Analyzing the Framing of 2020 Presidential Candidates in the News

Audrey Acken
Nueva School
audrey.acken@gmail.com

Dorottya Demszky
Stanford Linguistics
ddemszkystanford.edu

Abstract

It is well known that there is inherent bias in media, and that bias permeates the political news coverage in the United States. In this study, we apply NLP methods as a framework to learn about the biases in the framing of the 2020 Democratic Presidential candidates in news media. We use both a lexicon-based approach and word embeddings to analyze how candidates are discussed in news sources with different political leanings. Our results show significant differences in the framing of candidates across the news sources along several dimensions, including valence, arousal, dominance, power and agency, paving the way for a deeper investigation.

1 Introduction

Understanding the media coverage of presidential candidates is important in today's increasingly divided political world. Using articles written about the 2020 Democratic candidates, we use NLP to analyze the differences in candidate coverage across various news sources with different political leanings and answer two main research questions. First, do we observe a significant difference across sources in terms of how they discuss candidates? Second, are the differences in candidate framing consistent across different sources? If so, what are the main differences and similarities we observe, both across candidates and sources?

2 Related Work

Previous work has applied NLP to study political framing in the news media (Boydstun et al., 2013; Baumer et al., 2015; Field et al., 2018). Within the political domain, issues surrounding the presidential elections have been studied from many perspectives using NLP. This line of work includes analyses of the media portrayal of presidential debates (Tan et al., 2018), the language used in presidential debates (Prabhakaran et al., 2014; Wang et al., 2017) and social media coverage of the candidates (Jahanbakhsh and Moon, 2014). Even though NLP methods have been used in various contexts to study the framing of groups of people (Sap et al., 2017; Field et al., 2019), to our knowledge, there is limited research into the framing of presidential candidates in the news media.

3 Methods

Data. Our dataset consists of text from 12,464 news articles discussing 7 of the 2020 Democratic presidential candidates: Joe Biden, Pete Buttigieg, Kamala Harris, Amy Klobuchar, Bernie Sanders, Elizabeth Warren, and Andrew Yang. The data includes articles published from July 2019 to January 2020, across 22 different news sources, categorized into 5 groups based on their political leaning: left, left-leaning, center, right-leaning and right, based on the Media Bias Chart created by Ad Fontes Media.[1]

We collect the articles via the News API[2] and scrape them using Beautiful Soup.[3] We pre-process the

[1] https://www.adfontesmedia.com/how-ad-fontes-ranks-news-sources/?v=402f03a963ba

[2] https://newsapi.org/docs/client-libraries/python

[3] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

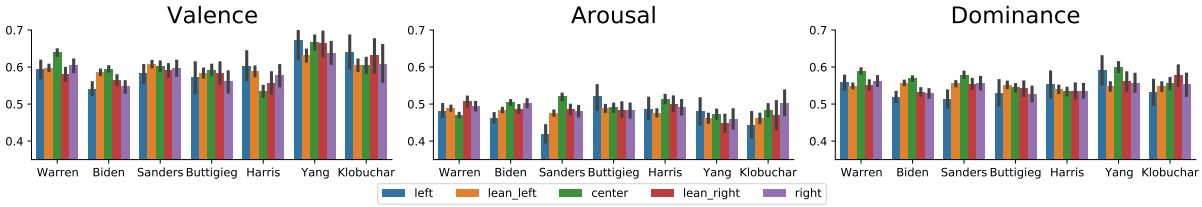

Figure 1: Valence, arousal and dominance scores by candidate and the news sources' political leaning.

data by standardizing the candidates' names to their last names and perform coreference resolution and dependency parsing to extract verbs and adjectives used to describe each of the candidates via SpaCy.[4]

Lexicons. Lexicon-based approaches can be used to understand the affective and social connotations of words. We apply the NRC Valence, Arousal, and Dominance lexicon (Mohammad, 2018) to verbs and adjectives and the Connotation Frames lexicon of Power and Agency to lemmatized verbs (Sap et al., 2017) to estimate core dimensions of word meaning relevant for understanding the framing of people. The metrics we chose (valence, arousal, dominance, power, and agency) are related to the primary affective dimensions identified in social psychology: potency (strong vs. weak), valence (good vs. bad), and activity (active vs. passive). As examples, a high valence adjective is *love*, a low arousal one is *asleep*, a high dominance one is *powerful*, a low power verb is *beckons*, and a low agency one is *obeys*.

