# Proposal for News-Bias Project

|  |  |
| --- | --- |
| **Problem Statement:** | |
| **Context:** Fewer people trust the news media than ever before. The increasing ease of adding to the media landscape via the internet, and with 24-hour news channels in ever greater competition, recent years have seen a massive uptick in opinion masquerading as journalism and marketing itself to niche-customers. Many people are worried about the prevalence of bias in their news, but do not have a trusted source of information on that bias itself. A purely quantitative, hands-off approach to measuring that bias could provide trustworthy and needed insight into the state of the news media. | **Constraints:**  Some people are already set in their beliefs on the bias of the news, and no amount of evidence will change them. |
| **Criteria for Success:** This project will establish a scale for measuring sentiment (bias) in news articles, allowing news sources to be ranked as more/less biased than others. | **Stakeholders:**  The news media itself is a stakeholder, as their bias determines the size and makeup of their audience.  The general public is a stakeholder, as the news media defines our national conversation.  Advertisers are stakeholders, as they make decisions based on the size and makeup of the audience to which their ads will be shown. |
| **Scope of Solution Space:** The usefulness of this project relies heavily on the way it is explained post-analysis. It must be explained to the public in an un-sentimental, non-biased, and trustworthy manner. It must also be explained so that even laymen can understand the quantitative, unemotional nature of the analysis. | **Data Sources:**  The data used comes from Kaggle at this URL:  <https://www.kaggle.com/snapcrack/all-the-news/>  It contains ~150,000 news articles from 2015-2017, from 15 different sources. They were scraped from the internet. The dataset contains the title, author, date, publisher, and test of each article. |

## Method

The articles are presented in 3 separate .csv files that will need to be combined into a single DataFrame. From there, the data can be explored based on author, publication date, and publisher. Then, the data will be copied to a new data frame and the text column will be dropped; this data frame will be used to analyze the headlines by themselves. The goal is to use sentiment analysis to measure the bias; the media is expected to be presented with a focus on facts, and so should not register strong sentiment. The text of the headlines may be split into single words, or it may be analyzed by sentence, or perhaps as a whole. The best way to analyze the sentiment will be determined by exploratory analysis.

Once each article has a rating for sentimentality, those ratings will be combined, probably by arithmetic mean, to find a sentimentality rating for the publisher. It would also be interesting to find the average sentimentality rating for the entire media (as represented by the sample of articles). Bayesian analysis may also be used to approximate the distribution of bias within articles from each publisher.

It will be important to create high-quality, readable visualizations to make the post-analysis explanation more accessible.

The analysis and findings will be explained in a report, and a slide deck/presentation will be prepared as well.

All related notebooks, the report, and the slide deck will be stored in a GitHub repo.