
Subreddit
Analysis
Corey J Sinnott

TOC

What is Reddit?

What is a Subreddit?

- Communities within Reddit
- Users share information and media related to the primary topic
- More than 2.2mil subreddits
    - ~130k currently active

Problem Statement

- How can we classify a post as belonging to one of two subreddits?

r/askaconservative

r/AskALiberal

Steps to solve

1. Obtaining the data.
2. Cleaning, organizing, and featurizing our data sets.
3. Building a model to analyze our data.
4. Visualizing and reporting the findings.

Project objective 1

Data Collection and Preparation

Data Collection

01

- Utilized Pushshift API
- Developed a function to workaround API's limits
    - Utilized sleep timers to overcome rate limiting
    - Pulled all posts, from each subreddit, in one iteration
- Pulled 8000 total posts
    - 4000 from each subreddit

pushshift.io

Data Preparation

02

- Retained the post body, post title, and subreddit for each
- Duplicates and empty posts removed
    - ~10% removed

pushshift.io

Project objective 2

Featurization

01

Featurization

- Created metrics for:
  - Post length
  - Post word count
  - Title length
  - Tite word count

02

Featurization

- Created metrics for:
  - Grammar errors
  - Grammar error %
- Utilized:

language-tool-python 2.5.1

- A wrapper for:

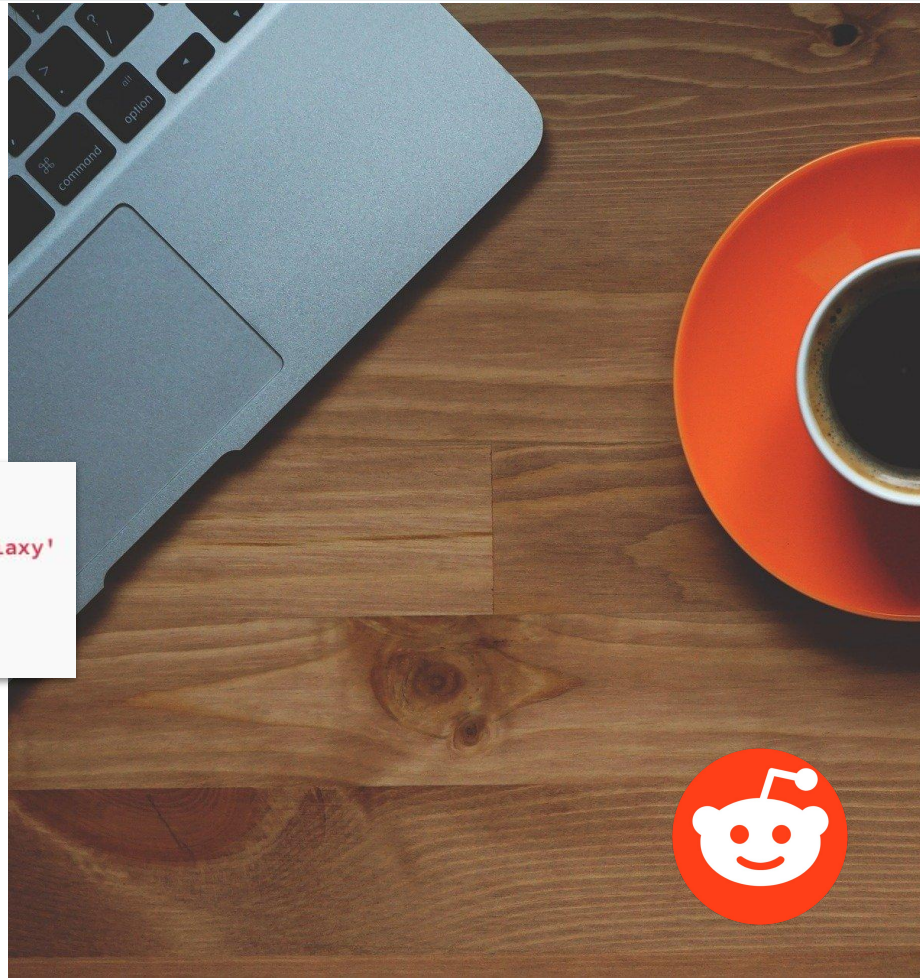
LanguageTool
Your writing assistant

02

Featurization

language-tool-python 2.5.1

```
>>> import language_tool_python
>>> tool = language_tool_python.LanguageTool('en-US')
>>> text = 'A sentence with a error in the Hitchhiker's Guide tot he Galaxy'
>>> matches = tool.check(text)
>>> len(matches)
2
```



03 Featurization

- Created metrics for:
  - Polarity
    - -1 to 1 range
      - -1 linguistically negative speech
      - 1 linguistically positive
  - Subjectivity
    - 0 to 1 range
      - 0: Objective
      - 1: Subjective
- Utilized Beautiful Soup Library

