JURIST Digital Scholars Program Report

Imad Rizvi, Advised by Alice Chen, Cai Chen, and Xiaoli Jin

Summer 2020

1   Introduction

The JURIST dataset contains thousands of articles written by students over the past two decades. With this project, we aimed to develop a better understanding of how sentiment analysis and bias detection techniques could be used to understand the dataset. Further, we hoped to explore whether there were any significant differences in sentiment or bias for articles written about domestic issues versus articles written about international issues.

The dataset consists of 48,775 news articles, along with 3,384 commentary articles. Articles were tagged based on the topics they discuss: various regions, countries, or themes were popular tags. There were 7,693 unique tags among news articles, and 3,175 unique tags among commentary articles. The most popular tags among news articles were US (12,300 articles), International (6,661 articles), and Domestic (6,260 articles). The most popular tags among commentary articles were United Sates (1,345 articles), Terrorism (570 articles), and Middle East (383 articles). The commentary articles tended to be much longer in length and represent more thorough analysis and opinions, whereas the news articles were generally much shorter and only included basic facts that described what had occurred, rather than much analysis. Therefore, for the purpose of understanding how sentiment may change across analysis of different regions, we used only the commentary pieces, as they contained more subjective components rather than just reporting. We used the news articles for other analyses such as the creation of word clouds.

2   Sentiment Analysis

The first challenge before beginning sentiment analysis was processing the data. Since it appeared in HTML format, each article had different types of HTML tags, and these formats were not consistent across articles. Therefore, it required a lot of manual scanning of many different articles to catch as many of the cases as possible.

2.1 Lexicon-based Sentiment Analysis

One approach to conducting sentiment analysis is a lexicon-based approach, which refers to using a lexicon or dictionary that maps words to a sentiment score, with several rules builtin as well. These rules may include negating a sentiment score for a sentence if the word "not" is added, for example. There are several different Python packages that contain lexicons that can be used for sentiment analysis. One lexicon is the Vader lexicon, which is specifically attuned to sentiments expressed in social media. It seems to have many heuristics built in that give the lexicon a solid performance compared to other methods (Ribeiro, Filipe N., et al). The TextBlob package also features a sentiment analysis lexicon. It seems to perform generally well in formal settings, whereas Vader has advantages in more informal settings (emojis, slang, etc).

Upon analysis, we found that there were not significant differences in Vader or TextBlob sentiment scores for articles about the US vs articles about the Middle East. We used the comparison of these two regions as a proxy for measuring how sentiment differs between domestic and international reporting, because these two were some of the most popular tags among commentary articles.

One concern of these approaches was that based upon manual inspection of how each method performed, it seems that lexicon based approaches tend to be very domain-specific — therefore, it would be difficult to use a lexicon-based approach that wasn't trained on JURIST data to evaluate JURIST data. Otherwise, by the standards of the social media data that the VADER lexicon was trained on for example, all the JURIST articles might tend to have similar scores.

2.2 Unsupervised Sentiment Analysis

Since there were clear difficulties with the lexicon-based approach, we wanted to attempt to make use of a method that would be more grounded in the JURIST data. One approach could have been supervised sentiment analysis conducted by first annotating the sentiment of a large number of the JURIST articles to use as training data. However, this would've required a lot of work and would have been difficult to do accurately without a broader understanding of sentiment analysis. Therefore, we wanted to attempt an unsupervised method. Our approach was largely based off of Rafał Wójcik, 2019. The basic premise of the method explained in the article was to cluster word embeddings into two groups, one of which would be positive, and one of which would be negative.

After implementing this approach on the JURIST data, the first issue encountered was that one of the clusters created tended to include predominantly author names. In order to counteract this problem, we tried creating three clusters, one of which would be mostly author names and represent a more neutral group. However, the "positive" and "negative" clusters were still not particularly accurate. Finally, we tried four clusters as well, in case this may lead to a positive, negative, neutral, and author name group. We judged the

accuracy of the clusters based on manual inspection of which words would fall under "most positive" or "most negative" and found no sentiment-based patterns. The words "good" and "bad" appeared in the same cluster. Upon further reflection, it seems not necessarily the best approach to assume that the word embeddings would translate particularly well to understanding differences in sentiment between words, as there is a difference between a word's sentiment and meaning.

3   Bias Detection

Compared to sentiment analysis, there has been a lot less research done on the topic of bias detection, likely because of the difficulties defining what makes a piece "biased." The most popular approaches seem to be lexicon-based, and more focused on recognizing whether something is presented more as fact or opinion, essentially, measuring the subjectivity or objectivity of text. Therefore, we decided to use the Objectivity/Subjectivity Analyzer from the TextBlob package that had also been used for sentiment analysis. However, once again, we found no significant differences in subjectivity scores between articles about the US and Middle East.

4   Word Clouds

The final part of the analysis was to generate word clouds and conduct general word frequency analyses. Initially, after removing "stopwords," the most popular words in articles were "law", "state", "court" and "war". However, we wanted to know for each category (US vs Middle East articles) which words were disproportionately represented rather than just their overall frequency. Therefore, we calculated a chi square test statistic for each word, which measured how much each word appeared compared to the expectation based on all the articles. For the Middle East articles, the words that appeared a disproportionate amount were "Iraq", "Saddam", "war", and "Israel".

5   Dashboard

Finally, we aimed to display some of the findings in an easier to use method. We decided to create an interactive dashboard using the Dash and Plotly packages added to a Flask Python webapp. The app features several interactive graphs that contain information on subjectivity and objectivity scores of JURIST commentary articles with various tags and in various time frames. Users can also enter in an article, and see how its subjectivity or objectivity score based on Textblob would have it compare to other articles in the JURIST dataset. It also will show users what the sentences are that contribute most to the subjectivity and objectivity score.

6 Conclusion

We did not find any evidence that there was any bias in reporting across different regions in the JURIST data. We were still able to explore many different textual analysis methods including sentiment analysis and bias detection. We found that there is a lot of difficulty adapting across domains for natural language processing tasks, and so future analyses of the JURIST dataset should involve training on JURIST or other news data in order to improve performance.

7 References

Rafał Wójcik. "Unsupervisedd Sentiment Analysis". (2019) https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483

Ribeiro, Filipe N., et al "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods." EPJ Data Science 5.1 (2016): 1-29.