**Dataset Metadata**

|  |  |
| --- | --- |
| **Metadata Field** | **To be completed by Editorial** |
| Dataset title | Using Word Frequencies to Analyze Political Language and Moral Focus |
| Final zip file name | dataset-word-frequency-analyse-political-language-moral |
| Author/contributor(s) | Erin M. Buchanan |
| SMART author ID(s) | 753327 |
| Author/contributor(s) | William E. Padfield |
| SMART author ID(s) | 753328 |
| Content type | Qualitative |
| Exemplar type  (qual only) | Text |
| Downloadable files? (binaries) | Yes |
| If yes, list downloadable file name(s) and type(s) | dataset-word-frequency-analyse-political-language-moral-exemplar1-news-articles.csv  dataset-word-frequency-analyse-political-language-moral-exemplar2-immigration.csv |
| Display/link text for each Downloadable file | Exemplar 1—US News Articles  Exemplar 2—Immigration and Political Party |
| Downloadable file(s) introductory text | Here you can download the two dataset exemplar spreadsheets. Exemplar 1 is a worked example of word frequency analysis, whereas Exemplar 2 provides an opportunity to try out the analysis yourself. |
| Abstract | These data were collected to explore the use of moral words, as formulated in the Moral Foundations Dictionary (Graham, Haidt, & Nosek, 2009; e.g., abuse, fair, sacred), in political news articles. The data were gathered from four popular U.S. news websites (*National Public Radio*, *The New York Times*, *Fox News*, and *Breitbart*) because of their known political affiliations. This example focuses on how a researcher can turn qualitative data, such as news articles, into a measurable outcome through word frequency analysis. By analyzing each source’s political language through word frequency, you can examine trends in many psychological topics. This example focuses on morality and moral language to provide examples of differences in political rhetoric across party affiliation. The dataset files are accompanied by a Teaching Guide and a Student Guide. |
| Dataset tab text | You can view and download the data exemplar(s) in this tab. |
| Teaching and Learning Material tab text | In this tab you will find guides on using this dataset. The Teaching Guide is designed for Faculty who are teaching research methods and statistics, with suggestions on how to use the dataset in lab exercises, in homework assignments, and as exam questions. The Student Guide introduces the method for students and can be used in teaching to provide students with an introductory overview of the method or test. |
| Teaching Guide text | These data were collected to explore the use of moral words, as formulated in the Moral Foundations Dictionary (Graham, Haidt, & Nosek, 2009; e.g., abuse, fair, sacred), in political news articles. The data were gathered from four popular U.S. news websites (*National Public Radio*, *The New York Times*, *Fox News*, and *Breitbart*) because of their known political affiliations. This example focuses on how a researcher can turn qualitative data, such as news articles, into a measurable outcome through word frequency analysis. By analyzing each source’s political language through word frequency, you can examine trends in many psychological topics. This example focuses on morality and moral language to provide examples of differences in political rhetoric across party affiliation. Full details about the dataset and more information about the original dataset can be found on the “Datasets Info” tab.  This dataset and the accompanying guides can be used in a classroom or can form the basis of an exam question or homework exercise. The data file is accompanied by a Student Guide, which explains the method and includes a worked example, and finishes with exercises and discussion questions that the student can do on their own. The Student Guide can be shared with students by e-mail using the “Share” button, can be embedded into your Learning Management System or Virtual Learning Environment, or can be downloaded and printed to use in the classroom. |

|  |  |
| --- | --- |
| **Metadata Field** | **To be completed by Production** |
| Product name | SAGE Research Methods Datasets Part 2 |
| Publication date | 2019 |
| Copyright line | © 2019 SAGE Publications Ltd |
| DOI | 10.4135/ |
| ISBN |  |
| Product Code |  |
| SPIN ID |  |
| URL |  |
| URI |  |

|  |  |  |
| --- | --- | --- |
| **Metadata Field** | **Description/explanation** | **To be completed by Contributor** |
| Author/contributor bio(s) |  | Dr. Erin M. Buchanan is an associate professor of Quantitative Psychology at Missouri State University, and her research focuses on computational linguistics, statistics, and scientific practice. She received her undergraduate degree from Texas A&M University and graduate degrees from Texas Tech University. Her work explores how to best model and calculate information gleaned from large quantities of text to understand the semantic memory system. Her research into statistics focuses on best practices, improvement of statistical literacy, and how to answer complex applied questions with simple statistical answers. You can learn more about her research at her website: <https://www.aggieerin.com/>.  William E. Padfield is a master’s degree candidate at Missouri State University. He earned his BS in Psychology at Missouri State University. His research largely focuses on the moral and linguistic aspects of political discourse, especially that of the media. Specifically, his research interests include moral foundations theory, attentional focus in congressional speeches, and semantic priming. |
| Discipline(s) | i.e. Those disciplines covered by dataset and guides. A dataset may have multiple subject areas. | Psychology [D3]  Political Science and International Relations [D8]  Communication and Media Studies [D13] |
| Data type | Please choose ‘Other’ if your data type is not listed, and add the new data type underneath | Website Posts |

