Project Elimi-‘Hate’

Chua Soon Poh

DSI-32

GENERAL ASSEMBLY

Table of contents

01

Problem

- Definition & Issue of Hate Speech

02

Dataset

- Project Overview

03

EDA

- Multi-Labels

04

Modelling

- Results
- Best Model

05

Streamlit API

- Demo
- Design Thinking

06

Conclusion

- Limitations
- Future Work

Problem

What is Hate Speech?

“**any kind of communication** in speech, writing or behaviour, that **attacks** or uses **pejorative** or **discriminatory** language with reference to a person or a group on the basis of **who they are**, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.”

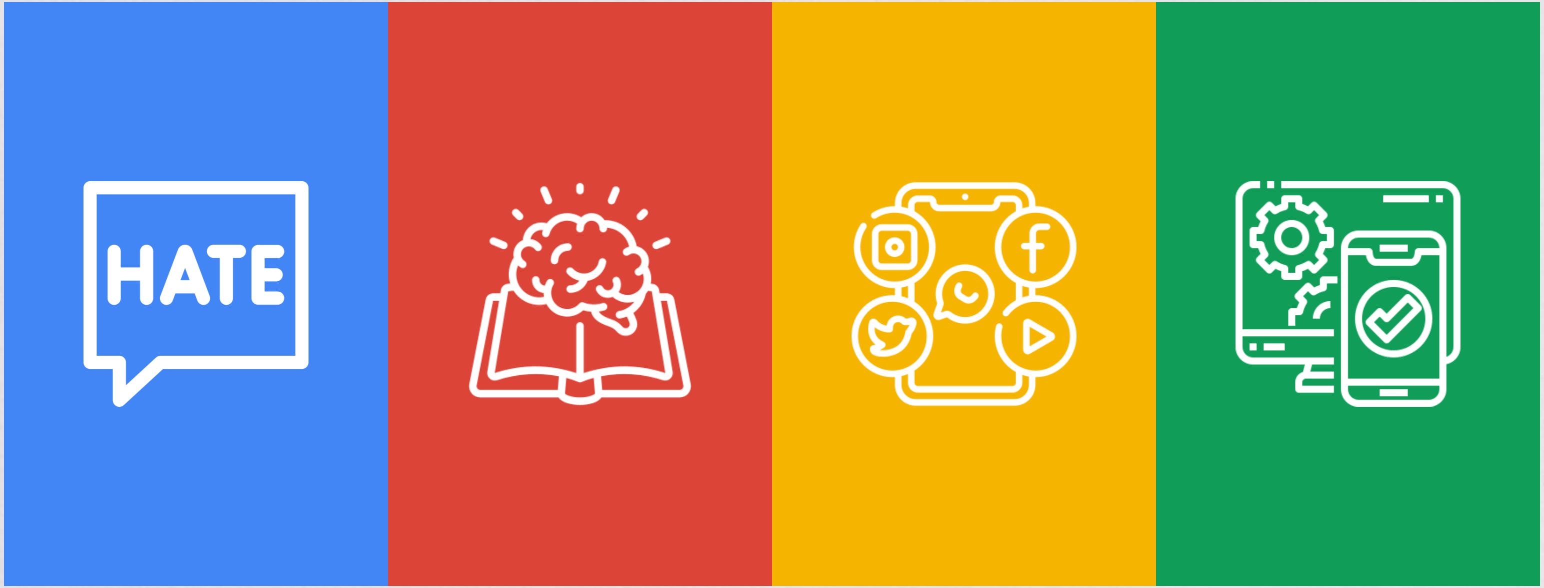
UNITED NATIONS

Problem

> “Social media provides a global megaphone for hate.”

