# Real Time Cyber Bullying tweets analysis

*Using Transformers (Hugging Face) and tweepy*

Pratham V K

# Intent and Inspiration

This project on Real Time Tweet Classification of Cyber Bullying is inspired from the recent acquisition of X (formerly Twitter!) by Elon Musk.

Some people feel that there has been an increase in Hate Speech on the Platform, some blame the restructuring of the workforce (since lot of employees who were working in censoring department were fired!).

But most people including Elon do not feel the same..

If there was really hate speech out there on any social platform censoring (or not is a different topic) but we should at least be able to identify some potentially harmful tweets!

# Insight!

So how do we approach this problem?  I may have a solution!

I do feel that when someone posts something online in this case tweets (or text) should go through a membrane (a safety net). A filter that can identify the potentially harmful tweets so that before the person can actually post the tweet, it should give a warning.

May be because of this warning the person might hesitate and not post the potentially harmful tweet, but if he feels it is not a harmful tweet or he does not care and want to still post it, is upto the sender..

But what is this safety net? A LLM that can identify the harmful tweet!

# Implementation

For this we need some text data (basically tweets) to train the LLM.

The Dataset - https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

More About the Dataset - The Data has a collection of 47000 tweets !! The tweets are either describing the bullying tweet or the bullying tweet itself. (This might show some harmful tweets pls explore to point you are comfortable)

The Dataset has been classified to 6 different labels - age, ethnicity, gender, religion, other type of cyber bullying or not cyber bullying. The data was preemptively balanced with each class approx. amounting to 8000 tweets.

# Implementation

In this project I have used Distilbert as my model by Hugging Face Transformers!

The Main reason I have used DistilBERT Model is, lighter than both BERT and RoBERTa. May be there is trade of in accuracy but in this implementation or use case speed is the priority!

Also some of the other technologies that I have used are as follows - Pytorch, nltk, tweepy, torch, tokenizer and more

So let's head on to the code! -

# Implementation

The Dataset in short -

| | cyberbullying_type | count |
|---|---|---|
| 0 | age | 7992 |
| 1 | ethnicity | 7961 |
| 2 | gender | 7973 |
| 3 | not_cyberbullying | 7945 |
| 4 | other_cyberbullying | 7823 |
| 5 | religion | 7998 |

Cleaning the Dataset -

| | tweet_text | cyberbullying_type | hashtags | segmented_hashtags | clean_text |
|---|---|---|---|---|---|
| 0 | In other words #katandandre, your food was cra... | not_cyberbullying | [katandandre, mkr] | kat and andre mkr | in other words , your food was crapilicious! |
| 1 | Why is #aussietv so white? #MKR #theblock #ImA... | not_cyberbullying | [aussietv, MKR, theblock, ImACelebrityAU, toda... | aussie tv mkr the block im a celebrity au toda... | why is so white? |

# Implementation

My Train, Val and Test Split -

```
train_df, temp_df = train_test_split(
    df,
    test_size=0.3,
    stratify=df['cyberbullying_type'],
    random_state=42
)

val_df, test_df = train_test_split(
    temp_df,
    test_size=0.5,
    stratify=temp_df['cyberbullying_type'],
    random_state=42
)
```

✓ 0.0s

My Input Sequence Length is 64 tokens -

```
train_encodings = tokenizer(
    train_df['clean_text'].tolist(),
    truncation=True,
    padding=True,
    max_length=64,
    return_tensors='pt'
)

val_encodings = tokenizer(
    val_df['clean_text'].tolist(),
    truncation=True,
    padding=True,
    max_length=64,
    return_tensors='pt'
)

test_encodings = tokenizer(
    test_df['clean_text'].tolist(),
    truncation=True,
    padding=True,
    max_length=64,
    return_tensors='pt'
)
```

