Detection & Modeling of Substances, Diagnoses, & Medications Reported by Prescription Stimulant Users on Reddit

Max Z. Grice, BS, Kirk Roberts, PhD, MS
McWilliams School of Biomedical Informatics, Houston, Texas

Abstract

As the only major controlled drug with increasing prescription rates, the popularity of stimulants like Adderall, primarily for the treatment of Attention Deficit Hyperactivity Disorder (ADHD), continues to prompt debate over appropriate use and long-term effects[1]. With discrepant reports on user numbers (anywhere from 5-35% on college campuses), alternative data sources are needed provide transparent accounts on stimulant use[2]. Providing one such source of data, 2,001 posts were extracted from r/adderall, a subreddit dedicated to the discussion of stimulants. Named Entity Recognition (NER) was then used to tag diagnoses as well as specific substances and/or medications taken by the user. After building, training, and evaluating three models, the support vector machine was found to be the best performing classifier. While many limitations suggest that additional work is needed to improve performance, these results are helpful in establishing an alternative framework for extracting information on stimulant users.

Introduction

With the capacity to stimulate the central nervous system and cause excitation, elevated mood, and increased alertness, psychostimulant use is expanding. While prescription stimulants can suppress appetite for weight loss or alleviate conditions like ADHD and narcolepsy, they are also commonly used recreationally[3]. This is particularly apparent in college settings, where 61.7% report diverting prescription stimulants and 36% believe stimulants like Adderall, are not harmful and can make them smarter[4]. While these accounts along with rising prescription rates have raised some alarm in the wake of a devastating opioid crisis, a protracted controversy surrounding stimulant use and impact remains. On the other side of this debate, studies point to the high efficacy of stimulants when prescribed properly and interpret the rise in prescription rates as correcting for the underdiagnosis of conditions like ADHD[5]. Regardless, it's clear that prescription stimulant use is on the rise, increasing 70% between 2011 and 2021[6].

To better understand this rise in use, it's important to understand not only *what* users are specifically taking (including other medications in addition to prescription stimulants) but also *why* they are choosing to take them. Due to the quasi-anonymous nature of r/Adderall, a subreddit dedicated to the discussion of ADHD drugs for recreational and medical use, users are generally transparent in their personal recollection of medications and/or substances consumed as well as their reason for use. Using a technique in Natural Language Processing (NLP) known as Named Entity Recognition (NER), my goal was to extract any specific mentions of medications, substances, or diagnoses used by the author of a given post to gain a more comprehensive picture of this *what* and *why* of prescription stimulant use. Designing a model to identify diagnoses (e.g. ADHD, depression…etc.), stimulant brands (e.g. Adderall, Ritalin, Vyvanse…etc.), substances (e.g. marijuana, alcohol…etc.), and other medications (e.g. SSRIs, antacids…etc.), could help to clarify popular stimulant use cases, motivating more targeted treatments and better patient outcomes.

Methods

To automatically detect substances, diagnoses, and medications reported by prescription stimulant users on reddit, there were three sequential steps that were undertaken, including data exploration, data annotation, and data modeling which are described in further detail in the sections below.

Data Exploration

Using the Python Reddit API Wrapper (PRAW), 6,417 total posts were initially extracted from r/adderall, each containing a unique post identifier, text, date posted, author username, and URL[7]. Upon visualizing the distribution of word counts in posts, I observed that the data had a positive skew which led me to apply a word count filter, removing posts with word counts below 20 and above 500 words. Due to the time constraints of manual annotation, I was only able to analyze 2,001 posts which I randomly selected as a subset of the filtered data. Next, I split the selected data into a training set (70%) and validation set (30%) which contained 1600 and 401 posts respectively. After creating a folder for the training data and a separate folder for the test data, the text of each post was saved to a separate text file in its corresponding folder after removing URLS, return carriages, and lowercasing. The result of this preprocessing left 2 folders, one containing all training data with 1600 text files and one containing all test data with 401 text files. Both folders were copied to the data directory of my local brat installation and the brat standalone server was used to access the data in perform manual annotation using the annotation schema described below[8].

