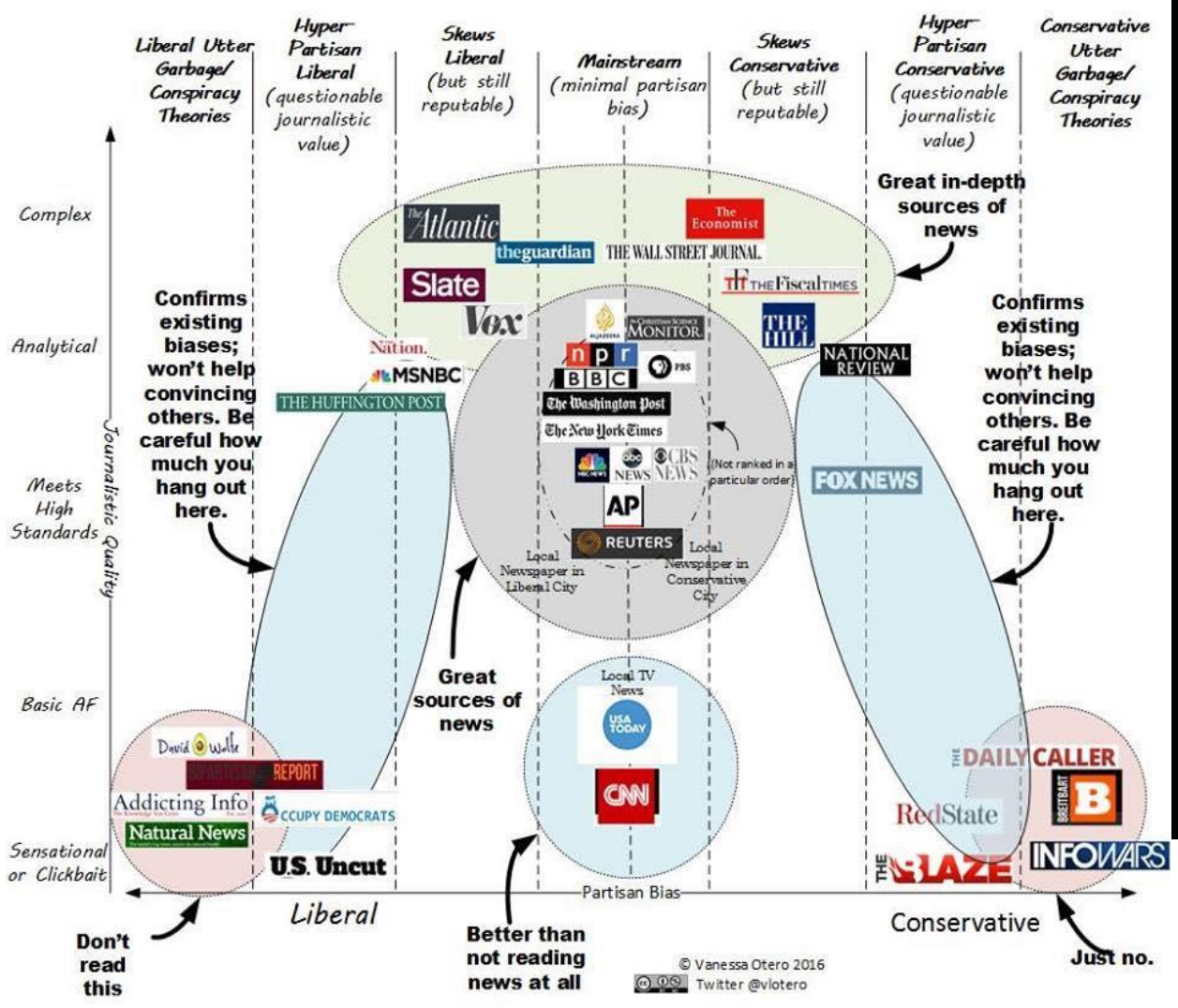
Fake News

Grant Wilson

Problem Statement

- Problem
  - Can we fit a model that accurately differentiate posts from News and Worldnews subreddits?
- Model Types (Classification)
  - Column Transfer: Logistic Regression-TFIDF
  - Bernoulli Naive Bayes: Stemmed
- Success
  - Able to successfully classify posts
  - High Accuracy Score
- Why is this Important?
  - We might rely on subreddits for the news
    - It is important where news comes from
      - Credible sources
      - Political leanings/biases



