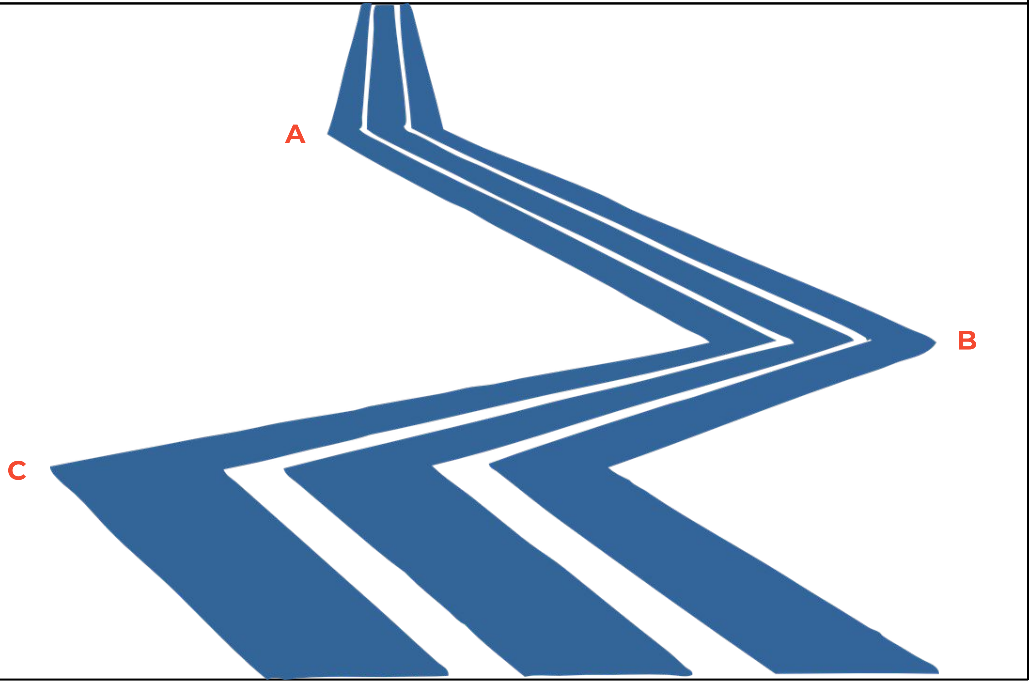
Reddit Scraping:
Testing NLP for Targeting

Elliot Richardson, Junior Applied Data Scientist

targetsmart

Outline

- ◎ Introduction
- ◎ Methodology
- ◎ Initial findings
- ◎ Modeling
- ◎ Conclusion
- ◎ Questions



targetsmart

Introduction

◎ **Elliot Richardson**, Jr Applied Data Scientist

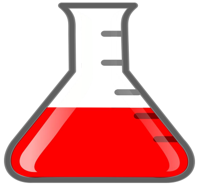
◎ **Problem Statement:**

- Digital outreach increasingly important for voter engagement
- Difficult to connect social media users to voter file & therefore scores
- Identifying receptive targets manually is slow and laborious

