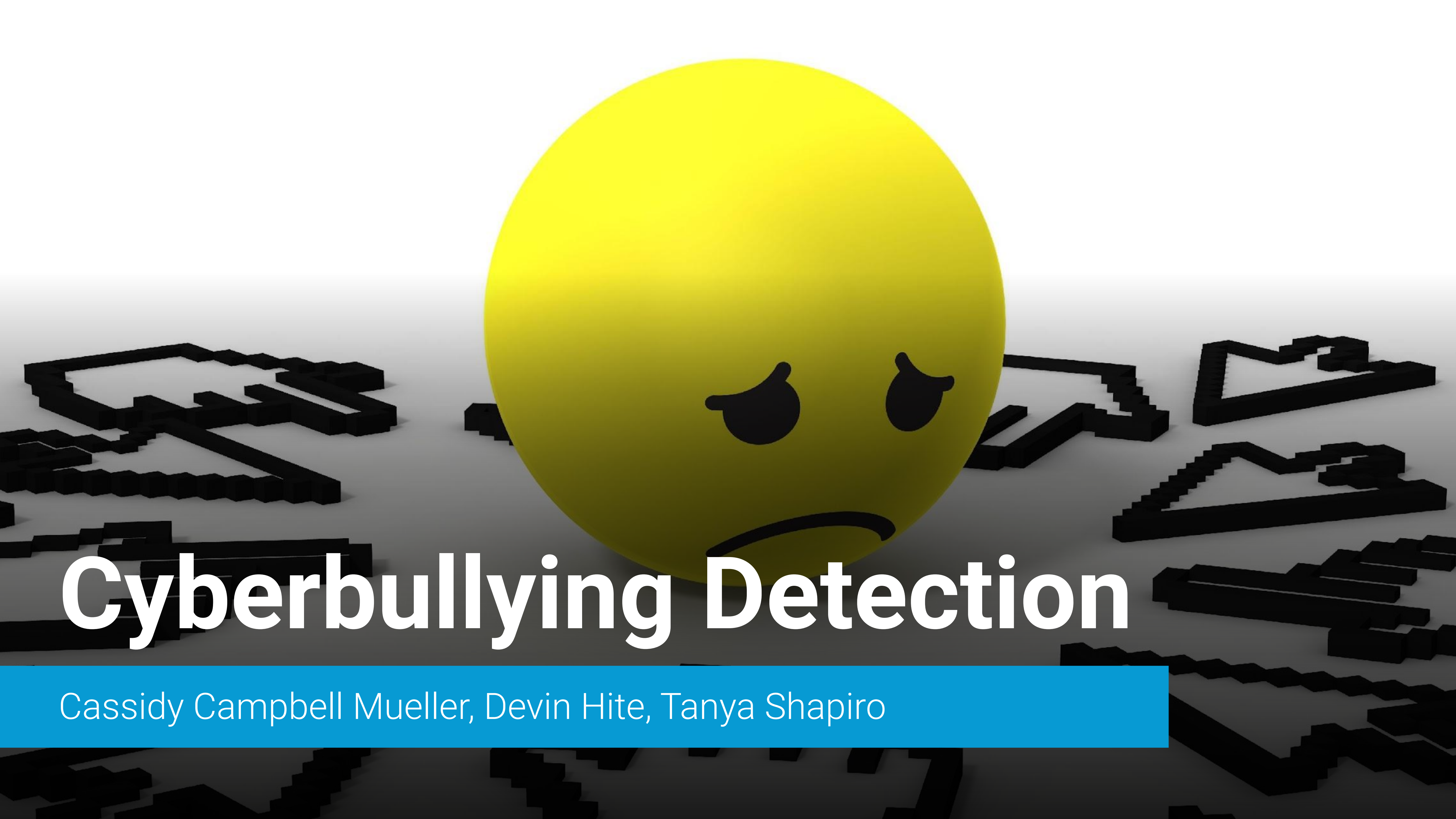


# Cyberbullying Detection

Cassidy Campbell Mueller, Devin Hite, Tanya Shapiro

# Social Impact

- Cyberbullying Definition: “the use of digital technology to inflict harm repeatedly or to bully” *(Wang, Lu, & Fu 2020)*
- Over a third of middle school & high school students have felt cyberbullied *(Wang, Lu, & Fu 2020)*
- Cyberbullying has gotten worse. Reported 70% increase in cyberbullying since COVID lockdowns *(Digital Trends 2020)*
- Children and young adults who are victims of cyberbullying are more than twice as likely to self-harm and enact suicidal behavior, according to a study. *(Swansea University, Wales, UK 2018)*

# Problem Statement

Twitter's CEO, Parag Agrawal, is invested in making Twitter a safe space for users to express themselves and connect with others. In the past decade, there has been an increase in offensive tweets.

Agrawal has asked you and a team of Data Scientists to develop an algorithm to detect inappropriate Tweets and flag them for cyberbullying.

The objective is to not only classify which tweets are offensive, but what type of cyberbullying (gender, race, age, etc). Although the baseline accuracy is ~17%, Agrawal will not implement the algorithm unless it can classify tweets with at least 70% accuracy.

