Political Leaning Identification Based on Article Text

Samuel Olausson

Desire to be a part of it: 7/10

Description

When viewing news media in the modern world, understanding the impact of bias on the author's views is an essential skill. Because understanding the individual bias of a specific writer is a pretty difficult task when one reads a large amount of news, it is more typical to evaluate bias based on the political leanings of the source of the paper. I want to see if a machine learning algorithm could be trained to identify which political leaning the publisher specific article is based solely on the text/content of the article.

Features

I hypothesize that an algorithm could be trained to do this if it was provided with the correct features. One category of features to look at would be the usage of certain vocabulary. For example, I am pretty confident that certain politically-charged terms would be used by only one political leaning. In addition, it's possible that the names of certain common political issues would be present in only certain leanings. Going another way, I would also want to test features such as the length of the article: it seems possible that publishers with moderate views would write articles of different length than publishers with more extreme views. Finally, the number of quotations included could also be potentially relevant.

Ex:

#"woke"	"pro life"	Article Length	#Quotations	**TARGET:Publisher Political Leaning**
3	1	1000 words	5	Conservative

Gathering Data

The data could all be gathered from the internet, likely using web scraping. There might be some difficulty in getting large numbers of articles from just one news source if they have protections against scraping, but the large number of potential sites to scrape from would hopefully alleviate any difficulties in that. The data could be labeled by comparing the source (publisher) of the article to a given set of rankings of the political leanings of all major news sources. Then, the data would need to be processed to gather the features. Most of this processing would be pretty easy, as it would mainly consist of searching for specific strings or patterns in the article text. It might also be important, however, to strip out the name of the author and any other identifying information in the article.