FLAGGING CYBERBULLYING ON TWITTER FOR BE KIND ORG.

SABINA BAINS

AUGUST 2022

BUSINESS OBJECTIVE

- Changes in Twitter leadership could lead to a **rise in hate speech and misinformation** online

- BeKind Org. will create a **browser extension** to flag potentially harmful tweets

- A **predictive model** is needed to **accurately flag instances of cyberbullying**

OUR SOLUTION

- Utilize **real tweets** that have already been manually flagged by type of cyberbullying
- Process tweets such that a predictive model can understand words the same way human beings can
- Train dataset using **Neural Networks**
  - An algorithm inspired by the biological neural networks that constitute the human brain

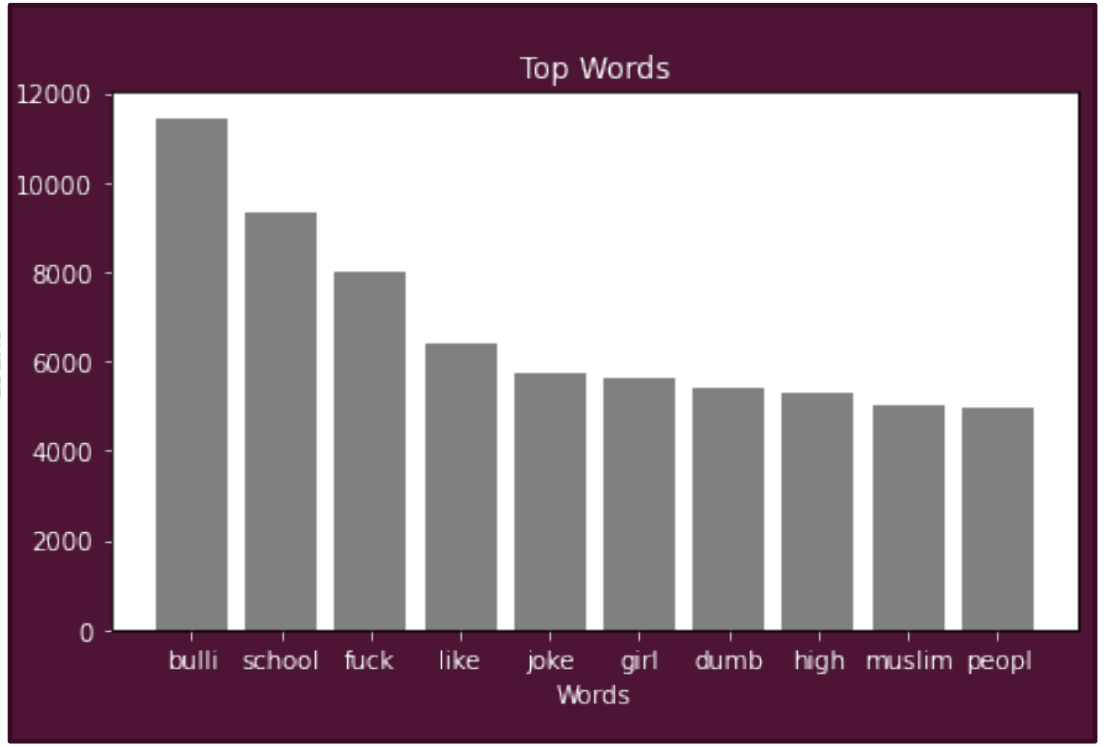
Data source: https://www.kaggle.com/code/anayad/classifying-cyberbullying-tweets/data

DATA EXPLORATION AND PROCESSING

- Data used for prediction consists **of 47K tweets** labeled according to the **class of cyberbullying:**
  - Age
  - Ethnicity
  - Gender
  - Religion
  - Other types of cyberbullying
  - Not cyberbullying

