Subreddit Classification

r/SeriousConversation vs. r/CasualConversation

The Problem

- How well can posts about serious conversation topics be identified from posts about casual conversation topics?
- If they can be successfully identified…
  - What are the main unique identifiers?
  - What are the most correlated words to serious and casual conversations?

Approach

- Data Collection
- Data Cleaning
- Exploratory Data Analysis
  - Posts: Title & Body
  - Comments
  - Sentiment
- Model
  - Testing
  - Selection
  - Improvement
  - Evaluation

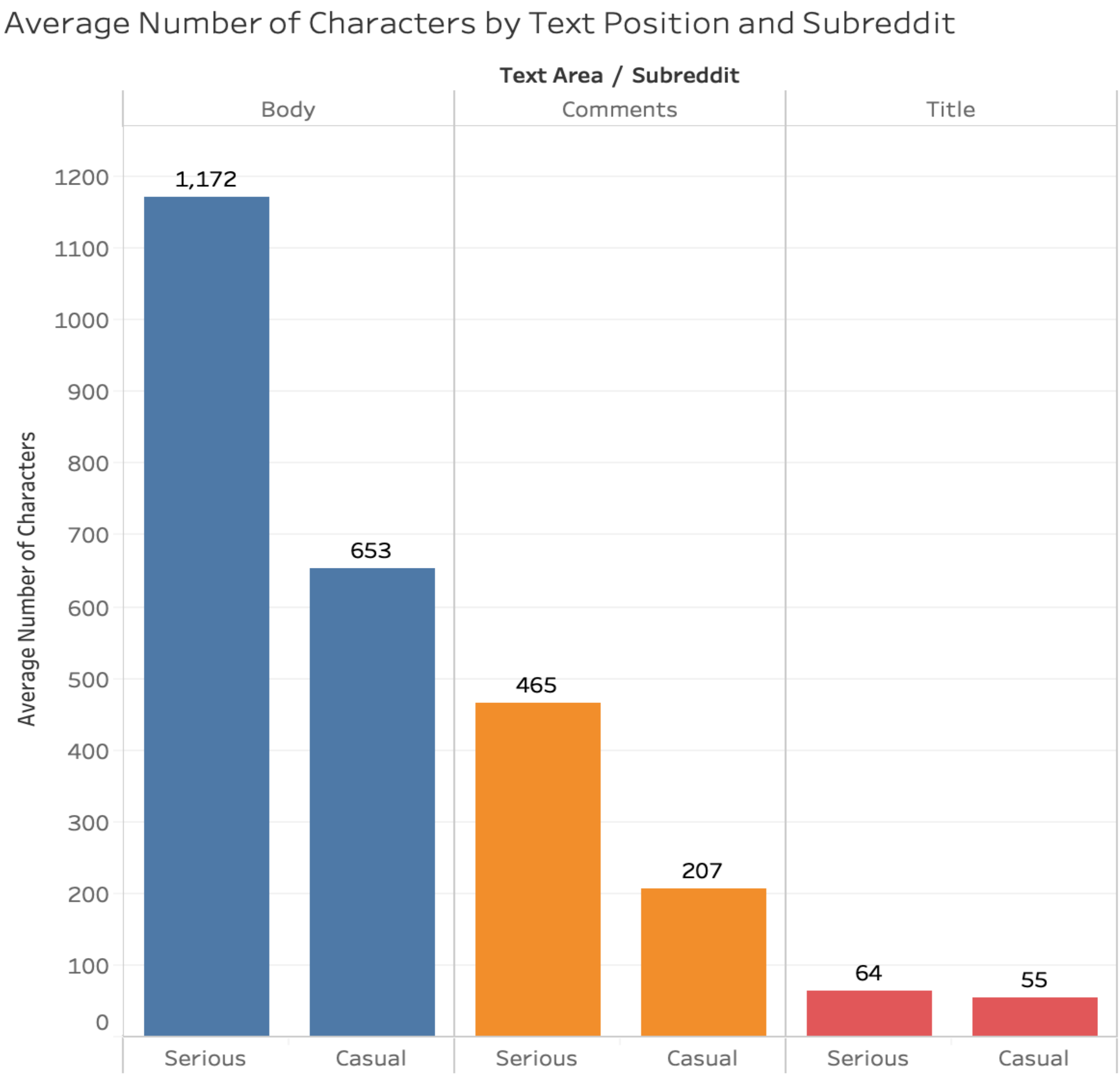
The Data

- Data Collection: Reddit API
- Total Post Observations Collected: 1,694
  - r/CasualConversation: 797
  - r/SeriousConversation: 897
  - Information: author, subreddit, title, body, comments, etc.
- Total Comment Observations Collected: 17,476
  - 1st and 2nd level comments
  - Maximum of 30 comments per post
  - Average of 10 comments per post
  - 39/1694 posts without comments

By Definition

- casual: relaxed and unconcerned
- serious: demanding careful consideration or application