ANTÓNIO GUTERRES, *United Nations Secretary-General, 2021*

Problem



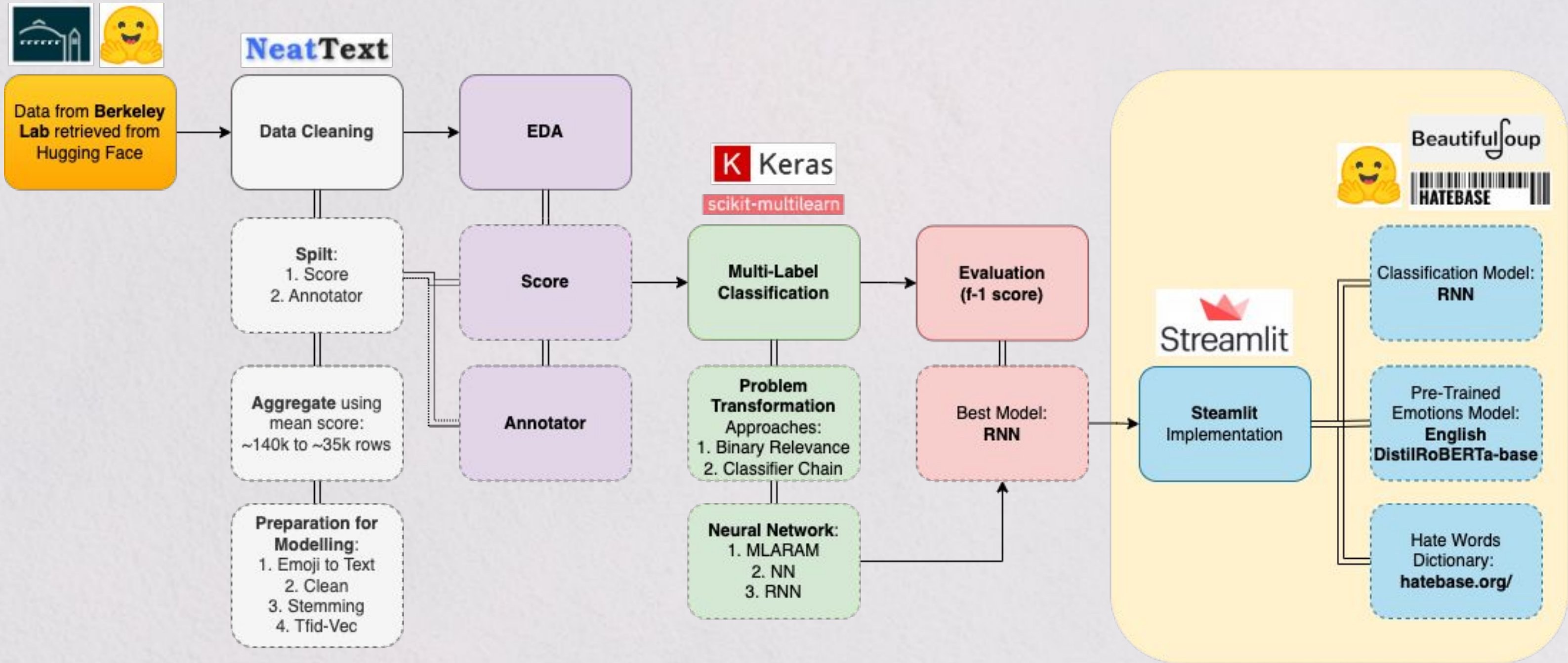
Problem

How?

Where?

So?

Project Overview



NeatText

Keras

scikit-multilearn

Beautiful Soup

HATEBASE

Streamlit

Data from **Berkeley Lab** retrieved from Hugging Face

Data Cleaning

EDA

Spilt:
1. Score
2. Annotator

Score

Annotator

Aggregate using mean score:
~140k to ~35k rows

Preparation for Modelling:
1. Emoji to Text
2. Clean
3. Stemming
4. Tfid-Vec

Multi-Label Classification

Problem Transformation Approaches:
1. Binary Relevance
2. Classifier Chain

Neural Network:
1. MLARAM
2. NN
3. RNN

Evaluation (f-1 score)

Best Model: **RNN**

Steamlit Implementation

Classification Model: **RNN**

Pre-Trained Emotions Model: **English DistilRoBERTa-base**

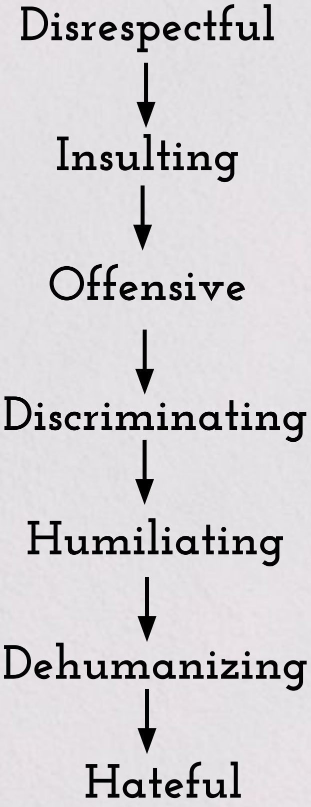
Hate Words Dictionary: **hatebase.org/**

