

# TONE

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

# Our Team



**Noor Gill**

Project Manager  
SME & EDA Lead



**Evan Phillips**

Machine Learning Lead  
Technical Product Manager



**Alejandro Pelcastre**

Lead Data Engineer  
Co-Web Developer



**Jaypal Bhatia**

Product Manager  
Data Analyst



**Elias Saravia**

Web Developer  
UI Designer

# Presentation Overview



The Problem

1

Product Demo

2

Technical Approach

3

Potential Next Steps

4

# The Problem



## Increasing Personal attacks

Created an environment on social media where others are targeted for their race, sexual orientation, disabilities, and much more.

---

Twitter tries to identify offensive content

but struggles to distinguish between the types of personal attacks, experimented with features

# The Problem in Numbers



**3.8 Million**

Tweets removed for offensive content policy  
July 2020-August 2020

---

**1.1 Million**

Twitter users disciplined for offensive content

# Benefits to Solving This Problem

## **Less Attacks. More Empathy**

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

## **Stop Targeting to Stop Declining Mental Health**

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

## **Expand Twitter's Current Stance**

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





# Benefits to Solving This Problem

## Less Attacks. More Empathy

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

## Stop Targeting to Stop Declining Mental Health

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

## Expand Twitter's Current Stance

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



# Benefits to Solving This Problem

## **Less Attacks. More Empathy**

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

## **Stop Targeting to Stop Declining Mental Health**

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

## **Expand Twitter's Current Stance**

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





# Twitter's Current Stance

“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.”



# A More Expansive Stance

Twitter and Tone work together to identify the specific type of harmful content you may be tweeting. You'll receive a breakdown of your tweet and the categories of offensive content present in your tweet.



# Current and Future Impact

## Social Media Beyond Twitter

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

## Negative Tweets Still Growing

July 2020-August 2020, 3.8 million tweets removed for Twitter's offensive content policy.

## Beta Feature to Edit Tweets

Twitter's current venture to edit offensive tweets shows movement and need for expansion regarding offensive content



# Current and Future Impact

## Social Media Beyond Twitter

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

## Negative Tweets Still Growing

July 2020-August 2020, 3.8 million tweets removed for Twitter's offensive content policy.

## Beta Feature to Edit Tweets

Twitter's current venture to edit offensive tweets shows movement and need for expansion regarding offensive content



# Current and Future Impact

## Social Media Beyond Twitter

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

## Negative Tweets Still Growing

July 2020-August 2020, 3.8 million tweets removed for Twitter's offensive content policy.

## Beta Feature to Edit Tweets

Twitter's current venture to edit offensive tweets shows movement and need for expansion regarding offensive content