Data Annotation

Because there was uncertainty on how many posts would contain mentions of specific entities, I created a hierarchical annotation scheme where the top-level entities contained many more detailed entity subsets. This would allow me to choose the level of annotation detail based on how many posts were annotated with each entity (since a given entity needed to be annotated in a fairly large number of posts for that entity to perform well during training and testing). As depicted in Figure 1, the broadest category of entities were *Diagnoses*, *Prescription Stimulants*, and *Substances/Medications*. The *Diagnoses* category contained all instances in which a user mentioned having a specific mental condition. The three major types of diagnosis categories found were *Depression*, *ADHD* and *Anxiety*. Other diagnoses that were less common included *Eating Disorders*, *Narcolepsy*, *OCD*, and *Bipolar*, all of which were grouped together under the category of *Other Diagnoses*. The *Prescription Stimulant* entity contained all mentions of prescribed stimulant medications which I defined as the brand names *Adderall*, *Vyvanse*, *Ritalin*, *Dexedrine*, *Concerta*, *Modafinil*, and *Focalin*. Users referring to their *meds* or *medications* in posts were also classified as *Prescription Stimulants* under the same broad entity. *Substances & Medications* could be further classified into three main categories. The first two categories were focused on medications, specifically, *Prescription Medications*, which included *Anti-Depressants*, *Benzodiazepines*, and *Antipsychotics*, as well as *Over the Counter Medications (OTCs)* which included *Supplements*, *Sleeping Pills*, and *Antacids*. The last category, *Substances*, was concerned more with possibly addictive but commonly used substances, including mentions of *Alcohol*, *Nicotine*, *Marijuana*, and *Caffeine*.

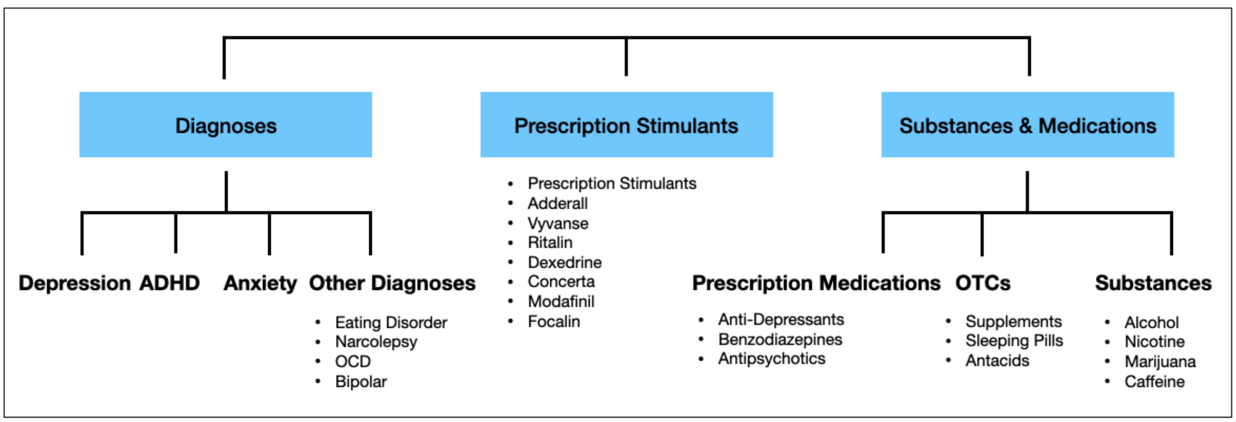

Figure 1: Hierarchical Annotation Scheme for Entities and Sub-Entities

For all entities described above, I only annotated when it was made explicitly clear that the user was talking about the entities in relation to themselves rather than a general observation with no explicit reference to a person. For example, the text '*do not take Xanax when you have ADHD and are on adderall'* would contain no annotations because there is no explicit reference to the person taking Xanax or Adderall or having ADHD. On the other hand, if the text were to read '*I was also taking Xanax after being prescribed Adderall to treat my ADHD*,' then all three entities (Xanax=benzodiazepine, Adderall, ADHD) would be annotated. Additional examples of appropriate annotations are given and explained with more detail below.

Example Annotation Key Diagnoses Prescription Stimulants Substances & Medications

"taurine. it calms me down better than a xanax and keeps my cognitive functions on point. i usually get anxiety late in the day when the adderall starts to wear off. 1-2 grams of taurine soothes me almost too perfectly."

"i have been smoking marijuana for a while now and usually smoke at night right before bed to help me fall asleep. i have recently been prescribe 10mg of adderall xr that i take in the morning, once a day."

"i have dealt with depression and suicidal thoughts my whole life. Late in life i got diagnosed bipolar and put on lithium. life became stable, but just stable. almost three months ago they added Adderall to the lithium."

