Bias in the News

Analysis of semantic bias and sentiment across major American news outlets.

https://github.com/ibury08/news_bias

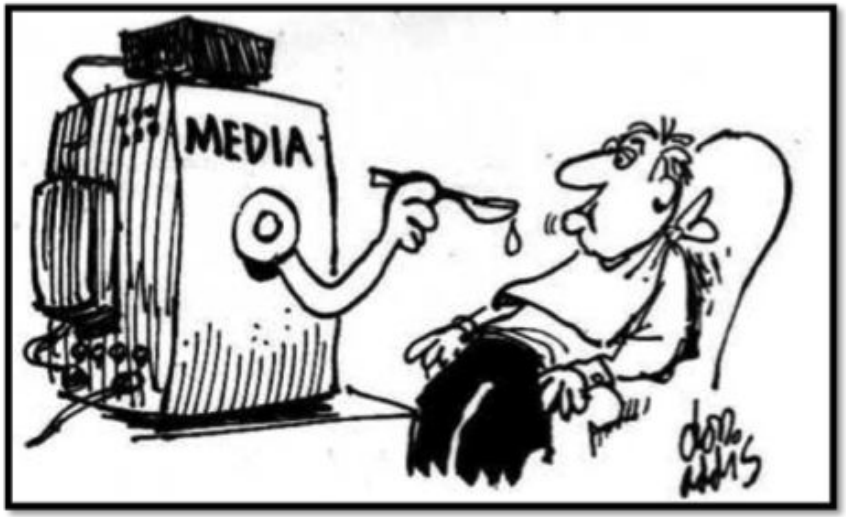
Contents

Introduction

- 45% of Americans source news from social media[1]

- FCC Fairness Doctrine (requiring news broadcaster to present contrasting views on issues) was repealed in 1987

- As a result, American news may be increasingly politically-biased. Content is also more abstracted from its source via social media posts.



[1] https://marketingland.com/pew-research-center-says-45-americans-get-news-facebook-228001

Goals

1. Increase awareness of the nuance in presentation of the same events and issues in major U.S. news sources

2. Provide a framework to evaluate news content in terms of bias and sentiment

3. Contextualize the content of a given article by comparing its bias to similar articles from other major publications and suggest related articles with less bias

Data

- Raw data sourced from Kaggle
- Format matches what could be quickly scraped from an online news website
- ~140,000 articles, from 15 major U.S. news sources
- Articles dating roughly from 2013 to 2017

Data Attributes:

- Basics columns include title, publication, author, date, and article text.
- No target column -> unsupervised learning problem with regard to this data

	id	title	publication	author	date	year	month	url	content
53293	73471	Patriots Day Is Best When It Digs Past the Her...	Atlantic	David Sims	2017-01-11	2017.0	1.0	NaN	Patriots Day, Peter Berg's new thriller that r...
53294	73472	A Break in the Search for the Origin of Comple...	Atlantic	Ed Yong	2017-01-11	2017.0	1.0	NaN	In Norse mythology, humans and our world were ...
53295	73474	Obama's Ingenious Mention of Atticus Finch	Atlantic	Spencer Kornhaber	2017-01-11	2017.0	1.0	NaN	"If our democracy is to work in this increasin...