**About This Dataset info**

|  |  |  |
| --- | --- | --- |
| **Meta Data Field** | **Description/explanation** | **To be completed by Contributor** |
| Data source citation |  | Padfield, W. E., & Buchanan, E. M. (2018, September 4). Moral foundations of U.S. political news organizations. Retrieved from osf.io/5kpj7. doi:10.17605/OSF.IO/5KPJ7 |
| Full title of originating dataset |  | Moral Foundations of U.S. Political News Organizations |
| Data author(s) and affiliations |  | William E. Padfield and Erin M. Buchanan |
| Dataset source website address |  | Taken from:  <https://www.npr.org/>  <https://www.nytimes.com/>  <http://www.foxnews.com/>  https://www.breitbart.com/ |
| First publication date |  | This dataset is published online, and we are working on the paper associated with it. |
| Data Universe |  | N/A |
| Funding sources/suppliers | Names of funding bodies, including grant numbers and related acknowledgements | N/A |
| Sample/sampling procedures | If readily available | We used web scraping to collect the data. |
| Weighting | If appropriate and available | N/A |
| Data collection dates |  | February 2018 to May 2018 |
| Time frame of analysis |  | May 2018 to current |
| Unit of analysis | Technical term for who or what is being studied | Word frequency analysis |
| Location covered by data | Location from which the data was gathered | USA |
| Other sources | If dataset is drawn from other secondary sources | N/A |
| Links to SRM content |  | N/A |
| List of variables | Where available, practical and accessible | Source, URL, Text, Processed Text, Word Count, Moral Foundations Percentages |
| Abbreviations, conventions or notation devices |  | MFD = Moral Foundations Dictionary  MFT = Moral Foundations Theory |

Data Exemplar

Exemplar 1 – U.S. News Articles

**Data collected by:** Erin M. Buchanan and William E. Padfield

This dataset consists of gathered text from four notable U.S. news sources compiled into a spreadsheet for further processing and analysis. The sources included in this research include *The New York Times*, *National Public Radio (NPR)*, *Fox News*, and *Breitbart*. The researchers specifically scraped political news coverage and commentary from the denoted politics pages on each website, as political coverage was suspected to include discussion of moral topics. The dataset includes the website link for each article (**URL**), examples of the processed data (**Processed**), and word counts from the overall article (**WordCount**), and each of the moral foundations areas created by Graham, Haidt, and Nosek (2009; **HarmSum, FairSum, IngroupSum, AuthoritySum, and PuritySum**).

Moral Foundations Dictionary (MFD)

The MFD includes lists of words that are associated with each of the moral foundation areas. To create the **Sum** columns described in the exemplar dataset, you would need to know the list of words for foundation area. Table 1 includes the words used for each area in their unstemmed (i.e., full word) format. In the analysis, you would process these words like the text sources, so that you could match the root word stems in the processed text to the word stems for the dictionary. For example, you would count the number of times that *abus* appeared in the text source for the Harm category, as *abus* is the processed text for *abuse*. The number of times these words appeared would be included in the **Sum** columns in the dataset, and you would create a sum of the number of times all the words from each category appeared.

Table 1: Moral Foundations Dictionary Word Lists.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Harm** | **Fair** | **Ingroup** | **Authority** | **Purity** |
| abuse | balance | collect | abide | abstain |
| attack | bias | community | authority | adultery |
| benefit | constant | deceive | class | church |
| care | discriminate | desert | command | clean |
| cruel | equal | family | control | dirt |
| crush | equate | fellow | defect | disease |
| damage | even | foreign | defer | disgust |
| danger | exclude | group | defy | gross |
| defend | fair | individual | desert | innocent |
| destroy | favor | member | duty | modest |
| fight | honest | nation | faith | preserve |
| guard | impartial | side | father | promiscuous |
| harm | just | together | honor | pure |
| hurt | justify | trait | law | right |
| kill | prefer | unite | lead | ruin |
| preserve | prejudice |  | legal | sacred |
| protect | reason |  | mother | sick |
| ruin | right |  | obey | sin |
| safe | tolerate |  | oppose | trash |
| shelter |  |  | order | whole |
| spurn |  |  | permit |  |
| stomp |  |  | position |  |
| suffer |  |  | preserve |  |
| sympathy |  |  | protest |  |
| violent |  |  | refuse |  |
| war |  |  | respect |  |
|  |  |  | revere |  |
|  |  |  | serve |  |
|  |  |  | tradition |  |
|  |  |  | trait |  |