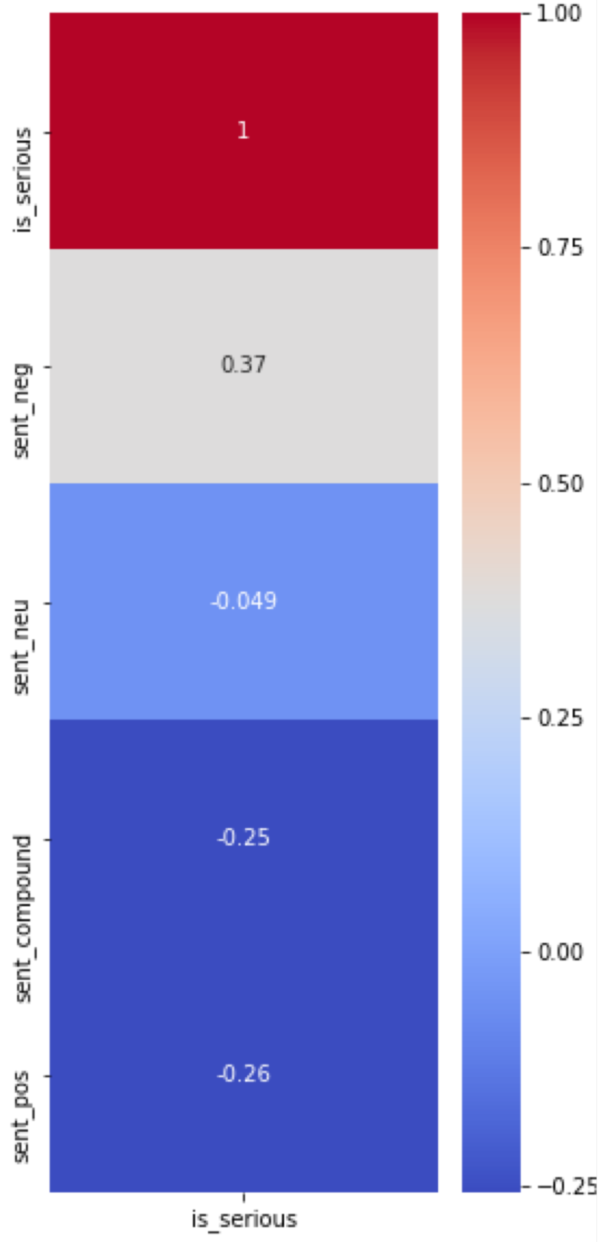
Analysis: Length of Text



All Text Character Count Correlation with Serious: 0.40

Analysis: Sentiment

- Serious: Negative Sentiment
  - Title: 0.22
  - Body: 0.30
  - Comments: 0.15
  - Combined: 0.37
- Casual: Positive Sentiment
  - Title: -0.08
  - Body: -0.21
  - Comments: -0.13
  - Combined: - 0.26



Highest Correlated Words

	Casual	Serious
Title	Day Today April Fool Favorite	Feel Death People Depressed Suicide
Body	Favorite Excited Song Today Nervous	Think People Know Don’t Feel
Comments	Haha Lol Congrats Song Cat	People Life Don’t Think Way

Casual: April Fool’s Day

- **April fools** really bothers me because it pretty much lasts for two days if you’re in Australia
- My friend and I may have pulled off our best **April Fools** prank yet.
- I've just ordered about £30 of Indian food for a friend as an **April Fools** Joke; what wholesome prank ideas have you done?
- Today I got fired ON **April fools** day, I just hope my boss tells me it was a joke tomorrow.
- Dad just got me with a good **April fools** joke. I think April Fools Day should be changed to Rick Astley Day! Who’s with me?
- Is anybody else super excited for Reddit on **April fool's** day? With things like Place and the button bringing back fond memories, I'm excited to see what happens this year!

Model 1: Logistic Regression

- Baseline Accuracy: 0.5295
- Baseline Logistic Regression Model with Sentiment Features: 0.6724
  - +24% increase from baseline

LOGISTIC REGRESSION

- TF-IDF, Grid Search and SVD

	Predicted Negative (Casual)	Predicted Positive (Serious)
Actually Negative	158	41
Actual Positive	40	185

- Accuracy: 0.8575
  - +62% change from baseline
- Sensitivity: 0.822
- Specificity: 0.794

Model 2: Naïve Bayes

- Baseline Accuracy: 0.5295

NAIVE BAYES

- Count Vectorize and Grid Search

	Predicted Negative (Casual)	Predicted Positive (Serious)
Actually Negative	160	39
Actual Positive	39	186

- Accuracy: 0.8449
  - +59% change from baseline
- Sensitivity: 0.831
- Specificity: 0.804

Conclusion

- The Problem: Can posts about serious conversation topics be identified from posts about casual conversation topics?
  - Yes, with ~85% accuracy
- Identifiers:
  - Text Length
  - Sentiment
  - Word Choice
- Correlated Words:
  - Casual: today, favorite, excited, song, haha, congrats
  - Serious: feel, people, think, death, life, depressed

Application

- Identify serious and casual posts and reviews
  - Determine the gravity of a post
- Flag serious posts that can then be further reviewed to determine if the post mentions potentially harmful behavior

Thank You!