◎ **Proposed Solution:** Use Reddit to test NLP as online targeting method

targetsmart

Methodology



◎ Scraped 2500+ posts from r/Socialism and r/Capitalism

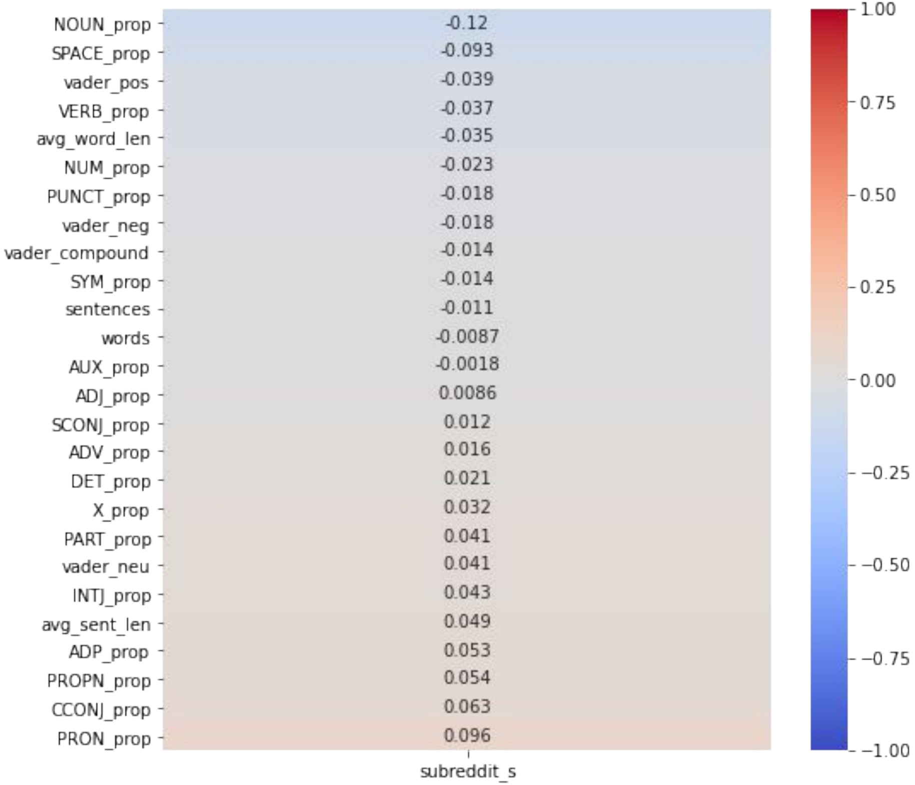
◎ Boiled text down to important words

◎ Calculated some other metrics (i.e. sentence length, etc)

◎ Utilized sentiment analysis to assign positive/negative scores

◎ Testing and assembling various models to find reliable patterns

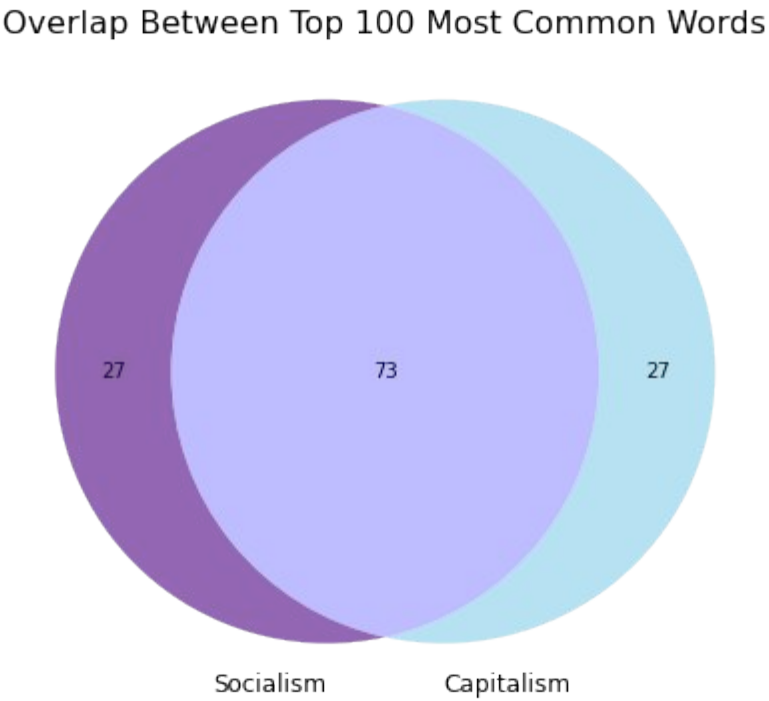
Initial Finding: Subreddits similar on surface



No strong correlations among the various calculated metrics

targetsmart

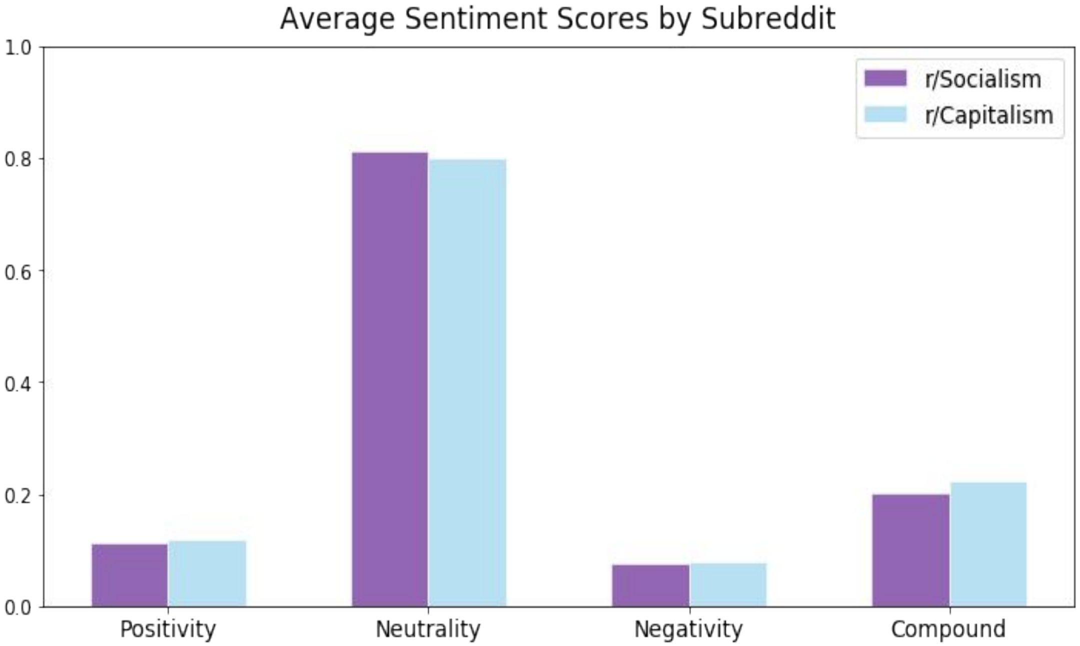
Initial Finding: Subreddits similar on surface



