Disability Trends: An examination of global disability-related terminology

Anonymous ACL submission

Abstract

This document contains the instructions for preparing a manuscript for the proceedings of ACL 2020. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1 Introduction

Intro text goes here.

2 Related Work

Related work goes here.

3 Our Approach

Despite the highly multidisciplinary nature of the project, the present work is primarily focused on the natural language processing aspect. The project was divided into three main parts: data acquisition and pre-processing, word vectorization, and sentiment analysis. The data acquisition and pre-processing part of the project involved scraping a large number of news articles from Nexis Uni, converting them to plaintext, and then lemmatizing the text. The word vectorization part of the project involved training a FastText model on the lemmatized text and then visualizing the results. The sentiment analysis part of the project involved training a sentiment analysis model on the lemmatized text and then interpreting the results. The project was conducted using Python and multiple libraries, including `Stanza`, `Gensim`, `TensorFlow`, and `Alibi`.

4 Experiments

There were multiple experiments run throughout the project. These include a massive scraping operation with subsequently large-scale data pre-processing, a word vectorization model, and an interpretable sentiment analysis model.

4.1 Data Sources

The data for this project was acquired using the export feature of Nexis Uni. Using this feature, a large number of news articles were downloaded as rtf files. These were then converted to plaintext using the `rtf to txt` Python library. Each article's content was then preprocessed using the lemmatization pipeline in the `Stanza` library. This pipeline tokenizes the text, removes stopwords, and lemmatizes the remaining words using a neural seq2seq model. The resulting data was then stored as a list and exported to JSON alongside article metadata.

4.2 Word Vectorization

The word vectorization model used was the Fast-Text model from the `Gensim` library. After training the model on the preprocessed data from a given decade, the model was used to generate a list of the most similar words to a list of selected words. The results were then visualized using a t-SNE plot. The t-SNE plot was generated using the `Plotly` library.

4.3 Sentiment Analysis

BRADY: In this section, discuss the sentiment analysis model. Include the model architecture, training data, and training results. For example, "Tensor-Flow was used to train a sentiment analysis model similar to the one proposed in this paper. It uses an RNN trained on this dataset, and achives an average accuracy of $x\%$."

4.4 Interpretable Sentiment Analysis

RAGHAV: In this section, discuss the interpretations of the sentiment analysis model. For example, how does Alibi work for generating those colors?

5 Results

Multiple visualizations and analyses were conducted on the data. These include a word vectorization model, a document-level sentiment analysis model, and an explanation of said sentiment analysis model.

5.1 Word Vectorization

Word vectorization models were trained on the data for each search term for each decade. This allows for the visualization of not only the most similar words to a given word in that snapshot, but also how the clusters of words shift over time. For example, figures **??** and **??** below show the tsne plots for selected words using the "disab" search term from the 1980s and 2010s, respectively.

1980s t-SNE plot goes here.

Figure 1: t-SNE plot of the word vectorization model using the "disab" search term from the 1980s.

2010s t-SNE plot goes here.

Figure 2: t-SNE plot of the word vectorization model using the "disab" search term from the 2010s.

6 Discussion

Discussion goes here.

7 Conclusion

Due to the short turn-around time of the project, avenues for future work are plentiful. These include a more centralized system for creating the analyses, a more robust sentiment analysis model, and a deeper review of these analyses from a critical disability studies and sociological perspective.

2