Word embeddings. To learn more about semantic similarities in our data, we create word embeddings. We learn our own embeddings as opposed to using pretrained ones, since we want to understand semantic patterns in our data. To obtain statistically robust measures of word similarity, we use bootstrapping (Antoniak and Mimno, 2018), which involves estimating the cosine similarity of words across several word embedding models trained on different subsets of the original dataset. We use the word2vec skip-gram model (Mikolov et al., 2013) to train separate word embedding models on each of the 22 sources.[5]

4 Results

Our results suggest a greater framing difference between candidates (*who* is being described) than between sources (*who* is describing) – see Figure 1. An ordinary least squares regression shows that, controlling for source, Harris is described with the lowest valence (0.58, $p < 0.001$) and Yang is described with the highest sentiment (0.66, $p < 0.001$). Interestingly, the ranking of candidates in terms of dominance is very similar to their sentiment ranking. Contrarily, Yang is described with the lowest arousal (0.45, $p < 0.001$), whereas Harris, Biden and Buttigieg are described with the highest arousal (0.48, $p < 0.001$). Looking at power and agency, Yang is described with the least power (0.59, $p < 0.001$) and lowest agency along with Klobuchar (0.81, $p < 0.01$), while Biden and Sanders are described with the highest power (0.67, $p < 0.001$) and Buttigieg with the highest agency (0.85, $p < 0.001$). However, the story is not very clear-cut for either of these dimensions, as there is a significant interaction between certain sources and candidates. Harris, for example, is described with significantly lower sentiment by center-leaning sources than by other media sources. While there are differences across news sources, there is no significant trend among left or right leaning news source groups. Instead, the fluctuations in Valence, Arousal, and Dominance between news sources seems dependant on the specific candidate.

Looking at the word embedding results gives insight into the differences in types of words used to describe candidates. For Biden, we find that the most similar words are associated with his political history – e.g., the phrase *former vice* is among the top 5 closest cosine similarities. For Warren, the highest cosine similarity words tend to be policy-related (*medicare for all*, *wealth tax*) and for Sanders, we see that the most similar words are more related to personality characteristics (*lie*, *honest*). We also

[4] `https://spacy.io/`

[5] Using contextualized word representation models (Devlin et al., 2018) was not an option, given that bootstrapping with those large models would have been very expensive computationally.

see trends across news source leaning groups. For right leaning sources, the words *gaining*, *slow* and *lie* are among the top 10 closest words to many candidates, but not for left-learning sources.

5 Discussion

Our research shows that there are significant differences in the ways in which presidential candidates are described in the news. We also find some differences in the ways sources from different leaning groups frame the candidates. The biases in sources may also may be a result of the economic reality that some news sources tailor their stories to appeal to their readers' biases. Expanding the size of the data set, both in terms of the number of articles and candidates, and including data beyond the date range we used could lead to deeper analyses. In addition, going beyond these lexicons and using contextual representations to better understand the framing of candidates would be fruitful avenue of future work. Finally, studying how these findings change temporally and in relation to debates or the impeachment trial could lead to interesting insights.

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482.

Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2013. Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop, London*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. In *2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online# metoo stories. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 158–169.

Kazem Jahanbakhsh and Yumi Moon. 2014. The predictive power of social media: On the predictability of us presidential elections using twitter. *arXiv preprint arXiv:1407.0622*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1486.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334.

Chenhao Tan, Hao Peng, and Noah A Smith. 2018. " you are no jack kennedy" on media selection of highlights from presidential debates. In *Proceedings of the 2018 World Wide Web Conference*, pages 945–954.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.