Project objective 3

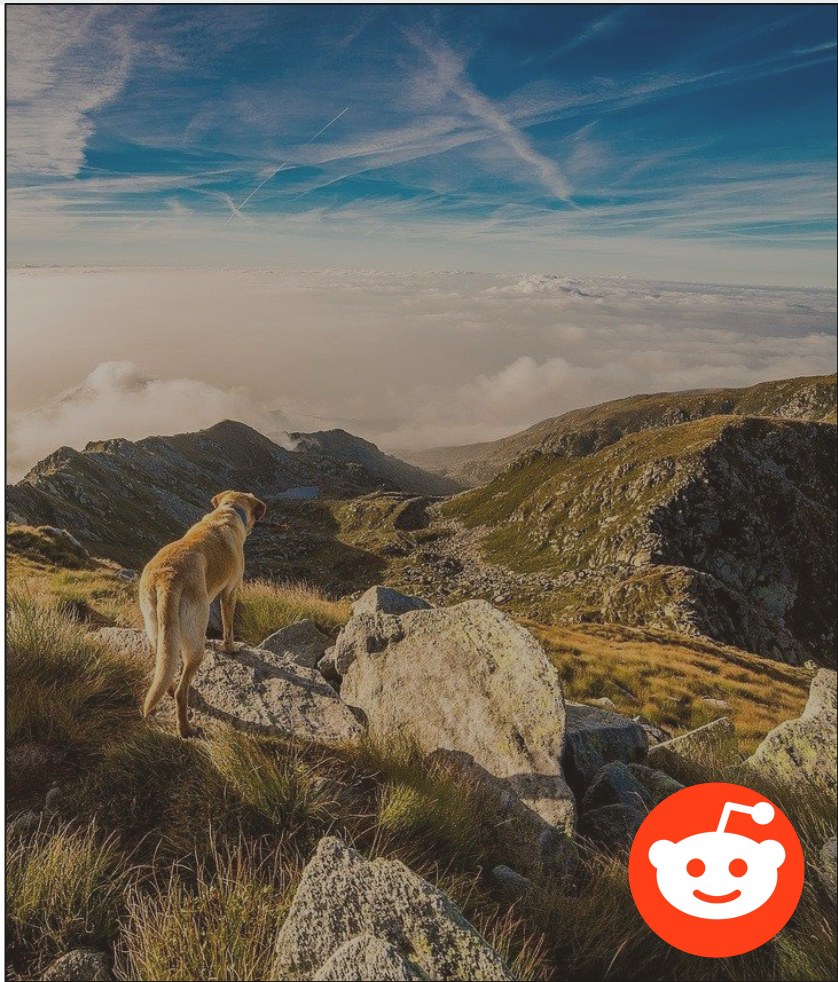
Exploring the Data

01

Exploring the Data

Polarity scores

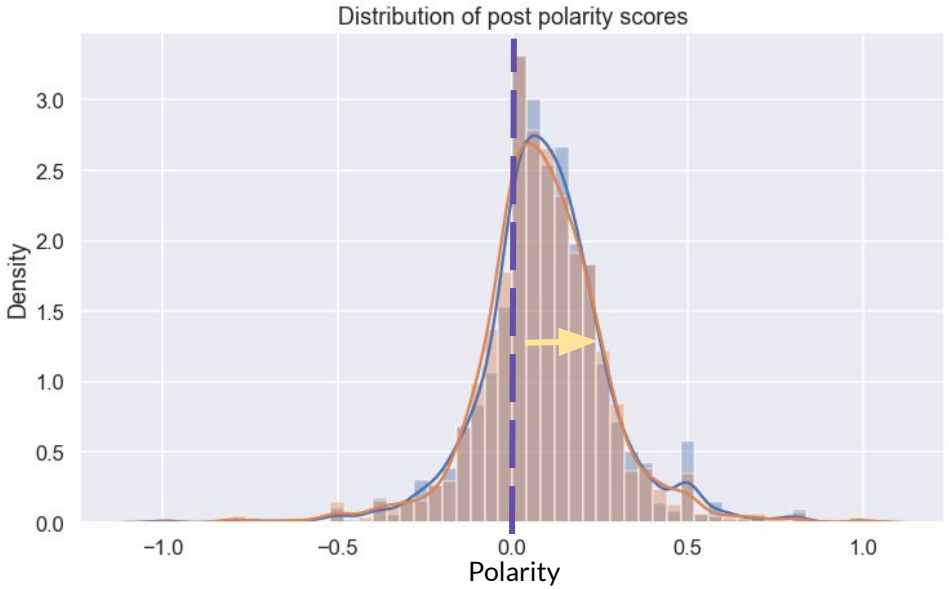
- r/askaLiberal
  - **0.064** average for posts
- r/askaConservative
  - **0.056** average for posts