The annotations above are classified by the broadest annotation category but some could also be annotated with more detailed annotation categories. For example, Xanax is a *Substance & Medication* but it specifically belongs to the *Prescription Medications* sub-group and even more specifically is a *Benzodiazepine*. Bipolar is also classified under the broader category of *Diagnoses* but specifically belongs under the grouping *Other Diagnoses*. In all posts given above, the user is referencing these substances, medications, and diagnoses in relation to themselves which is why they are correctly annotated.

Heuristic Data Modeling Approach

There were a few different approaches which I used to model the data which was annotated manually via brat tools (according to the entities and examples described above). The first, a rule-based heuristic modeling approach, was helpful in establishing baseline evaluation metrics which could later be compared to the other modeling approaches. Using a pretrained tokenizer from hugging face (*allenai/longformer-base- 4096*)[9], text from each of the 2,001 posts was tokenized into corresponding word pieces and mapped to integer ids. After importing annotation offsets from brat, offsets were aligned with each token so that the integer ids in every post could be categorized as some specific entity (with the specific id label to entity mapping shown in the table below) or as a non-entity token (i.e. WILD).

Table 1. Integer-Entity Dictionary Mappings

0: '*Prescription_Stimulants*': ['Adderall', 'Vyvanse', 'Dexedrine', 'Focalin', 'Ritalin', 'Concerta', 'Modafinil']
1: '*Perscription_Medications*': ['Anti-Depressants', 'Benzodiazepines', 'Antipsychotics']
2: '*OTCs*': ['Sleeping Pills', 'Supplements', 'Antacids']
3: '*Substances*': ['Caffeine', 'Alcohol', 'Marijuana', 'Nicotine']
4: '*Anxiety*'
5: '*Depression*'
6: '*ADHD*'
7: '*Other_Diagnoses*': ['Bipolar', 'Eating Disorder', 'Narcolepsy']
8: '*WILD*'

After this alignment, each post in the set of 2,001 posts could be defined by a list of corresponding integer id tokens (e.g. [1234, 123, 2312, 5932]) as well as the integer-entity label assigned to each integer id/token (e.g. [8, 8, 6, 8, 0]). Given these results, my hypothesis was that selecting the tokens associated with each entity id in the training set would work as a viable way to predict the entity ids of each token in the test set. In other words, a list of integer id tokens was mapped to each entity id (e.g. 0: [5932, 223, 109]….6: [2312, 34102]…etc.) so that, upon encountering each unlabeled integer id in the test set, I would check to see if it existed in any entity id list and use this to predict the corresponding id (or predict 8 if it did not exist in any list). After predicting the entity labels for each token in the test set of 401 posts, I evaluated the model by comparing the gold labels to the prediction labels, calculating the accuracy, precision, recall, F1-score, and support for each entity.

Traditional Machine Learning Data Modeling Approach

Like the heuristic modeling approach, I used the same *allenai/longformer-base-4096* tokenization and entity alignment scheme to create the training and test data with entity labels. However, rather than looking specifically at the integer ids associated with each entity, I combined all unique integer ids (associated with at least one entity) into one list. In total, there were 467 integer ids associated with at least one entity and each of these integer ids was used as a feature in my new dataset. This allowed me to create a numpy array with 1600 rows (one for each post in the training set) and 467 columns (one for each integer id found to be associated with an entity). If a given post (ie row) contained a given integer id (ie column), I would set that corresponding cell to 1 (otherwise its value was set to 0). For the corresponding label matrix, I created a numpy array with the same 1600 rows to represent each post but this time with 9 columns to represent each entity (including non-entities or the WILD id). After applying the same data transformation to the 401 posts in the test dataset, I created 8 different support vector machine models with a linear kernel, one to classify each entity, and fed in the corresponding column from the label matrix. The predictions of each model were compared to the gold labels and accuracy, precision, recall, F1-score, and support were calculated for each entity model.

