Matt Thompson

Dr. Klein

QTM 340

December 9, 2021

**Contextualizing ADHD Diagnostic Criteria using Natural Language Processing of r/ADHD**

## Introduction

ADHD or Attention-Deficit Hyperactivity Disorder is a neurodevelopmental disorder that causes symptoms related to inattention, hyperactivity, and impulsivity[[1]](#footnote-1). ADHD is often associated with the stereotypical image of a dumb little boy who can’t stop running around and causing chaos in school. This stereotypical image is often not representative of how this disorder presents, especially since ADHD is a disorder that occurs in men and women of all ages that effects nearly all domain areas (education, work, relationships, personal, home, etc) of that person’s life. ADHD is frequently underdiagnosed amongst women and high-functioning adults because of these faulty stereotypes.[[2]](#footnote-2) These stereotypes are also not helped by medical criteria that somehow are both too broad and vague whilst also too clinical, causing many people to think “everyone is a little ADHD” and for people that might have it, to not understand what the symptoms are referring to.

This project intends to illustrate and enrich these medical diagnostic descriptions by analyzing a corpus of personal pieces by those with ADHD using the subreddit r/ADHD. Using a continuous bag of words model, word associations are generated from user posts to help contextualize what diagnostic criteria are often really referring to. Using these word embeddings, this project helps provide context to everyday people what underlying issues are apart of ADHD and how it presents. Hopefully after reading this, you can better understand what the underlying issues are that cause ADHD symptoms and how they appear.

## Related Research and Works

One of the inspirations for this project was a paper by graduate student Maria Antoniak.[[3]](#footnote-3) In that paper she uses a subreddit dedicated to pregnancy called r/babybumps to provide a corpus to analyze birth stories.[[4]](#footnote-4) She analyzes birth stories to determine sentiment towards aspects of the stories and specific actors in the narratives. Similar to this project, that project uses posts from a subreddit as the corpus and also analyzes the stories to determine how parts of the birth stories relate to other factors. How they differ is the focus of the analysis since This project focuses on generating word embeddings based on keywords instead of finding sentiment scores based on keywords and actors.

This publication in the *Journal of Medical Internet Research*  examines how posters in multiple mental health focused subreddits responded to the COVID-19 pandemic. The project took a before and during approach to the pandemic by looking at posts from 2018 through 2020 and then analyzed text through sentiment analysis, topic modeling, unsupervised clustering, and various other methods.[[5]](#footnote-5) They found increased stress across communities, some of the highest being in the ADHD subreddit. These projects are similar by analyzing the r/adhd and using unsupervised methods to find associations, but they differ since my focus in on one subreddit and is focused primarily on finding word clusters connected to external sources and not sentiments.

This last paper tests multiple natural language processing methods to determine how to best detect depression from reddit user posts that are not explicitly related to clinical depression. This paper differs from mine by focusing on finding the best method to detect latent depression and not analyzing posts that explicitly connected to depression.[[6]](#footnote-6) This paper is similar to mine by using an external set of terms to guide the text analysis and natural language processing methods to help find latent words and meaning within the corpus.

Beyond works for methodology work, there are important sources to understand exactly what ADHD. Information on the definition and description of ADHD comes from the CDC and the DSM-5, the official diagnostic book for conditions related to mental health. However, to further enrich this essay, research from Dr. Thomas Brown help offer a more in-depth understanding of ADHD, but his ideas will be in the discussion section, but wanted to keep the surprise in your mind now 😊

## Corpus

The corpus was created by scraping the text from submissions to r/ADHD. This project originally intended to use the PRAW API, an API designated by Reddit for scraping. However, since PRAW places a limit at 1000 posts, this is not a large enough sample for using a continuous bag of words model. I attempted to generate a for-loop to mitigate this problem, but that approach failed. Due to those limitations, I switched to using PSAW, a wrapper for the Pushshift API. The Pushshift API allowed me to scrap just the text in the post based on searching by keywords and did not include a limit on the number of posts that can be collected.

While reddit is a convenient source for data, it has its flaws. Reddit is disproportionately used by middle-class to wealthy white English-speaking men, which may skew some of the experiences of symptoms and which symptoms are most frustrating amongst users. Also since this is a largely anonymous social media discussion board, it is not known if the posters truly have diagnosed ADHD or just suspect it and more significantly, if the poster is even a real person. Due to these factors, this corpus is definitively not representative of the entire ADHD population, especially those who are under-privileged, who are most likely to not be receiving proper treatment or support for their ADHD. Also, there is potential for internal groupthink to shift how some of the people discuss the disorder over time.