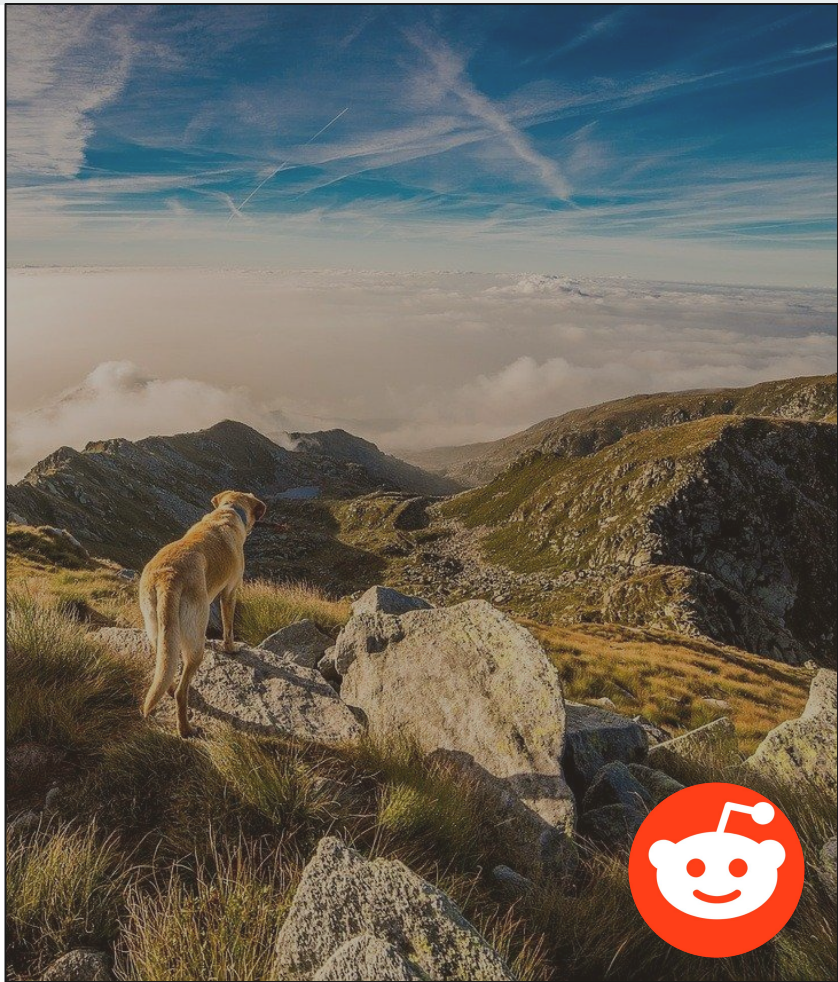


01

Exploring the Data

Polarity scores



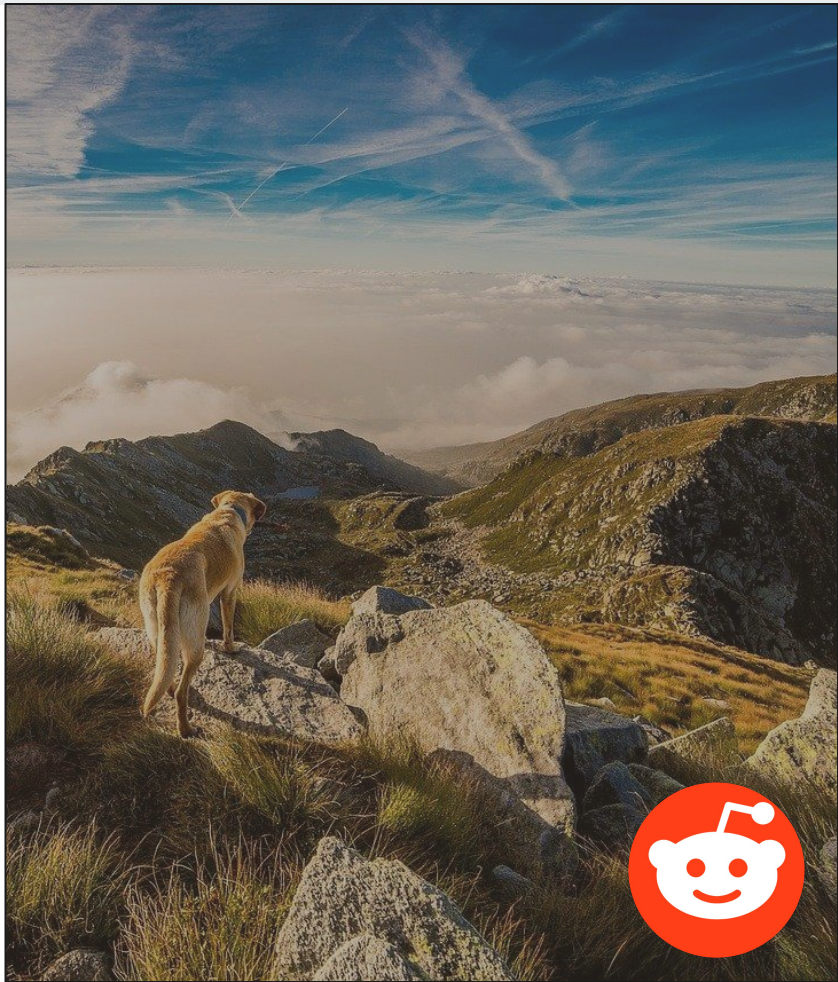


02

Exploring the Data

Subjectivity scores

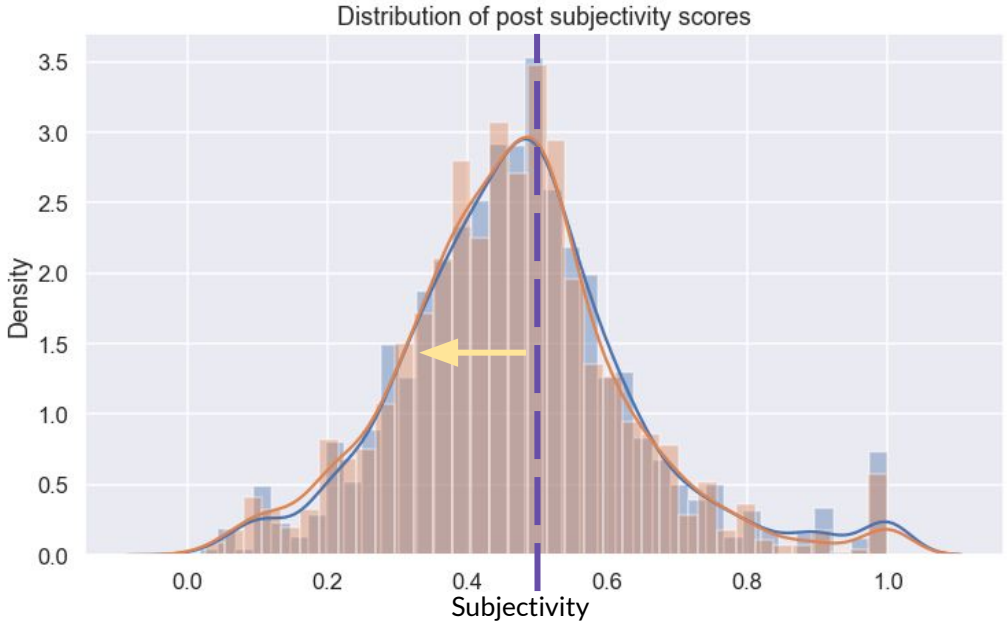
- r/askaLiberal
  - **0.322** average for posts
- r/askaConservativel
  - **0.297** average for posts