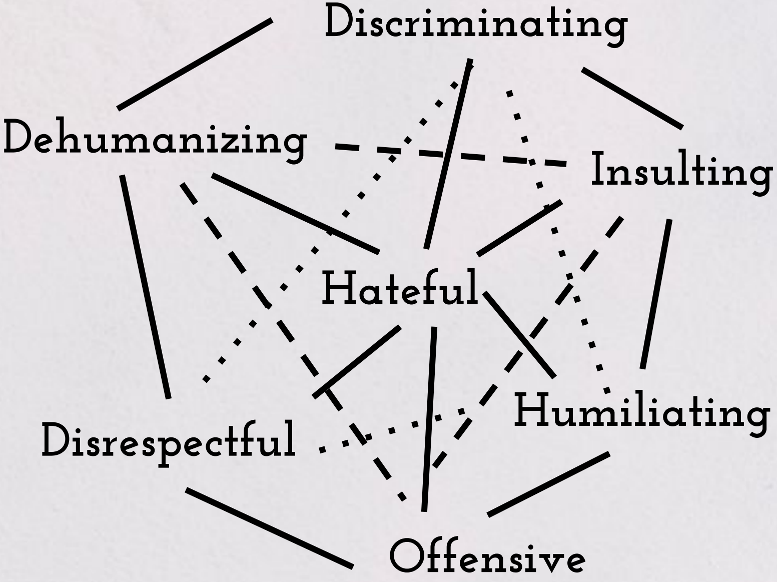
Dataset

1. Increase empathy towards others in the social media space
2. Cultivate better awareness of unhelpful written expressions

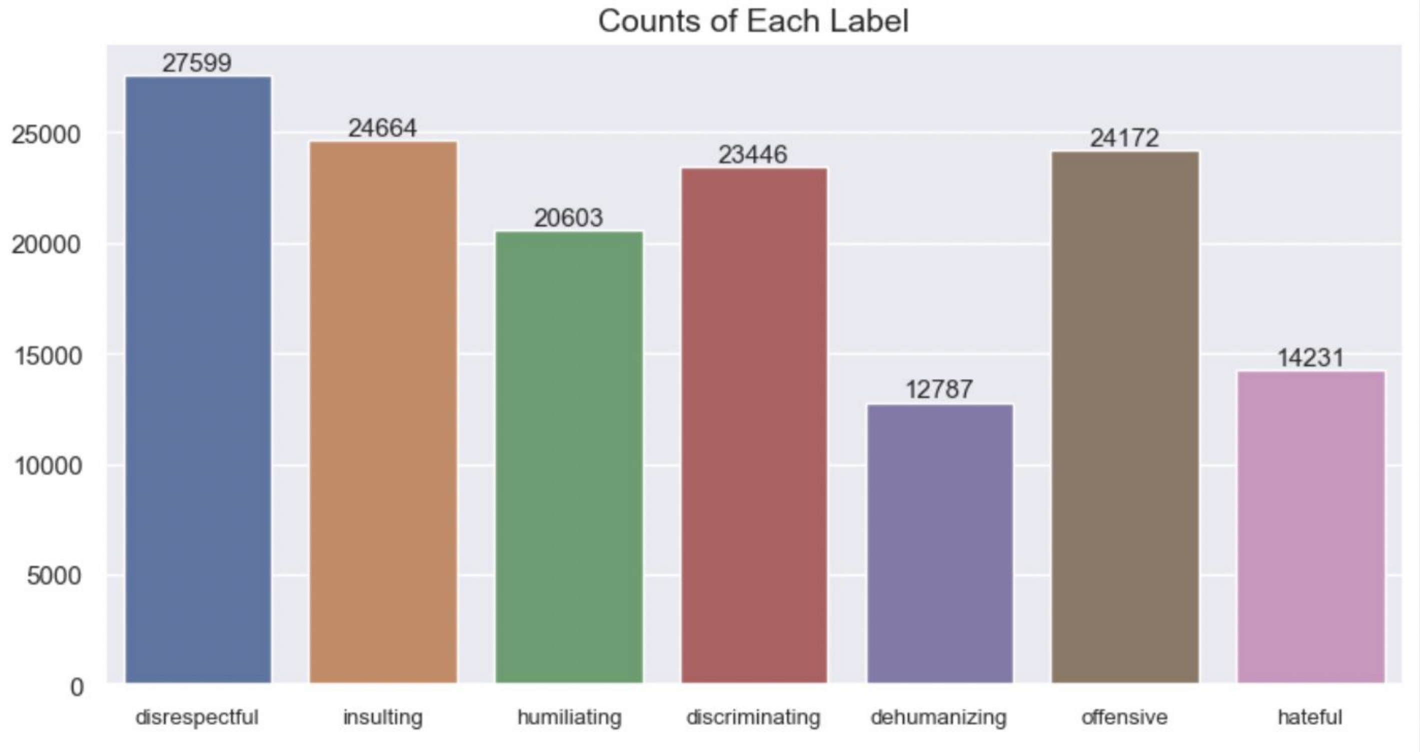
Multi-Labels?