# Background

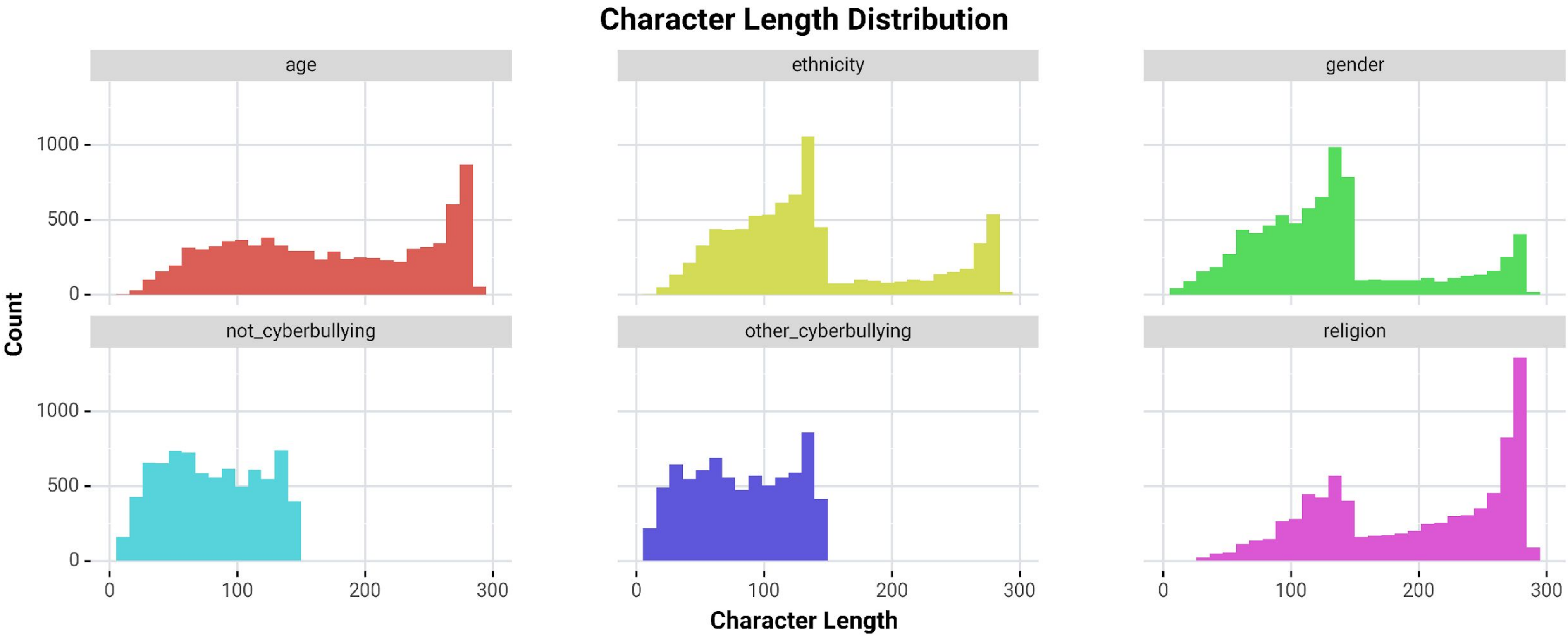
- Dataset from Kaggle
- Contains 47K text tweets
- Consists of 6 Classes

  religion, gender, age, ethnicity, other, & not cyberbullying
- Balanced dataset (8K per class)

# Verbosity : Character Length

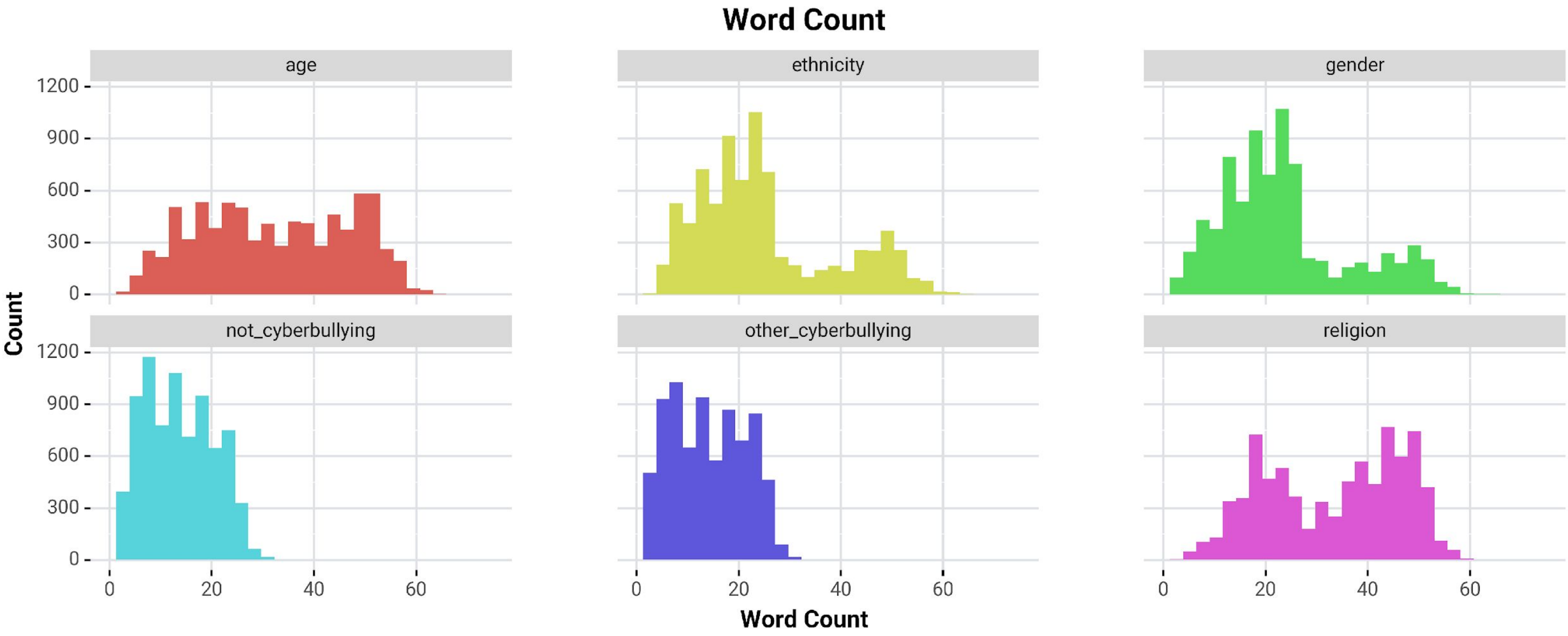
**Not Cyberbullying** and **Other Cyberbullying** tweets less verbose compared to tweets in other classes.