✓ 1.3s

# Implementation

Training - My Best Training Accuracy on the Third Epoch! Of 0.8269

```
Epoch 3/8 - Training loss: 131.2040
Validating Epoch 3: 100%|██████████| 224/224 [00:15<00:00, 14.79it/s]
Validation Accuracy: 0.8269
                    precision    recall  f1-score   support

               age       0.98      0.98      0.98      1199
         ethnicity       0.98      0.98      0.98      1194
            gender       0.84      0.87      0.86      1196
  not_cyberbullying      0.57      0.60      0.59      1192
other_cyberbullying      0.62      0.56      0.59      1173
          religion       0.96      0.96      0.96      1200

          accuracy                           0.83      7154
         macro avg       0.83      0.83      0.83      7154
      weighted avg       0.83      0.83      0.83      7154

  0%|          | 1/2087 [00:00<04:45,  7.30it/s]
🔸 Best model saved at epoch 3 with accuracy 0.8269
100%|██████████| 2087/2087 [04:36<00:00,  7.55it/s]
```

# Implementation

My Test Accuracy -

Final Test Accuracy: 0.8228

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| age | 0.98 | 0.98 | 0.98 | 1199 |
| ethnicity | 0.98 | 0.99 | 0.99 | 1194 |
| gender | 0.84 | 0.89 | 0.87 | 1196 |
| not_cyberbullying | 0.56 | 0.56 | 0.56 | 1192 |
| other_cyberbullying | 0.59 | 0.56 | 0.58 | 1174 |
| religion | 0.96 | 0.95 | 0.95 | 1199 |
| accuracy | | | 0.82 | 7154 |
| macro avg | 0.82 | 0.82 | 0.82 | 7154 |
| weighted avg | 0.82 | 0.82 | 0.82 | 7154 |

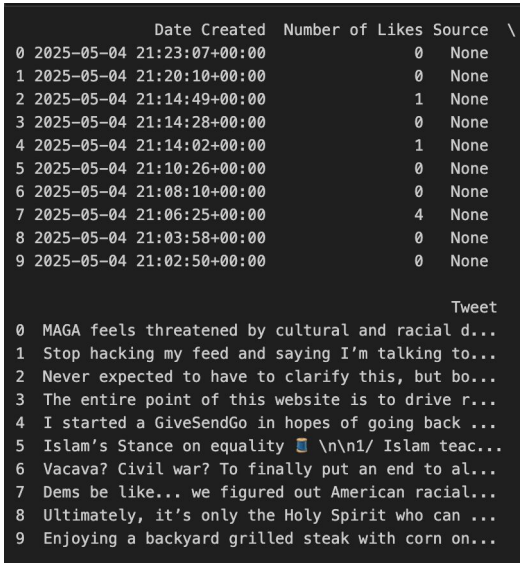
# Implementation

Implementation of the classification by pulling tweets in Real Time

```
import tweepy
import pandas as pd
from config import BEARER_TOKEN


client = tweepy.Client(bearer_token=BEARER_TOKEN)

search_query = "racial -is:retweet -is:reply -has:links lang:en"


response = client.search_recent_tweets(
    query=search_query,
    max_results=10,
    tweet_fields=["created_at", "public_metrics", "source", "text"]
)
```

# Implementation

Implementation of the classification by pulling tweets in Real Time



| | Date Created | Number of Likes | Source | Tweet |
|---|---|---|---|---|
| 0 | 2025-05-04 21:23:07+00:00 | 0 | None | MAGA feels threatened by cultural and racial d... |
| 1 | 2025-05-04 21:20:10+00:00 | 0 | None | Stop hacking my feed and saying I'm talking to... |
| 2 | 2025-05-04 21:14:49+00:00 | 1 | None | Never expected to have to clarify this, but bo... |
| 3 | 2025-05-04 21:14:28+00:00 | 0 | None | The entire point of this website is to drive r... |
| 4 | 2025-05-04 21:14:02+00:00 | 1 | None | I started a GiveSendGo in hopes of going back ... |
| 5 | 2025-05-04 21:10:26+00:00 | 0 | None | Islam's Stance on equality 🧵 \n\n1/ Islam teac... |
| 6 | 2025-05-04 21:08:10+00:00 | 0 | None | Vacava? Civil war? To finally put an end to al... |
| 7 | 2025-05-04 21:06:25+00:00 | 4 | None | Dems be like... we figured out American racial... |
| 8 | 2025-05-04 21:03:58+00:00 | 0 | None | Ultimately, it's only the Holy Spirit who can ... |
| 9 | 2025-05-04 21:02:50+00:00 | 0 | None | Enjoying a backyard grilled steak with corn on... |