# 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



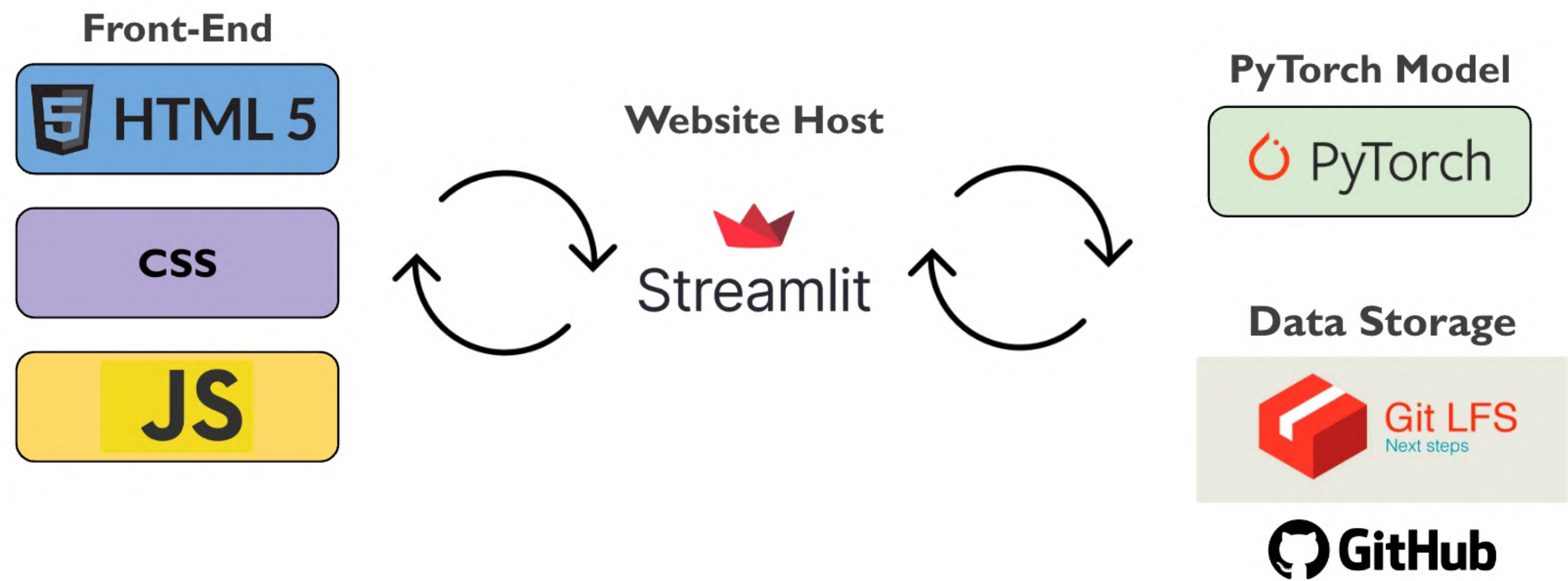
Sexism



LGBTQ+ Phobia



# Data Engineering Pipeline





# Pre-Processing Data

## Removal of duplicate tweets and NA values

- Deleted 836 tweets in total

---

## Removal of URLs, hashtags, usernames, emojis, numbers, and RT

- https://~