Lots of overlap between most common words

targetsmart

Initial Finding: Subreddits similar on surface

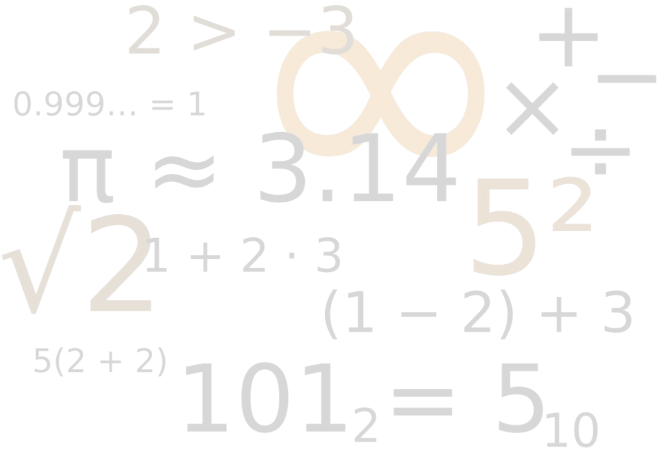


Very little difference in sentiment scores

Modeling

◎ Ensembled models performed best! Used Voting classifier to ensemble:

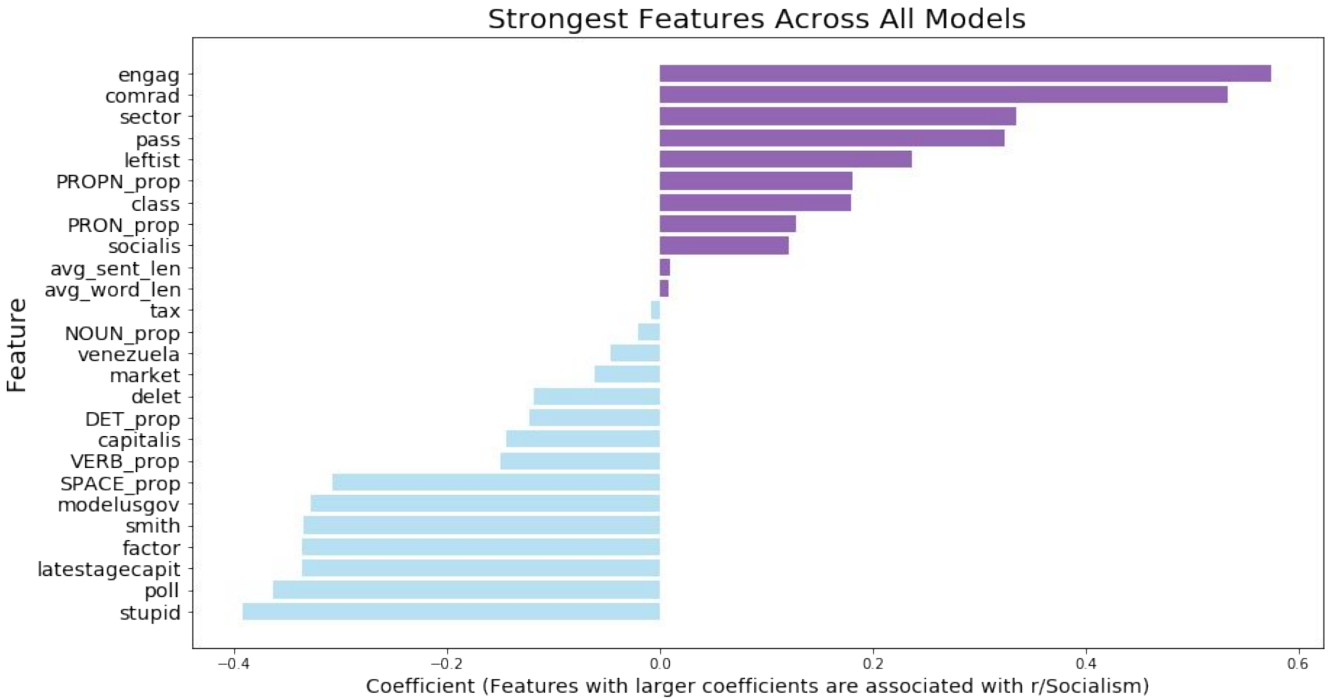
- Logistic regression
- Random forest
- Extra trees
- Ada boost
- Gradient boost



◎ 77.4% accuracy: outperformed baseline by 51.7%

targetsmart

Conclusion: Promising but needs more work!



targetsmart

Thank you!

Any questions?

targetsmart