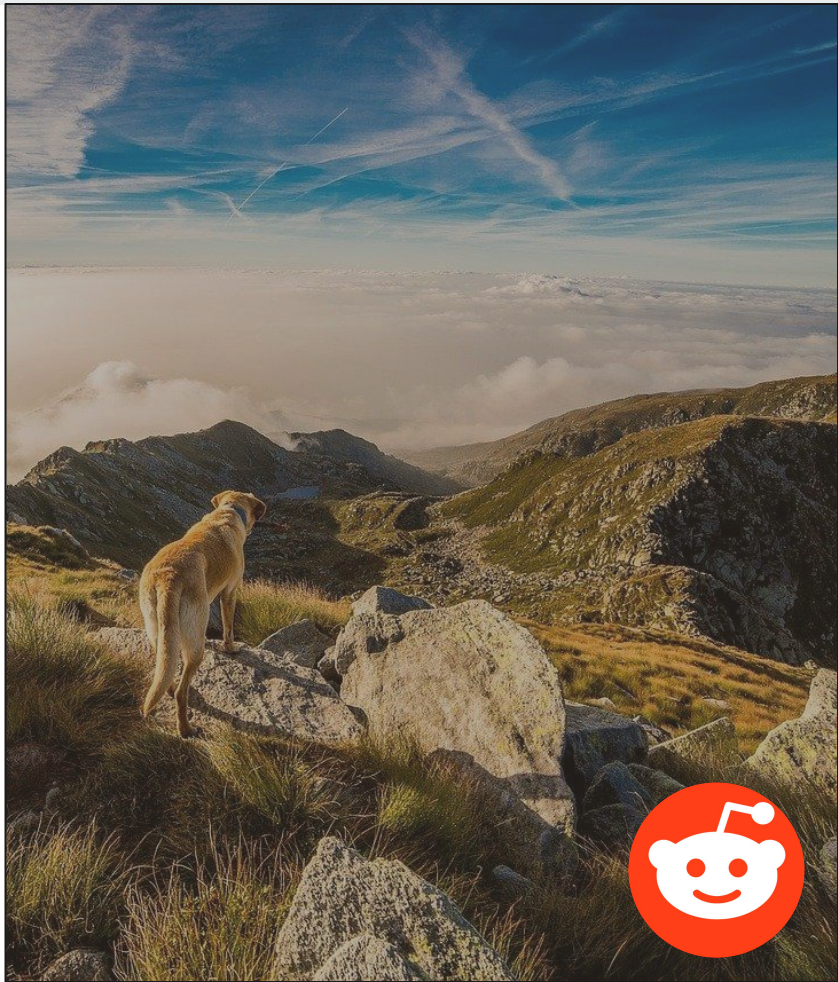


02

Exploring the Data

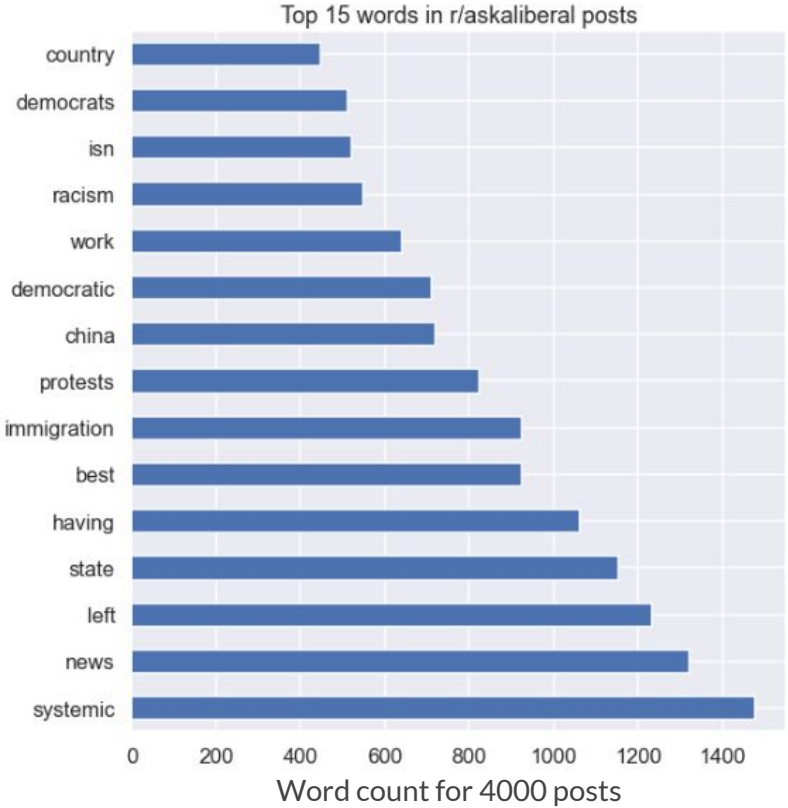
Subjectivity scores





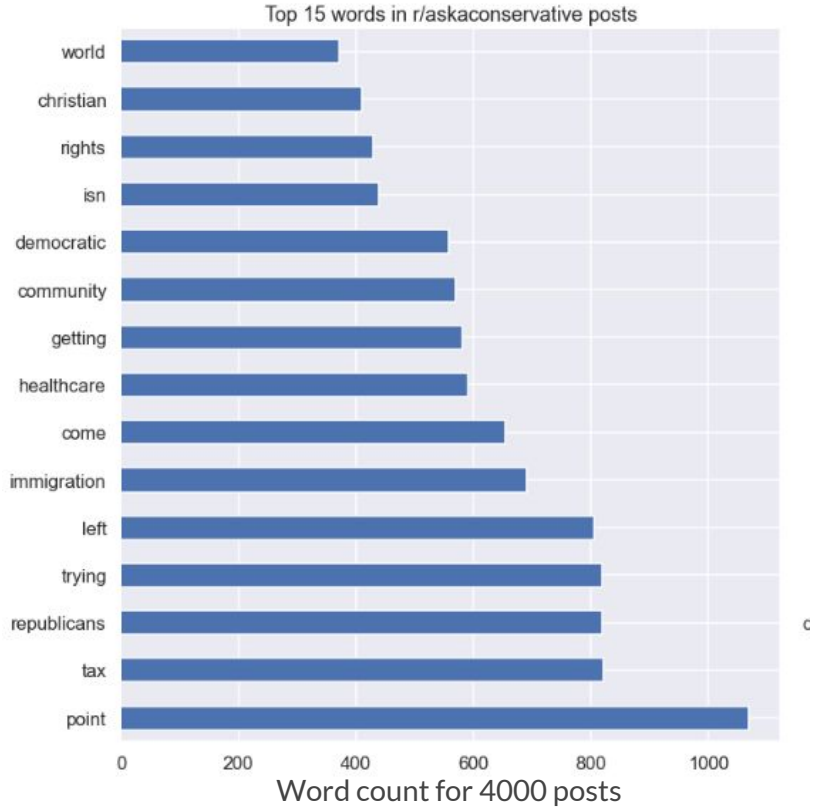
04

Exploring the Data



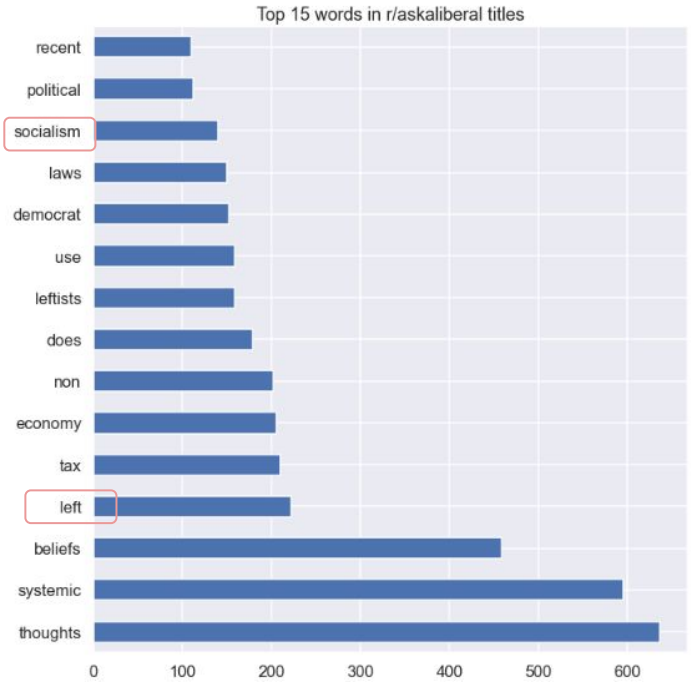
04

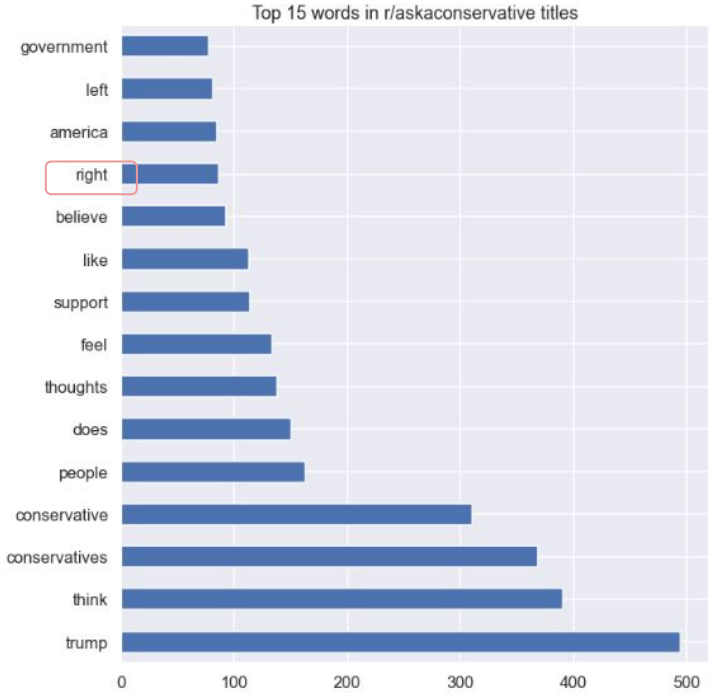
Exploring the Data



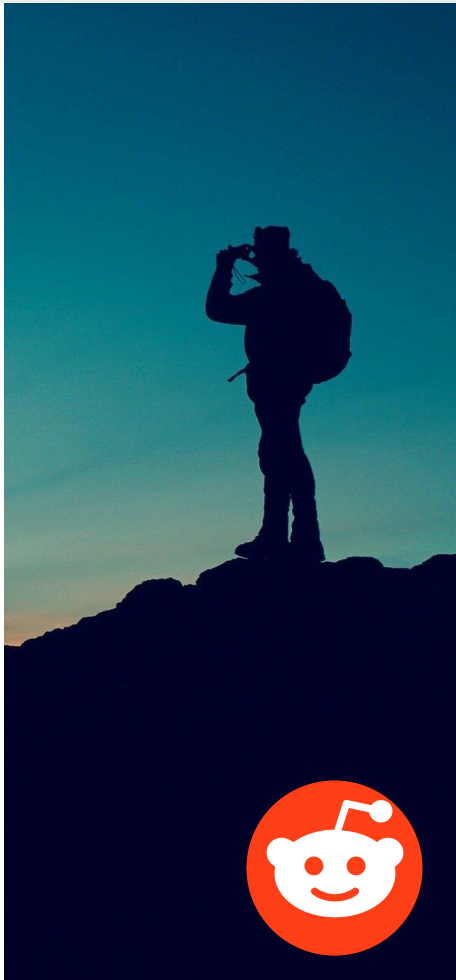
04

Exploring the Data



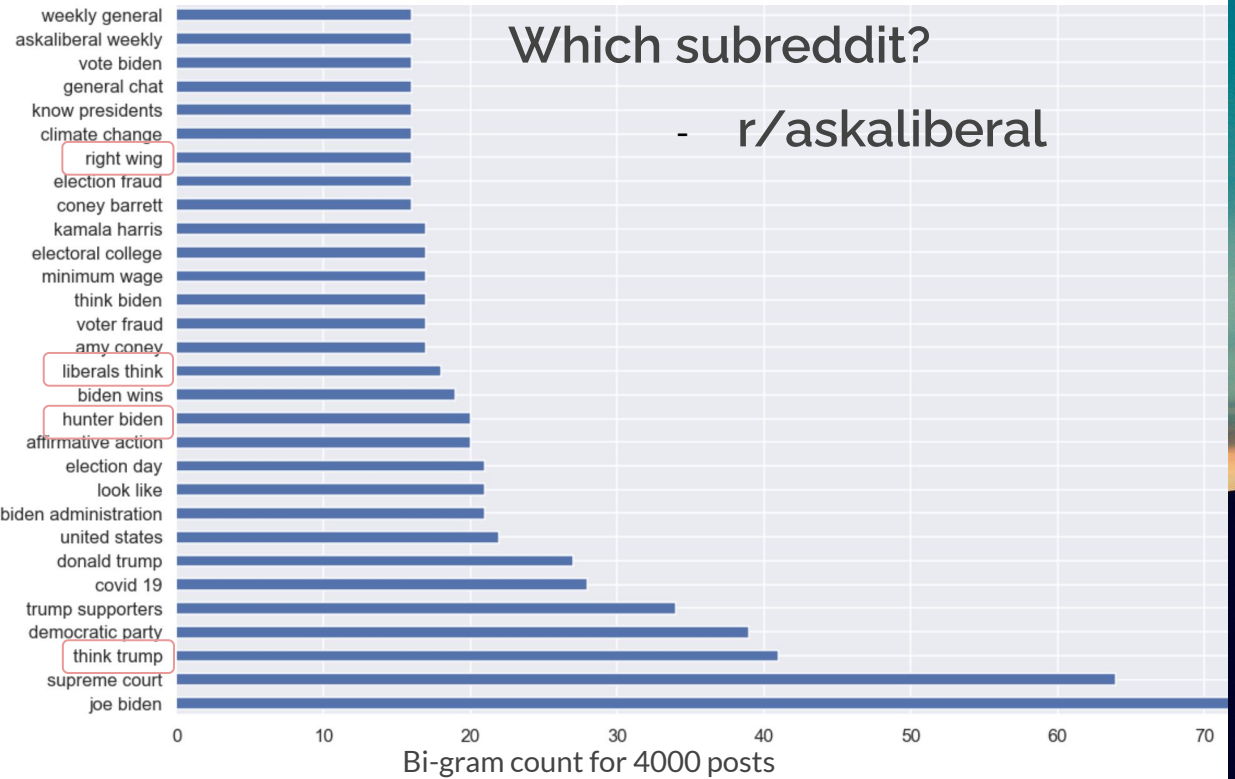


Word counts for 4000 posts



04

Exploring the Data

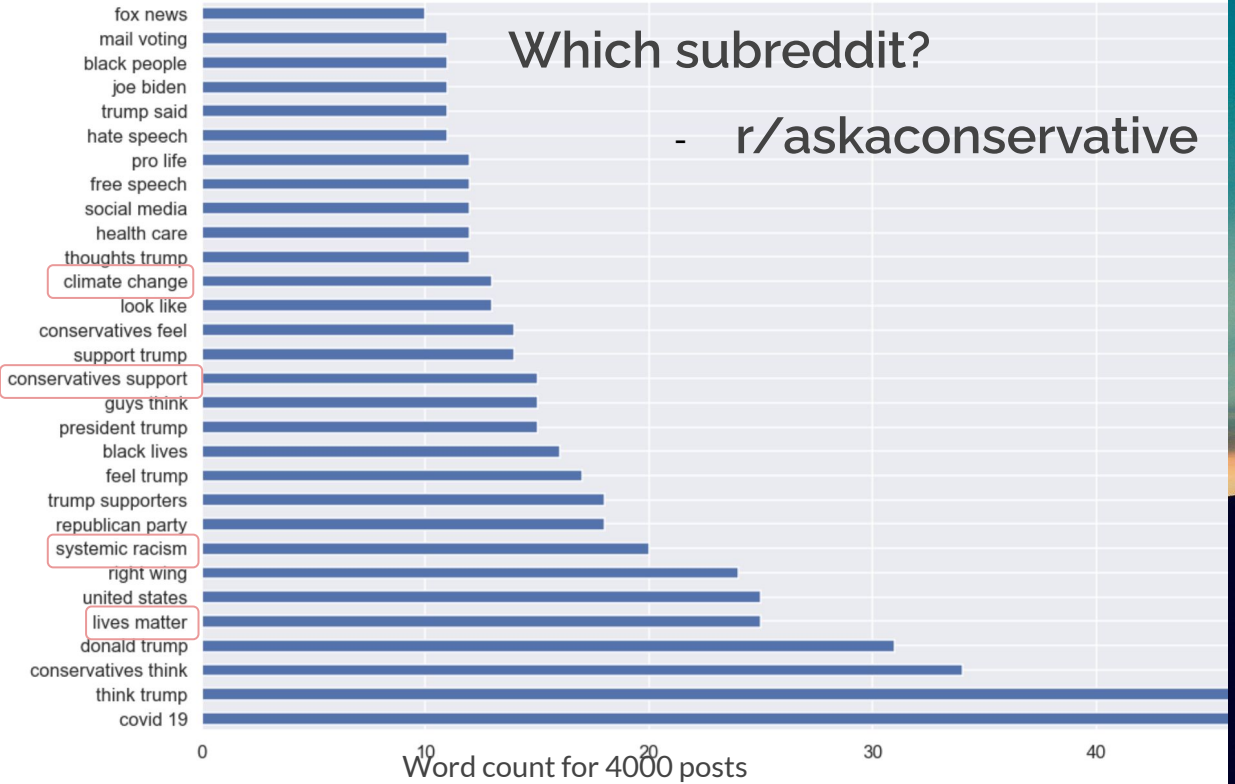


Which subreddit?

- r/askaliberal