- #...
- @...
- 1234...
- RT...
- 🤔

---

## Implemented a 90-5-5 train-validation-test split of our dataset

- Train size: 20,372 tweets
- Validation size: 1,132 tweets
- Test size: 1,132 tweets

---

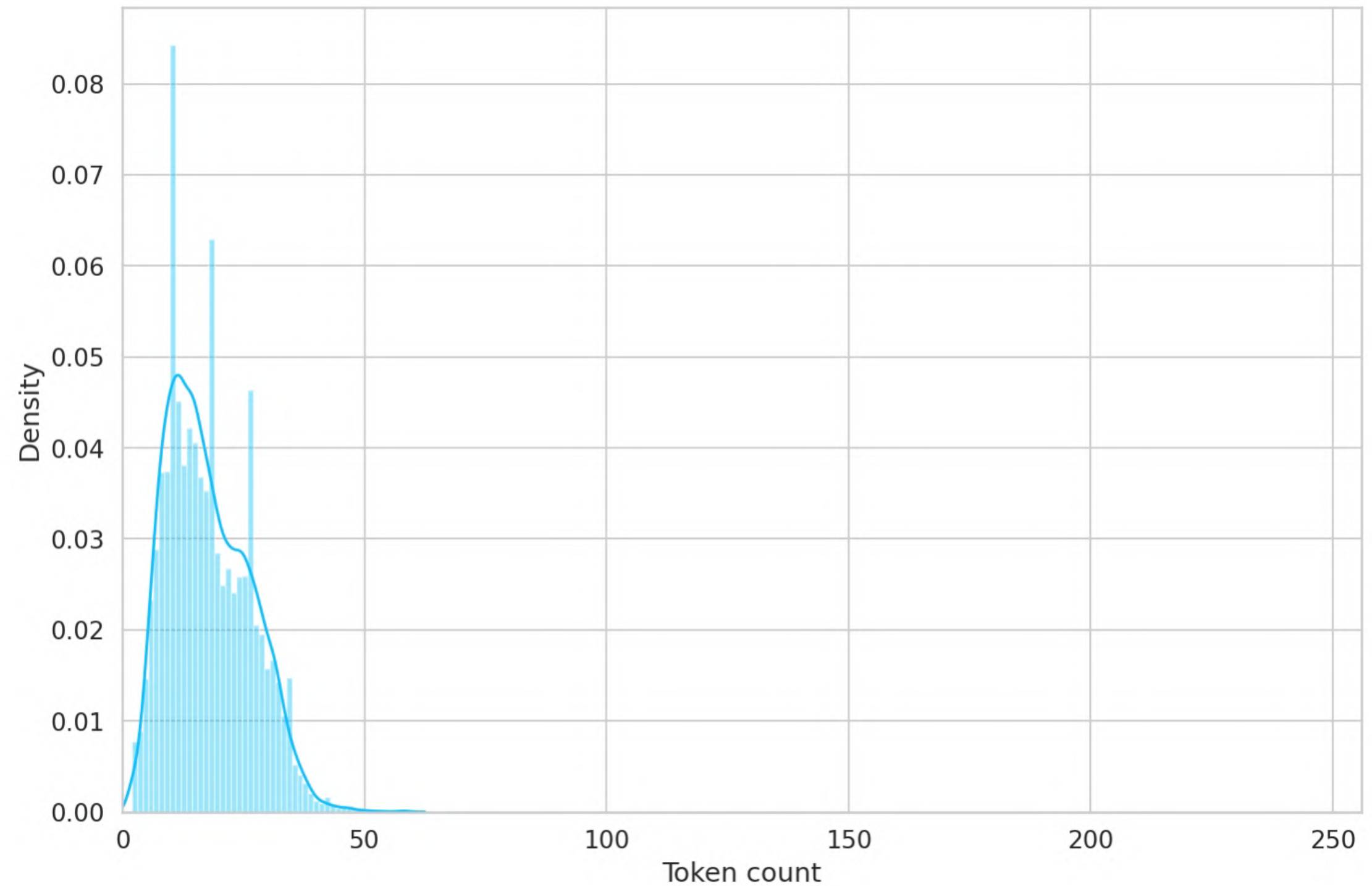
## Tokenization (cased-BERT tokenizer)