**Age** and **Gender** tweets more verbose compared to tweets in other classes.



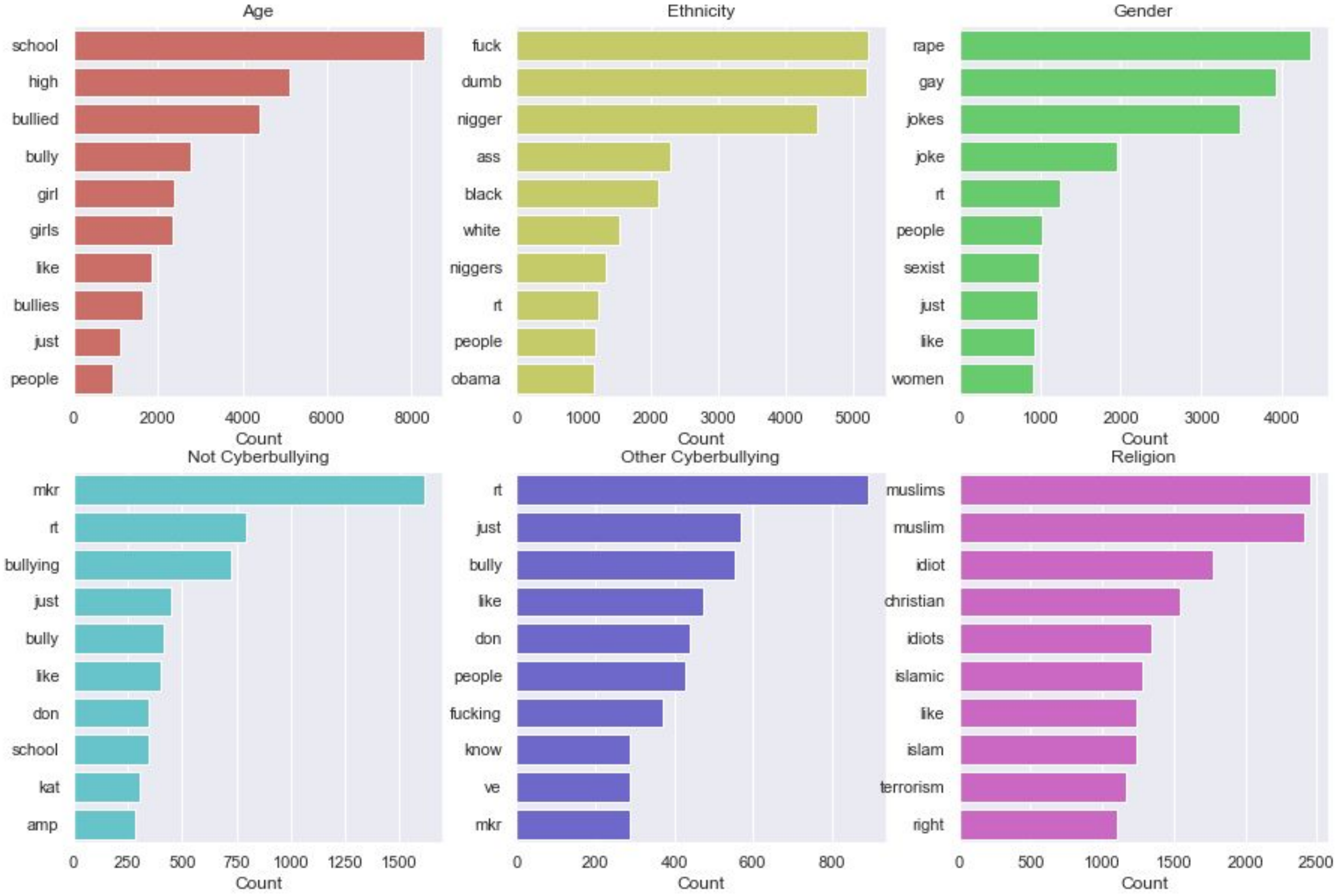
# Verbosity: Word Count

Word Count comparison supports similar observations noted about character length distribution.