Exemplar 2 – Immigration and Political Party

**Data collected by:** William E. Padfield

The dataset includes 30 congressional speeches covering the nature of immigration given to the House and Senate in more recent (2017 and later) months. These speeches were coded by political party, Republican, or Democrat, to allow for a similar moral analysis as described above, as political lean is expected to influence speech and writing patterns.

This dataset would allow you to apply moral foundations theory (MFT) to current political trending topics to determine whether there are differences in moral speech across political party. Immigration has always been a contested topic, and with the current political climate, this topic appeared to be an apt area to explore speech and discourse patterns for distinctions between Republican and Democratic Congress people. Given the research on MFT, we might expect to find that Democratic speakers to use more individualizing foundations of *harm/care* and *fairness/reciprocity*, while the Republican speakers to use more *binding* foundations such as *ingroup/loyalty*, *authority/respect*, and *purity/sanctity*.

The dataset provided gives you an opportunity to try a word frequency analysis to determine whether these differences exist between parties. The **Source** column includes the political party of the original speaker, and the **URL**column includes a link to the Library of Congress website where the data were found. The **Text** column is the unprocessed data from the Library of Congress that you would use to start your analysis.

Student Guide

Introduction

This example demonstrates the use of various techniques for the purpose of gathering, processing, and analyzing text from various news organizations’ websites in order to understand their moral content. Techniques include collecting text from the web (i.e., web scraping) using the *rvest* library in the R statistical package, word stemming with the *ngram* library for R, and word frequency analysis (Tausczik & Pennebaker, 2010). Libraries and packages are specialized plug-ins for R, which researchers have written to allow others to perform analyses, like the ones described here. While we used this specialized programming language, the data could be any text data, collected by copying/typing or otherwise, and the analyses shown here can be created in any software that allows for sorting and counting words, like Excel.

This qualitative analysis is intended to reveal the moral and political qualities of the news text in order to discover whether or not a U.S. news organization’s ideological lean (i.e., conservative or liberal) influences the endorsement of several moral foundations as described in moral foundations theory (MFT; Graham et al., 2011). Specifically, the researchers investigated whether or not news organizations of divergent political alignments tended to endorse differing moral foundations through their use of language in news text. William E. Padfield, a master’s degree candidate in psychology at Missouri State University, and Dr. Erin M. Buchanan, associate professor of psychology at Missouri State University, conducted this research.

Moral Foundations Theory

At its core, MFT attempts to explain the totality of different people’s moral alignments. Specifically, MFT seeks to illuminate the differences between political conservatives’ and liberals’ morals (Graham et al., 2011). These differences are established through the measure of individuals’ endorsement of five moral foundations.

The first two foundations—*harm/care* and *fairness/reciprocity*—represent concern for individual-focused social justice and equality. These two foundations can be conceptualized as the *individualizing* foundations. The following three—*in-group/loyalty*, *authority/respect*, and *purity/sanctity*—represent perceptions of right and wrong from a group-level perspective and can be thought of as the *binding* foundations (Haidt & Graham, 2007). Research (described below) indicates that political liberals tend to endorse the two individualizing foundations above all others, while conservatives tend to endorse all five foundations with greater endorsement of the binding foundations (Graham, Haidt, & Nosek, 2009).

Graham et al. (2009) developed the moral foundations dictionary (MFD) in order to determine endorsement of the five foundations in speech and text. The MFD consists of roughly 50 words per foundation that exemplify their meaning. For example, words such as *abuse* and *protect* indicate endorsement of the *harm/care* foundation. Graham et al. (2009) then validated the MFD dictionary lists by analyzing the speech content of liberal and conservative church sermons. They found liberal sermons endorsed the individualizing foundations and conservative sermons endorsed all five moral foundations.

Data Source

In an era in which political divides appear to run deeper, news is obtained from more sources than ever, and perceptions of the truth seem to follow ideological lines. Thus, it becomes important for the research community to discover and communicate the nature of the news that people consume. The extraordinary nature of the current political landscape in the US, and the vastly divergent political stances assumed by certain news outlets, drew the researchers to this project.