Source: https://twitter.com/vlotero/status/808696317174288387

Source: https://www.allsides.com/media-bias/media-bias-chart

Importance

News:		World:	
trump	302	u	650
u	296	trump	577
say	222	syria	562
new	195	turkey	438
n	177	say	407
china	163	hong	395
amp	161	kong	395
syria	161	china	356
news	157	turkish	239
police	152	protest	221

1. Accuracy can hint at subreddit differences
    a. Input: Titles
        i. Subreddits with similar/same titles are harder to classify
        ii. Higher accuracy scores signify more easily separated communities
    b. Ex: /r/books and /r/videogames
2. Tool for self-examination
    a. Examining the relationship between news subreddits you follow can point out echo chambers
    b. Emphasis on getting a well-rounded view of the news
3. Counting the frequencies of titles can give insights
    a. What the subreddit focuses on
        i. Ex: Police vs Protesters
    b. Common posts

Data Collection

1. API calls
    a. News subreddit
        i. 19.1 million members
    b. Worldnews subreddit
        i. 22.3 million members
2. 5000 most recent posts in each
    a. Sleep call at the end of loop
3. Predictors
    a. Title
    b. Score
    c. # of Comments
    d. Domain
    e. Gilded

r/worldnews

Data Cleaning and EDA

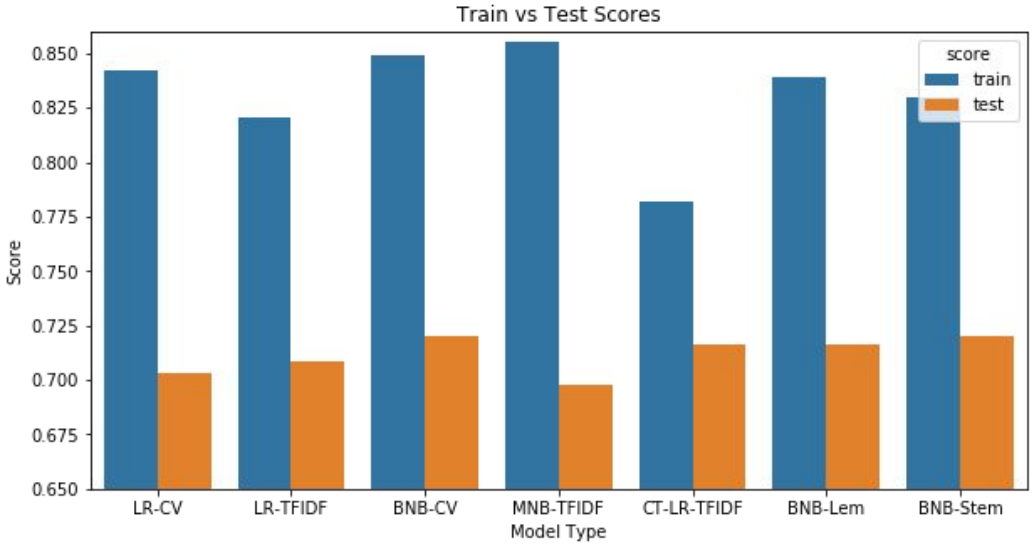
- No missing values within the data
  - API was pretty clean with retrieval
- Primary focus was the text analysis of titles
- Feature examination
  - Most recent posts were not gilded
  - Most scores were fairly low
  - Domains were too unique to be helpful
    - Consistent between subreddits
  - Models were tested with Stemmed, Lemmatized, and full post titles

```
Shared Top 15 domains:

0.6
```

Model Fitness

- Baseline score
  - Positive Class
    - 50.03%
- Model scores
  - Column Transfer: LR TFIDF
    - Train
      - 78.21%
    - Test
      - 71.64%
- Metric
  - Accuracy



Weaknesses

1. Similar Subreddits
2. News constantly shifting
    a. Topics of today may not generalize
    b. Ex: Impeachment, Protests in China, Syria/Turkey
3. To improve generalizability and discriminant ability, need to have a much longer timespan

Conclusions and Recommendations

- Key Takeaways
  - For two subreddits that cover the same topics, we did remarkably well
  - We need more information about each subreddit's political tendencies to better discriminate
  - Post titles are the best discriminant for subreddit classification
- Recommendations
  - Expand model to non-binary classifier tool
    - Create a "biases tracker"
      - Analyze content you receive and consume from your news sources
      - Get recommended news outlets to round out news consumption
  - Get data for longer span of time to train model towards the nuances in political leaning and coverage