The corpus itself is 18 separate documents, with each document matching with one of the symptoms of ADHD. The search term used by PSAW were either words directly from the ACDS (Adult ADHD Clinical Diagnostic Scale)[[7]](#footnote-7) or slightly paraphrased versions of those words. The below table shows the ACDS description, the label of the txt file, and the number of sentences per document. Documents with lower sentence counts had a smaller number of posts. When scraping submissions with the api, each query was limited to 5000 posts, with some documents having over 4000 posts and some having less than a thousand.

|  |  |  |
| --- | --- | --- |
| Symptom | Keyword | # Sentences |
| Avoidance of tasks with sustained mental effort | Avoid | 41363 |
| Blurts out answers | blurts | 11411 |
| Careless/Sloppy | careless | 24317 |
| Easily distracted | distracted | 56592 |
| Fidgets/Squirms | fidget | 32259 |
| Forgetful in daily activities | forgetful | 56017 |
| Difficulty following instructions (finishing) | instructions | 35801 |
| Interrupts or intrudes | interrupts | 26131 |
| Doesn’t listen | listen | 102072 |
| Loses things | misplaces | 128237 |
| Runs/Climbs excessively/inappropriately | Moves | 27728 |
| On the go/Driven by a motor | onthego | 37789 |
| Difficulty organizing tasks/activities | organization | 48913 |
| Difficulty sustaining attention | payingattention | 75098 |
| Difficulty playing quietly | Quiet | 55045 |
| Difficulty remaining seated | seated | 95135 |
| Talks excessively | Talks | 18820 |
| Difficulty waiting turn | waiting | 19604 |

## Process and Methods

The above table shows how each of the document was tokenized to the sentence level. Before tokenizing, all documents were forced to encode as utf-8 to keep formatting consistent. Then stopwords were then removed and all sentences were converted to lower case. Since Pushshift produced surprisingly clean documents, minimal cleaning was needed for the model. The documents were then tokenized. The above table shows that in the data collection process, some symptoms had more results to collect from than others.

The first phase of this project was to generate a continuous bag of words model (CBOW) to determine word associations with each of the 18 diagnostic symptoms as well as word associations with the official clinical presentation types (predominantly inattentive, predominantly hyperactive/impulsive, and combined). For those who enjoy staring at code and mild self-torture, this model was generated using the following packages genism, nltk, glob, os. A CBOW model was viewed as the best approach since there would be a large set of inputs due to the large documents across multiple search terms and we are only looking for a set of specific word embeddings per symptom. Using word2vec, a model was generated from those 18 documents. Since ADHD symptoms tend to overlap or manifest together, analyzing the documents collectively was considered the best approach. Using the most similar function within word2vec, 15-word associations or word embeddings were generated for each of the symptoms.

This project intended to generate topic models after the first phase of exploratory analysis, but the topic models generated from the reddit data had too many extraneous words that would need to be filtered out by a more granular process than time allowed. Luckily, due to common themes within the CBOW model, a second round of the word embeddings were generated based on the themes from the initial results. From there, those second round results were contextualized with outside research and description.

## Results – Inattention Symptoms

A first glance at this table might feel overwhelming since there are so many words so I qualitatively color-coded by theme so that its now overwhelming by words and colors! Actually the colors do represent some common themes or trends between the words based on a qualitative analysis. Green represents personal descriptors that a person uses to describe how they feel related to the symptoms. Often these words signify internalized frustration, irritation, and boredom. Blue-Grey represents symptoms related to forgetfulness and difficulties with memory recall. These symptoms seemed to occur across multiple symptom areas. Gold represents symptoms related to difficulties with attention, focus, and concentration. Again, this color occurs in multiple areas of the table. Pink represents symptoms related to difficulties with implementing a process or following through with steps. Also once again, seems to be occurring across multiple symptoms. Purple represents symptoms related to starting a task. And in a total shock to you I’m sure, this issue seems to occur across multiple symptoms. Dark Blue describes symptoms related to organization and prioritizing, which surprise surprise is happening across multiple symptoms. Lastly, grey represents words associated commonly associated with hyperactivity and impulsivity.