- **ex:** ['[CLS]', 'For', 'the', 'record', 'No', '##H', '##omo', 'but', 'don', '##t', 'care', 'who', 'is', 'unless', 'I', 'gotta,']
- "bad" vs. "BAD"

# Base Framework

## Key Takeaways:

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

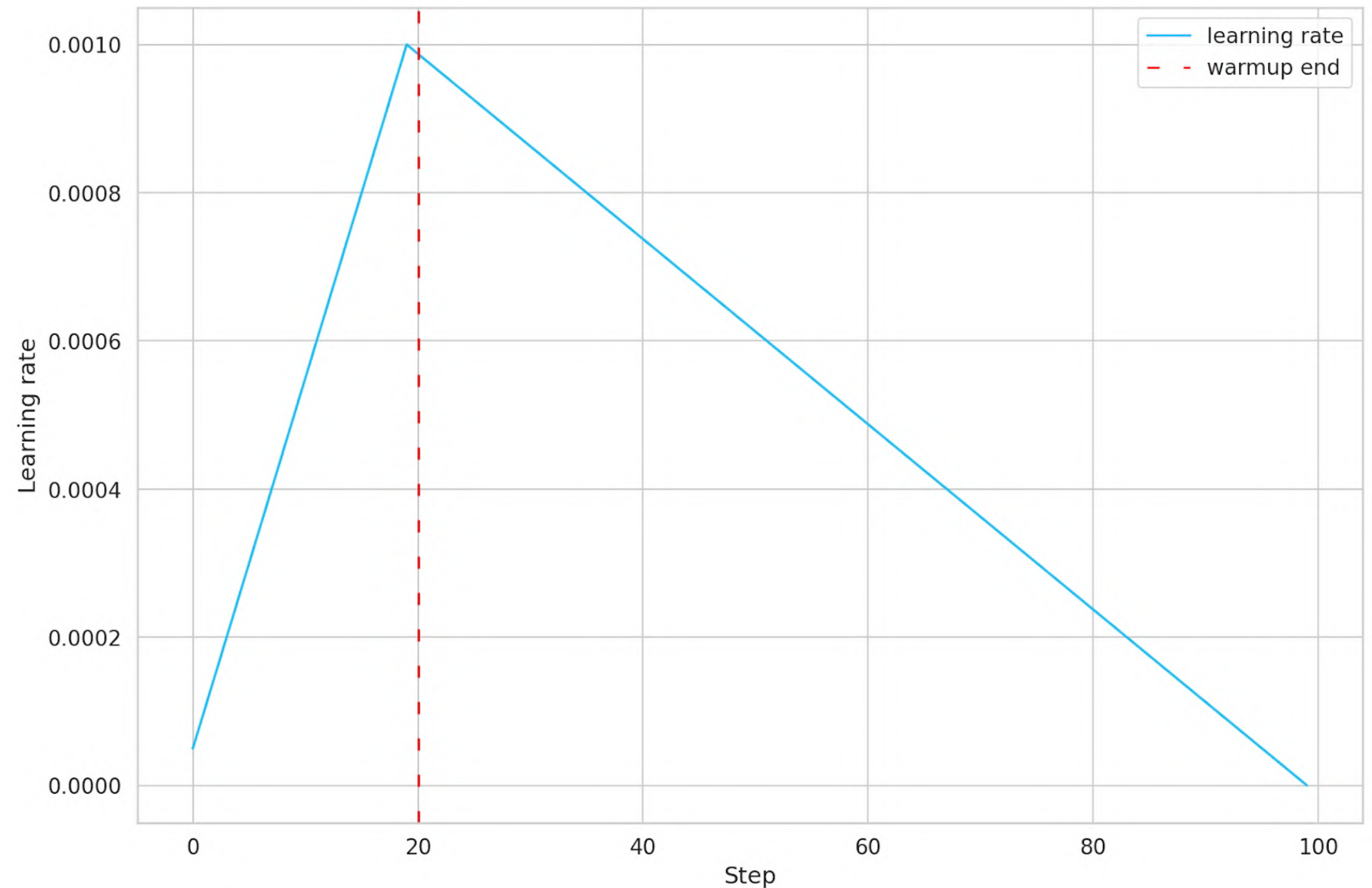
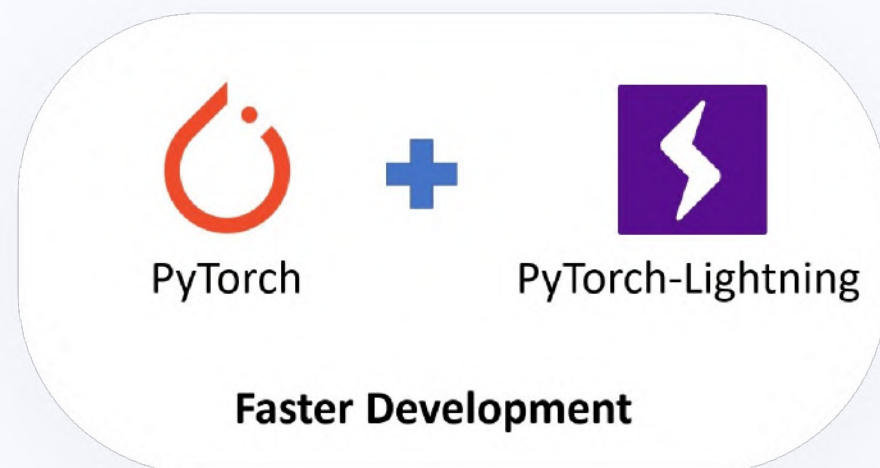


Token Length Distribution

# Dynamism of our Model

PyTorch Lightning:

- ModelCheckpoint
- Simulate 100 training steps
- Optimal number of 2 epochs

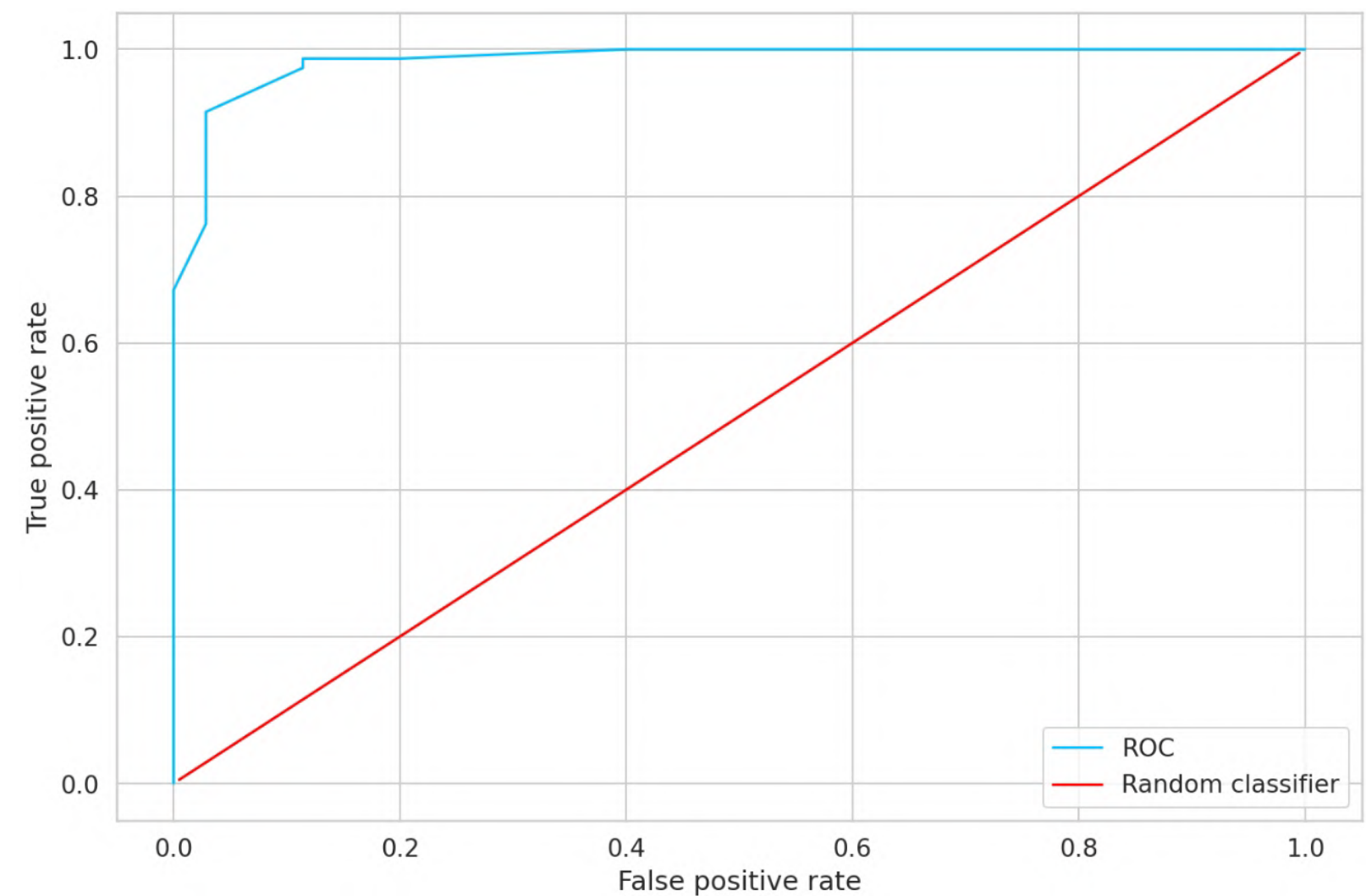


Learning Rate vs. Step per Epoch

# Model Evaluation

Key Takeaways:

- 98.4% Accuracy
- Sigmoid, not Softmax Output Layer
- ROC fit depends on class balance



	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

# What's Next for **TONE**?

## **Increase Accuracy by 15%**

Optimization + Collect More Training Data

---

## **Improve Front End Development**

New Host + Improve User Experience

---

## **Optimize Data Pipeline Infrastructure**

Leverage AWS for data streaming + storage

---

## **Conduct Usability Testing + Feedback**

Collect user interviews, surveys, and focus groups

---

## **Implement New Highlight Feature**

Highlighting / Detecting Hate Speech Words

---

## **Research Contextualization and Labeling**

Communicating with experts to improve product

---

# 5 Key Takeaways

Product Must Reflect the Source of the Data



NLP + ML Model Development, Bias, and Optimization



Conducting Research + Talking to Experts is Important



How to Create a Startup + Its Components



The 4 Cs: Consistent Communication, Contribution, and Collaboration



# WE ARE TNE

## **Our Mission:**

To promote empathy among Twitter Users, in order to reduce offensive content that harms the wellness of others.