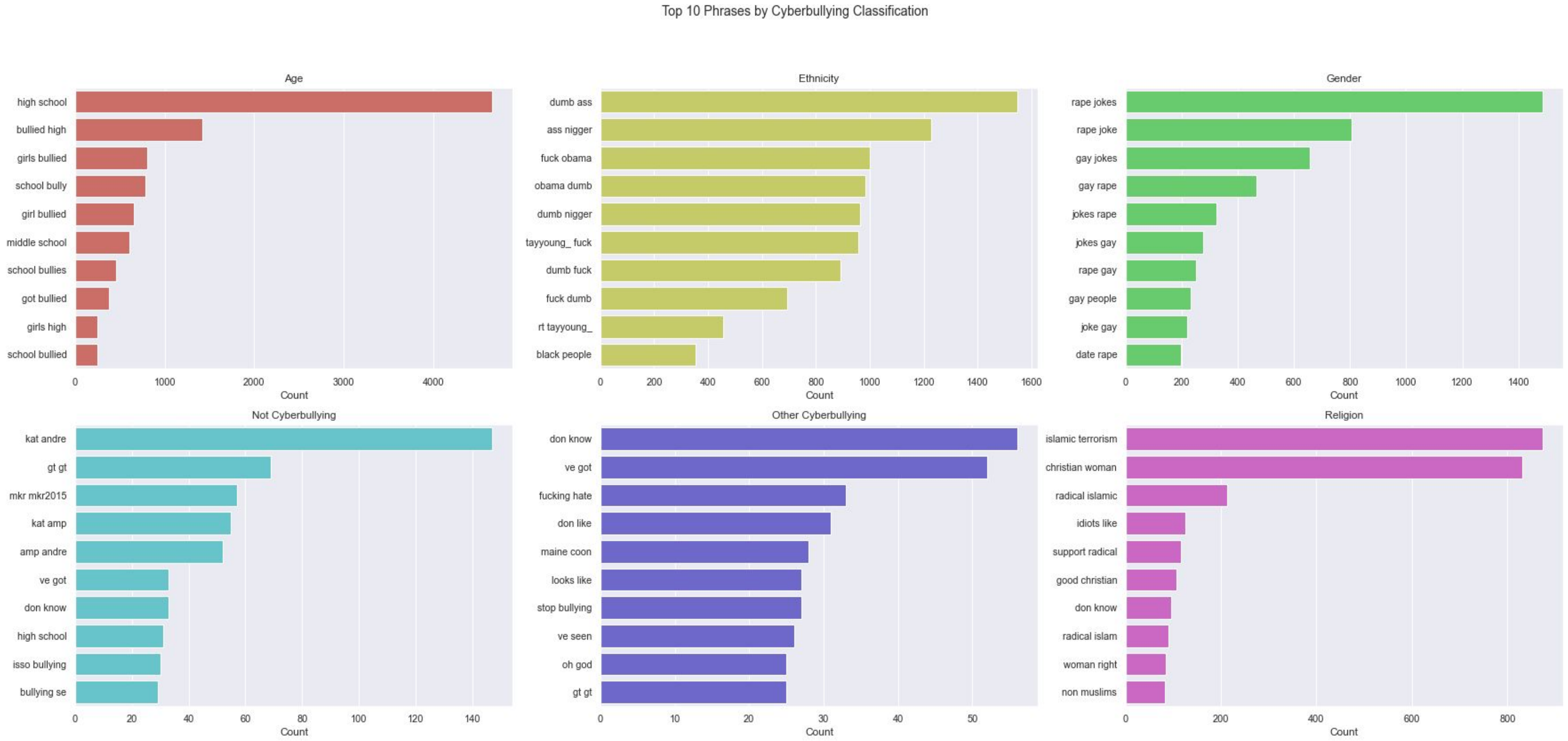


# Common Words

- **Age** language specific to schools
- **Ethnicity** top words contain several offensive racial slurs aimed at African Americans
- **Gender** tweets contains words that are also homophobic
- **Religion** cyberbullying tweets targeted mostly at Islamic communities
- Difficult to discern patterns for **not** or **other cyberbullying**



# Common Phrases



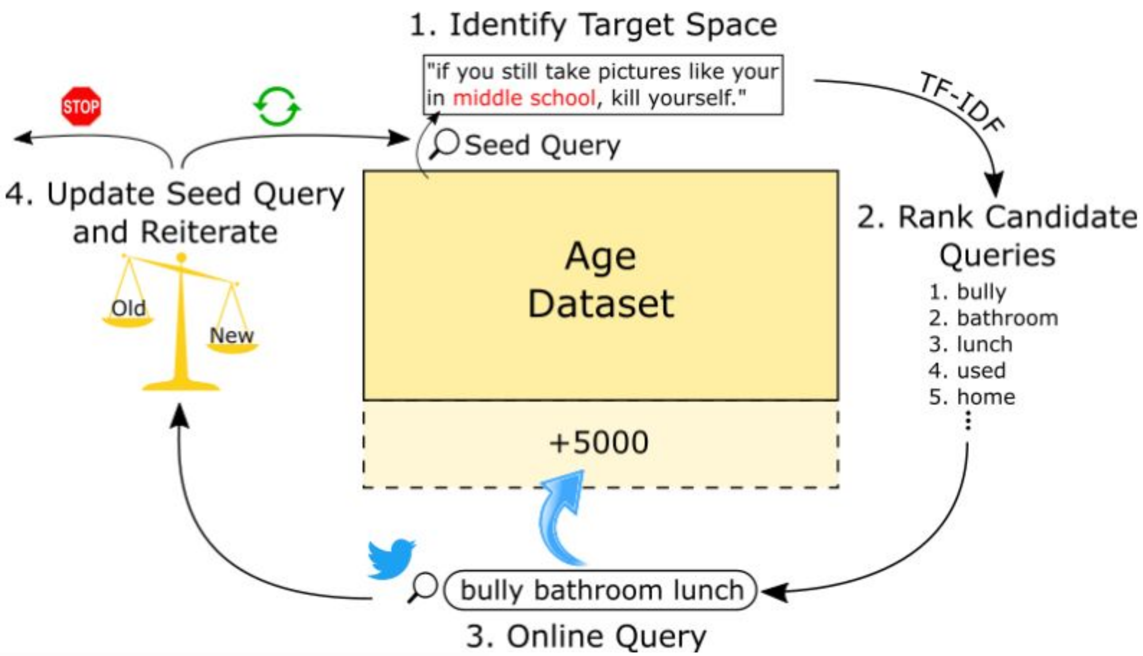
# Modeling Methods

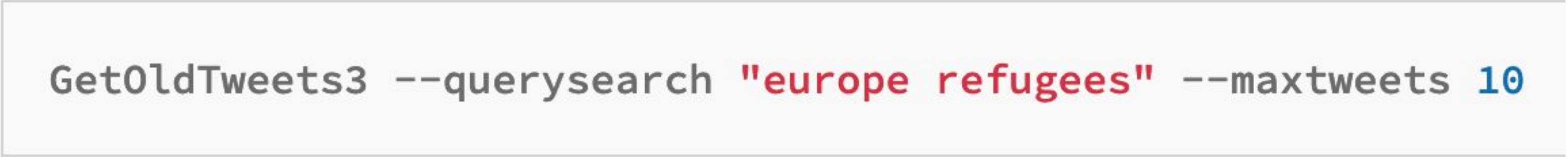
- Oversampling technique
- Text preprocessing
- Types of models
- Metrics
- Misclassified tweets

# Tweet scraping method used by Wang *et al.* (2020) to create balanced dataset

Dynamic Query Expansion was to mine tweets by class

GetOldTweets can scrape tweets older than 1 week. The twitter API doesn't do this.





```
GetOldTweets3 --querysearch "europe refugees" --maxtweets 10
```

Above: (Figure 2, Wang et al. 2020)
Left: (pypi.org/project/GetOldTweets3/)