04

Exploring the Data



Word count for 4000 posts

Which subreddit?

- r/askaconservative

04

Exploring the Data

When people say 'healthcare is a human right,' what does that mean functionally?

r/AskALiberal

Why don't minimum wage workers deserve at least the bare minimum of a liveable wage?

r/askaconservative

How do you respond to conservatives that say the Democrats spent the past 4 years saying the 2016 election was stolen by Russia but call republicans crazy for saying 2020 was stolen?

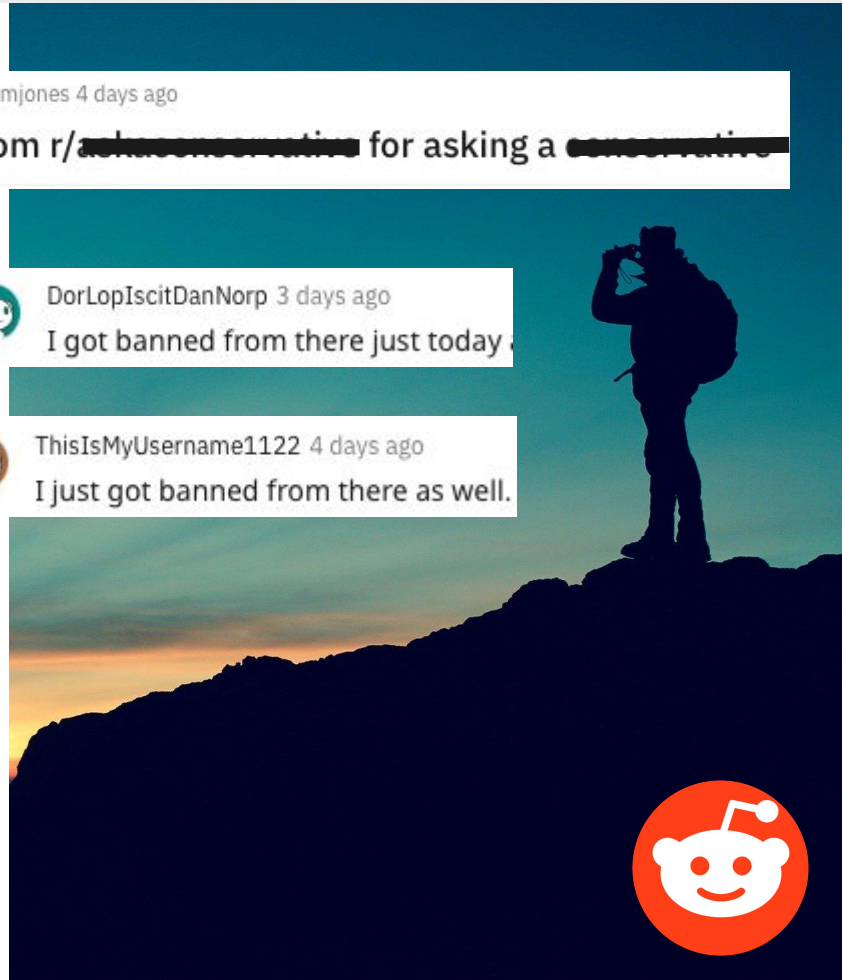
r/AskALiberal

Why is it considered "liberal" or "leftist" if you want to protect the environment?

r/askaconservative

04

Exploring the Data



Differences in moderation lead to posts that look nearly indistinguishable between the two subreddits

- Over the 8000 posts pulled, 10% were “removed by moderator”

- 100% of those posts belonged to r/askaConservative
  - This is a removal rate of 20%
  - Result is both subreddits having “conservative” verbiage
- Not a condemnation of moderation or censorship, but a serious challenge to analysis

