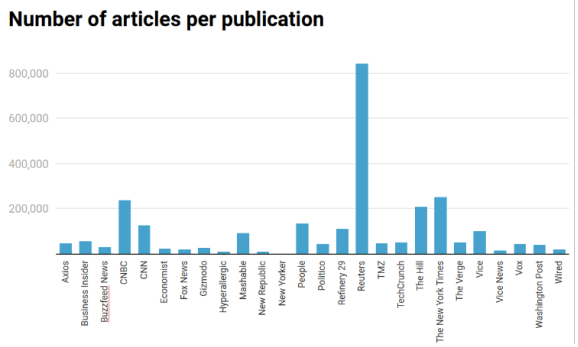
Political bias identification in internet language

ML 2020 R/W 2 Unsupervised

Shea Molloy

A real-world application of machine learning I would love to implement examines and identifies patterns in political bias in digital text online. In my research, I found other projects that are also interested in identifying political biases with machine learning that looked at the source of the content as a factor and goal in their algorithms' predictions, but I would like to focus specifically on using natural language processing to find patterns in words that articulate bias. Eventually, I would like to use this method to predict internet source reliability via a browser extension that would alert a reader to the political leanings and credibility of the web site they are on in real-time.

To do this, I propose scraping keywords from a number of political/news websites, and clustering the keywords to identify patterns of terminology used by various publishers. I could begin my research with the free data set All the News 2.0, published in 2017, which consists of 2.7 million articles from 26 different American publishers. As this data set includes article date, author, title, text, URL, section (if applicable), and publication name, it would allow us to begin by processing the article text.



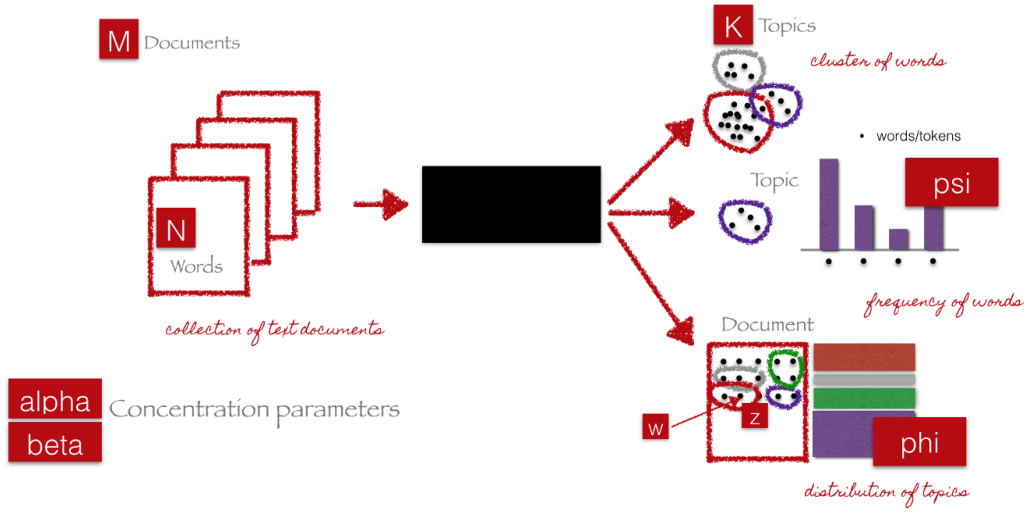
	date	year	month	day	author	title	article	url	section	publication
0	2016-12-09 18:31:00	2016	12	9	Lee Drutman	We should take concerns about the health of li...	This post is part of Polyarchy, an independent...	https://www.vox.com/polyarchy/2016/12/9/138983...	None	Vox
1	2016-10-07 21:26:46	2016	10	7	Scott Davis	Colts GM Ryan Grigson says Andrew Luck's contr...	The Indianapolis Colts made Andrew Luck the h...	https://www.businessinsider.com/colts-gm-ryan-...	None	Business Insider
2	2018-01-26 00:00:00	2018	1	26		Trump denies report he ordered Mueller fired	DAVOS, Switzerland (Reuters) - U.S. President ...	https://www.reuters.com/article/us-davos-meeti...	Davos	Reuters
3	2019-06-27 00:00:00	2019	6	27		France's Sarkozy reveals his 'Passions' but in...	PARIS (Reuters) - Former French president Nico...	https://www.reuters.com/article/france-politic...	World News	Reuters
4	2016-01-27 00:00:00	2016	1	27		Paris Hilton: Woman In Black For Uncle Monty's...	Paris Hilton arrived at LAX Wednesday dressed ...	https://www.tmz.com/2016/01/27/paris-hilton-mo...	None	TMZ

Sources and data in the All the News 2.0 data set.

We would first take the text and clean punctuation, special characters, and stopwords from the article text. We would also transform the text by stemming it with PrterStemmer in order to group words with the same roots together in an attempt to "normalize" similar words. We would further transform the text by tagging the parts of speech and then using TFIDF Vectorizer to create a usable vocabulary.

From there, we would use several feature methods to interpret meaning and patterns within the language intention. Firstly, we would also look at topic model summarization, specifically, Latent Dirichlet Allocation (LDA). An LDA framework would allow us to examine clusters of words as topics, and to compute the frequency of individually occuring words within those topics. The outputs of an LDA framework are a document-term matrix and a topic-term matrix.

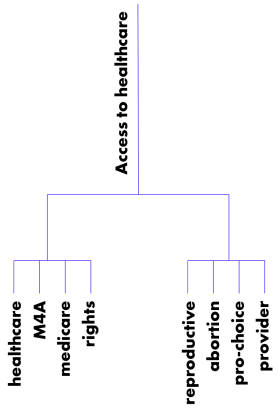
The document-term matrix features we could build include homogenous_lang i.e. how varied the document language is, divisive_terms i.e. frequency of words that are particularly one-sided on a political spectrum and terminology_divergence i.e. what is the deviation of a document's language. The topic matrix could produce such features as political_weight i.e. the frequency of topic that carry more weight in one political party over another, and topic_variance, i.e. the deviation of topics covered by a publication.



End-to-end LDA framework courtesy of C. Doig, Introduction to Topic Modeling in Python via DJ Sarkar, Towards Data Science.

Another feature we would implement involves a hierarchical agglomerative clustering algorithm to identify when certain words feed into the usage of other words (i.e. the word "patriot" or "patriotic" feeding into a cluster using "liberty"). This would allow us to identify inter-text colloquial patterns within different bias bubbles. The clusters generated from this method would be used in a political_spectrum feature, which would align the largest clusters at the topic of the hierarchy with "left" and "right" political values. This feature would honestly have to be developed after seeing the clusters and would hopefully organize terminology in an objective way.

To test the accuracy of this clustering approach, we would feed the algorithm several sets of text from additional sources from outside of the original database (manually scraped). Since this is currently only a human-intuitive metric, we would have to self-identify if these articles are assigned to the right political_spectrum clusters.



In total, the features homogenous_lang, divisive_terms, terminology divergence, topic_variance, political_weight and political_spectrum

would all factor in as part of obtaining a rank on a binary political spectrum that would identify how strongly politically biased a document and a publication is. Future iterations of this project could include cross-checking citations to other publications to reinforce how strong bias or possible neutrality may be. By examining large swaths of textual data and their relationship to each other within the political news spectrum, I hope to articulate the patterns in online discourse that are so quickly identifiable to digital critics online.