All 3 are rabid hindu hating bigots. @SpeakerPelosi **@JoeBiden @TomPerez @DNC** get ur act together &amp; not let these 3 be the face of ur party. U already hv bigots @RepJayapal @tedlieu @IlhanMN @aoc @BernieSanders who also love **&amp;** islamic terrorism

All 3 are rabid hindu hating bigots. get ur act together not let these 3 be the face of ur party. U already hv bigots who also love islamic terrorism

# Text Preprocessing

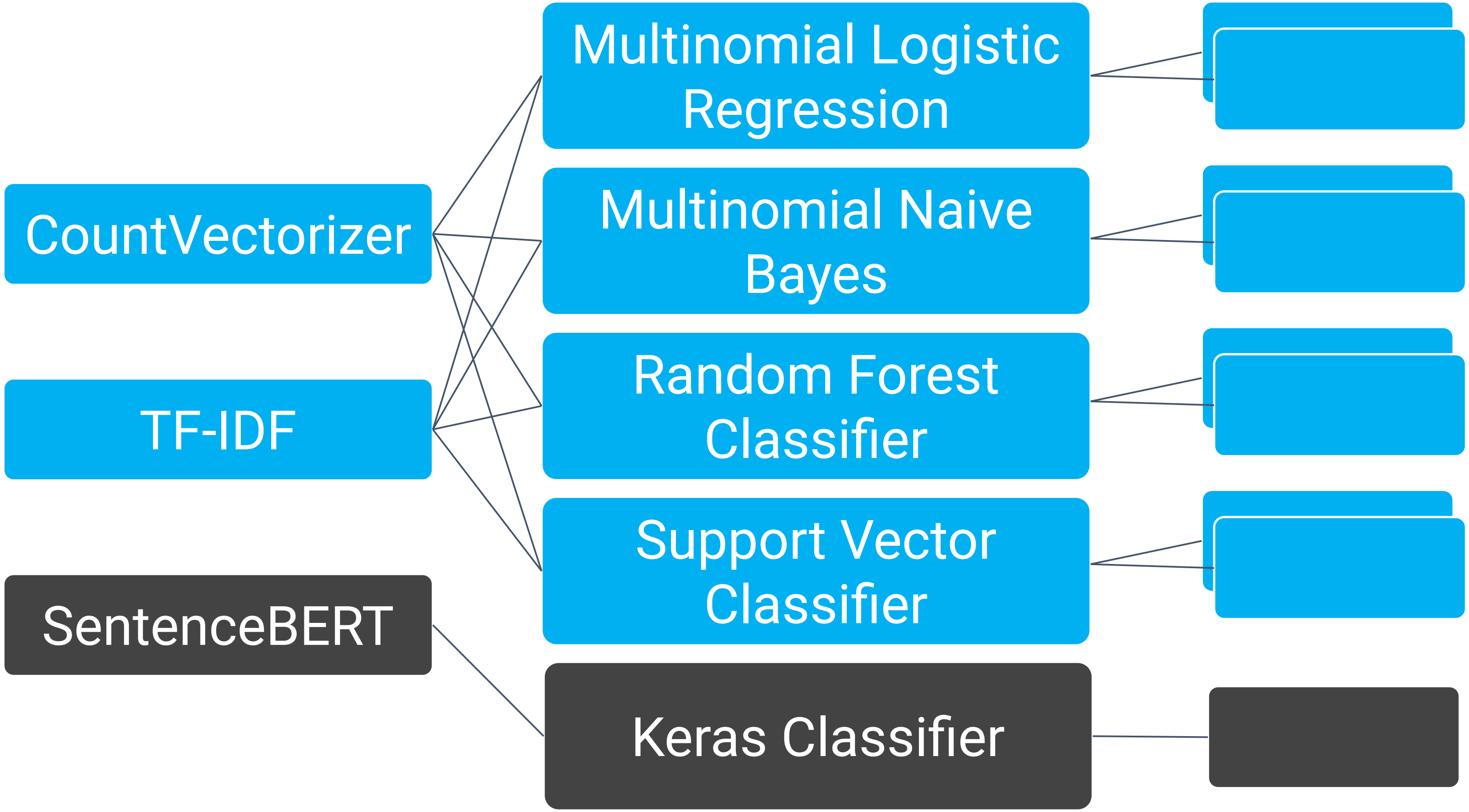
Removed

- URLs
- Usernames and anything after @ symbols
- Stop words ‘english’
- HTML artifacts

