How Subjective are Senators' Press Releases? An analysis of 16 Senators and their Press Releases

Elliot Chang

West Virginia University

DSCI 301 Project 3 Report

April 26, 2023

1 Database

The *senators.archive* database is a *mongoDB* NoSQL database containing 23 collections, which each correspond with one Senator and their respective press releases. These press releases were collected as a group effort by the Spring 2023 DSCI 301 class, where each student scraped the *senate.gov* websites of 3 senators of their choice (without overlap among students) for their press releases. Each student submitted 3 JSON files, one for each of their senators, and our instructor combined all of the files into a single *mongoDB* database. *senators.archive* was accessed using pymongo with *mongosrestore* while additional data of the Senators was provided from a *senators.csv* table created earlier in the semester by the same group of students.

1.1 Database Construction: Web Scraping

My responsibility in the database construction was retrieving the press releases of Elizabeth Warren, Rand Paul, and James Lankford. *BeautifulSoup* from the *bs4* library was used to scrape

press releases from the following websites for the respective Senators:

https://www.warren.senate.gov/newsroom/press-releases (Elizabeth Warren)

https://www.paul.senate.gov/ (Rand Paul)

https://www.lankford.senate.gov/news/press-releases (James Lankford)

The *url's* for each article were extracted from each website, and each *url* was individually scraped for the *Article Title* and *Article Content*. Finally, three dictionaries were made, one for each senator, where each dictionary contains *index: document* pairs and *document* is a dictionary with the following structure:

```
{
    "name": "First Last"
    "url" : "Article URL"
    "title": "Article Title"
    "content": "Article Content"
}
```

While the list of *url's* for each Senator was long (In the magnitude of $\sim 10^2 \sim 10^4$ entries each), SSL errors reduced the number of actual documents retrieved down to 18 for Senators Warren and Lankford and 32 for Senator Paul.

1.2 Data Cleaning

After gaining access to the data using pymongo and *mongorestore* on *senators.archive*, some data cleaning took place. Since this database was "crowdsourced", the first task was to normalize the format of storage. Since my analysis required *Senator Name* and *Article Content*, any collection where not every document fit the following structure was removed.

{

”name”: ”First Last”

...

”content”: ”Article Content”

}

As a result, 7 of the 23 collections were removed from the analysis pool, some due to incorrect naming (”name” referred to article title and there was no senator name field) and others due to empty contents.

1.3 Subjectivity Calculation

My goal from the analysis was to see if there were any trends with Senators' press release subjectivity. Subjectivity Calculation was done using the TextBlob library (in-class presentation incorrectly said the spaCy library was used; spaCy was the original plan but TextBlob worked better). TextBlob makes *blob* objects from an input string that has *subjectivity* and *polarity* properties; *subjectivity* was the focus of this analysis. The *subjectivity* score is a floating point number in the interval of $[0, 1]$ where a score of 0 represents very objective text and a score of 1 represents very subjective text. A new DataFrame was created with a *name* column for Senators' names and a *scores* column containing the list of scores of the senators' press releases. This table could then be used for Analysis by using joining methods on *Senator Name* to create specific tables for plotting.

2 Analysis

In order to get a better understanding on the overall subjectivity of Senators' press releases, I made a stacked box-and-whisker plot to compare between individual Senators. The boxes were colored based on the Senators' party affiliation.

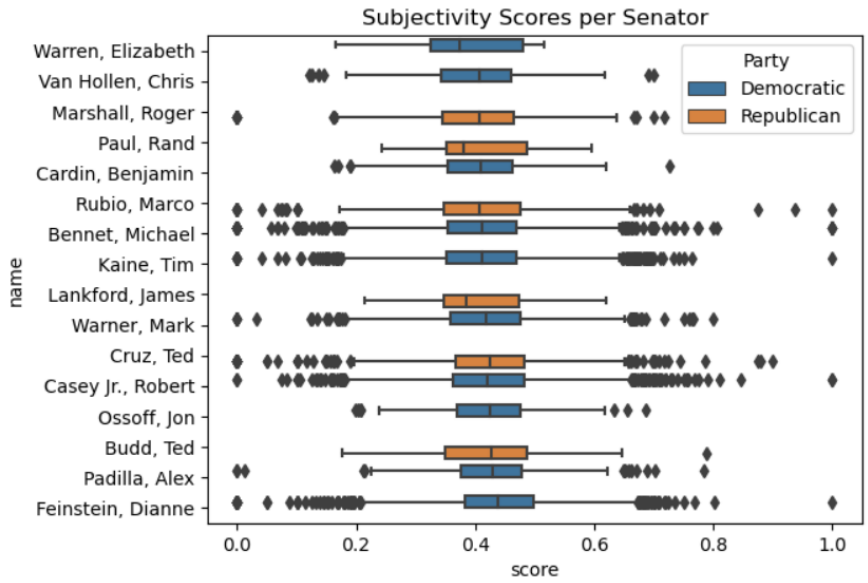

Figure 1: Per-Senator Subjectivity

Figure 1 shows a general trend of Democrats having larger ranges while neither party seemed to have different means. This is then confirmed in Figure 2, which depicts the per-Party Average Means and Variances of Subjectivity Scores:

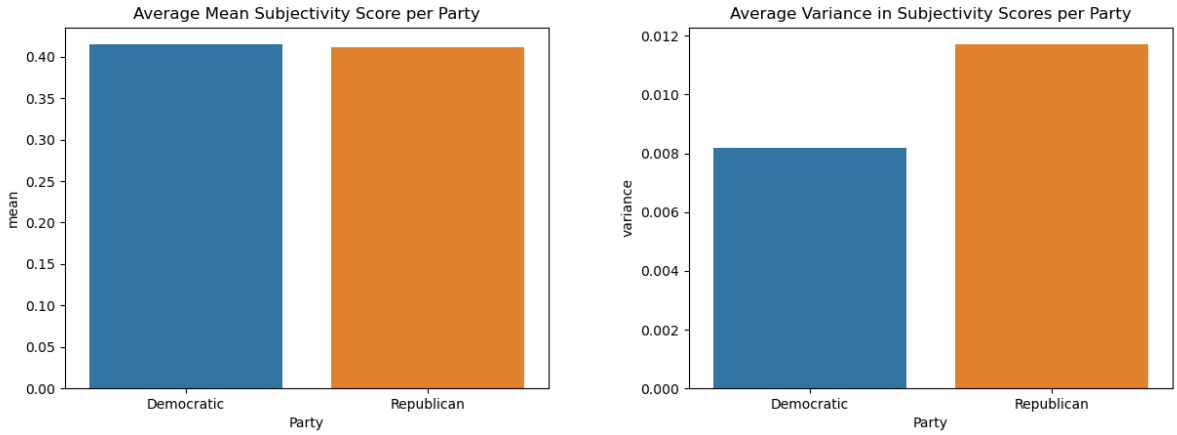

Figure 2: Average Mean and Variance of Subjectivity Scores per Party

In Figure 2, we can clearly see that the means do not differ but the variances do between parties. However, since the goal of this project was to conduct an Exploratory Data Analysis, no hypothesis testing was performed. With the small sample sizes and small differences, it is likely that these findings are not significant.

Another possible factor in subjectivity could be due to Senators' age, but a quick look at the Mean/Variance of Subjectivity Scores per Age in Figure 3 shows no trend.

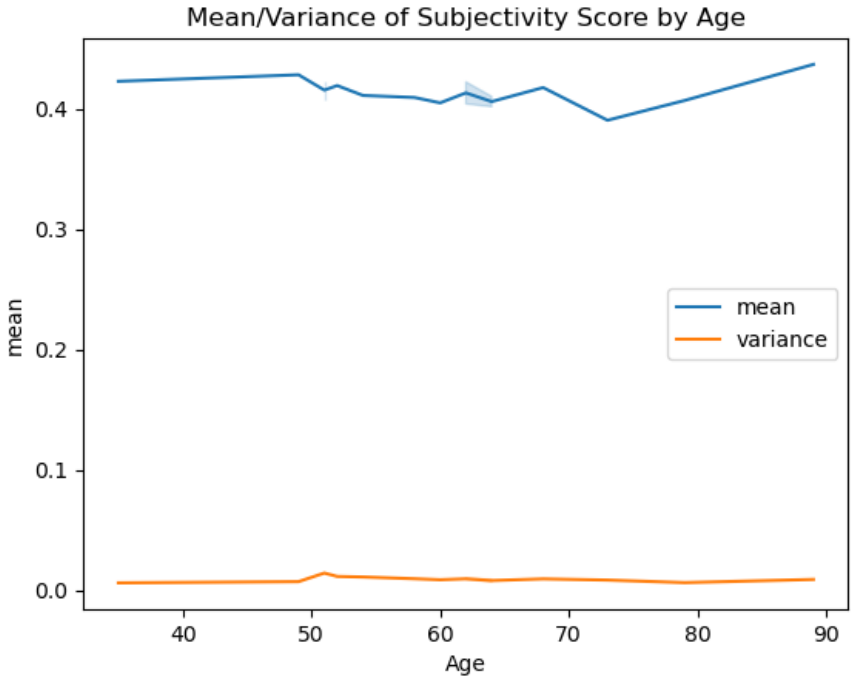

Figure 3: Average Mean and Variance of Subjectivity Scores by Age

3 Future Work

In the future, I would like to recreate this project on a larger database. Since we only had access to 23 senators' press releases (only 16 of which were useful to this analysis), and not every senator had a sufficiently large number of documents, it would be very interesting to see whether or not the same trends hold on a much larger set of data. Then, we could also perform hypothesis testing on these trends to see if they are meaningful. Another direction for further analysis could be to explore other Sentiment Analysis measures, such as Polarity, Emotion, or Intensity.