https://www.kaggle.com/snapcrack/all-the-news

Methodology & Results

1. Data Wrangling
2. Data Analysis
3. Topic Modeling
4. Sentiment / Subjectivity Analysis
5. Application Development

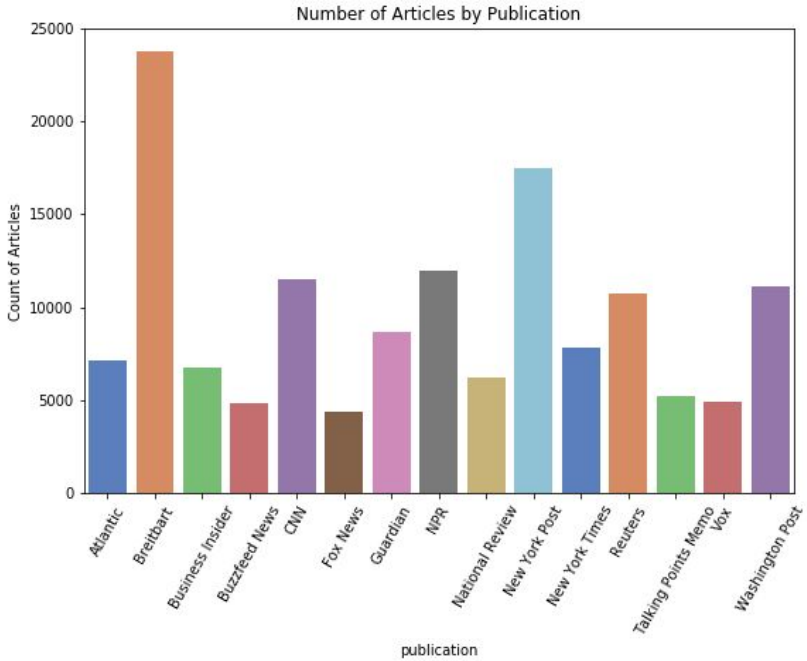
1. Data Wrangling

1. Process raw text by removing punctuation and stopwords, tokenizing the cleaned text, and finally lemmatizing the tokens for better matching against existing corpora. *(nltk)*

2. Convert clean text to a tf-idf vector array, to account for varying lengths of articles. *(sklearn)*

3. (Poor results, not in production) Apply SVD of resulting feature matrix to reduce dimensionality. This feature is was removed due to low **explained variance vs. number of features.** *(sklearn)*

2. Data Analysis - Skewing

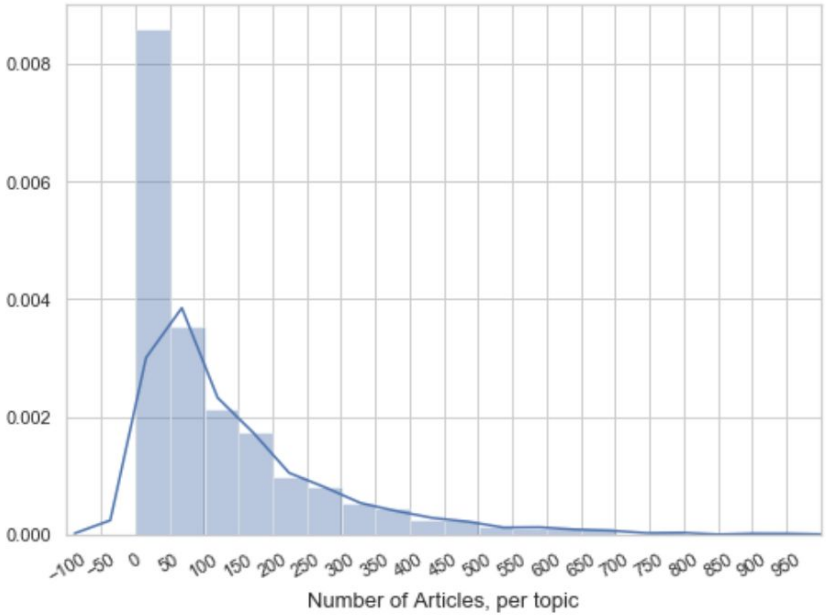


Fairly even distribution of number of articles by each publication.

3. Topic Modeling

1. Use an efficient unsupervised clustering algorithm (KMeans) to group similar articles into topics by word prevalence (tf-idf). (sklearn)
2. Tune *n* clusters to minimize inertia. Given time complexity for KMeans is *O(n *K *I* d)*, training higher values for *n* had prohibitively high compute costs.
3. Apply MiniBatch KMeans (same concept as Kmeans, smaller initialization set, vastly more efficient, slightly less accurate) for values of *n* 50 -> 5,000
4. Tune number of clusters to minimize inertia. (*n*=1,000)
5. Apply LDA to determine top 5 tokens (keywords) for each topic (cluster). (sklearn)

3. Topic Modeling - Results



Most topics have <100 articles, but a long tail introduces skew, as some topics have 100s