# Implementation

The real time tweets results -

| | Date Created | Number of Likes | Source | Tweet | Prediction | Label |
|---|---|---|---|---|---|---|
| 0 | 2025-05-04 21:23:07+00:00 | 0 | NaN | MAGA feels threatened by cultural and racial differences - evil when it hurts others for no reason beyond childish misdirected hate. Billionaires ripped the guts out of America with their greed - and will do it again. Not fault of migrants or Muslims - fault of men like those | 5 | religion |
| 1 | 2025-05-04 21:20:10+00:00 | 0 | NaN | Stop hacking my feed and saying I'm talking to you ninjas. I dunno anybody and I'm to arms length for comfort for you to know me so please don't do the uh clout chase thing . If I date in the future it will only be middle eastern and white appearing men or MULTI RACIAL no bi | 1 | ethnicity |
| 2 | 2025-05-04 21:14:49+00:00 | 1 | NaN | Never expected to have to clarify this, but both murder and directing racial slurs at a child are wrong.\nAlso, murder is worse.\nAny conversation thereafter should begin by recognizing those two truths. | 0 | age |
| 3 | 2025-05-04 21:14:28+00:00 | 0 | NaN | The entire point of this website is to drive racial hatred it really seems. | 1 | ethnicity |
| 4 | 2025-05-04 21:14:02+00:00 | 1 | NaN | I started a GiveSendGo in hopes of going back to school for IT to provide a better life for my kids, but I didn't call anyone a racial slur or stab anyone so I guess fuck me. | 0 | age |
| 5 | 2025-05-04 21:10:26+00:00 | 0 | NaN | Islam's Stance on equality 🧵 \n\n1/ Islam teaches that all humanity descends from Adam, uniting us as one family. The Quran (49:13) states: "We made you into nations and tribes to know one another." Racial distinctions are for recognition, not division. Equality is divine. | 5 | religion |
| 6 | 2025-05-04 21:08:10+00:00 | 0 | NaN | Vacava? Civil war? To finally put an end to all this prejudice on both sides of the racial divide? | 1 | ethnicity |
| 7 | 2025-05-04 21:06:25+00:00 | 4 | NaN | Dems be like... we figured out American racial harmony, let me introduce you to a concept and slogan we call ALL LIVES MATTER | 1 | ethnicity |
| 8 | 2025-05-04 21:03:58+00:00 | 0 | NaN | Ultimately, it's only the Holy Spirit who can bring about society's transformation. It's the Spirit who brings unity, breaking down divisions of gender, race & social position. Those indwelt by the Spirit should be fighting for gender, racial & social equality. @nickygumbel | 5 | religion |
| 9 | 2025-05-04 21:02:50+00:00 | 0 | NaN | Enjoying a backyard grilled steak with corn on the cob and a baked potato with the wife of 43 years watching college baseball with no out of control players, at commercial go to golf also no racial drama. With decent politicians it is the golden era. | 0 | age |

# Challenges

Some of the Challenges that I faced are in while doing this assignment are -

- Accuracy in dealing with the categories other_cyberbullying and not_cyberbullying.
- The X API (or twitter API) has limited restrictions.
- Some of the tweets are hard to classify credit -  Sarcasm (more than estimated)
- Where do you draw the line?
- Increase in the General use of Foul Language..

# Conclusion

I had great time exploring the not so fancy side of social media (X to be precise).

This is “Extreme Censoring” and I know it. People might feel overwhelmed and may not be inspired to put their opinions out there. These interactions might not be approach or intent of the general public but we still cannot ignore any online hate or bullying for that matter!

Thanks to my Professor Binil Starly for giving this opportunity to work on this project.

No to Online Hate!!