**Title**

**Using Word Frequencies to Analyze Political Language and Moral Focus**

**Abstract**

This data was collected to explore the use of moral words (e.g. abuse, SOMETHING, SOMETHING) in political news articles. The data was gathered from four popular news websites, *National Public Radio*, *The New York Times*, *Fox News*, and *Breitbart*, because of their known political affiliations. This example focuses on how one can turn qualitative data, such as news articles, into a measurable outcome through a word frequency analysis. By analyzing each source’s political language this way, 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 data file is accompanied by this teaching and student guide.

**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 web scraping using the *rvest* library in the *R* programming language, word stemming with the *ngram* library, and word frequency analysis (CITE, 20XX). While we used a 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 news organization’s ideological lean (i.e., conservative or liberal) influences the endorsement of several moral foundations as described in Moral Foundations Theory (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, Moral Foundations Theory (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, *ingroup/loyalty*, *authority/respect*, and *purity/sanctity* represent perceptions of right and wrong from a group-level perspective. These three can be thought of as the *binding* foundations (Haidt & Graham, 2007). Research 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.

The researchers 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) validated the MFD wordsets by analyzing the speech content of liberal and conservative church sermons. They found liberal sermons endorsing the individualizing foundations and conservative sermons endorsing all five.

**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, it becomes incumbent upon the research community to discover and communicate the nature of the news people consume. The extraordinary nature of the current political landscape 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 US 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 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* lean more conservative. The researchers specifically scraped political news coverage and commentary, as more general or human-interest stories were believed to lack the moral perspectives of interest.

**Analysis:**

**Word Frequency Analysis**

**Stage 1: Data Collection**

A key component to understanding the way human’s 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 for 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 this data is 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. This part of the dataset constitutes the raw discourse that we used for word frequency analysis.

**Stage 2: Stemming, Counting, Creating Percentages**

One unique problem with analyzing language is that each concept or word has multiple forms, such as *walk, walked,* and *walking*. Often we want to strip these affixes off the text data collected to be able to combine words with similar meanings for analysis, and this process is called stemming. We stemmed the data to create the **Processed** column. 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. As you can see if you try the example option on the text processing webpage, 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 Moral Foundations Dictionary 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,* etc. 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

* Then we counted up all the words
* Then we figured out how many of those words were in the dictionary
* That creates percentages that you can see in column X Y Z.

**Stage 3: Data Visualization**

What do we want to do here? Simple ANOVA of all four? Just create some graphs? What?

* Put in panel graph or some graph thing here (could separate or panel by lean)

**Summary**

It is always useful to finish the dataset with a section that sums up what the user has learnt and ties everything together.

**Reflective Questions (3-4 questions) [H1]**

Please include 3-4 questions that allow students to reflect on issues brought up in the exemplar. As much as possible, these questions should also have a ‘review’ function (i.e. they draw attention to the purpose of the example and then ask students to think further about these methodological questions). Datasets are a practical teaching tool, so these questions must also challenge the student to attempt the method or analysis using the data you have provided.

**Further Reading (6-8 references, please don’t be afraid to advertise your own work here) [H1]**

**Data Exemplar [H1]**

The data you provide will appear on the platform alongside your narrative on the analysis – it will also be available to download. For example, depending on your research this could be - photographs, a skype conversation or an interview transcript, a video, archive material, ethnographic field notes or documents. This data is the illustrative example that used to provide a description of how one might go about analysing in the Analysis section on the dataset.

If possible, alongside this we would like some ‘clean’, unanalysed extra data which users can have a go at analysing after reading the exemplar as part of the questions in the reflective question section. For example, “Considering all the data available, what themes can you see emerging from the data?”

The data should be presented as follows:

**Data collected by:** [insert name of researcher here]

[A brief explanation of presented data so the user knows what they are looking at (<100 words) e.g: “The data reproduced here is the transcript of an elite interview investigating XYZ”]

[Present the data. Please use headings and sub-headings to separate sections of data if necessary or if aids in reference from the narrative e.g:]

**Exemplar 1 [H2]**

**Sub-heading [H3 heading level]**

**Extra Data for Use With Reflective Questions [H2]**

**Sub-heading [H3]**

**METADATA**

|  |  |  |
| --- | --- | --- |
| **Metadata Field** | **Description/explanation** | **To be completed by Contributor** |
| Author /Contributor biographies | Brief (<150 words) academic biography of EACH author |  |
| Discipline(s) | i.e. Those disciplines covered by dataset and guides. A dataset may have multiple subject areas. | Psychology  Political Science and International Relations  Communication and Media Studies  Choose an item.  Choose an item. |
| Data type | Please choose ‘Other’ if your data type is not listed, and add the new data type underneath | Website Posts |
| Prerequisites | Quant only |  |

**About This Dataset info**

|  |  |  |
| --- | --- | --- |
| **Meta Data Field** | **Description/explanation** | **To be completed by Contributor** |
| Data Source Citation |  |  |
| Full title of originating dataset |  |  |
| Data author(s) and affiliations |  |  |
| Dataset source website address |  |  |
| First publication date |  |  |
| Data Universe |  |  |
| Funding sources/suppliers | Names of funding bodies, including grant numbers and related acknowledgements |  |
| Sample/sampling procedures | If readily available |  |
| Weighting | If appropriate and available |  |
| Data collection dates |  |  |
| Time frame of analysis |  |  |
| Unit of analysis | Technical term for who or what is being studied |  |
| Location covered by data | Location from which the data was gathered |  |
| Other sources | If dataset is drawn from other secondary sources |  |
| Links to SRM content |  |  |
| List of variables | Where available, practical and accessible |  |
| Abbreviations, conventions or notation devices |  |  |