Dataset



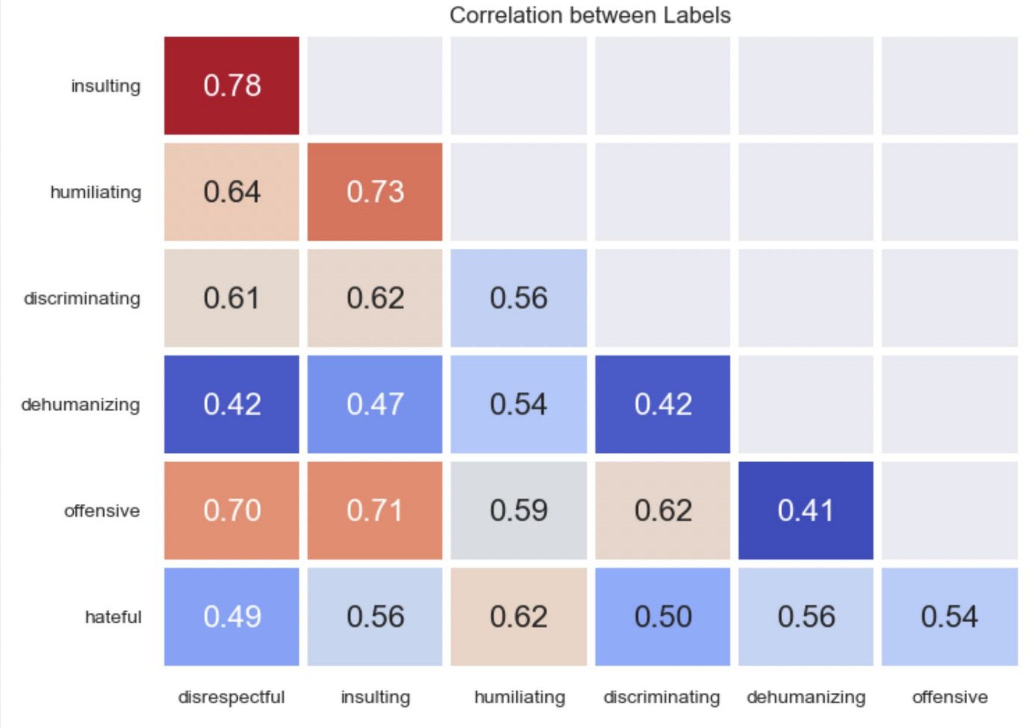


Multi-Labels



EDA

Multi-Labels



Correlation between Labels

	disrespectful	insulting	humiliating	discriminating	dehumanizing	offensive
insulting	0.78					
humiliating	0.64	0.73				
discriminating	0.61	0.62	0.56			
dehumanizing	0.42	0.47	0.54	0.42		
offensive	0.70	0.71	0.59	0.62	0.41	
hateful	0.49	0.56	0.62	0.50	0.56	0.54