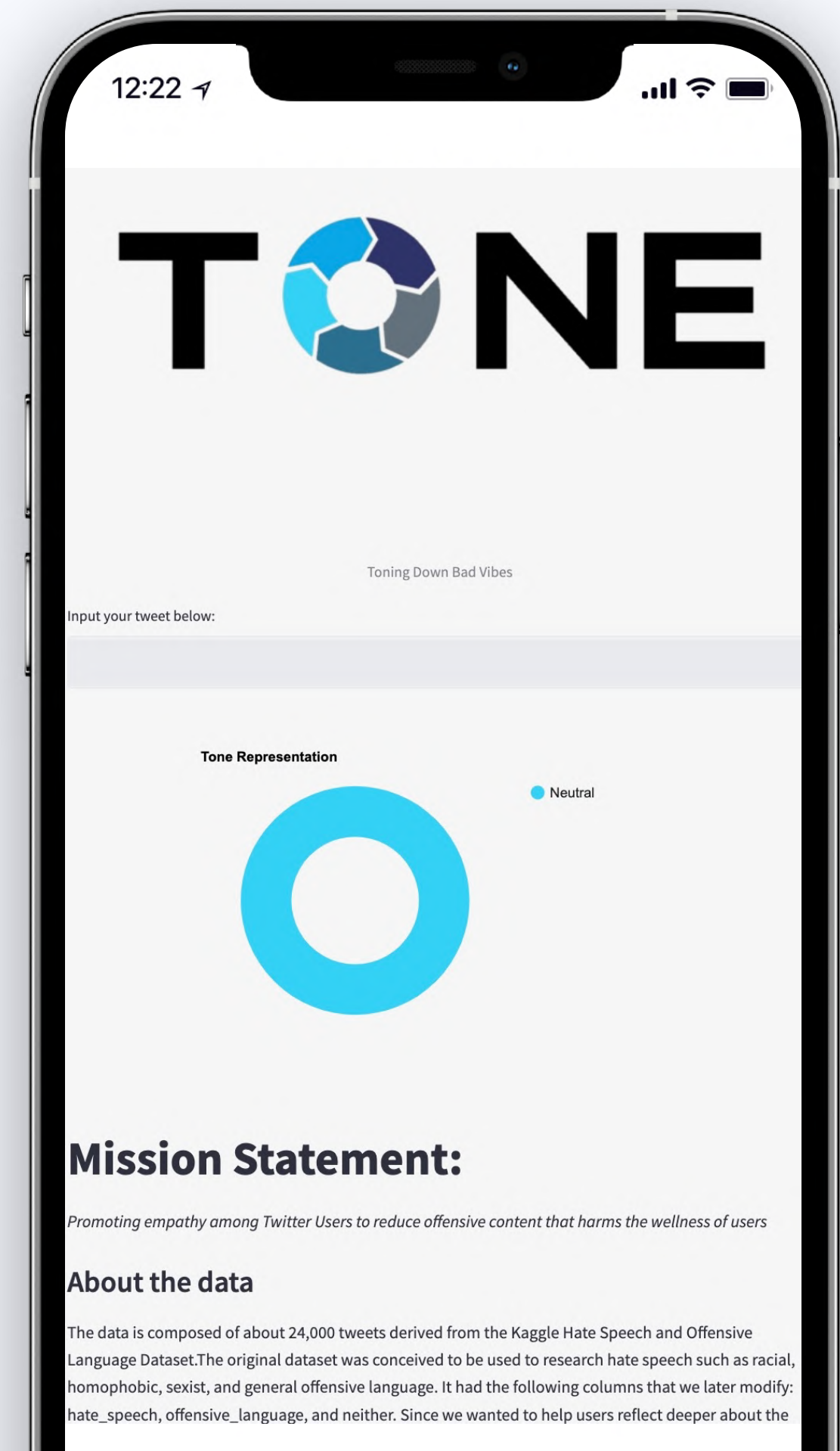


Company Name : Date  
Acme IncAugust 2024

Contact:  
email@acme.com

# Thank You!

## Any Questions?



# 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