When formatting this, I *almost* created a second table with everything grouped by theme to ease the readers mind, but the cacophony of color does a good job demonstrating how these groupings seem to be occurring across all the symptoms and how there is significant overlap between symptoms, also the almost chaotic nature of how this disorder feels and is described. Almost every column has multiple colors, signifying many symptoms are related to multiple themes according to these user posts. Also, different users seem to relate symptoms to different struggle areas with a different thematic area and at times relates to a word in near polar opposite ways. Apparently, difficulties with listening can manifest by interrupting or by ignoring, while someone might associate being distracted with feeling very bored or very overwhelmed.

## Results – Hyperactive/Impulsive Symptoms

Again, another jumble of words, but color-coded to ease reader inconvenience or maybe create an aura of chaos! Green indicates words related to visible movement or visible behaviors. These behaviors seem to indicate lots of energetic movement, be it fidgeting in a seat or running around. Blue refers to a behavior that is internal and not something someone would notice from the outside most likely. A fair amount of user posts associate hyperactivity symptoms with an internal behavior or action. Often these symptoms are connected to either passively observing and caring or passively not doing that. Pink represents words associated with social interaction, of which are often disruptive or associated with verbal discussion. Red refers to how the user feels, which seems to fluctuate between feeling calm and shy, feeling very energetic and irritable, and a variety of physical feelings, which are very different ways of someone describing feeling hyperactive or impulsive. Gold represents words that would seem to be associated inattentive symptoms. Similar to the above table, most of these themes seem to occur across the table, with some symptoms of hyperactivity being very external, some seemingly very internalized, and at times many of them of occurring within social interactions. In general though, it seems there is significant exhaustion and frustration associated with hyperactivity and impulsivity symptoms, one of many things this table has in common with the other table.

## Results – Generalized Symptoms

While there were trends within each predominant presentation type, there was significant overlap between them as well. The inattention word associations documented multiple associations with impulsivity. Also, the hyperactive/impulsive words mentioned words related to concentration, preparing, scheduling, which seem to be more connected to inattentive symptoms. Both tables also showed that users had almost opposite feelings towards certain themes, with people feeling very strongly associated with it or very opposite of that theme.

 Based on the common associations within each set of predominant symptom presentations, I re-ran the CBOW model with a new set of symptom-related words but with the top 10 terms. These words were the following: organize, prioritize, initiate, follow-through, manage, forget, emotional, regulate. The words organize, prioritize, initiate, follow-through, manage, forget were selected to represent the common themes from the inattention table while the words emotional and regulate were picked to represent the symptoms from the hyperactivity table.

Woo-hoo, another color collage of chaos for the reader, but to save you reading time, the labels are included in the table! While the table is less splotchy, there is still significant overlap in symptoms and the trend of opposite words being associated is still apparent. How can someone struggle to concentrate but also “hyper-focus”? How can someone with the same disorder be forgetful but also overthink? How come users associate the word regulate with the need to both suppress, but also to express, and to just sustain??? These users seem to have lot of issues with these mental functions, but perhaps not in the same way…

