Noticing Word Variance And Frequency Distributions In Reporting the Dakota Access Pipeline Controversy

How Language Effects Current and Future Readers

Creating Effective Visualization Tools for Distant Reading and Learning

Hood College

HUM560K/CS599H

Emily Hardy (M.A., Humanities)

Olaoluwa Komolafe (M.S., Computer Science)

Minghua Liu (M.S., Computer Science)

**Narrative**

**Introduction**

Today’s current events are tomorrow’s history. As the old saying goes, “the winner writes the history books;” how do we figure out who the winner is when the lines are not clean-cut?

The role of the media is to provide information to the general public. It is a well-known fact that media biases exist across political lines. This project seeks to find how the type of language used in the reporting of the Dakota Access Pipeline (DAPL) controversy varies between liberal and conservative news sources. The effects of these variances inevitably has an influence on those who consume different media; if someone only receives his or her information from one media outlet, his or her view on a subject would be vastly different than someone who uses another source.

We chose to analyze DAPL news stories based on the overlapping controversies and angles that can form people’s opinions on the subject. On one hand, the DAPL can be seen as an infringement on the rights of indigenous people and an intrusion on their protected land and sacred grounds. On the other, the protests that erupted in light of the controversy can be viewed as a roadblock to American industrial project and a gross overreach of governmental oversight into the dealings of a private business entity.

DAPL is a partisan issue and decisive current event, which is why it is a perfect case study for this type of collaborative analysis. A thorough understanding of the reporting methods can help future researchers understand concurrent views and struggles. For this project, we collected news articles from the Internet across a variety of sources to find the most common words and phrases presented by the medium. The ultimate goal is to discover whether the topics discussed by the sources are indicative of the opinions of the political affiliations of their readership.

**Collection Methodology**

We collected articles from the websites of The New York Times, Washington Post, Huffington Post, Slate, Fox News, and Breitbart. Op-eds, letters to the editor, and similar solely opinion-based publications were omitted in favor of traditional journalistic news reporting, though an analysis of such work would be interesting for future research.

The sources were chosen because of where they each lie on the political spectrum. Slate and Huffington Post are generally considered to be liberal/leftist publications, The New York Times and Washington Post are centrist (outside of op-eds, which have obviously liberal and conservative commentators, which is why such forms of writing were omitted), and Fox News and Breitbart which are right-leaning and alt-right, respectively. Links to the articles were collected into Excel spreadsheets, labeled by their date.

**Technical Methodology**

Our goal is to create a tool that can be used to pull create word clouds of most used words from various news sites to be used for sentiment analysis. For this, we used the Natural Language Toolkit (NLTK) provided by the Python programming language.

The process of analyzing the word use of a news source is broken into the following steps:

1. We collect a list of internet links for articles written about the DAPL protests in a tab- delimited text file with the date posted and link.
2. The links are read into a Python list. The articles are separated by the news source from which they were acquired.
3. For each element in the link, we read the web page source code. Depending on the source, we have a different method of pulling and parsing the language. Some sources provided errors initially that required additional steps to successfully mine the text
4. The parsed language is then tokenized into the word stems which are saved into a master stem list for the current source. This eliminates overlapping words based on verb tense and other such occurances.
5. With this list of words, we create a frequency distribution that counts the use of word stems across articles for the sources. Additionally, NLTK provides a database of “stop words” so that common linguistic articles, prepositions, and conjuctions are not included within the word pool.
6. With this list of word stems and counts we create a word cloud visualization. Larger texts within the word cloud designate more commonly-occuring words, whereas smaller texts are less frequent.

The goal of developing this tool is to create a streamlined process for digital humanistic analysis that will allow any set of URLs to be parsed for research and analysis. A word count and distribution tool can allow professionals in the humanities a quick snapshot of important and common themes within a canon of work. This type of analysis is truly only possible using a distance-reading approach.



Example of word cloud visualization output

**Limitations**

While we have made great progress on this project, there are certain limitations and issues we have yet to successfully overcome that require additional time and monetary investments.

We encountered issues with the New York Times due to the limits on the number of articles we could access per month without a subscription. Without paying for a subscription to the publication, we could only view ten articles per month. In order to perform a more thorough analysis of New York Times articles, a paid monthly subscription is necessary.

In addition to having issues with limited access due to a lack of paid subscription, we have been encountering some persistant errors with accessing and mining data from the Washington Post. Specifically, we repeatedly receive 403 (16/23) errors from a few stubborn links. While we did face similar error messages from Huffington Post and New York Times, we were able to create a patch that eliminated approximately 90% of the problem. Further work must be done to achieve the goal of a more streamlined tool with more efficient accessability.

**Future Goals and Further Research**

We have several exciting plans for a more in-depth look into DAPL reporting using our word visualization tools. They involve expanding the amount of national news sources, incorporating local and community sources, and analyzing trends on social media.

While we selected a variety of common news sources for our initial project, we omitted several mainstream and alternative media outlets. There are a few reasons for this. Since we were aiming to find general trends in liberal versus conservative sources, the current selections work as a good base. Adding additional national news sites will create a larger pool and help us pinpoint common talking points. A great deal of mainstream media sites, and also a good portion of independent journalists, incorporate videos into their online news articles. This is an extraordinarily common phenomenon as we surely missed out on some interesting data using a text-only approach. For contining the project, we would have to look into methods for transcribing videoes and running their dialogue through our programs.

As previously stated, we had some road blocks due to inability to gain access to unlimited New York Times and Washington Post articles. This is also a problem when looking at local newspapers; local and community-based media outlets are often highly restrictive with online access to their articles without payment. It would be interesting to compare articles regarding DAPL on the national stage compared to articles written on the local level. This would require additional funds to pay for subsciptions to local newspapers in the parts of North Dakota, South Dakota, Iowa, and Indiana directly affected by pipeline construction and the consequent protests. We would also need to find a way to access any printed media from the Standing Rock Sioux and other Native American tribes to compare the language between mainstream and indigenous publications.

We did not include op-eds and letters to the editor in our initial collection, however, public opinion on controversial matters is important and interesting to consider when looking into general trends on language. Twitter is often useful in gauging public opinions on topics due to its diverse user base. Twitter hashtags are also useful for catergorizing and locating topical trends. We could easily create a Twitter Archiving Google Sheet (TAGS) (<https://tags.hawksey.info)> to collect data about public opinions on DAPL topics.

**Conclusion**

Though we focus on DAPL for this project, the tool we are developing could have several practical applications beyond this topic. For example, a historical news event could easily be run through the program for a fresh look at media biases that could have been long forgotten or overlooked; this could prove to be useful and relevant as more archives become digitized. This tool could also be used for a wide array of current events to track language and reporting trends. This project is only the beginning of collaborative research into media language trends across the political spectrum.