps

With more resources and time, we would like to achieve the following to improve our product:

- Model Optimization
- Front-End Development
- Revisit Data Pipeline + Storage
- User Interviews + Feedback
- Implement New Feature
    - Highlighting Target Words

Demo Example Limitations

This is where we would show a bad tweet where our model does not detect certain things. With further work and research, we believe we can make it better

TONE Live Demo

[redacted] · Jul 13

Replying to @HackneyAbbott

this **fat** retarded black **bitch** thinks you should be forced to feed and house a bunch of violence foreign invaders. i strongly disagree.

[redacted] · Jul 13

@HackneyAbbott Piss off you disgusting useless **fat bitch!** You're a parasite alien looking to silence native people for your power.

We Are TONE

Our Mission: To promote empathy among Twitter Users, in order to reduce offensive content that harms the wellness of users.

Acknowledgements

We would like to give special thanks to the following individuals for helping us out with our development:

- Prof. Joyce Shen
- Prof. Zona Kostic
- Prabhu Narsina
- Kevin Hartman
- Robert Wang (AWS)
- UC Berkeley 5th Year MIDS Cohort 2022

THANK YOU!

Any Questions?