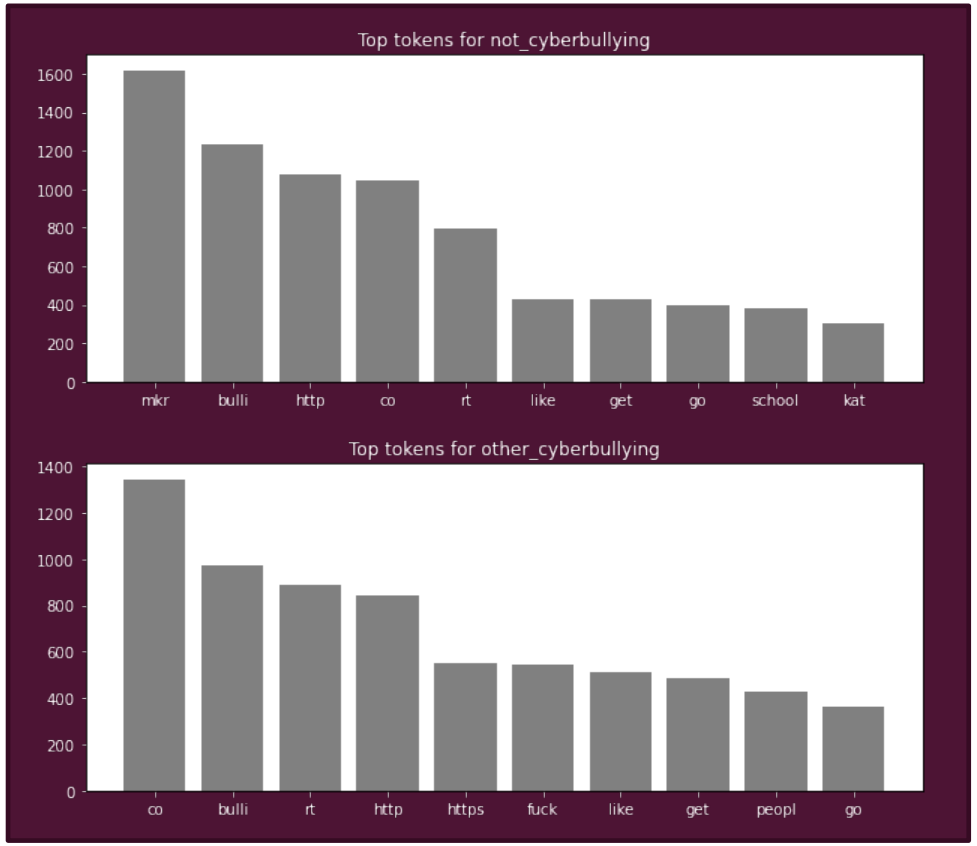
- Standardization and elimination are performed for model efficiency
  - "Bullies" and "Bullied" become "bulli"
  - Words such as "the", "a", "and" etc. will be removed

Frequency Distribution Of Top Words Used in Cyberbullying Tweets



Data source: https://www.kaggle.com/code/anayad/classifying-cyberbullying-tweets/data

DATA EXPLORATION AND PROCESSING

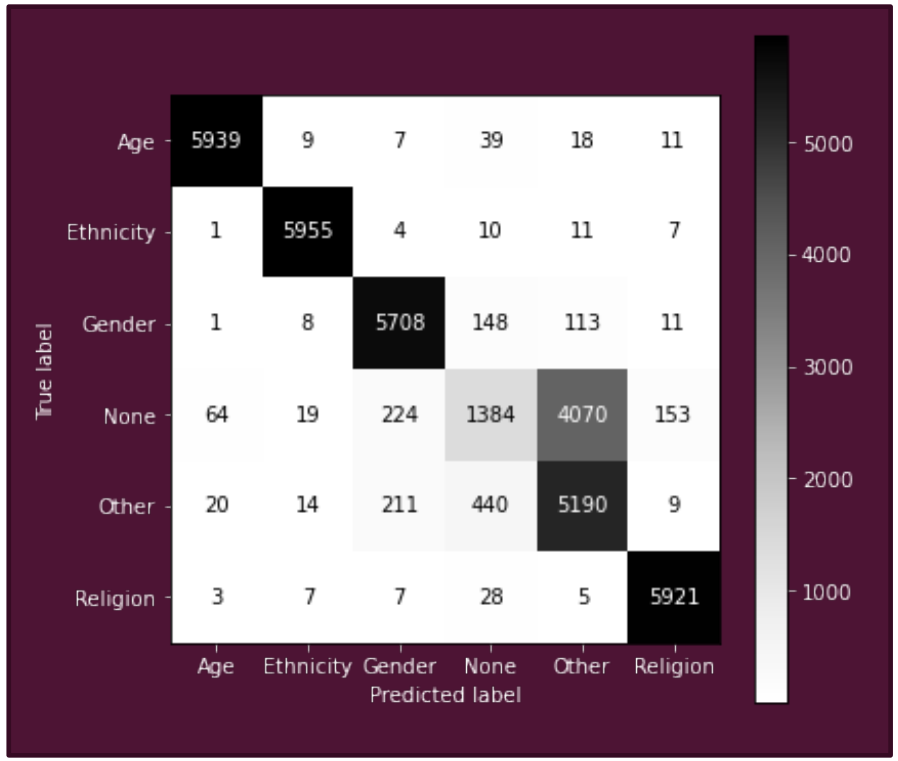


- The most common words for most Cyberbullying categories are **unique**
- However, "Not Cyberbullying" and "Other Cyberbullying" had a **high overlap** in word usage
- Neural Networks are needed so our model can understand the **order of words** used

Data source: https://www.kaggle.com/code/anayad/classifying-cyberbullying-tweets/data

EVALUATION OF FINAL MODEL

Matrix of True and Predicted Values



- Our model **predicted the correct label 82%** of the time
- The model was less efficient classifying "other" types of cyberbullying with "non-bullying"
- This model **correctly predicted non-bullying tweets 32%** of the time

Data source: https://www.kaggle.com/code/anayad/classifying-cyberbullying-tweets/data

RECOMMENDATIONS / NEXT STEPS

Implement this model as a starter for the browser extension and continue to gather data to improve accuracy in models.

Remove "other" classification of cyberbullying as it is too vague.

Include a flag for tweets with misleading or incorrect data articles, as the rise in these types of tweets can also negatively affect users.

Sabina Bains

Email: Sabinabains3@gmail.com

LinkedIn:

https://www.linkedin.com/in/sabina-bains-a58645a6/

THANK YOU