EDA

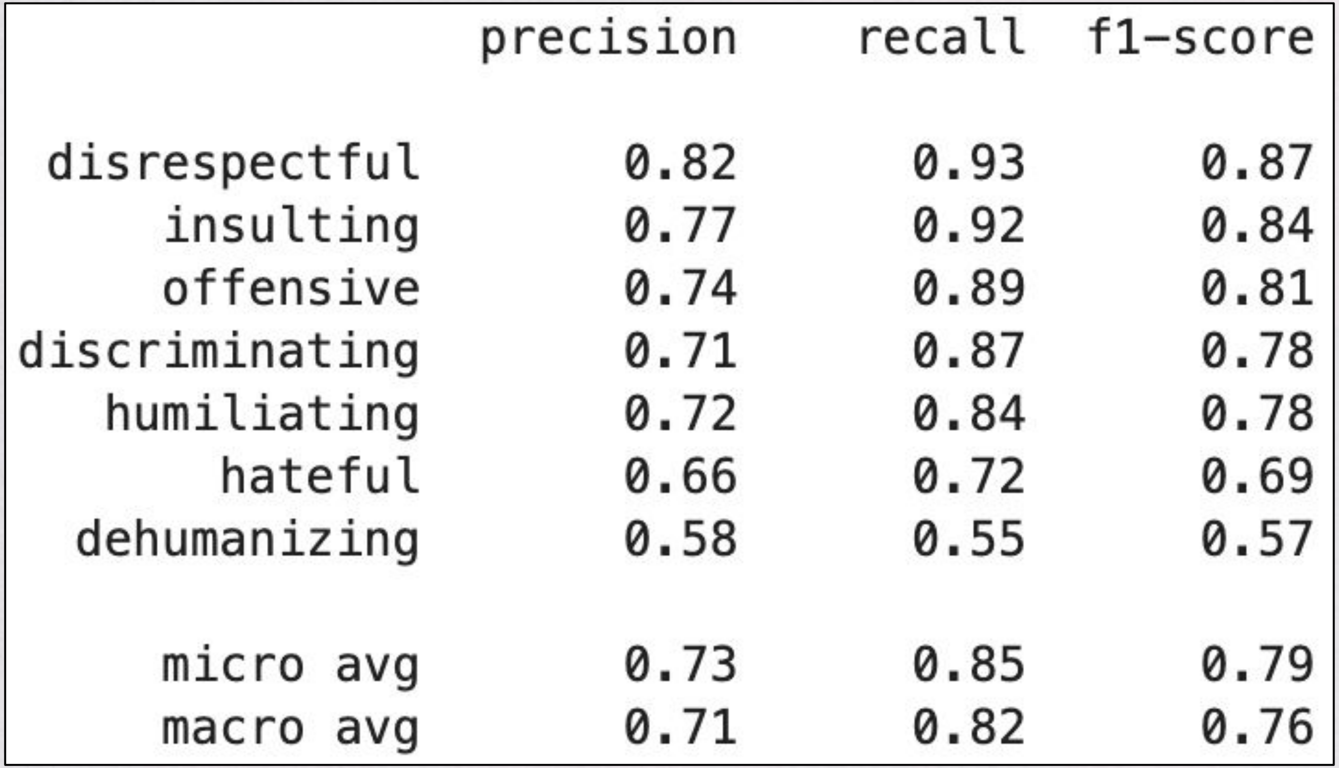
Modelling

Results

Model	Precision Macro Avg	Recall Macro Avg	F-1 Macro Avg
Binary Relevance & Naives Bayes	0.73	0.74	0.70
Classifier Chain & Naives Bayes	0.62	0.93	0.74
MLARAM	0.62	0.51	0.53
LSTM NN with GLOVE Embedding	0.71	0.82	0.76

Best Model: LSTM NN with GLOVE Embedding

Modelling



```
                 precision    recall  f1-score

 disrespectful        0.82      0.93      0.87
     insulting        0.77      0.92      0.84
     offensive        0.74      0.89      0.81
discriminating        0.71      0.87      0.78
   humiliating        0.72      0.84      0.78
       hateful        0.66      0.72      0.69
  dehumanizing        0.58      0.55      0.57

     micro avg        0.73      0.85      0.79
     macro avg        0.71      0.82      0.76
```

Demo Part 1

Streamlit API

https://tinyurl.com/Elimi-Hate-Demo

Write your post below to check if:

1. The negative emotion that your post may contain;

j-hartmann
/emotion-english-distilroberta-base

2. Your post is potentially disrespectful, insulting, offensive, discriminating, humiliating, hateful or dehumanizing towards others;

Best Model: LSTM NN with GLOVE Embedding

3. Your post contains any hate words from https://hatebase.org.

HATEBASE

Demo Part 2

Streamlit API

DSI-32

General

Jonathan

Coon

Salad

Wink

Rube

Assembly

Gimps

Gooks

Pain

Lubra

Ho

Snowflake

Libtard

Pikey

Suntan

Thirty-Six

Deep

Yoga

Demo Part 2

Streamlit API

Jonathan

DSI-32

Coon

General

Salad

Rube

Assembly

Wink

Gooks

Pain

Gimps

Lubra

Ho

Snowflake

Libtard

Pikey

Thirty-Six

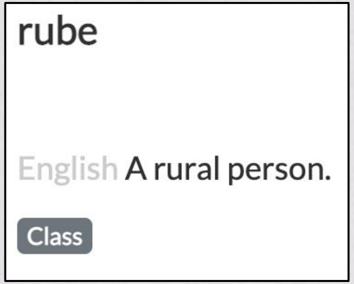
Yoga

Deep

Suntan

Results from HateBase

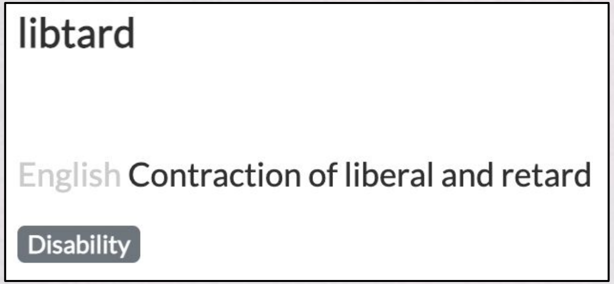
Streamlit API



rube

English A rural person.