Transformer Data Modeling Approach

The last modeling approach made use of a Transformer model to detect the entities within the text. For this model, I used a different pretrained token classification model from hugging face, *distilbert/distilbert-base-uncased*, because the long-former model was unable to correctly classify any entity during evaluation using the *allenai/longformer-base-* tokenizer. Unlike the previous approaches, I also choose to classify tokens into one of the broader 3 entity categories rather than the 8 sub-categories because there did not seem to be enough annotated data for the model to correctly identify any of the 8 entities within the test data. Consequently, there were now 4 possible labels (0 for non-entities, 1 for prescription stimulants, 2 for substances & medications, and 3 for diagnoses) which could be assigned to each token in the training and test data. After training the distilbert token classification model on truncated posts (max length of 512 per model specifications) with 5 epochs, a batch size of 4 and a learning rate of 1e-3, the resulting model for each entity was saved and evaluated, recording the accuracy, precision, recall, F1-score, and support metrics.

Results

After tokenizing the data with pretrained tokenizer models from hugging face and aligning the brat annotation offsets with each token to create corresponding token classification labels, it was important to check how many tokens were actually classified into each entity type. In Figure 2, I compiled a table with the number of unique tokens and the total number of tokens classified as each entity type. While each row in the table represents the most detailed sub-entity level, the table cell coloring is used to represent the broader entity types (i.e. *Prescription Stimulants, Prescription Medications, OTCs, Other Substances, Anxiety, Depression, ADHD*, and *Other Diagnoses*). The highest (broadest) entity level is represented by the 3 different highlighter colors on the right-hand side of the table, with blue denoting *Prescription Stimulants*, green denoting Substances & Medications, and yellow denoting *Diagnoses*.

Entity	Total Labeled Tokens	Total Unique Tokens	
Prescription_Stimulants	140	43	**0: Prescription Stimulants**
Adderall	10283	65	
Vyvanse	2614	25	
Ritalin	291	12	
Dexedrine	184	23	
Concerta	55	5	
Modafinil	32	4	
Focalin	30	2	
Anti-Depressants	574	76	**1: Prescription Medications**
Benzodiazepines	79	26	
Antipsychotic	36	21	
Supplements	771	71	**2: Over The Counter Medications**
Sleeping_Pills	719	46	
Antacids	32	4	
Caffeine	216	56	**3: Other Substances**
Marijuana	85	21	
Nicotine	66	28	
Alcohol	56	19	
Anxiety	252	10	**4: Anxiety**
Depression	151	12	**5: Depression**
ADHD	1146	32	**6: Attention Deficit Hyperactivity Disorder**
Eating_Disorder	45	15	**7. Other Diagnoses**
Narcolepsy	34	13	
OCD	32	7	
Bipolar	13	6	
WILD	1,052,464	9078	**8. Non-Entities**
Total Entity Tokens	**17,936**	**642**	
Total	**1,070,400**	**9720**	

Figure 2: Number of Unique and Total Tokens Associated with each Entity and Sub-Entity, Colored by Grouping

Once the data was tokenized and entity label classifications were assigned to each token in training and set sets, every model was evaluated using accuracy, precision, recall, F1-score, and support metrics for each entity class. These evaluation metrics are recorded in Table 2 below which is split into three sections to show the metrics for each approach (heuristic, traditional, and transformer). While most scores are far from ideal, higher scores (above 0.70) for specific modeling approaches are colored in green.

Table 2. Evaluation Metrics (Accuracy, Precision, Recall, F1-Score, & Support) for each Modeling Approach

Entity	Accuracy	Precision	Recall	FI-Score	Support
Modeling Approach 1 - Heuristic Modeling Evaluation Metrics					
Prescription Stimulants	0.094	0.10	0.09	0.10	3,278
Prescription Medications	0.402	0.18	0.40	0.24	112

OTCs	**0.750**	0.12	**0.75**	0.20	56
Substances	**0.763**	0.11	**0.76**	0.19	80
Anxiety	0.139	0.36	0.14	0.20	36
Depression	**0.834**	0.62	**0.83**	**0.71**	36
ADHD	0.037	0	0.04	0	136
Other Diagnoses	0.440	0.04	0.44	0.07	25
Modeling Approach 2 - Traditional Machine Learning Modeling Evaluation Metrics					
Prescription Stimulants	**0.833**	**0.90**	**0.89**	**0.89**	318
Prescription Medications	**0.955**	0.57	0.40	0.47	20
OTCs	**0.973**	**0.71**	0.67	0.69	18
Substances	**0.948**	**0.78**	0.58	0.67	36
Anxiety	**0.975**	**0.72**	**0.91**	**0.81**	23
Depression	**0.978**	**0.85**	**0.74**	**0.79**	23
ADHD	**0.934**	**0.80**	**0.81**	**0.80**	63
Other Diagnoses	**0.970**	0.29	0.22	0.25	9
Modeling Approach 3 - Transformer Modeling Evaluation Metrics					
Prescription Stimulants • *Adderall, Vyvanse...Etc.*	0.441	0.65	0.44	0.53	2,731
Medications & Substances • *Prescription Medications* • *OTCs, Substances*	0	0	0	0	282
Mental Diagnoses • *Anxiety, Depression* • *ADHD, Other Diagnoses*	0.378	**0.75**	0.38	0.50	241