Kept

- Hashtags
- Vulgar language
- Emojis

# Compared 3 text embedding methods and 5 multiclass classifier models



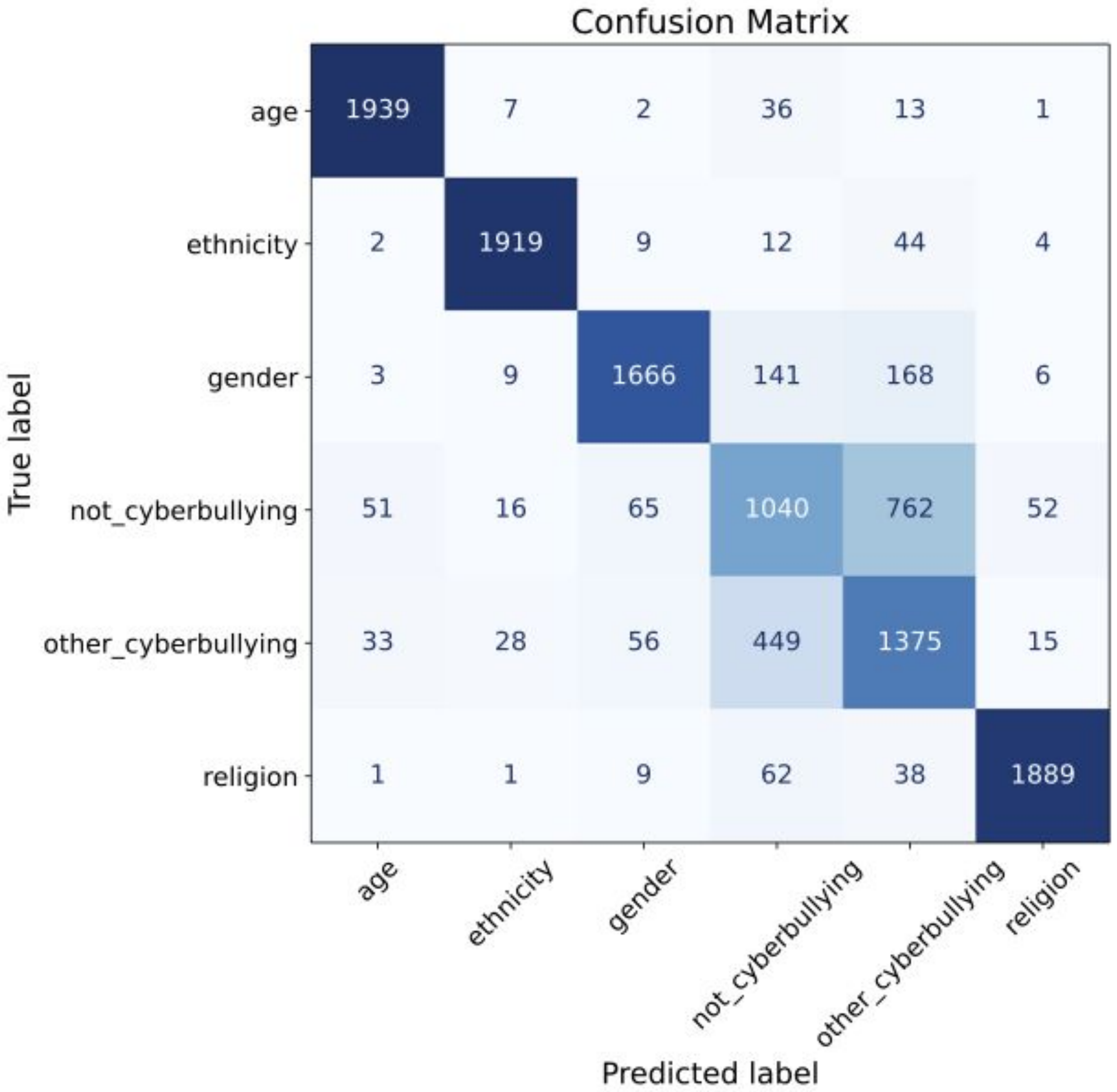
9 models were tested

2 were interpretable

8 were white-box models

# TF-IDF text embedding with logistic regression had the highest accuracy score

| Text Embedding | Model | Train accuracy score | Test accuracy score |
|---|---|---|---|
| CountVectorizer | Multinomial Logistic Regression | 0.920 | 0.820 |
| TF-IDF | Multinomial Logistic Regression | 0.888 | 0.824 |
| CountVectorizer | Multinomial Naive Bayes | 0.822 | 0.775 |
| TF-IDF | Multinomial Naive Bayes | 0.825 | 0.761 |
| CountVectorizer | Support Vector Classifier | 0.907 | 0.769 |
| TF-IDF | Support Vector Classifier | 0.990 | 0.789 |
| CountVectorizer | Random Forest Classifier | 0.994 | 0.804 |
| TF-IDF | Random Forest Classifier | 0.994 | 0.796 |
| SentenceBERT | Keras Classifier | 0.827 | 0.776 |



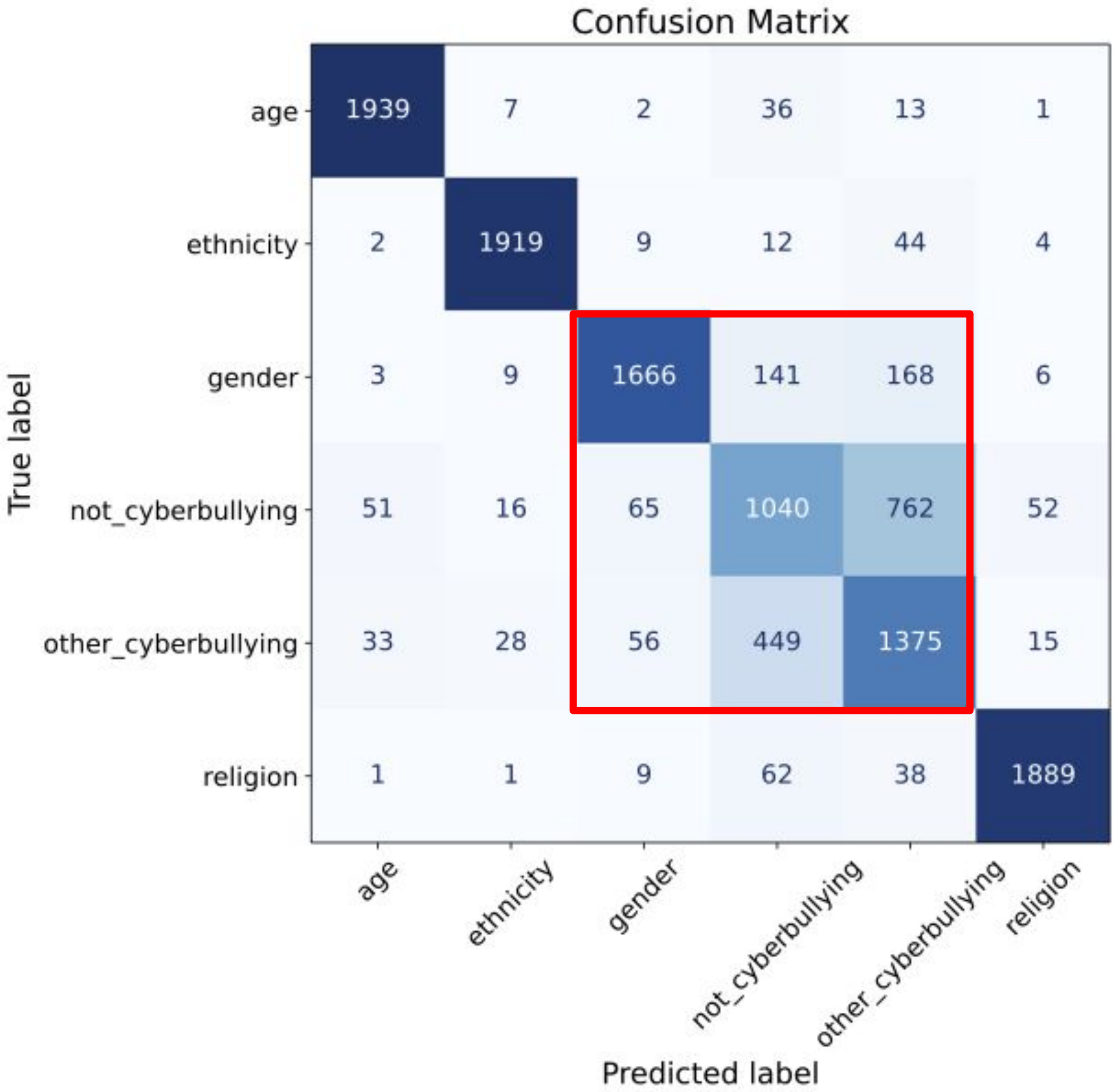
| | age | ethnicity | gender | not_cyberbullying | other_cyberbullying | religion |
|---|---|---|---|---|---|---|
| age | 1939 | 7 | 2 | 36 | 13 | 1 |
| ethnicity | 2 | 1919 | 9 | 12 | 44 | 4 |
| gender | 3 | 9 | 1666 | 141 | 168 | 6 |
| not_cyberbullying | 51 | 16 | 65 | 1040 | 762 | 52 |
| other_cyberbullying | 33 | 28 | 56 | 449 | 1375 | 15 |
| religion | 1 | 1 | 9 | 62 | 38 | 1889 |