3. Topic Modeling

Current Shortcomings

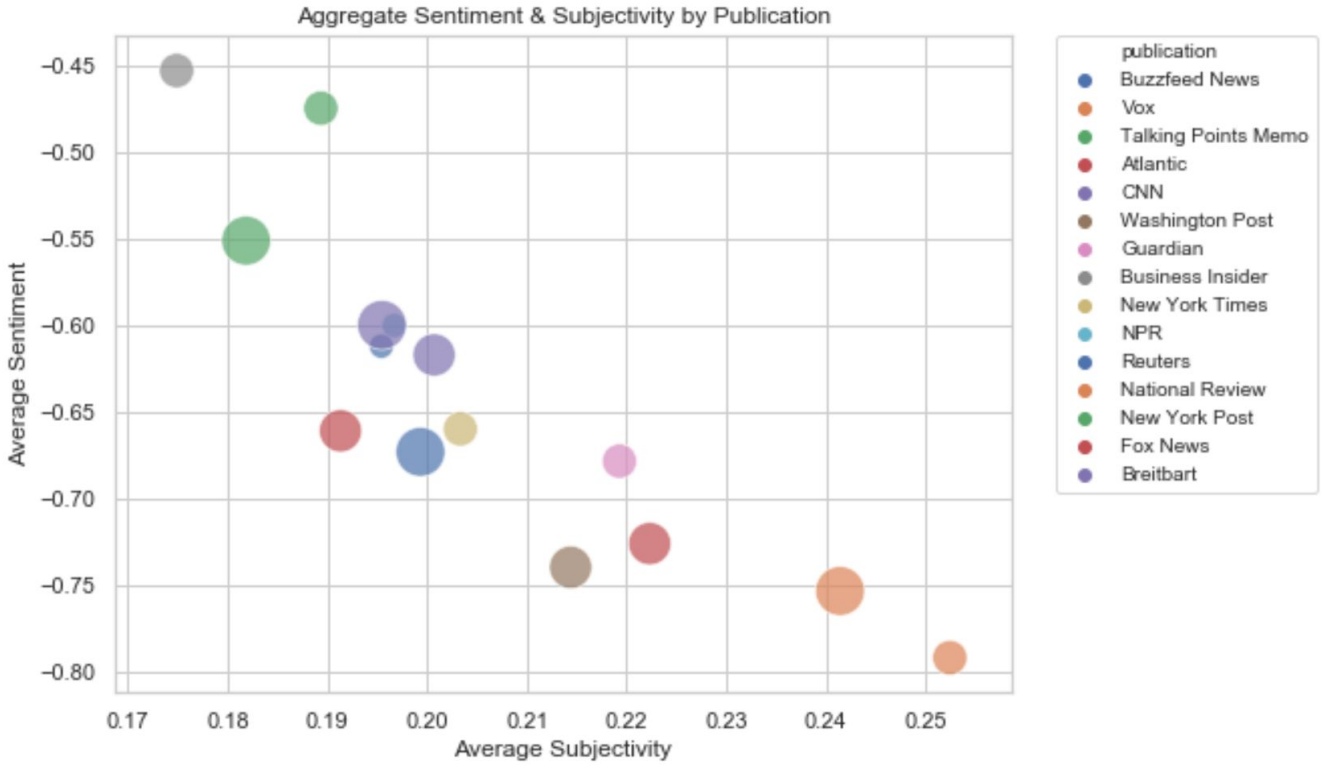
A. Sheer size of dataset (140k observations x 200k features) restricted modeling only most efficient clustering algorithms. More computationally efficient versions of non-parameterized clustering algorithms, e.g. HDBSCAN, may determine more optimal number of clusters than KMeans.

B. By adding n-grams or named-entity recognition to LDA when resolving topic keywords, it may be possible to clean results. For example, “trump”, “donald”, “clinton”, “hillary”, “say” -> “donald trump”, “hillary clinton”, “say”, *keyword4*, *keyword5*.