r/askaconservative

r/AskALiberal

Project objective 4

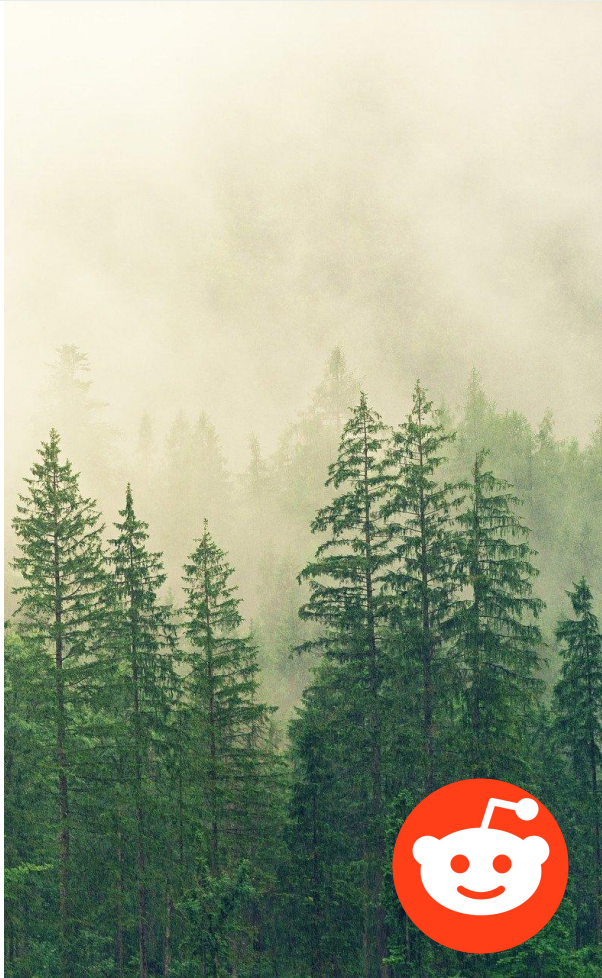
Modeling and Classification

04

Testing Models

Several models were fit, including:

- Naive Bayes
- Logistic Regression
- K-Nearest
- MLP
- AdaBoost
- Bagging
- **Random Forest**
- Neural net