# Confusion matrix

The model predicted age, ethnicity, and religion well (F1 > 0.95)

Tweets that were labeled “not” had the lowest F1 score (0.56)

Tweets that were labeled “other” had the second lowest (0.63)



Confusion Matrix

| True label \ Predicted label | age | ethnicity | gender | not_cyberbullying | other_cyberbullying | religion |
|---|---|---|---|---|---|---|
| age | 1939 | 7 | 2 | 36 | 13 | 1 |
| ethnicity | 2 | 1919 | 9 | 12 | 44 | 4 |
| gender | 3 | 9 | 1666 | 141 | 168 | 6 |
| not_cyberbullying | 51 | 16 | 65 | 1040 | 762 | 52 |
| other_cyberbullying | 33 | 28 | 56 | 449 | 1375 | 15 |
| religion | 1 | 1 | 9 | 62 | 38 | 1889 |

The model had a hard time distinguishing between language labeled “gender”, “other” and “not”

weird teen twitter is really bad with it. I've been seeing a lot more gay slurs and rape jokes

# Misclassified tweets

## predicted ‘gender’ / actual ‘not cyberbullying’

### Totally wrong

RT This tweet deserves more love. It's a good point.

RT 3 followers till 1000!

### Could actually be cyberbullying

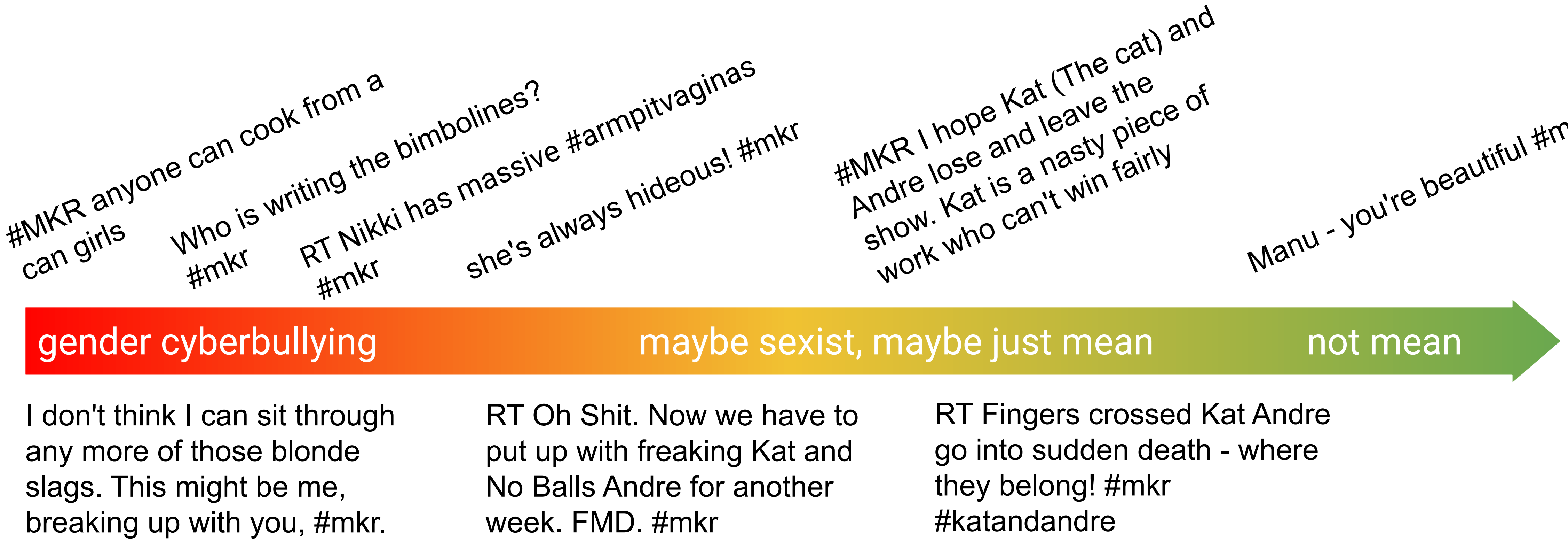
Don't count any chickens...most of the GOP candidates suck and could well lose to uneducated "history" vagina voters.

