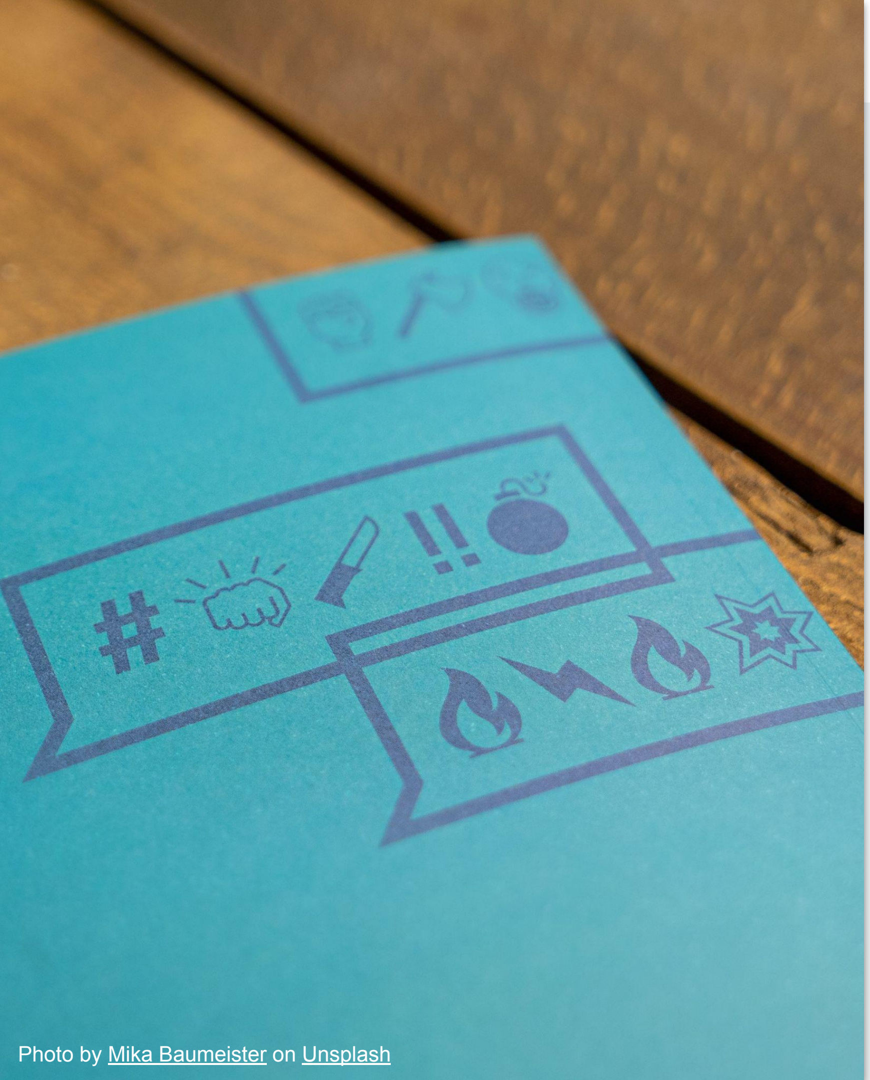


Cyber Bullying Detection

Word Warriors:
Christian Normand, Julia Rowe,
Nati Marcus, Rebecca Patterson

Outline

- Problem statement
- Background
- Data
- Methodology
- Modeling
- Results
- Conclusions
- Future research
- Streamlit app

Problem Statement

Use machine learning to predict the likelihood that an online comment contains a personal attack, toxic language, or aggressive tone.

Background

- 41% of Americans have experienced online harassment [1].
- People who are Muslim, Hispanic or African-American, and those who identify as LGBTQ+ are more likely to be harassed because of their identity [2].
- Online harassment is a threat to healthy discourse, because people who experience harassment are less likely to participate in online discussions [3].

Cyberbullying is bullying that takes place over digital devices like cell phones, computers, and tablets [4]

[1] Pew Research: State of Online Harassment
[2] Anti-Defamation League: Hate and Harassment Online
[3] Google Jigsaw Report on Toxicity
[4] StopBullying.gov

Data

- Data for this project comes from Wikipedia Detox - a project of the Wikimedia Foundation.
- Three datasets consisting of comments from 2001-2015
  - Aggregation
  - Attacks
  - Toxicity
- Each comment annotated by ~10 annotators
- Roughly 300,000 annotated comments in total

WIKIMEDIA
FOUNDATION

By Logo and trademark of the Wikimedia Foundation

Methodology

Definitions: the following criteria was given to reviewers and used to define the type of comment:

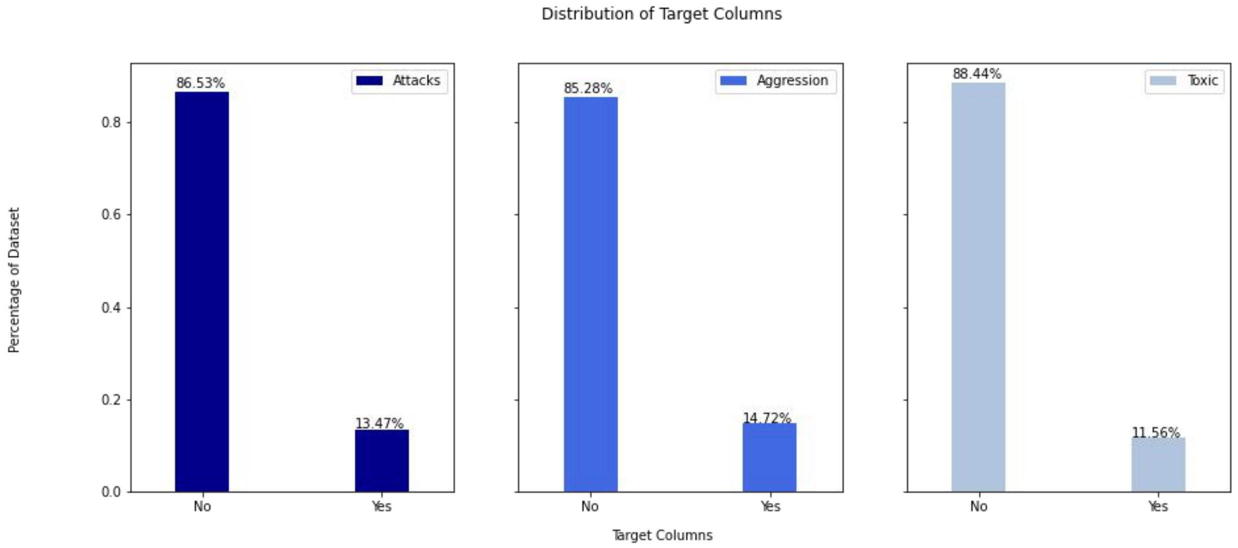
- Attack:
  - Does the comment contain a personal attack or harassment? (Targeted at the recipient of the message, a third party or an attack that is reported or quoted)
- Aggression:
  - How friendly or aggressive is this comment?
- Toxic:
  - Rate the comment from: very toxic (very hateful, aggressive or disrespectful comment that is likely to make you leave a discussion) to a very healthy contribution (likely to make you stay in the discussion)

Scoring: each comment was reviewed by 10-20 reviewers who evaluated it as 1 or 0 for the above criteria

- We averaged the review across reviewers and then evaluated the score to 1 or 0 based on the percentage of reviewers scoring it negatively (an average of 0.5 or higher was set to 1)

Methodology

Train/test split: our data had unbalanced classes so we created a custom train/test split method that would allow us to train our models on a 50/50 split of negative to positive comments



Methodology

Modeling methodology: we evaluated all three topics (attacks, aggression, toxicity) on four different models to assess which type of model would score best with our data sets:

- XGBoost
- Naive Bayes
- Logistic Regression
- SVC

Modeling - Accuracy scores

Model	Topic	Train Score	Test Score
XGBoost	Attacks	**0.906**	**0.906**
Naive Bayes	Attacks	0.892	0.891
Logistic Regression	Attacks	0.960	0.885
SVC	Attacks	0.787	0.716
XGBoost	Aggression	**0.898**	**0.894**
Naive Bayes	Aggression	0.891	0.869
Logistic Regression	Aggression	0.959	0.871
SVC	Aggression	0.778	0.720
XGBoost	Toxicity	**0.916**	**0.917**
Naive Bayes	Toxicity	0.905	0.886
Logistic Regression	Toxicity	0.966	0.897
SVC	Toxicity	0.829	0.791

Modeling - Recall Scores

XGBoost, Naive Bayes, and Logistic Regression had better scores than SVC models.

topic	model	recall_score
aggression	XGBoost	0.768278
aggression	Naive Bayes	0.763561
aggression	Logistic Regression	0.829403
aggression	SVC	0.619104
toxicity	XGBoost	0.796047
toxicity	Naive Bayes	0.836119
toxicity	Logistic Regression	0.861276
toxicity	SVC	0.777358
attack	XGBoost	0.785344
attack	Naive Bayes	0.778859
attack	Logistic Regression	0.838954
attack	SVC	0.639429

Results

Best model: XGBoost

Aggression: Baseline score: 85.28% → **89.4%**
Toxicity: Baseline score: 88.4% → **91.7%**
Attack: Baseline score: 86.53% → **90.6%**

Count Vectorizer:

- max_df = 0.95
- max_features = 5,000 or 6,000
- min_df = 2 or 3
- n_gram range = (1,1)
- stop_words = english
- strip_accents = ascii
- token_pattern = \\w+|[A-Z]\\w+ (maintains uppercase letters)

XGBoost:

- colsample_bytree = 0.6 - 0.75
- n_estimators = 250

Conclusions

- In all of our models, we improved upon the baseline accuracy by around 5%

- This model can be used by any platform to flag and/or warn users if their comment or post may violate the platforms’ code of conduct

Future Research

- More robust text pre-processing
- Narrow definitions of cyber-bullying (more specific)
- Data collection and modeling that incorporates proper context
    - Accounts for tone/sarcasm
    - Different dialects of English
    - Common idioms and other similar phrases

[Check out our app!]