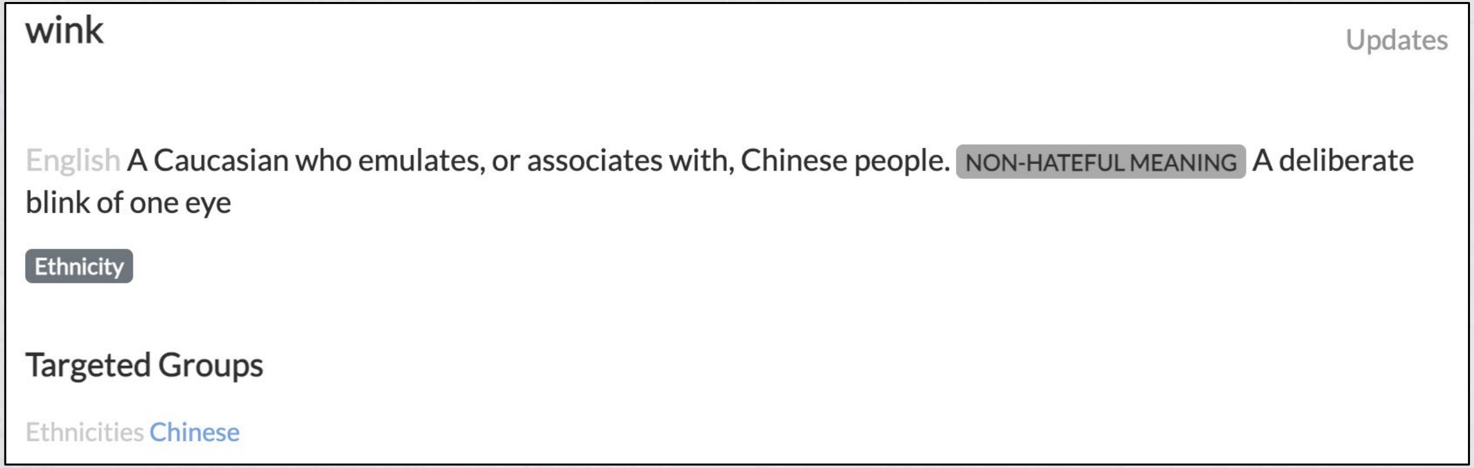
Class



libtard

English Contraction of liberal and retard

Disability



wink

Updates

English A Caucasian who emulates, or associates with, Chinese people. NON-HATEFUL MEANING A deliberate blink of one eye