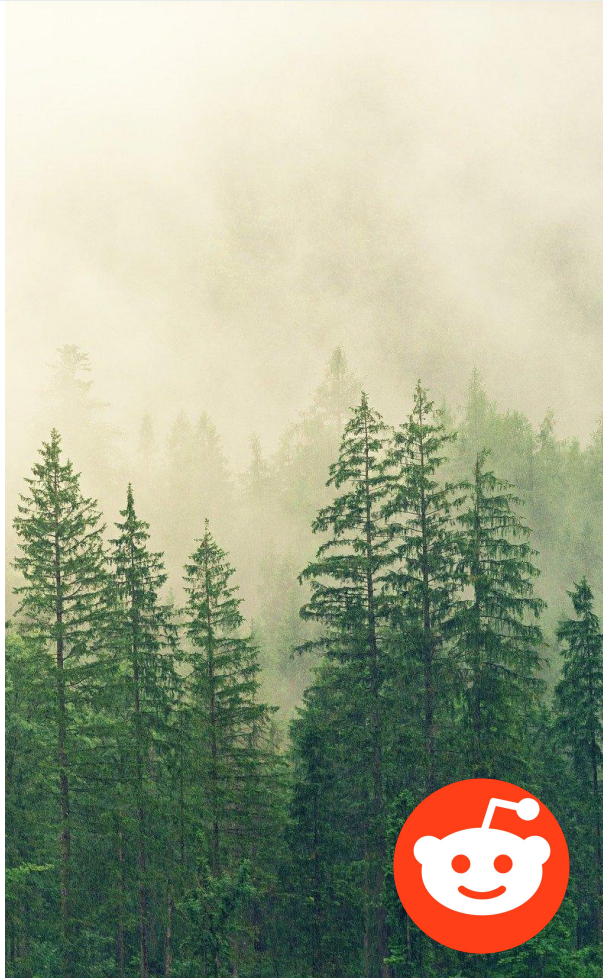


04

Testing Models

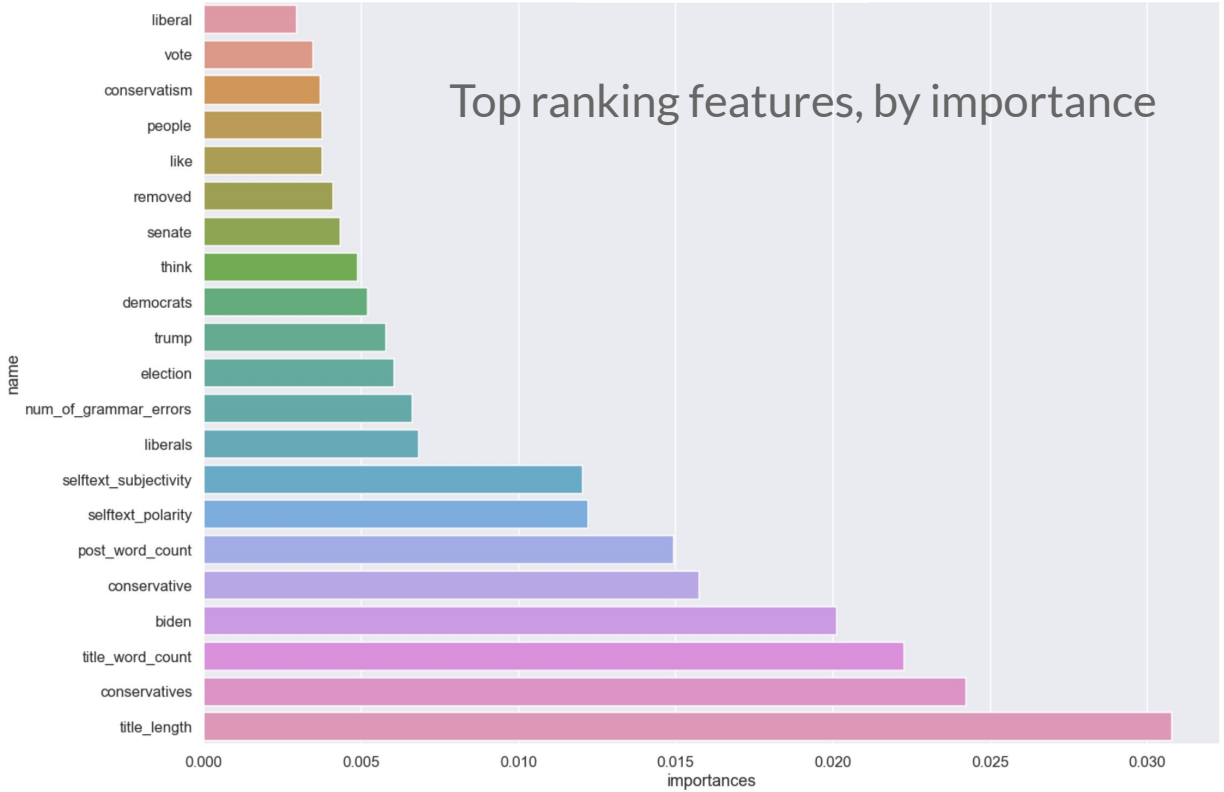
Most models averaged ~65% accuracy

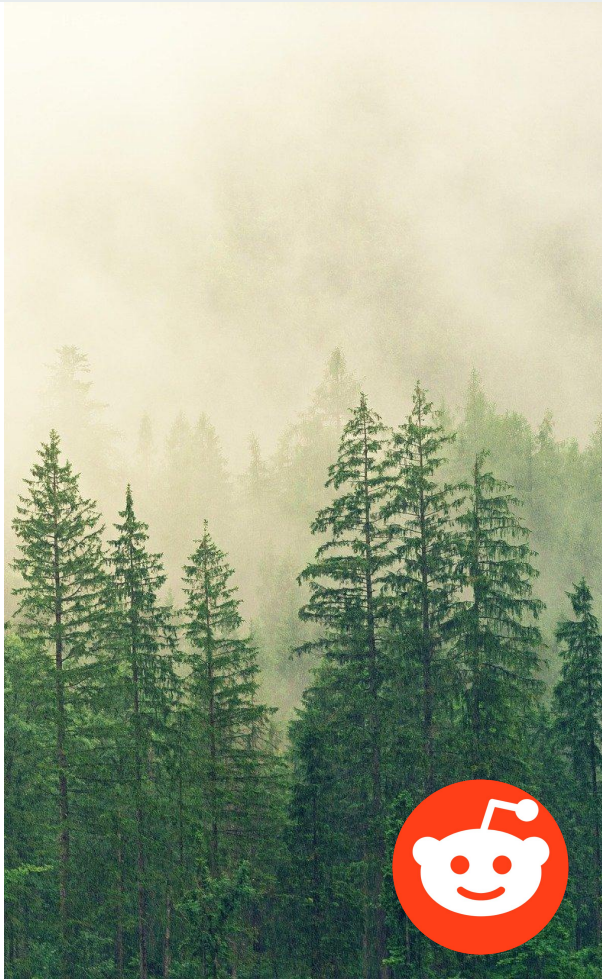
- **Random Forest** performed highest
  - Notable parameters:
    - Minimum samples leaf: 1
    - N_estimators: 500
    - Criterion: entropy
    - CountVectorizer max features: 2500
    - CountVectorizer n_grams: (1, 1)
  - Accuracy: **74%**
  - Precision: **75%**
  - Recall: **62%**
  - F1 Score: **0.68**
  - AUC = **0.82**



04

Testing Models



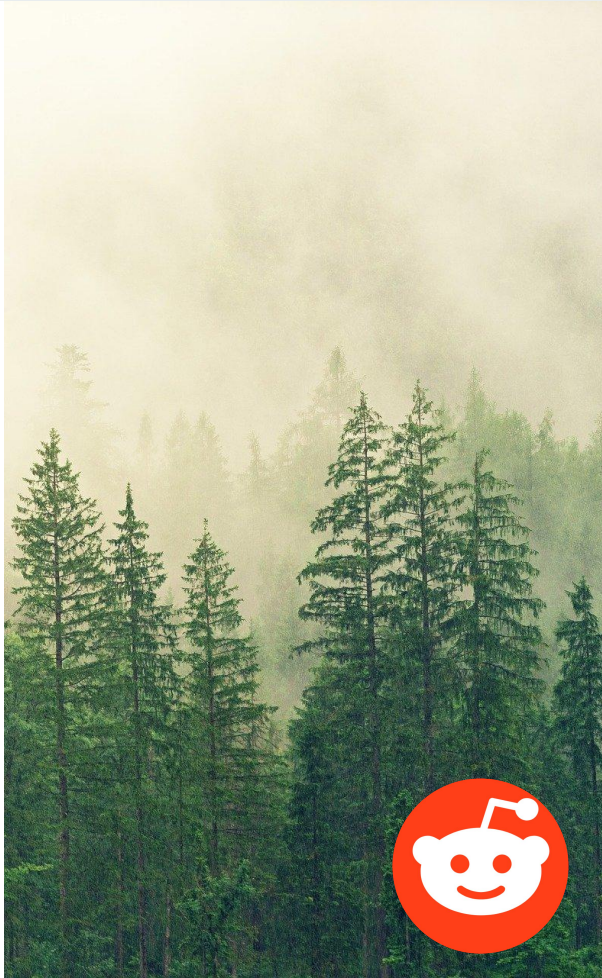


04

Testing Models

Limitations

- As discussed, posts were extremely similar
- Most posts could be interchanged with either subreddit

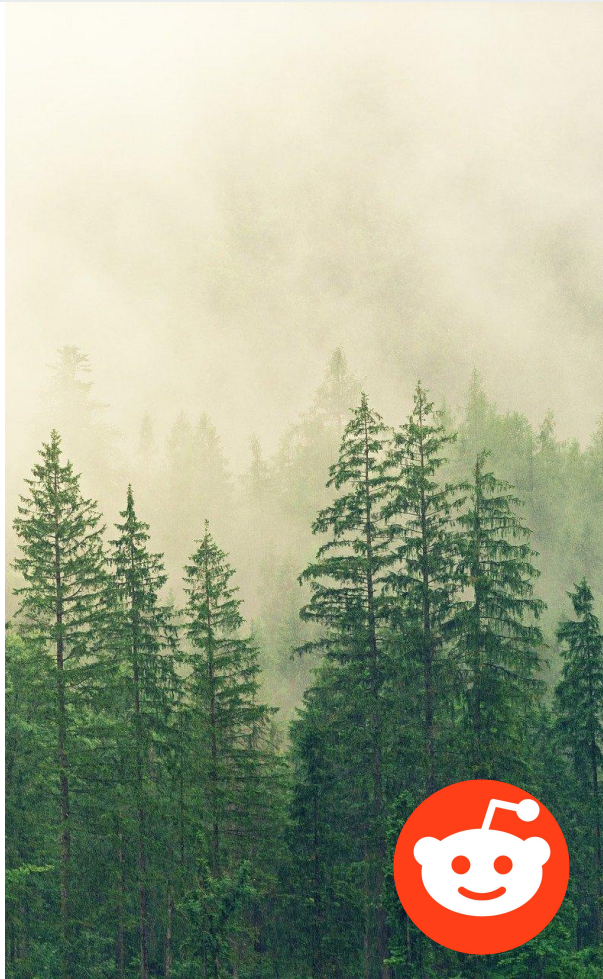


04

Testing Models

Recommendations

- Obtain user data and search for crossposting
- Creating a function that, rather than relying on a timestamp, scrapes posts as they are posted
    - Allowing for the analysis of balanced, unmoderated posts.



Conclusion

- Reddit is one of the largest social networks on internet, with millions of subreddits.
- Currently able to classify whether a post was made to r/askaLiberal or r/askaConservative with 74% accuracy.
- Antagonistic cross-posting and imbalanced post removal is a hurdle that needs further exploration.

Any Questions?

All photos courtesy of pixabay.com