Identifying Political Bias in the Media

Introduction

The media is a service industry and has a responsibility to be factual and without bias. Every news-writer and news agency is bound by a certain set of ethics and unwritten laws that are based on this responsibility.
The first and foremost is the responsibility of accuracy in reporting. The facts have to be precise, specific, and clear. Reporting is about stating facts without the injection of opinions.
Although these rules and regulations are well understood, the deviation from them is sometimes very easily evident, i.e. inclination towards and away from some group can be noticed. This is particularly true when it comes to politics.
This is where we (the team) would like to assess how the major news providers are performing, i.e. we would like to see if the major news providers of the country are biased towards or away from certain political figures.

Data Selection

News Providers

The selection of news provides were based on their popularity and the number of unique visitors the news provider receives on their website. The ranking was based on each news providers Alexa Rank. The list was altered as drawing data from some of the news providers was infeasible. The final list is as follows.

1. New York Times
2. Washington Post
3. Fox News
4. USA Today and
5. New Yorker

Political Figures

The political figures picked were based on the popularity of individual figures based off of polls conducted by https://www.crowdpac.com/. Two from each of the leading political parties were picked and for an added perspective, the two political parties were included as well.
The final list of the political figures are as follows.

1. Hillary Clinton
2. Bernie Sanders
3. Democratic Party
4. Donald Trump
5. Ben Carson and
6. Republican Party

Data Collection

Data collection involved scraping data manually from individual pages of news articles as a single repository of the data was not available. The cost of simply purchasing the data was not feasible. Scraping a web page is a method by which single elements of data are extracted from the webpage. The single elements are identified by their html tags.

A list of all URLs to be scraped was created using a search engine optimization tool knows as SEOQuake. The tool interacted with the google search engine to return relevant results, i.e. results having a certain key word from specific websites within a given time period.

The job of extracting the specific text body from all the URL's was done using the tool Kimono Labs. Once specific web elements were selected, Kimono labs was capable of creating a template and applying it across multiple webpages to return a flat file containing the extracted data for a chosen political figure from a specific news provider. 30 such flat files were generated; each containing a certain combination of the political figure and news agency.

Data Cleaning

The flat file containing the extracted data contained the following attributes.

1. Date of the article.
2. Text of the article
3. Index (Row identifier)
4. URL - URL of the article

Microsoft Excel

Each article had several rows of data (each paragraph in the article is assigned a different row). The different rows of data for a single article was concatenated into one single cell. This was done using the MoreFunc add-in in Microsoft Excel.

Another Microsoft excel add-in "PowerQuery" was used to collate the 30 different files containing data from the different news providers and for different political figures.

R

R was used for finer cleansing of the data. Abbreviated words and irrelevant text (such as "Mr.", "Mrs.", Quotes etc.) were removed from the all the article text using R.

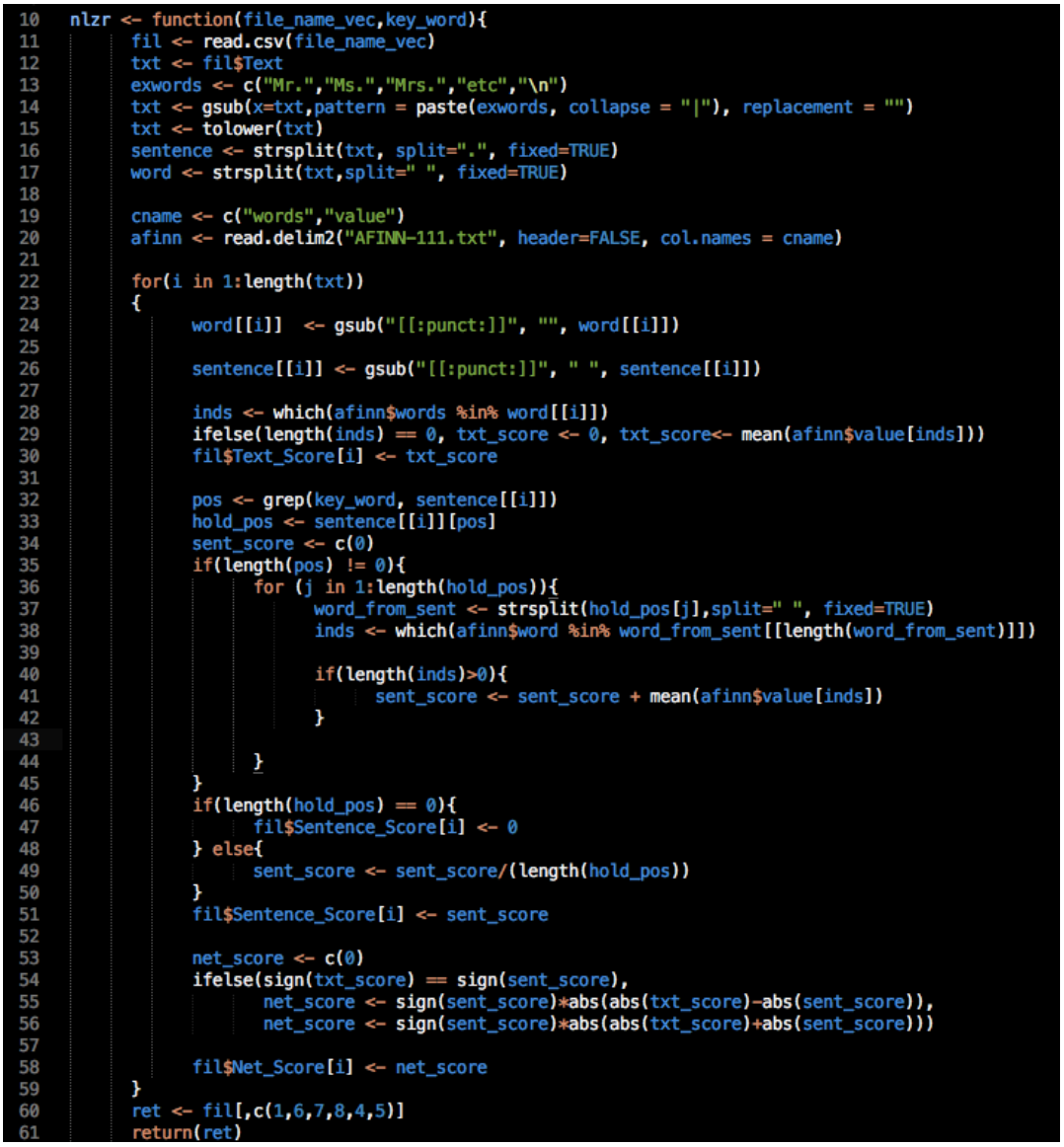
Analyzing the data

Sentiment analysis was conducted on the text to identify bias. A positive or negative sentiment on a certain political figure does not indicate bias but a consistent sentiment can be used to identify bias. Another way to identify bias is to spot consistence in associating the political figure with a certain sentiment irrespective of the sentiment of the article itself.

Application of the NLP to build sentiment scores.

The following steps describes how the model is designed to function:

Step 1: Read data from the flat csv file that contains the news article, the date and URL.
Step 2: Break the article down to individual sentences and words.
Step 3: Use the AFINN lexicon to calculate the mean of all the sentiments of the words from the article. This is the text score of the article. (AFINN is a dictionary of approximately 2500 words that contain the strength of the positive and negative sentiment associated with words)
Step 4: Use the AFINN lexicon to calculate the mean of all the sentiments of the words from the sentences that mention the political figure. This is the sentence score of the article.
Step 5: Calculate the intensity of the conflict in the sentence score and the text score for every article. This is the net score. This measures the intensity of bias towards the mentioned political figure in the article.



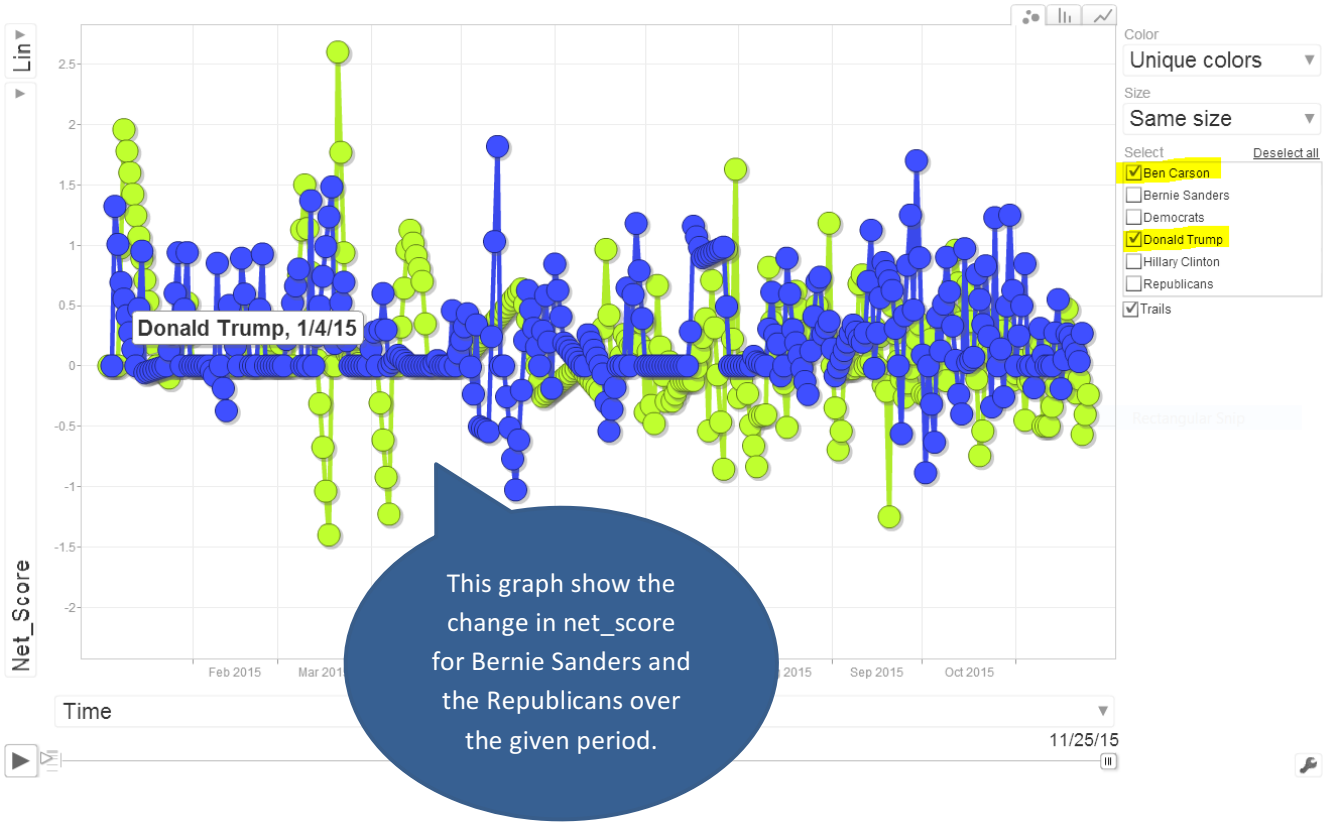
```
nlzr <- function(file_name_vec,key_word){
      fil <- read.csv(file_name_vec)
      txt <- fil$Text
      exwords <- c("Mr.","Ms.","Mrs.","etc","\n")
      txt <- gsub(x=txt,pattern = paste(exwords, collapse = "|"), replacement = "")
      txt <- tolower(txt)
      sentence <- strsplit(txt, split=".", fixed=TRUE)
      word <- strsplit(txt,split=" ", fixed=TRUE)

      cname <- c("words","value")
      afinn <- read.delim2("AFINN-111.txt", header=FALSE, col.names = cname)

      for(i in 1:length(txt))
      {
            word[[i]]  <- gsub("[[:punct:]]", "", word[[i]])

            sentence[[i]] <- gsub("[[:punct:]]", " ", sentence[[i]])

            inds <- which(afinn$words %in% word[[i]])
            ifelse(length(inds) == 0, txt_score <- 0, txt_score<- mean(afinn$value[inds]))
            fil$Text_Score[i] <- txt_score

            pos <- grep(key_word, sentence[[i]])
            hold_pos <- sentence[[i]][pos]
            sent_score <- c(0)
            if(length(pos) != 0){
                  for (j in 1:length(hold_pos)){
                        word_from_sent <- strsplit(hold_pos[j],split=" ", fixed=TRUE)
                        inds <- which(afinn$word %in% word_from_sent[[length(word_from_sent)]])

                        if(length(inds)>0){
                              sent_score <- sent_score + mean(afinn$value[inds])
                        }

                  }
            }
            if(length(hold_pos) == 0){
                  fil$Sentence_Score[i] <- 0
            } else{
                  sent_score <- sent_score/(length(hold_pos))
            }
            fil$Sentence_Score[i] <- sent_score

            net_score <- c(0)
            ifelse(sign(txt_score) == sign(sent_score),
                   net_score <- sign(sent_score)*abs(abs(txt_score)-abs(sent_score)),
                   net_score <- sign(sent_score)*abs(abs(txt_score)+abs(sent_score)))

            fil$Net_Score[i] <- net_score
      }
      ret <- fil[,c(1,6,7,8,4,5)]
      return(ret)
```

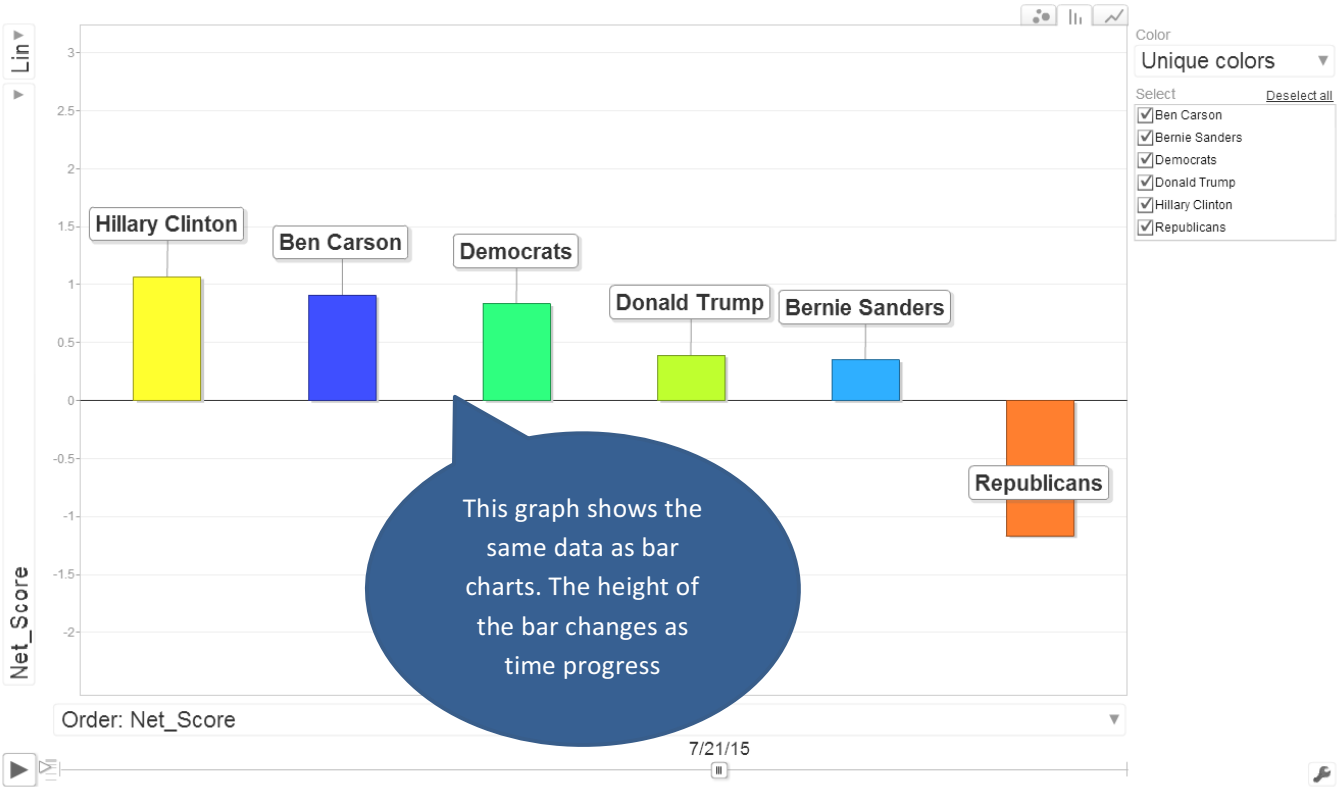
Conclusion

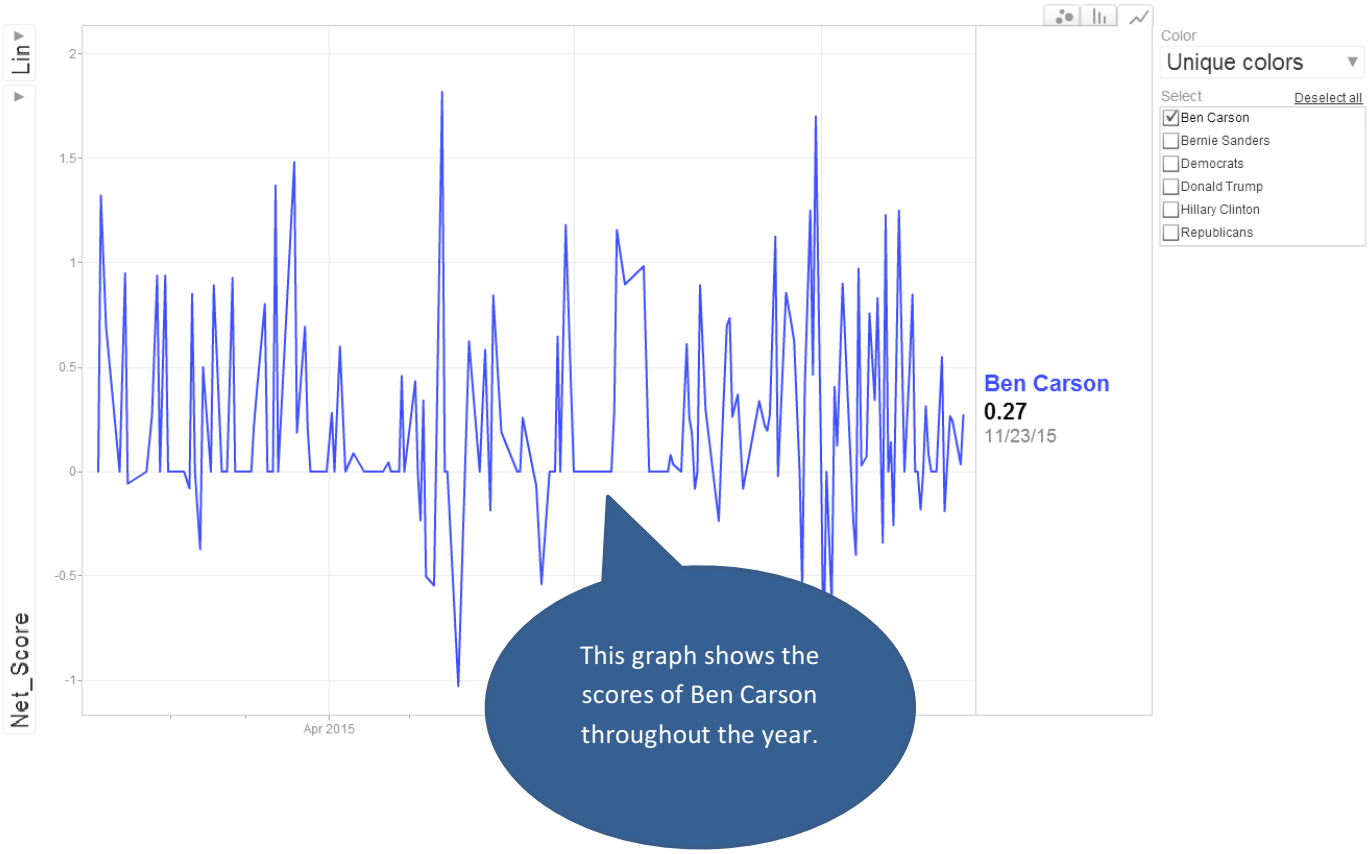
A stagnant or single sided text score indicates bias. The same applies to the sentence score. A very high net score (above a chosen threshold) indicates bias. This score needn't be consistent to be an accurate indicator of bias.
The complete R Code is attached separately.

Visualization of the result

The GoogleVis package in R is used to represent the relevant information. The package served as a convenient tool to visualize multi dimensional data. A set of sample visualization images are attached as below:




Lin
Net_Score
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
Hillary Clinton
Ben Carson
Democrats
Donald Trump
Bernie Sanders
Republicans
This graph shows the same data as bar charts. The height of the bar changes as time progress
Order: Net_Score
7/21/15
Color
Unique colors
Select
Deselect all
Ben Carson
Bernie Sanders
Democrats
Donald Trump
Hillary Clinton
Republicans


Lin
Net_Score
2
1.5
1
0.5
0
-0.5
-1
Apr 2015
Ben Carson
0.27
11/23/15
This graph shows the scores of Ben Carson throughout the year.
Color
Unique colors
Select
Deselect all
Ben Carson
Bernie Sanders
Democrats
Donald Trump
Hillary Clinton
Republicans