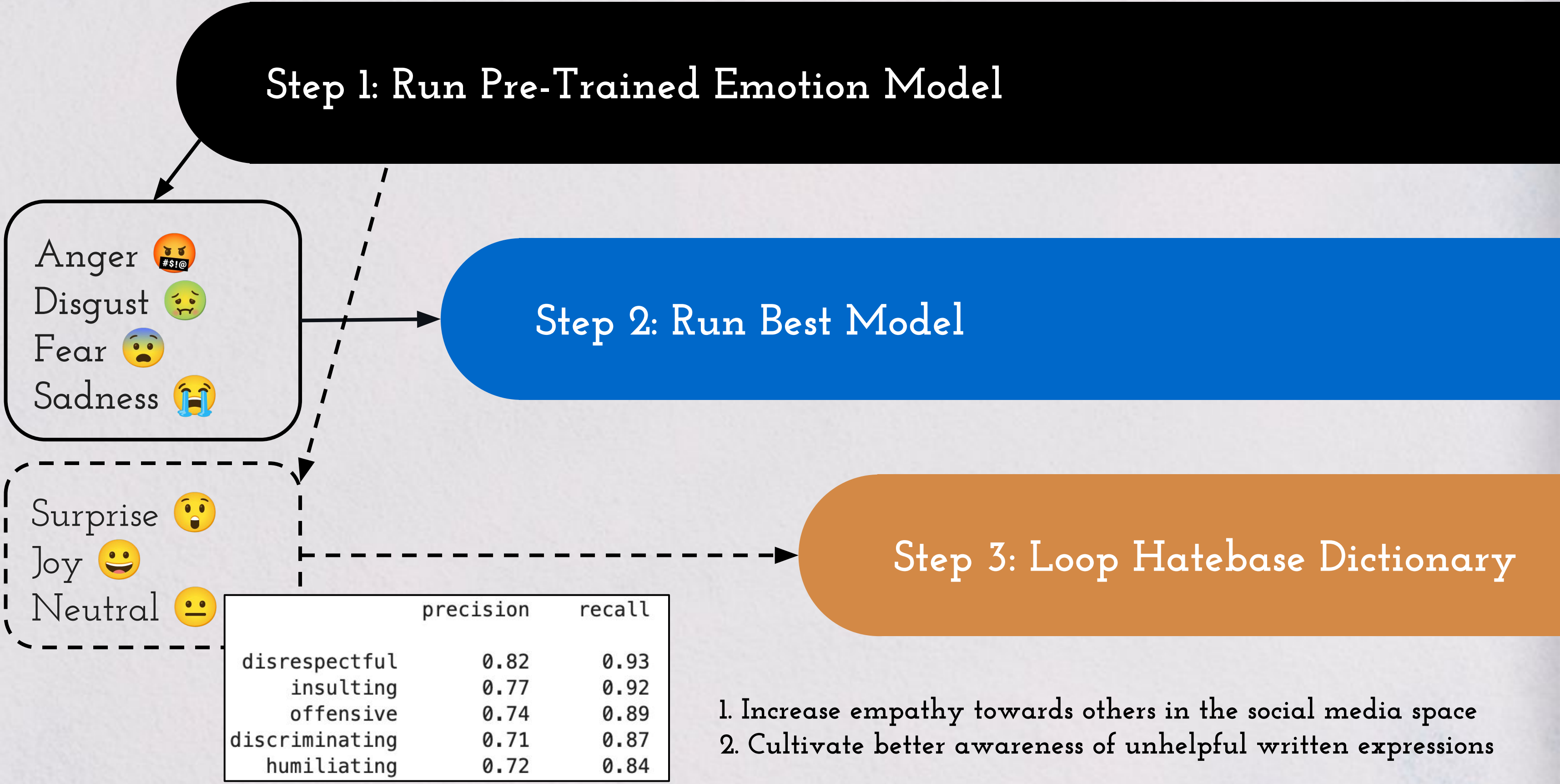
Ethnicity

Targeted Groups

Ethnicities Chinese

Design Thinking

Streamlit API



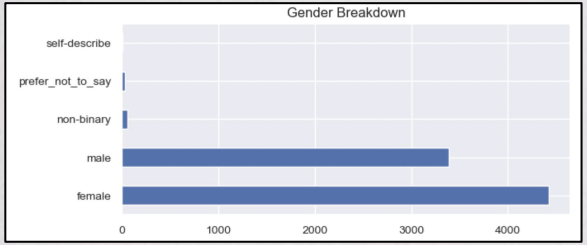
```
                precision    recall

disrespectful        0.82      0.93
    insulting        0.77      0.92
    offensive        0.74      0.89
discriminating       0.71      0.87
  humiliating        0.72      0.84
```

1. Increase empathy towards others in the social media space
2. Cultivate better awareness of unhelpful written expressions

