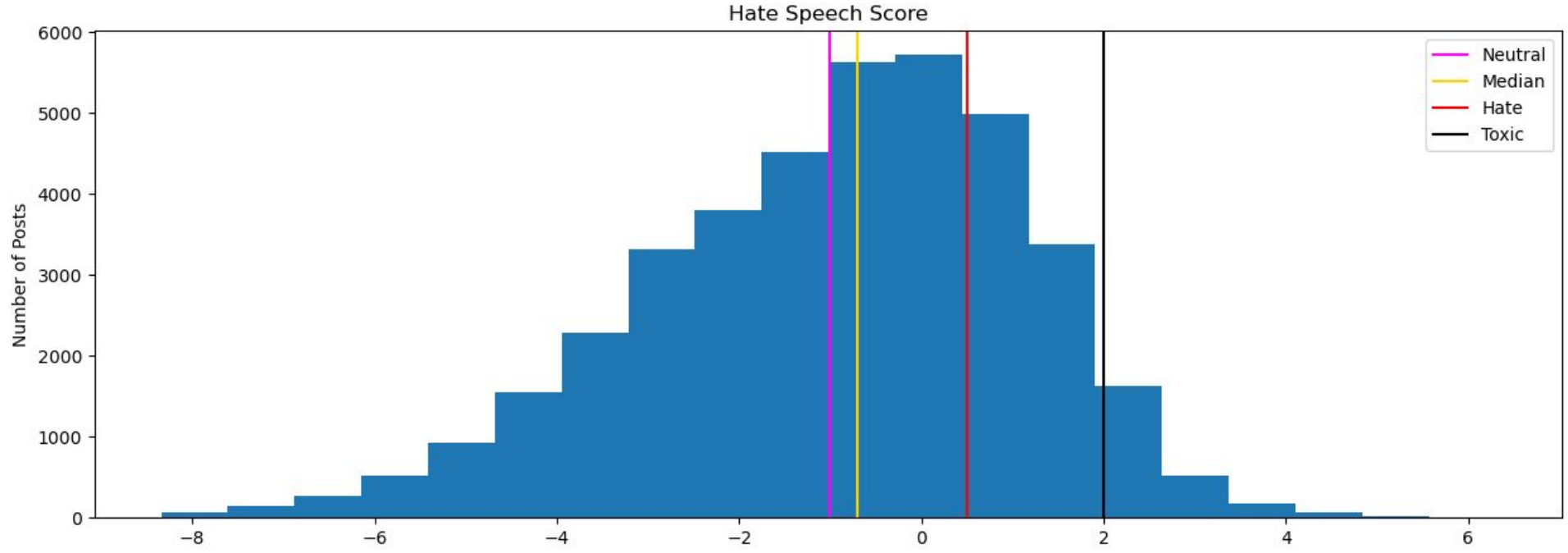
Malevolence Prevention

Now or Never, Propositions in Preventing Hate Speach

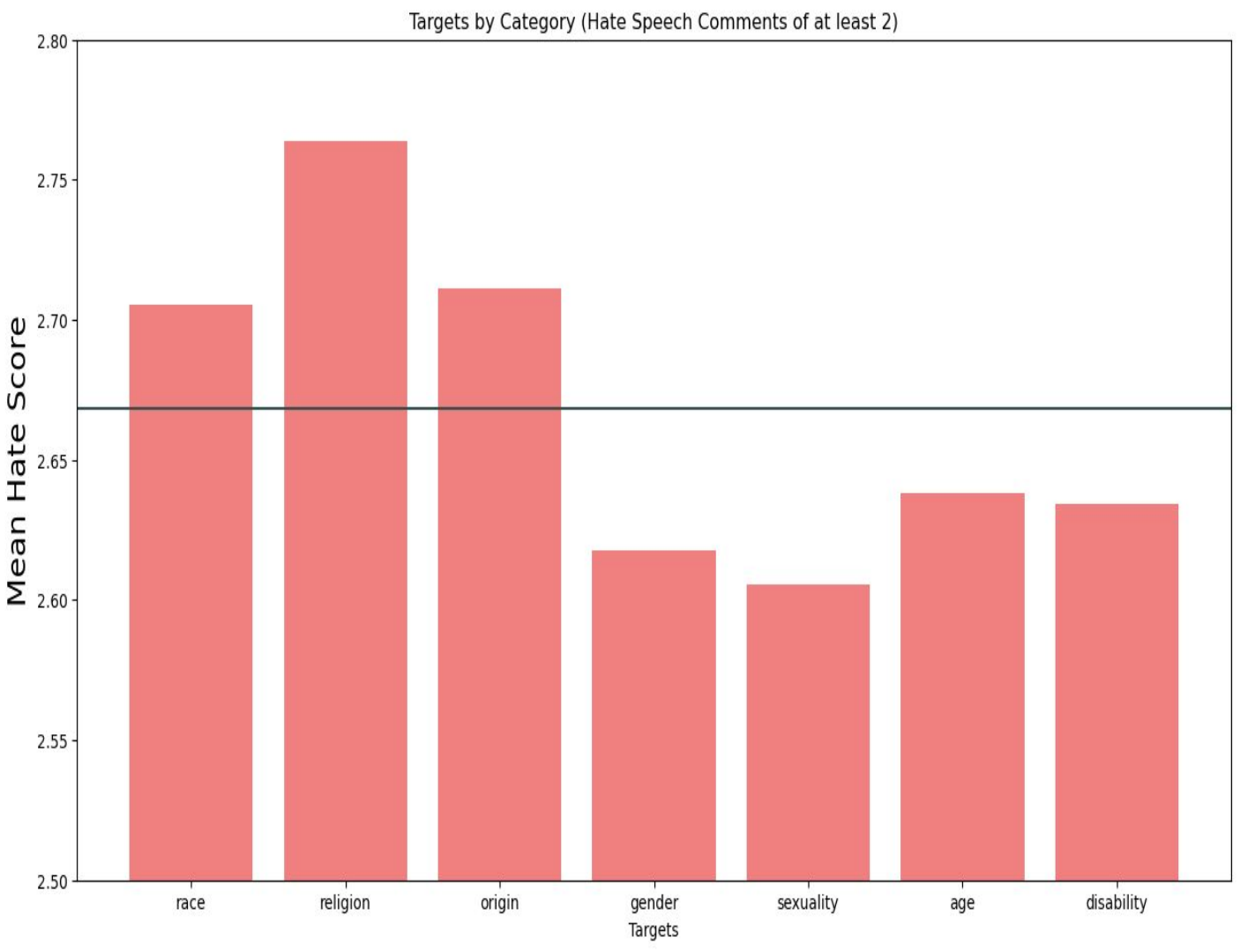
Problem Statement:

- Hate speech is a malevolent thing that must be monitored, stopped, and demonstrated how terrible it is.
- A local school, unable to read all of their students' social media accounts, reach out to you to make a make a predictive model to identify different levels of hate speech posts to have their staff focus on manually reading.
- Their expectation is to have multiple staff members review the 'toxic' hate speech and speak to the student in front of their parents; 'standard' hate speech would just be reviewed by one teacher and talk to the student.
- Data used is from a Berkeley 2020 study, gathered via Twitter, Reddit, and Youtub's APIs.

- Below is the distribution of hate speech scores from Berkeley, with a distribution of comments on the Internet being slightly skewed to the right (average being -.94, so approximately at the border between defensive and neutral comments).
- Real life hate speech might vary; our data and project is focused online.



- Relatively, extra care should be focused on race, religion, and origin based-hate speech comments.
- Pragmatically, targets in the ‘minority’ should be treated more seriously.
- Regarding prevalence: Almost half of all toxic hate speech is gender based.

