NLP Project Write-up

Understanding Autism Through NLP of Reddit Autism Forums

Abstract

Autism spectrum disorder (ASD) is identified in females at a substantially lower rate than in males. Research suggests that current diagnostic procedures may fail to capture how ASD manifests in girls and women (as well as in some boys, men, and non-binary folk). In order to better understand and support these under-diagnosed people with autism, I think it would be helpful to gather more information about their shared experiences.

My goal was to capture alternate ways that autism presents in humans, so that people with ASD can more fully accept, understand, and work with their unique challenges and gifts.

Design

This project involved NLP analysis of posts on forums where people go to seek support on autism-related interests and issues. I scraped several subreddits that were both specific to women with autism and general autism forums. I then cleaned and performed text pre-processing on the text data. Finally, the text was analyzed using NLP techniques such as topic modeling and sentiment analysis. I also created a basic recommender system inside a Jupyter Notebook, which demonstrated how Cosine Similarity of the text data might be used to search for relevant posts.

Data

The dataset was obtained via scraping of several subReddits using PRAW. Two of the subReddits are specifically for women with autism (r/aspergirls and r/AutismInWomen) and two subreddits are general autism forums (r/autism and r/aspergers). I scraped 5 times over the course of a month. The final dataset contains over 13k unique text posts, totaling about 400k words.

Algorithms

Data Preprocessing

1. Initial data cleaning was done with Pandas
2. Regex was used to clean up the text (punctuation, web links, emoji, etc.)
3. SpaCy was used for parts of speech tagging and lemmatization

Topic Modeling

4. LSA analysis: had a lot of overlap in words of the different topics. I added the top 20 most common words to the stop words list, which improved the model a bit.
5. LDA, and NMF analysis: both better than LSA, but many of the topics were still unclear, even after trying several different numbers of topics.
6. CorEx topic modeling using the LDA results as a guide for the anchor words. This model produced the best results, with words following the anchor words being highly relevant

Sentiment Analysis

7. I used exclusively VADER for sentiment analysis, which provided only extremes in sentiment.

Recommender System

8. I built a simple proof of concept recommender system in a Jupyter Notebook that was based on Cosine Similarity analysis of each Reddit post's text.

Tools

- Reddit API PRAW for scraping
- Numpy and Pandas for data cleaning
- NLTK and SpaCy for text pre-processing
- LSA, NMF, LDA, and CorEx for topic modeling
- VADER for sentiment analysis
- Tableau for interactive visualizations

Communication

All code and project deliverables are available on my public Github: https://github.com/ericajstevenson/METIS_projects/tree/main/NLP%20Autism%20in%20Women