## Discussion

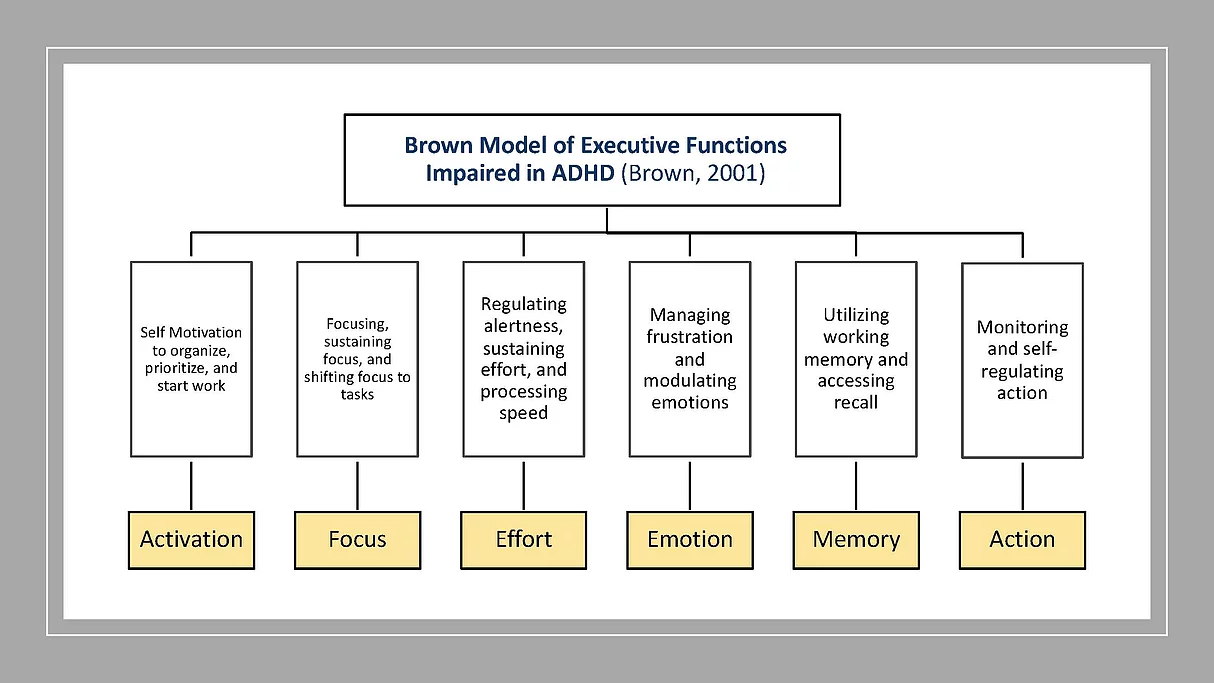
At this point after using the CBOW model a few times and narrowing down symptoms, it seems the users describe difficulties with organization, prioritizing, remembering, focusing, following through, regulating oneself, and emotional frustrations and that these difficulties seem interconnected. And these frustrations seem to manifest in various ways for different people, but how can one disorder manifest these issues in significantly different ways

Let’s say one theoretical person, Lauren, who is very intelligent and responsible working professional but has predominately inattentive ADHD. She is so absent-minded that she keeps forgetting to go to important meetings for her job and forgot about her friend’s birthday party last Sunday. Since she can’t seem to get organized, she keeps falling behind in her work and can’t keep track of her social commitments. She might get to work on time but can’t focus during her work meeting, struggles to organize when to work on various tasks throughout the day, and keeps missing steps within the procedures she is supposed to follow. After work, she goes to the grocery store, though she forgot her grocery list she did remember to make alongside her the grocery bags. She haphazardly meanders through the store and buys only half of what she needs since she can’t concentrate. At this point she is really tired from doing a day of things consistently full of tiny mistakes and errors. She feels frustrated and overwhelmed by her inability to “get her shit together” – a technical term.

Say there is another person, Matt, who struggles with many hyperactive/impulsive symptoms. He is so mentally restless and constantly jumping between tasks and activities that he loses track of time and misses a work meeting and arrived very late to his friend’s party after work since he couldn’t find his work in a timely manner. Since he struggles to filter his words, he ends up starting an argument at the party, embarrassing himself, though he jumps between enough conversations with different people he forgets about the argument. After the party, he tries to swing by Walmart to buy some essentials but ends up impulsively buying junk food he doesn’t need and forgets the thing he did need to buy, which was simply toilet paper. He gets home feeling irritable from such a chaotic day and sits on the couch sit still feeling internally frenetic wishing he could “keep his shit together” – another technical term.

These two randomly named vignettes, seem to line up with someone who is primarily inattentive and someone who is primarily hyperactive. However, upon closer inspection of these vignettes, you’ll notice some common problems. Lauren and Matt both seem to struggle with keeping themselves organized, with starting a task and/or following through with it, with sustaining their focus on whatever they are supposed to be doing, with remembering important details in the moment, and both seem to be deeply frustrated by these difficulties. It seems that while these two people might present very differently, they are both dealing with similar core issues.

At this point though, you might be thinking I’m on track to announce some grand breakthrough in understanding of a complicated neurological disorder because of natural language processing and reddit and I wish! These difficulties with managing core issues, has been referred to as executive dysfunction by psychologists for a few decades now, even if they aren’t officially in the DSM-5, which is used to diagnose people.