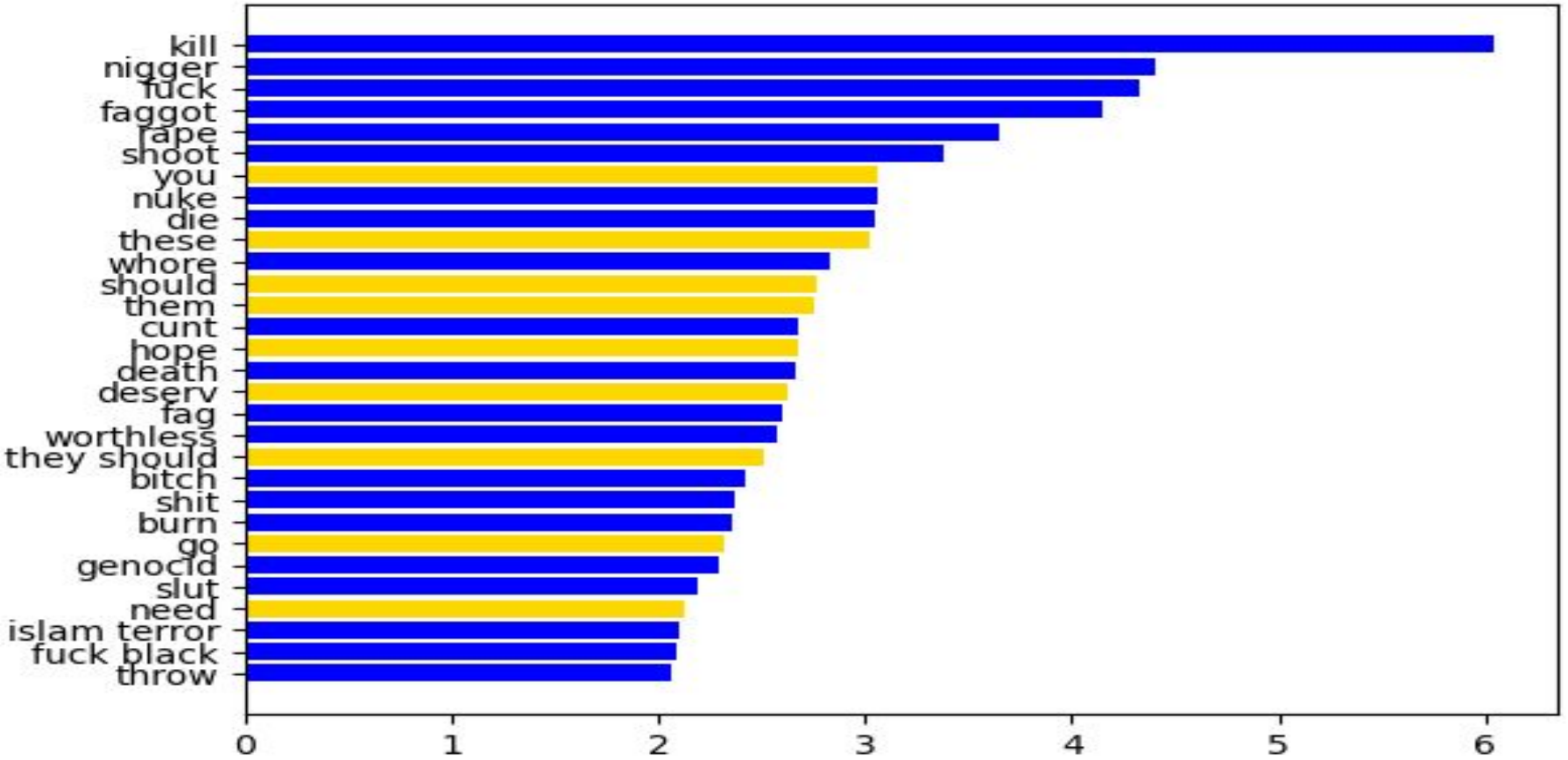


Methodology

- The original dataset was condensed and grouped by comment.
- After EDA, a hate speech score of 2 was chosen as the threshold to separate ‘normal’ from toxic hate speech.
- After cleaning, random replacement was needed to make up for the discrepancy in classes (74, 21, and 5% correspondingly).
- A custom sensitivity-weighted metric was used when grid searching through the many models and combinations of parameters.
- In short, modeling this data was not easy.
- Ultimately, the final model chosen was a TF-IDF Logistic Regression model, with a custom sensitivity of 63% (94% general sensitivity if this would be a mere binary classification).


WARNING !
SOME VIEWERS MAY FIND THE
FOLLOWING VIDEO DISTURBING
VIEWER DISCRETION IS ADVISED
Adobe Stock | #230486049

TF-IDF’s 30 Most Frequent Terms of Toxic Hate Speech



Conclusions

- Afterwards, the model, focusing on individual comments, was applied to a Streamlit application to process social media accounts' posts.
- Upon identifying one significantly confident prediction of a toxic post or three of a 'standard' posts (their respective median prediction confidence, .58 and .94*), appropriate measures would be taken by the school.
- A sliding scale was added to the system to accommodate for less confident predictions, deducting linearly the thresholds needed to call the teachers' attentions.

Future Ideas

- Model with SVM and Recurrent Neural Networks as well as any other algorithm that allows for weighted focus.
- Discover a way for the application to re-confirm past predictions, lowering the risk of false negatives.
- Construct new targets outside of what was already accounted for, such as body image.
- Explore other features that would not be noticed on a pure textual analysis: Initial post, urls, emojis, etc.
- Get more data!