Autism in Women – NLP Project Proposal

Introduction:

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 proposal is to do NLP analysis on forums where people go to seek support on autism-related interests and issues. There are several subReddits and Facebook groups that are specific to women with autism, which can be scraped and analyzed using NLP techniques such as topic modeling and sentiment analysis.

My goal is 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.

Question/need:

- What is the question behind your analysis or model and what practical impact will your work have?
  What are the common topics that are discussed on online forums for women with autism? How do these compare to general autism forums?
  Things to look for:
  - Areas that women struggle in as compared to the general population
  - Interests and skills that may differentiate women with autism
  - Traits and coping mechanisms that may differentiate women with autism
- Who is your client and how will that client benefits from exploring this question or building this model/system?
  My clients are the therapists, family, friends, partners of autistic women, as well as the women themselves. A deeper understanding of the topics most relevant to women with autism will allow for better diagnosis and support.

Data Description:

- What dataset(s) do you plan to use, and how will you obtain the data? Please include a link! (The link can be to the dataset you're downloading, the site you're scraping, etc.)
  I'm scraping 4 subreddits using the Reddit API, 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). Since the Reddit API limits scraping to 1000 entries per scrape, I'll scrape 3 times over the course of a month. Based on initial scraping results, I expect to get around 2500 unique entries per subreddit for each scrape.
- What is an individual sample/unit of analysis in this project?
  An individual sample will be a single Reddit post, which includes title text and usually a paragraph or more of body text.
- What characteristics/features do you expect to work with? title text, body text, author username, submission time / date, Reddit score, upvote ratio, number of comments
- If modeling, what will you predict as your target?
  I don't have plans to build a supervised model for this. I may use the Reddit score and/or number of comments to rank what posts are most interesting and relevant to people in these subreddit communities.

Tools:

- How do you intend to meet the tools requirement of the project?
  I will be using the Reddit API to pull in my data. Data Cleaning and EDA will be done with Python Pandas, sklearn and other text-specific preprocessing techniques. I plan on trying NLTK/gensim/spaCy/CorEx for topic modeling and VADER/DeepMoji for sentiment analysis. For topic modeling, I will try LSA, NMF, and LDA and include the best in my final presentation.

MVP Goal:

- What would a minimum viable product (MVP) look like for this project?
  The goal for the MVP of my project is to have at least half of my data scraped, to have a text preprocessing pipeline established, and some basic NLP techniques applied to the data. I plan to have my data fully cleaned and have done initial topic modeling, which will allow me to further process the data for intuitive topic modeling and classification.