### Discussions of gender issues

Kristen Tate says we choose to take maternity leave

# Misclassified tweets

## predicted ‘gender’ / actual ‘not cyberbullying’



#MKR anyone can cook from a can girls

Who is writing the bimbolines? #mkr

RT Nikki has massive #armpitvaginas #mkr

she's always hideous! #mkr

#MKR I hope Kat (The cat) and Andre lose and leave the show. Kat is a nasty piece of work who can't win fairly

Manu - you're beautiful #mkr

gender cyberbullying

maybe sexist, maybe just mean

not mean

I don't think I can sit through any more of those blonde slags. This might be me, breaking up with you, #mkr.

RT Oh Shit. Now we have to put up with freaking Kat and No Balls Andre for another week. FMD. #mkr

RT Fingers crossed Kat Andre go into sudden death - where they belong! #mkr #katandandre

**The majority of gender misclassified tweets were related to #mkr or My Kitchen Rules**

# Applying our model to a different corpus: Trump Tweets

- When running a corpus of 43,000+ tweets from Donald Trump through our best model:
- The score was .909 train/.800 test
- 21,152 (~50%) were found to be class “not cyberbullying”

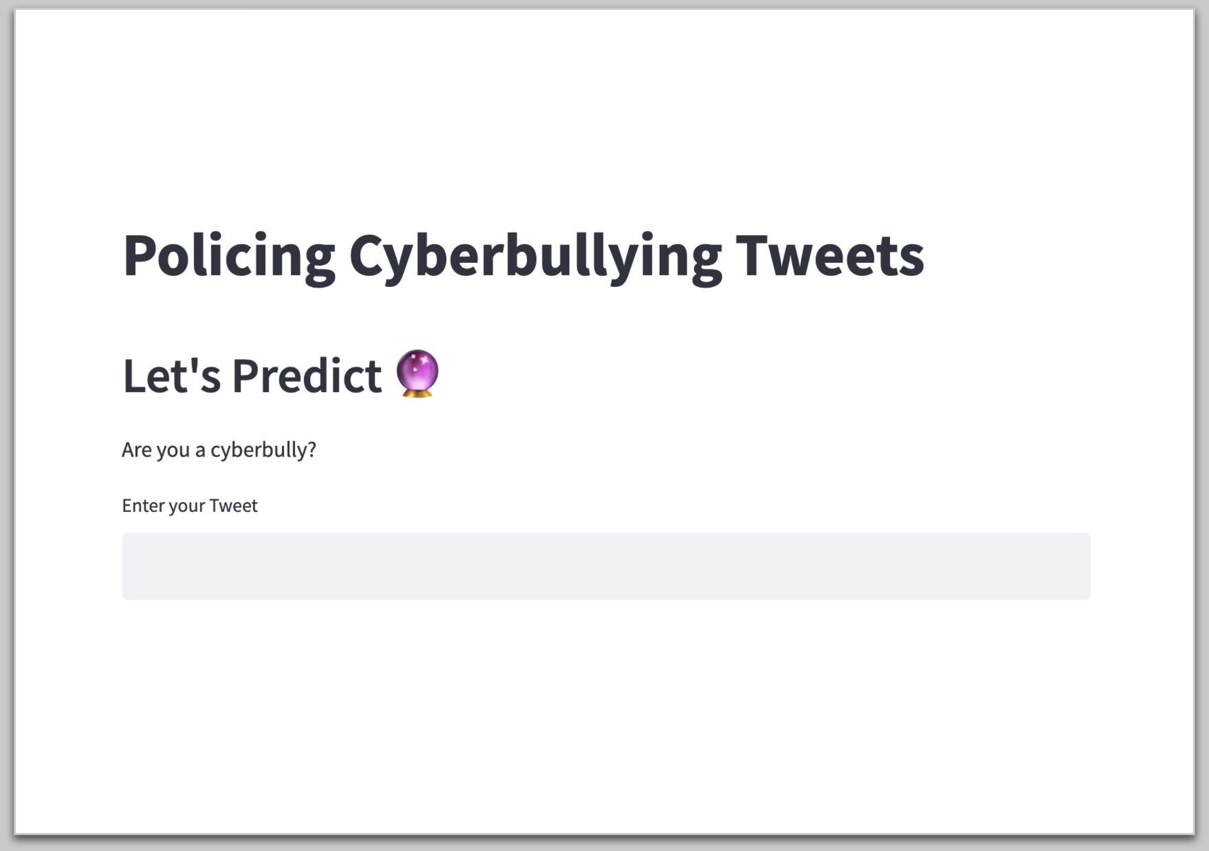
**Cyberbullying Breakdown**

| Class | # of Tweets | % of Tweets |
|---|---:|---:|
| Religion | 1500 | 7.1% |
| Gender | 421 | 2.0% |
| Ethnicity | 396 | 1.9% |
| Age | 184 | 0.9% |
| Other | 19700 | 93.8% |

# Conclusions

- Our best model, TF-IDF with multinomial logistic regression, classified tweets with 82% accuracy
- The model still struggled with tweets that were sarcastic or that used vulgar language but were against cyberbullying
- Classifying language as bullying is a grayzone (subjective)
- Future recommendation to apply this study to texts from other social platforms, e.g. Facebook, Instagram

# Proof of Concept



## Policing Cyberbullying Tweets

### Let's Predict 🔮

Are you a cyberbully?

Enter your Tweet

# Works Cited

- Wang, J., Fu, K., Lu, C. SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. Proceedings - IEEE International Conference on Big Data (Big Data): 1699-1708, 2020.
- GetOldTweets3. 2019. https://pypi.org/project/GetOldTweets3/
- https://www.kaggle.com/austinreese/trump-tweets