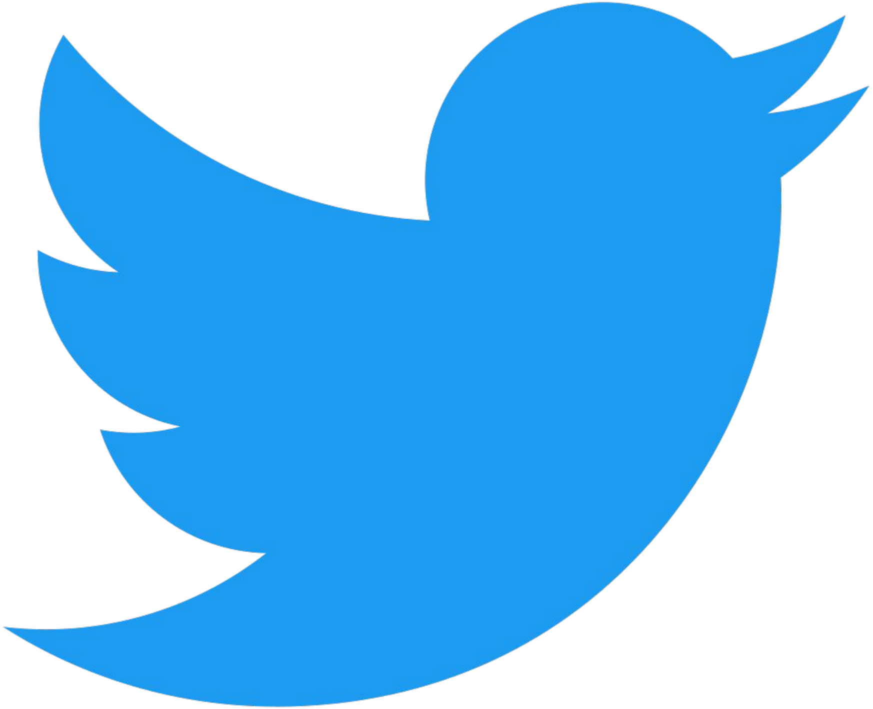
Capstone: Part I

Anna Jobsis

Idea 1: Racist Tweet Detector



Problem Statement

Twitter has a racism problem. Over the last few years, toxic tweets have become an increasing problem. Being able to identify these tweets has become a key problem for data scientists. Twitter datasets could be used to build models that predict hate speech. This would facilitate with helping to identify and take down offensive tweets.

Constraints

- Focus on the english language

Potential Audience

Possible consumers include:

- Twitter
- Southern Proverty Law Center
- ACLU
- Other data scientists

Goals

- Analyze Bag of Words vs TDIDF
- Train multiple models to detect twitter hate speech
- Explore lemmatization and stemmers

Success Metrics

Will use metrics such as:

- Sensitivity
- Specificity
- Accuracy
- Precision

Data Source

- https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech

Idea 2: Predicting Useful Yelp Reviews

Problem Statement

Yelp is a platform that allows users to make reservations, leave reviews, and find businesses. In addition, users can mark which reviews they find useful. Reviews can be both very helpful and very harmful for businesses. Being able to determine what makes a useful review can help businesses create better, more targeted listings.

Potential Audience

- Yelp stakeholders
- Restaurants
- Other data scientists

Goals

- Predict which reviews are useful
- Explore lemmatization and stemmers

Success Metrics

Will use metrics such as:

- Sensitivity
- Specificity
- Accuracy
- Precision

Data Source

https://www.yelp.com/dataset

Idea 3: Predict Who Shot First

Problem Statement

Star Wars fans have long argued over who shot first in the iconic scene between Han Solo and the bounty hunger, Greedo. A survey of Star Wars fans was undertaken by fivethirtyeight that asked, among other things, who shot first. For context, the scene was originally filmed showing that Han shot first, while a subsequent scene update makes the question more ambiguous. Analyzing which features are useful predictors of the “who shot first” will help with marketing the films to specific deomgraphics.

Potential Audience

- Star Wars fans
- George Lucas

Goals

- Determine which characteristics are the best predictors of the “who shot first” question

Success Metric

Will use metrics such as:

- RMSE
- R2

Data Source

- https://data.world/fivethirtyeight/star-wars-survey/workspace/file?filename=StarWars.csv