Dr. Thomas Brown frames ADHD as a disorder related to six impaired executive functions: activation, focus, effort, emotion, memory, and action, with more detailed description in the diagram[[8]](#footnote-8). As described in the above model diagram, these executive functions are neurological functions that are used throughout day-to-day tasks and activities in one’s life, be it in school, a personal project, social relationships, or just one a person is alone. With ADHD, there is significant impairment in the ability to use these functions to manage and regulate oneself, though how they manifest is dependent on that individual person. In fact, if you take a another gander at some of the color collages, you’ll notice that many of the symptoms are related to school, social interaction, work, and just general life. It seems the model based off of reddit users describing ADHD was able to highlight similar connections as Dr. Brown did at least 20 years ago.

When looking at the two examples, both individuals struggle with their ability to utilize multiple executive functions throughout their daily lives, even though they might act very differently. Regardless of how they appear, they both have similar underlying problems. This inability to use and regulate these functions also can demonstrate why throughout the data seems to have contradictory pairs. One disorder does not encompass the entirety of a person, other personal and environmental factors, likely explain how these symptoms manifest in different manners.

## Conclusion

So, you’ve read through lots of words so far, and might be curious why all this matters. ADHD is often viewed as either something made-up or a disorder only for dumb and rambunctious little boys, which is evidently not true. Viewing ADHD exclusively as a disorder of chaotic boys excludes the vast majority of the population whom have ADHD, like someone who is internally hyperactive but otherwise is introverted or someone who seems very personally responsible but can’t seem to get organized to save their life. Pervasive stereotypes such as these are likely a significant factor as to why women are frequently underdiagnosed with ADHD, since due to differences in biology and socialization, it is common for women to not present ADHD symptoms like the chaotic little boy. By using real-world data such as in this paper, it can help clarify what ADHD symptoms are and how they manifest, which hopefully can help lead to more people receiving proper treatment.

Future research should attempt to use a more properly cleaned corpus, especially one that eliminates repetitions of base word variations, like the multiple variations of attention. With that cleaner corpus, network styles of research could be used to map out if there are more distinctive clusters beyond identifying common areas of executive dysfunction, which could help provide an alternative approach to current clinical presentations.

Works Cited

Antoniak, M. (n.d.). *Narrative paths and negotiation cornell ... - Maria Antoniak*. Retrieved December 9, 2021, from https://maria-antoniak.github.io/resources/2019\_11\_12\_cscw\_birth\_stories\_presentation.pdf.

*The brown model of ADD/ADHD: Brown ADHD clinic: United States*. Brown ADHD Clinic. (n.d.). Retrieved December 9, 2021, from https://www.brownadhdclinic.com/the-brown-model-of-add-adhd.

Centers for Disease Control and Prevention. (2021, September 23). *Symptoms and diagnosis of ADHD*. Centers for Disease Control and Prevention. Retrieved December 9, 2021, from https://www.cdc.gov/ncbddd/adhd/diagnosis.html.

*Detection of depression-related posts in Reddit Social Media Forum*. IEEE Xplore. (n.d.). Retrieved December 9, 2021, from https://ieeexplore.ieee.org/abstract/document/8681445.

Low DM;Rumker L;Talkar T;Torous J;Cecchi G;Ghosh SS; (n.d.). *Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: Observational study*. Journal of medical Internet research. Retrieved December 9, 2021, from https://pubmed.ncbi.nlm.nih.gov/32936777/.

NIH. (n.d.). *Adult ADHD Clinical Diagnostic Scale (ACDS)*. Retrieved from https://datashare.nida.nih.gov/sites/default/files/studydocs/276/MDS0007\_CRF.pdf.

Sigler, E. (2021, May 5). *ADHD looks different in women. here's how - and why.* ADDitude. Retrieved December 9, 2021, from https://www.additudemag.com/add-in-women/.

1. CDC 2021 [↑](#footnote-ref-1)
2. Sigler 2021 [↑](#footnote-ref-2)
3. Antoniak, 2019 [↑](#footnote-ref-3)
4. Antoniak, 2019 [↑](#footnote-ref-4)
5. Low, 2021 [↑](#footnote-ref-5)
6. IEEE Explore, 2021 [↑](#footnote-ref-6)
7. NIH 2021 [↑](#footnote-ref-7)
8. Brown 2021 [↑](#footnote-ref-8)