Analyzing Cognitive Networks in Autism Spectrum Disorder:
A Comparative Study Using Reddit Data

Ilanit Sobol, Emmy Abitbul

Data Overview:

The data we will be working with consists of Reddit posts written by both autistic and neurotypical individuals.

Our research question aims to understand – "how the cognitive linguistic network of autistic individuals differs from that of neurotypical individuals, specifically in terms of structure, centrality, and communities?".

Data collection pipeline

The pipeline will be as follows:

1. Identify subreddit – e.g. **r/AutisticAdults.**
    - *For and about adults on the autistic spectrum. This is a relaxed discussion group, welcoming autistic people, non-autistic people seeking to learn, and people who believe they are or might be autistic.*
2. Identify ASD users by posts submitted to the above subreddit.
3. Extract ASD users' posts to different subreddits (not related to autism).
    - Due to memory and time efficiency, 100 latest posts per user extraction
    - Drop posts that don't contain text or that were removed from Reddit
4. Aggregation over subreddits.
5. Filter top subreddits according to the following rules:
    - More than 5 unique ASD users have been posting in the subreddit
    - The subreddit name doesn't have a direct connection to ASD.
6. Selecting subgroup of subreddits fr our analysis and research:
    - **r/depression**
        1. *"Peer support for anyone struggling with a depressive disorder."*
    - **r/raisedbynarcissists (short: rbn)**
        1. *"This is a support group for people raised by abusive parents (with toxic, self-absorbed, or abusive personality traits, which may be exhibited by those who suffer from cluster B personality disorders). Please share your stories, your questions, your histories, your fears, and your triumphs. Significant others and friends are all welcome."*
    - **r/AskDocs**
        1. *"Having a medical issue? Ask a doctor or medical professional on Reddit! All flaired medical professionals on this subreddit are verified by the mods."*
    - **r/Anxiety**
        1. *"Discussion and support for sufferers and loved ones of any anxiety disorder."*

7. Extract neurotypical and ASD users' posts' in the above subreddits.
    - For each subreddit, extract 15 users who are not in the ASD users list.
    - These users will be classified as neurotypical.

In addition to utilizing the PRAW Python package to retrieve raw posts from Reddit, we also collected general metadata provided by Reddit's API. This metadata includes supplementary information associated with the posts, allowing for a more comprehensive analysis of the data.

From post to network format:

After collecting the data, we have one CSV file with the following columns:

- Subreddit
- Post time
- Username
- Post Title
- Post Text
- Num comments
- Score
- Ups
- Upvote ratio
- URL
- ASD – true or false.
    - If true, the post originated from an ASD user = treatment group
    - If false, the post originated from a neurotypical user = control group

Our goal is now to convert each post, of each group, of each subreddit, from raw text to network.
There are different ways to convert text to a network. In the end, we have decided to convert each post to a network.

Meaning we will have $|C| \cdot |G| \cdot \sum_{c \in C, g \in G} n_{c,g}$ total networks:

i. C – number of subreddits. In our case – 4
ii. G – number of groups. In our case – 2 (ASD, notASD)
iii. $n_{c,g}$ – number of posts of each group and each subreddit

The next stage is text preprocessing:

1. Remove punctuation.
2. Split each post into semantic units (using nltk sentence tokenizer)
3. Split each semantic unit into words (using nltk word tokenizer)
4. Lemmatize, stem and lowercase each word

As each user can have more than one post and we want IID assumptions, we use only the first post of each user.

The networks setting is as follows:

- Nodes – processed words
- Edge – an edge exists between word x and word y if they appeared in the same semantic unit together.
- Undirected – we do not care for the order of appearance, i.e. the edge (x, y) and (y, x) are symmetrical.
    - For implementation purposes, we ordered the tuple (x, y) in alphabetical order.
- Weight – the number of times, 2 words appeared together in the same semantic unit, for each semantic unit in a post.
    - If 2 words appeared more than once in the same semantic unit, we will only count this as once.
    - Currently we haven't decided if we want to use this attribute so for the time being we kept it in the visualization.

Filtering settings:

To create meaningful networks, we propose 2 filtering settings that we may examine along our project:

- Filter out stop words - this is the baseline which we use in our analysis
- Stop words only - we may dwell on this in one of the subreddits.

Network analysis:

Our goal is to compare the networks' metrics distribution, between each subreddit, and between each group. In total, we will have multiple comparison problems at hand.

We want to compare the following metrics:

1. Average degree-
    a. It helps determine the overall level of connectivity in the cognitive linguistic network of autistic and neurotypical individuals.
    b. Differences in average degree can indicate variations in the extent or focus of their network connections.
2. Average path length:
    a. It provides insights into the efficiency of information flow within the network.
    b. Comparing average path lengths can reveal differences in how information spreads or is accessed by autistic and neurotypical users.
3. Centrality: Closeness
    a. It assesses the ability of individuals to access information quickly and efficiently.
4. Modularity:
    a. Helps us understand the network structure. We may be able to connect the modularity score with other psychological studies which assess how ASD affects the semantic structure of individuals.
5. Community detection:
    a. Clustering coefficient and Louvain algorithm.
    b. We can identify groups or clusters of words within each semantic network, revealing patterns of word associations and potentially uncovering differences in the organization of semantic networks between the two groups.

Bonus analysis:

If we have time, we also want to use LIWC (Linguistic Inquiry and Word Count), which is a text analysis software that categorizes words into predefined linguistic and psychological dimensions. It provides insights into language use by counting word occurrences in each category. LIWC is used in research to study emotions, cognitive processes, social dynamics, based on text analysis.

Summary:

By examining these metrics, we can gain a deeper understanding of how the cognitive linguistic networks of autistic individuals differ from those of neurotypical individuals in terms of structure, centrality, and communities. Additionally, integrating LIWC analysis can provide further insights into the linguistic dimensions and word usage patterns, enhancing your understanding of cognitive and linguistic differences between the two groups.

Basic analysis over the network:

All words without stop words:

Subreddit	r/depression		r/raisedbynarcissists		r/AskDocs		r/Anxiety	
Group	ASD	notASD	ASD	notASD	ASD	notASD	ASD	notASD
Number of networks	13	43	16	42	18	45	16	42

Link to notebook report with visualization – contains distribution of number nodes, edges, networks, etc..

Network visualization:

- After experimenting with different visualization settings, we decided to visualize a random network of each subreddit and each group.
- We believe that by adding the words on top of the nodes, it is much easier to understand the context.
    - This decision was based on the understanding that adding labels to the nodes significantly enhances the context and comprehension of the network.
    - By including the words associated with each node, we provide additional information about the nature of the connections and the content represented by the nodes themselves.
    - This labeling approach empowers viewers to quickly grasp the essence of the network and identify key elements within it.
- We observed that visualizing the graph as a circle enhances the visibility of clique structures within the network.
    - By arranging the nodes in a circular layout, we create a visual arrangement that highlights clusters or groups of tightly interconnected nodes. Cliques, or densely connected subgroups, become more apparent in this circular representation.