Reddit Language Analysis: Categorizing Users by Generation

Jack Frantz

Problem Statement

- Working for advertising agency
- Desire to better target their ads by age
  - Don’t always have correct age data
- Want a way to predict generation based on language used

Data Collection

- Subreddits
  - r/GenZ
  - r/Millennials

- Collected title and content of posts

Initial Analysis

- ~25% of posts lacked data for “selftext” (body of post)
- Generated new feature, “All Text”
  - Combined title and body of post
- Even split in data
  - Baseline accuracy: 50.2%

Models

- Consistency
  - All models used similar parameters
  - Tested multiple formatting techniques

- Preferences
  - Would rather classify as Gen Z than miss out on an opportunity
  - Scoring highly is more important than overfitting

Top Performing Model

- Count Vectorizer Linear Regression
  - Ignored common english words
  - Ignored infrequent data
  - More false positives than false negatives
- Performance
  - Testing Score: 0.7979
    - Highest of all models
  - Accuracy: 75.22%
    - Outperforming baseline of ~50%

Recommendations

- Run text through this model to determine generation
  - Serve ads to the generation you wish to target
- Collect more data to make the model more accurate
  - Model with CountVectorizer and Linear Regressions
- Use model to tailor advertising language

Future Applications

- Age restricted content
  - Modify model to recognize if a user may be too young

- Generating advertisements
  - Data from similar sources
  - Use text generation

Questions?

Thanks for listening!