CSCI 404: Final Project Proposal

Noah Duggan Erickson Raghav Vivek Brady Deyak

21 May 2024

Faculty Acknowledgements

We plan on working closely with Dr. Kristen Chmielewski and other WWU ICDS[1] faculty both in the development and execution of this project.

1 Problem Statement

In this project, we aim to use an assortment of NLP techniques to analyze the frequencies, contexts, and sentiments of various disability-related topics over time. The headline example of this inquiry is to investigate the shifts in terminology from "handicapped" to "disabled" to "differently abled" and so on. We will then use these analyses to identify the contexts in which these terms are used, the sentiments associated with them, and the impact of these terms on the disabled community. We will also investigate the impact of the Americans with Disabilities Act (ADA) on the frequency and sentiment of disability-related topics in the media.

2 Impact

Prior to the adoption of the ADA, there was a significant amount of pushback from transportation companies such as Greyhound [1]. This project aims to identify other similar connections.

3 Learning Goals

Due to the inquiry-heavy nature of this proposal, there is a certain amount of "the journey is the destination" in this project.

[1] Institute for Critical Disability Studies

4 Data Set

The dataset for this project will be a large collection of news sources and other media, such as newspaper articles, press releases, and transcripts of radio and television broadcasts. This data will be acquired via a massive scraping operation to extract plaintext from NexisUni's batch-export of rtf files using broad search terms related to the terms of interest (e.g. `disab*` for "disabled," "disability," "disabilities," etc.).

5 Methodology

Following the acquisition of data as described in Section 4 and subsequent cleaning and preprocessing operations, we intend to use a variety of NLP techniques to create our analyses. The major components of this analysis will be (but not limited to) the following:

1. **Word Embeddings:** We will use a tool such as Word2Vec or FastText to create word embeddings for the terms of interest. This will allow us to analyze the contexts in which these terms are used. Furthermore, by splitting the data into time periods, we can analyze how these contexts change over time by measuring the cosine similarities between relevant terms over time.

2. **Common Theme Clustering:** We will use a clustering algorithm such as K-Means to identify common themes in the data. This will allow us to identify the contexts in which the terms of interest are used and how these contexts change over time. We then plan to assign labels to these clusters in order to better illustrate these themes.

3. **Sentiment Analysis:** We will use a sentiment analysis tool such as VADER to analyze the sentiment of the terms of interest. This will allow us to identify the sentiment associated with these terms and how it changes over time. We will also investigate the impact of the ADA on the sentiment of disability-related topics.

6 Evaluation

References

[1] Calgary H. ALLYSON JEFFS. *Access for disabled costly, say bus lines: [Final Edition]*. English. Copyright - (Copyright The Edmonton Journal; Last updated - 2024-03-08. 1992 May 20 1992/05/20/. URL: https://ezproxy.library.wwu.edu/login?url=https://www.proquest.com/newspapers/access-disabled-costly-say-bus-lines/docview/250571090/se-2.