Conclusion

Limitations

1. Nature of Dataset

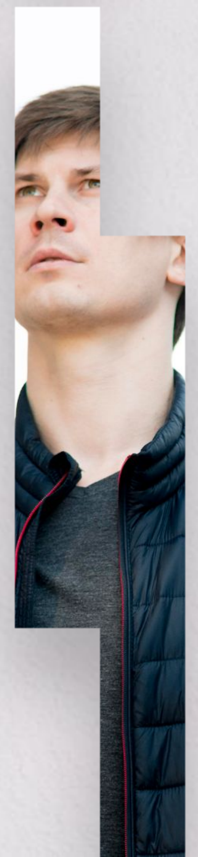


2. Nature of Language

3. Testing of Streamlit API

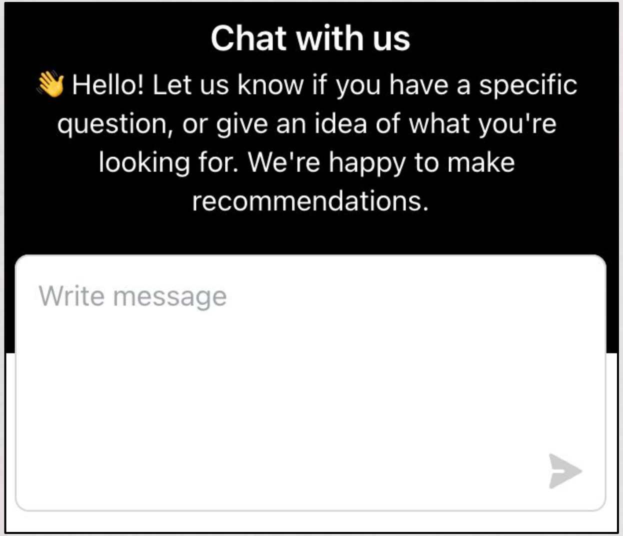
¯_(ツ)_/¯

Conclusion



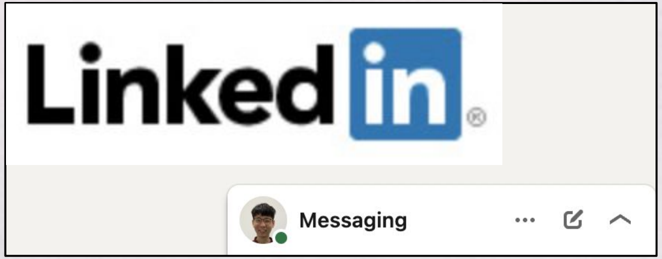
Future Work

1. Make the API more accessible

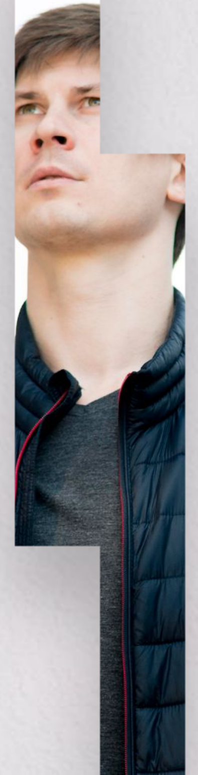


Chat with us

👋 Hello! Let us know if you have a specific question, or give an idea of what you're looking for. We're happy to make recommendations.

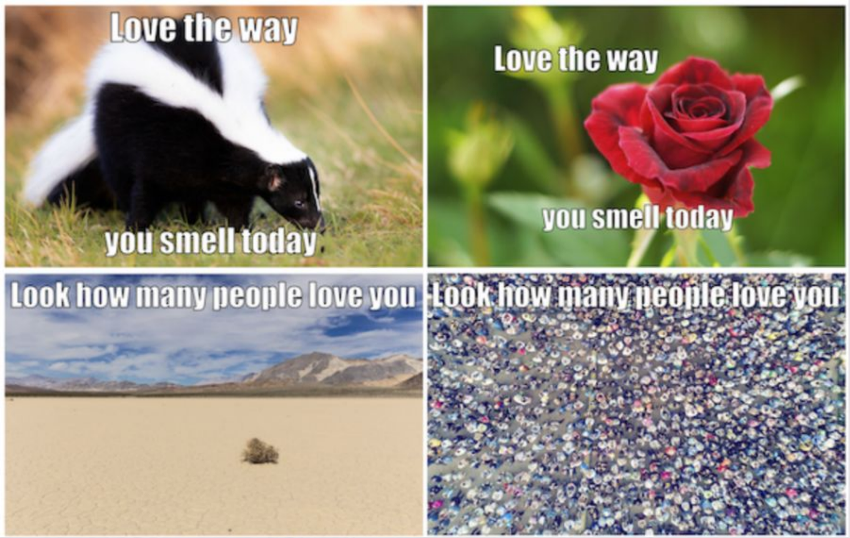


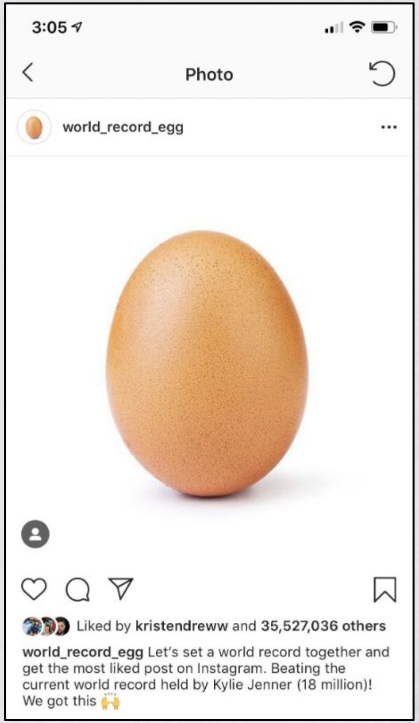
Conclusion



Future Work

2. Incorporate Images





world_record_egg

Liked by kristendreww and 35,527,036 others

world_record_egg Let's set a world record together and get the most liked post on Instagram. Beating the current world record held by Kylie Jenner (18 million)! We got this 🙌

Deep Thanks!

Chua Soon Poh
DSI-32

GENERAL ASSEMBLY

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik