Analysis of Mixed Quote Representation in News Sources

Mary Gibbs, Binbin Wu, Charles Garrett Eason, and Mihir Gadgil

Introduction

Can media bias be objectively measured? Probably not, but realizing that some conditions for political bias can be measured might get us part of the way there. For example, while media bias can take on many forms (intentional or not), mixed or partial quoting represents a unique situation where there exists some form of measurement between a full quote and some deviation of that representation (e.g. a partial quote omitting material). This measurability in turn implies an objective approach to assessing some base differences between media reporting is possible.

To illustrate, Breitbart reports a quote from President Donald Trump, "**The Radical Left Democrats** have failed on all fronts . . ." (Caplan, 2019). While on the same day, The Washington Post reports a partial version of the same quote, "Trump reacted on Twitter, writing that **Democrats** 'have failed on all fronts . . .'" (Marimow & Fahrenthold, 2019). Regardless of intention, these two examples display a clear route for measurement of the omitted material and, given the particular omissions, it seems plausible that these differences were due to political leaning. Thus, we hypothesize that similarities/differences among mixed quote reporting should square up against similarities/differences in political leaning.

Little research has been done on this topic from our perspective on bias. The literature that does touch on what we focus on here mainly looks at automated systems for quote assignment or content identification. These systems include algorithms that identify indirect/mixed quotes in text and/or assign speakers from a corpus of quotes (O'Keefe, Pareti, Curran, Koprinska, & Honnibal, 2012; Pareti et al., 2013; Pavllo, Piccardi, & West, 2018). They also include systems that compare news sources to mine for additional information given a single initial source (Iacobelli, Nichols, Birnbaum, & Hammond, 2012). Therefore, there is a gap in the literature concerning partial/mixed quote identification and cross news source comparison to objectively detect necessary conditions for political bias.

Objective

To develop a mixed quote comparison algorithm to analyze differences/similarities in how mixed quotes are reported by various news sources.

Implementation

We implemented a custom pipeline for collecting the data for this project (Figure 1). First, we used News API to collect news articles from a selected news source, in this case Fox News, that were published within a given period of time, in this case between October and November 2019. Second, we extracted all of the mixed quotes from these news articles using regular expression

matching operations. Afterwards, we performed a search using the extracted mixed quotes and Google API, specifying domains, such as www.bbc.com or www.cnn.com, to find news articles containing similar mixed quotes that were published between October and November 2019. We decided to extract the top two most relevant news articles from six different news sources, including Associated Press (ap), British Broadcasting Corporation (bbc), Breitbart (bb), Cable News Network (cnn), Fox News (fox), and The Washington Post (wp). Then, we used custom-built web scrapers to obtain article text and then extract mixed quotes from the article text. Finally, we utilized the clustering technique k-means clustering and the similarity metric Jaccard similarity to evaluate differences/similarities between the mixed quotes found within the six news sources.

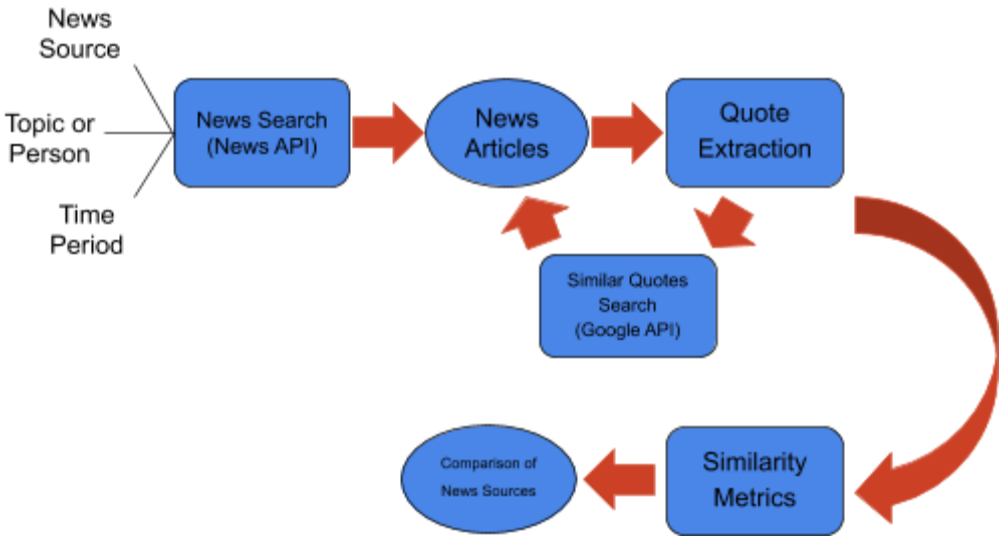
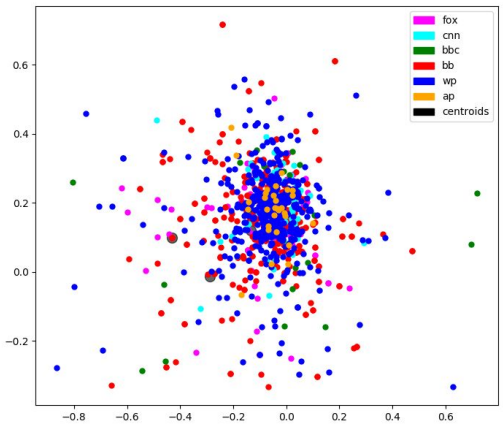


Figure 1. A diagram illustrating the project pipeline.

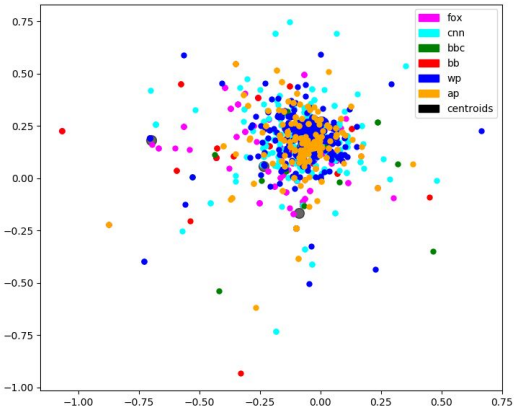
Evaluation

We decided to evaluate our algorithm on two different sets of news articles. The first set of news articles were published between October 31, 2019 and November 2, 2019 when relatively insignificant political events occurred. The second set of news articles were published between November 16, 2019 and November 21, 2019, which was the time period in which George Sondland testified in the President Donald Trump impeachment proceedings.

First, we decided to try a clustering technique to assess the differences/similarities between the mixed quotes found within the six news sources. We used Word2Vec to generate word embeddings for the mixed quotes found within each news source. Then, we created a k-means clustering model using six clusters to observe if the mixed quote word embeddings cluster together based on news source (Figure 2).



October 31, 2019 to November 2, 2019



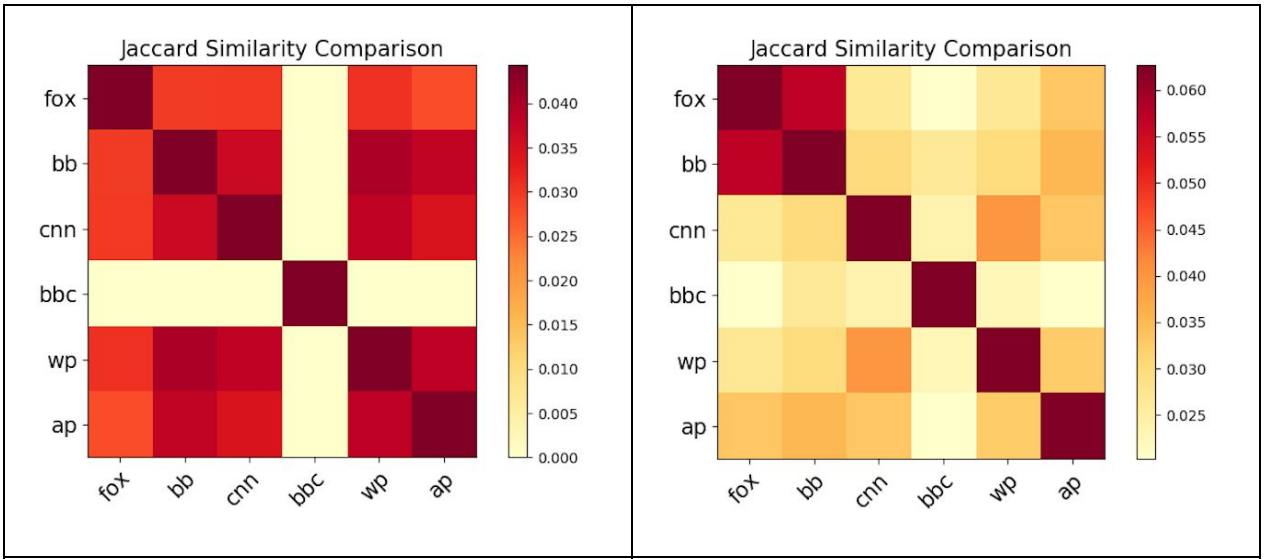
November 16, 2019 to November 21, 2019

Figure 2. k-means clustering results.

We did not observe any distinct clusters for the mixed quotes obtained from the six news sources for any of the date ranges. This may be due to the fact that the mixed quotes come from articles that are about the same topic, making them too similar to observe differences/similarities by using clustering techniques.

Second, we decided to try the similarity metric Jaccard similarity to assess the differences/similarities between the mixed quotes found within the six news sources. We performed tokenization and stemming followed by stop word removal for the mixed quotes within each news source. Then, we calculated the Jaccard similarity between mixed quote token sets from the news sources (Figure 3).

When relatively insignificant political events occurred, the news sources appear to be very similar to each other as indicated by the red shading between each news source comparison in the heat map for October 31, 2019 to November 2, 2019. However, when Gordon Sondland testified in the President Donald Trump impeachment proceedings, the news sources formed two distinct groups in regards to mixed quoting similarity. Fox News and Breitbart appear to be very similar to one another as indicated by the red shading between these two news sources, while the rest of the news sources appear to exhibit less similarity as indicated by the orange/yellow shading in the heat map for November 16, 2019 to November 21, 2019. Of note, the BBC results from Google API came back as unusually spurious; hence, the low Jaccard similarity scores for BBC across both time periods.



October 31, 2019 to November 2, 2019	November 16, 2019 to November 21, 2019

Figure 3. Jaccard similarity results.

Figure 4 provides an example from our results of how mixed quote differences correlate with political leaning on days with a significant news event, in this case between November 16, 2019 and November 21, 2019 when Gordon Sondland testified in the President Donald Trump impeachment proceedings.

"In his earlier deposition behind closed doors, Sondland had said that when he asked President Trump explicitly what he wanted from Ukraine, the president told him, 'Nothing,' and reiterated that there was no **'quid pro quo.' Sondland also testified that the president told him that he simply wanted Zelensky to do what he ran on, namely fighting corruption**." - Pollak, 2019 | Breitbart

The president challenged the "fake news" to report that he told the ambassador late in the game that "**I want nothing, want no quid pro quo, tell Zelensky to do the right thing**." - Kurtz, 2019 | Fox News

"**'I want nothing. I want nothing. I want no quid pro quo,'** the president told a U.S. diplomat in the September phone call. Questions were swirling by then about Trump's motivations for holding up U.S. security assistance for Ukraine as he pressed the country's leaders to investigate his political rivals. '**Tell Zelenskiy — President Zelenskiy to do the right thing**,' Trump added in his conversation with his ambassador to the European Union, Gordon Sondland". - Colvin, 2019 | Associated Press

"Sondland, the president noted, also made clear that Trump had told him he wanted 'nothing' from the Ukrainians and that there was no **'quid pro quo'**". - Zurcher, 2019 | BBC

Figure 4. Mixed quote examples from news sources between November 16, 2019 and November 21, 2019.

The conservative-leaning Breitbart reported that Sondland recalled President Trump stating that there was no "quid pro quo" along with the positive addition that Trump wanted the Ukrainian President Zelensky to fight corruption. Similarly, the conservative-leaning Fox News reported a little more of the original quote with a different positive addition that Trump wanted Zelensky to "do the right thing". Associated Press, a wire service, essentially reported the direct quotes and

information. Whereas, the liberal-leaning BBC only reported Trump claiming that there was no quid pro quo, omitting Trump's more positive statements.

Conclusion

Overall, we find that our approach can be used to identify mixed quoting differences/similarities in news sources on days in which significant events that dominate news coverage take place. On these days, news sources of similar perceived political bias tend to correlate together, while news sources of differing perceived political bias tend to diverge in terms of mixed quoting similarity.

Limitations

Several limitations shaped the way that we structured our approach to analyzing mixed quote representations in news sources. The first limitation that we encountered involved the project pipeline. Initially, we decided to create this pipeline by finding a ground truth quote that was reported across multiple news sources. However, we determined that this is not a common occurrence. Therefore, we decided to adjust our approach and built the pipeline that was previously discussed in the implementation section of this report. The second limitation that hindered our work involved News API and Google API usage restrictions. For example, News API only allows you to search for news articles that were published within the past month unless you have a paid subscription and Google API limits the amount of requests that you can make within a given period of time. Additionally, another limitation we encountered was the quality of the results generated by using Google API to search for similar quotes. To illustrate, Google API returned news articles related to hair products when searching for news articles on www.bbc.com using the keywords "Trump and impeach." Moreover, we had planned on evaluating mixed quote similarity using cosine similarity. However, the conversion of the text to vectors with spaCy took too long for this similarity metric to be feasible, resulting in its exclusion from this project.

Future Work

If we obtain more funding, we could greatly improve upon this project. First, we would use the paid version of News API to simplify the project pipeline and obtain more news articles. Second, we would implement parallel processing to significantly improve processing speed. To expand upon this project, topic modeling would be of interest to further assess if there are specific topics that result in mixed quoting differences/similarities between various news sources.

References

Caplan, J. (2019, October 7). Judge Rules Trump Must Hand Over Tax Returns to NY Prosecutors. Retrieved from https://www.breitbart.com/politics/2019/10/07/federal-judge-rules-trump-must-hand-over-tax-returns-to-new-york-prosecutors/.

Colvin, J. (2019, November 21). Trump's 'no quid pro quo' call open to interpretation. Retrieved from https://apnews.com/8dcdacf5a02245e8a57bc9aec00a798c.

Fahrenthold, D., Marimow, A., (2019, October 7). Federal judge rules Trump must turn over his tax returns to Manhattan DA, but Trump has appealed. Retrieved from https://www.washingtonpost.com/politics/federal-judge-rules-trump-must-turn-over-his-tax-returns-to-manhattan-da-but-trump-has-indicated-he-will-appeal/2019/10/07/29e1fda6-e8a4-11e9-85c0-85a098e47b37_story.html.

Iacobelli, F., Nichols, N., Birnbaum, L., & Hammond, K. (2012). Information Finding with Robust Entity Detection: The Case of an Online News Reader. Studies in Computational Intelligence Human-Computer Interaction: The Agency Perspective, 375–387. doi: https://doi.org/10.1007/978-3-642-25691-2_16.

Kurtz, H. (2019, November 21). Sondland declares quid pro quo, pundits call testimony damaging to Trump. Retrieved from https://www.foxnews.com/media/sondland-declares-quid-pro-quo-pundits-call-testimony-damaging-to-trump.

Marimow, A. E., & Fahrenthold, D. A. (2019, October 7). Federal judge rules Trump must turn over his tax returns to Manhattan DA, but Trump has appealed. Retrieved from https://www.washingtonpost.com/politics/federal-judge-rules-trump-must-turn-over-his-tax-returns-to-manhattan-da-but-trump-has-indicated-he-will-appeal/2019/10/07/29e1fda6-e8a4-11e9-85c0-85a098e47b37_story.html.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., & Honnibal, M. (2012). A Sequence Labelling Approach to Quote Attribution. Association for Computational Linguistics, 790–799. Retrieved from https://www.aclweb.org/anthology/D12-1072/?CFID=171106781&CFTOKEN=464fe32061b6c068-8C44A7AB-A71F-2821-0404FC7E347B2A2E.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., & Koprinska, I. (2013). Automatically Detecting and Attributing Indirect Quotations. Association for Computational Linguistics, 989–999. Retrieved from https://www.aclweb.org/anthology/D13-1101/.

Pavllo, D., Piccardi, T., & West, R. (2018). Quootstrap: Scalable Unsupervised Extraction of Quotation–Speaker Pairs from Large News Corpora via Bootstrapping. Association for the Advancement of Artificial Intelligence. Retrieved from https://arxiv.org/abs/1804.02525.

Pollak, J. B. (2019, November 20). Gordon Sondland: There Was a 'Quid Pro Quo' -- for White House Meeting. Retrieved from https://www.breitbart.com/national-security/2019/11/20/gordon-sondland-there-was-a-quid-pro-quo-for-white-house-meeting/.

Zurcher, A. (2019, November 20). Impeachment inquiry: A bombshell for President Trump. Retrieved from https://www.bbc.com/news/world-us-canada-50495289.