For a period of several weeks, the researchers gathered text from four notable U.S. news sources and compiled it into a dataset for further processing and analysis. The sources included in this research include *The New York Times*, *National Public Radio (NPR)*, *Fox News*, and *Breitbart*. The researchers decided to analyze these sources owing to their widespread recognition among the general American public as well as the fact that they are easy to categorize in accordance with perceived political lean. According to popular belief, *The New York Times* and *NPR* are often perceived as more liberal leaning, while *Fox News* and *Breitbart* are likely seen as being more conservative (Mitchell, Gottfried, Kiley, & Matsa, 2014). The researchers specifically scraped political news coverage and commentary from the denoted politics pages on each website, as more general or human-interest stories were believed to lack the moral perspectives of interest.

Word Frequency Analysis

Stage 1: Data Collection

A key component to understanding the way humans talk to each other is collecting samples of discourse or large amounts of text. Our research hypothesis focused on how people writing for specific political audiences would alter their language to fit within the moral foundations that those audiences should want to read about. Therefore, we picked four well-known news websites that were either conservative (*Breitbart*, *Fox News*) or liberal (*New York Times*, *NPR*) to explore their discourse. Over the course of a month, we downloaded every article in their specific news sections that focused on political coverage from U.S. news to foreign policy. We used specialized software to help with this process, but often these data are collected by simply cutting and pasting each document into text format that you can use later. In the provided data, you can see the **Source** of the data, the link of the article we used (**URL**), and the full **Text** of that article. (The full text has been omitted for the purposes of this SAGE Research Method Dataset, but you can use the URLs to find the original articles.) This part of the dataset constitutes the raw discourse that we used for word frequency analysis.

Stage 2: Stemming, Counting, and Creating Percentages

One unique problem with analyzing language is that each concept or word has multiple forms, such as *walk*, *walked*, and *walking*. A researcher may want to strip these affixes off the text data collected to be able to combine words with similar meanings for analysis, a process called stemming. We stemmed the data to create the **Processed** column found in the attached dataset. You can try stemming any document at <https://text-processing.com/demo/stem/>, which is a website devoted to different options available for creating root word forms. We used the English option under the Snowball stemmer when stemming our data. However, we knew that automatic stemming is not a perfect process. For example, *scientist* is a person who studies science and likely should be combined into a global *science* word form. If you try the example option on the text processing web page, this word does not change when processed through the stemmer. We handled these unique word forms in the next stage of analysis by making sure all word forms were included in our dictionary.

As mentioned earlier, we used the MFD to know what words to look for within the text that we collected from the news websites. For example, in the *harm/care* category, we looked for concepts such as *attack* and *protected*, while the *fairness/reciprocity* category included words such as *equal* and *bias*. Our original data source was stemmed, and therefore, in this step, we stemmed the dictionary words, so they would match the **Processed** data. Because we knew that stemming was not a perfect procedure, we added all forms of each root word to make sure to capture that concept such as *equal*, *equate*, *equals*, *equality*, and so on. This procedure sometimes produced duplicate dictionary stems, so those were discarded to only look for each concept once.

Now that both data sources (the **Processed** data and the dictionary list) were stemmed, we proceeded to create a frequency count of all words in each text source. The total **WordCount** is included to help control for the differences in article length within and between publications, as some articles were short summaries, while others were longer in form. The words used in each article were then compared to the dictionary words, and only those words were selected out of the larger set. For example, *equal* was included for the total words for *fairness*, while other words such as *another*, *the*, and *over* would be ignored. Last, we created a percentage of each foundation area found in the **Harm, Ingroup, Purity, Fairness,** and **Authority** percentage columns. The sum columns are also provided as the intermediate step between calculating the total words in each category before creating the percentages. Figure 1 provides an example of the dataset to help visualize how the data were laid out and processed during analysis.

Figure 1 shows you an example of the columns in the dataset. The **Text** column includes the original text of the article before it was stemmed, as described above. The **Processed** column is the text after stemming. The **WordCount** column includes the total number of words in the article, which is used to create the Percent column for each moral foundation. For example, the first article used two words in the Harm category (**HarmSum**), and therefore, the **HarmPercent** was calculated by dividing two by the **WordCount** column.

Figure 1: Example of the Columns in the Dataset.

[insert Figure 1 here]

Stage 3: Data Visualization

At this point, we have turned qualitative discourse data into a quantifiable percentage. These percentages represent the portion of text’s words devoted to each moral foundation. Given that the text data have now been converted to number data, we could use traditional quantitative statistics, such as the analysis of variance or ANOVA, to determine whether there are differences in percentages across the sources and foundation areas. We focused on visualizing the data, shown in Figure 2, since this analysis was exploratory.

In examining the data for the *Harm* foundation, we see that the *New York Times* uses the least amount of *harm words*, while the other three have very similar means. The fairness foundation showed that *NPR* likely used the least *fairness words*, while the in-group foundation showed that the *New York Times* represented the largest amount of in-group foundation discourse. The authority foundation showed very similar scores for the conservative news sources, while the liberal sources used fewer of these words. Finally, the purity foundation was the least used foundation, with only a few words found for each of the news sources and liberal sources being somewhat lower than conservative sources. However, in each of these comparisons, we found that the error bars (which the typical range of the percentages in each article) mostly overlapped for each news source, which indicates that they are likely not very different from each other since each news source tended to use around the same amount of moral words.

Figure 2: Percentages for Each of the Conservative and Liberal News Articles for the Harm, Fairness, Ingroup, Authority, and Purity Moral Foundations.

[insert Figure 2 here]

In Figure 2, the *X*-axis indicates the news source, along with their political lean, and these bars are arranged by moral category. The *Y*-axis, or height of the bars, indicates the percentage of the article that included the specific moral words. The error bars, or the capped lines on the bar, represent how much these percentages typically vary across articles.

Summary

In summary, we first collected traditional qualitative type data by saving news articles from four popular online news web pages. In order to understand political discourse, we connected the MFT and the MFD to these news sources. The discourse data were transformed into word frequency data by firstly stemming the data to combine similar words together into one related concept, and secondly, counting similar words. We then narrowed down the data into only words found in the MFD to guide our analysis, calculating the percentage of moral words in each category for each article. Finally, we created a bar graph to visualize whether there were differences in moral language for the political news sources. Some small differences occurred, but the overall percentage of words used in each category was very low. Future research could explore the nature of the MFD and see whether related concepts could be added to expand the dictionary to create a fuller picture of words used across a range of types of discourse.

Reflective Questions

1. Word frequency analysis was applied here to political texts using a framework of MFT, and therefore, only specific “moral” words were examined. What other ways might you use word frequency analysis to elucidate differences in writing styles?
2. Word stemming was a crucial concept in this analysis, as it allows you to combine words with the same meaning into one term for analysis. In what situations would you want to use each word in its raw form?
3. In this analysis, political articles from four news sources aimed at a general audience were gathered. What other kinds of text sources would be appropriate for analyses such as this? Are there situations in which you would want to utilize more niche or extreme sources?
4. This analysis takes place firmly at the intersection of moral and political psychology. In what other domains of inquiry would you be able to utilize word frequency analysis, web scraping, and word stemming?

**Further Readings**

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology,* *96*(5), 1029–1046. Retrieved from <https://doi.org/10.1037/a0015141>

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(03), 267–297. Retrieved from <https://doi.org/10.1093/pan/mps028>

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research,* *20*(1), 98–116. Retrieved from <https://doi.org/10.1007/s11211-007-0034-z>

Kertzer, J. D., Powers, K. E., Rathbun, B. C., & Iyer, R. (2014). Moral support: How moral values shape foreign policy attitudes. *The Journal of Politics, 76*(3), 825–840. Retrieved from <https://doi.org/10.1017/S0022381614000073>

Koleva, S. P., Graham, J., Iver, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality, 46*(2), 184–194. Retrieved from <https://doi.org/10.1016/j/jrp.2012.01.006>

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review, 32*(2), 132–144. Retrieved from <https://doi.org/10.1177/0894439313506837>

Slatcher, R. B., Chung, C. K., Pennebaker, J. W., & Stone, L. D. (2007). Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality, 41*(1), 63–75. Retrieved from <https://doi.org/10.1016/j.jrp.2006.01.006>

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology,* *29*(1), 24–54. Retrieved from <https://doi.org/10.1177/0261927X09351676>

**References**

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046. Retrieved from <https://doi.org/10.1037/a0015141>

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*(2), 366–385. Retrieved from <https://doi.org/10.1037/a0021847>

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20*(1), 98–116. Retrieved from <https://doi.org/10.1007/s11211-007-0034-z>

Mitchell, A., Gottfried, J., Kiley, J., & Matsa, K. E. (2014). Political polarization & media habits. *Pew Research Center: Journalism & Media*. Retrieved from <http://www.journalism.org/2014/10/21/political-polarization-media-habits/>

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. Retrieved from <https://doi.org/10.1177/0261927X09351676>