4. Sentiment / Subjectivity Analysis

- Used vaderSentiment library, chosen for its variety of tagged texts in its corpora, including one set of 500 articles from the New York Times.
- Updated the corpora with more domain-relevant, human-annotated text from MPQA Opinion corpus, contains ~1500 texts from a variety of news sources.
- Analysis returns **positive**, **negative**, and **neutral** scores for each text, based on prevalence and polarity of sentiment-tagged words.
- **Sentiment** is defined as the weighted, normalized composite score; **Subjectivity** is calculated as (1 - neutral_score)

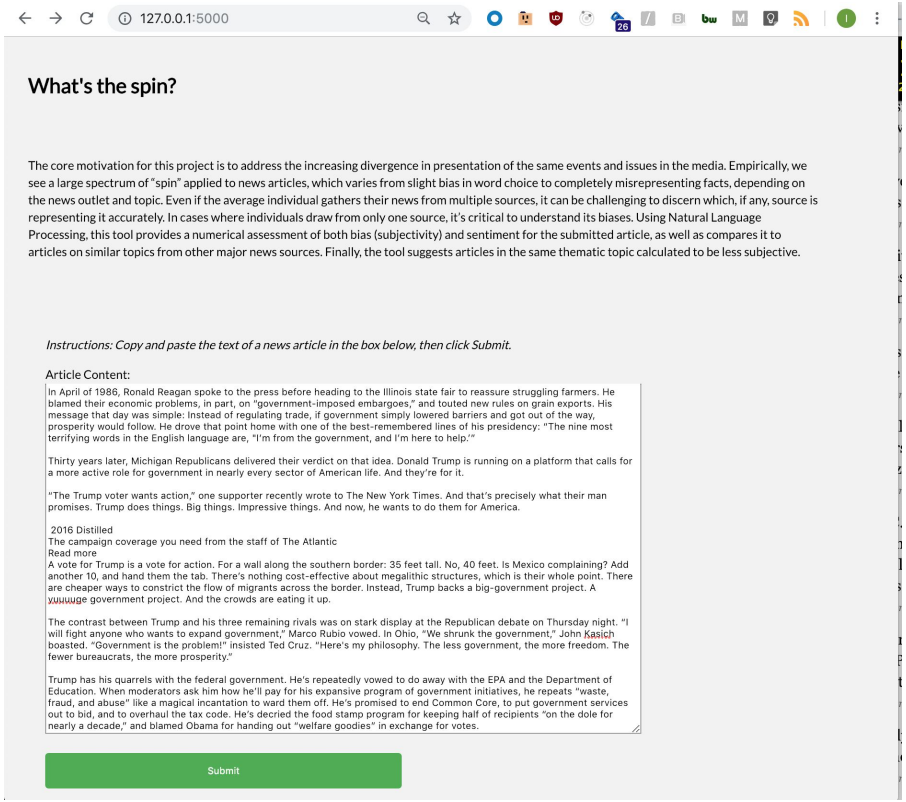
4. Sentiment / Subjectivity Analysis Results



5. Application Development

- Flask app loads resources:
  - Trained KMeans (topic modelling)
  - Training dataset (related articles)
  - Trained vectorizer (processing submitted article text)
  - Sentiment analyzer with updated corpora

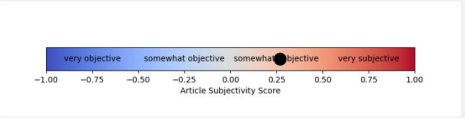
- User submits article text -> cleaned, vectorized, transformed and analyzed!



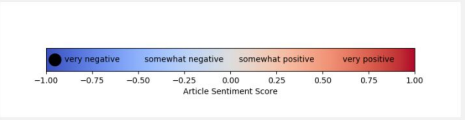
5. Application Development

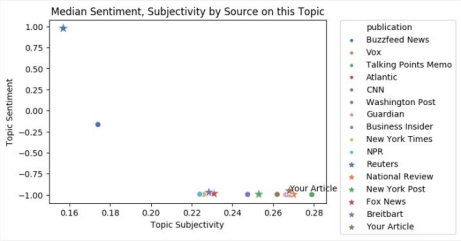
- Analysis returns subjectivity / sentiment scores
- Plots scores vs. other articles in the same topic by other major news outlets
- Suggests articles in the same topic with lower subjectivity scores

Subjectivity: 0.268



Sentiment: -0.953





Topic id: 77

Related articles from more objective sources

Who's Winning The Presidential Election, According To YouTube?
Sentiment Score: -0.166, Subjectivity Score: 0.083

Clinton's 'deplorables' gaffe touches off merch, meme frenzy
Sentiment Score: 0.816, Subjectivity Score: 0.108

The BuzzFeed Buzz Saw: Why Campaigns Should Fear These Four 20-Somethings