Discussion

Returning to the original goal of this investigation which was to extract specific mentions of diagnoses, medications, substances, or stimulant brands in relation to the user, three different modeling approaches were designed and tested on the 2,001 posts extracted from the subreddit r/adderall. By far, it seems as though the support vector machine model performed best out of all three approaches upon comparing the evaluation metrics depicted in Table 2. While the heuristic approach did alright in terms of accuracy when classifying OTCs, Substances, and Depression, the precision was still incredibly low for these entities while the other entities performed poorly in all metrics. I believe this is because I chose to filter tokens in the data that were common between more than one entity. For example, if one token was found to be classified as some entity (e.g. Adderall) but the same token id in a different post and text position was also found to be classified as a different entity (e.g. ADHD), then I was unable to use that token in my heuristic model. This is because my rule-based approach relied on the assumption that each token could only be associated with one entity. As a result, it makes sense that entities with more interrelated or similar topics (e.g. prescription stimulants and prescription medications are similar and most likely shared word-piece tokens) would perform worse since those tokens would not be available to assist in classification. On the other hand, entities with language that is more unique to that entity would perform better with this approach which might explain why the entities OTCs, Substances, and Depression showed a slightly better performance.

The poor performance of modeling approach 3 (the transformer model) was surprising given that transformer models are generally state-of-the-art for various unstructured natural language analysis problems. After trying out different tokenization schemes and testing out many different parameters on the transformer model to no avail, I concluded that there may have been some limitations in the data and annotation process which needed to be addressed and are explained in more detail below.

Limitation 1 - Lack of Annotated Data & Language Ambiguities

The first limitation which may have contributed to the poor performance of the transformer model and decreased the performance of other approaches was the lack of annotated data for each entity and also the uncertainty surrounding a single annotator. While I tried to develop an annotation scheme with edge cases and lots of detail, I still encountered many ambiguous cases during the annotation process, for example, a user implicitly referring to a diagnosis or describing stimulant use habits in relation to a friend or relative. Since posts on reddit are more casual in nature and often contain misspellings or incomplete sentences, I would spend much more time accounting for these uncertainties if I were to remake my annotation guidelines. I would have also annotated all posts at least twice (or recruited a second annotator) so that I could compare the results to validate whether my annotations were consistent.

Limitation 2 - Imbalanced Data

Another limitation which may have affected the performance of these models was the imbalance of entity versus nonentity tokens. Going back to Figure 2, one can see that while there were 17,936 tokens that were classified as some type of entity token, the remaining 1,052,464 tokens were classified as non-entities. While some imbalance of data classes should not disrupt training too much, the fact that the number nonentity tokens was over 58 times the number of entity tokens might have led the model to simply categorize everything as a non-entity token (since this would actually lead to a reasonable level of accuracy). In the future, it will be critical to find ways to balance the data during training so that the total number of non-entity tokens is more roughly comparable to the total number of entity tokens.

Limitation 3 – Post Truncation during Tokenization

A third and final major limitation, particularly in the transformer modeling approach, was the use of truncation which was required for the distilbert tokenizer (max length was 512 tokens). By truncating portions of posts, this approach may have removed parts of sentences with annotations that were critically helpful for training and testing. Due to the already lacking number of annotations, truncating posts may have further impacted the model's ability to correctly classify entities by removing even more annotated data. In the future, it would be helpful to test out different methods of accounting for this token limit such as applying a more stringent word count filter before selecting posts or increasing the number of manually annotated posts.

Conclusions

In total, 2001 posts were extracted from the subreddit r/Adderall, split into training and test sets, and then annotated using the brat tools standalone server according to the annotation guidelines specified in the methods section. After importing the annotation offsets into python, three different modeling approaches were designed to detect specific entity mentions in the posts, including a heuristic-based modeling approach, a traditional machine learning approach, and a transformer modeling approach. While the transformer model performed very poorly on all evaluation metrics, the heuristic approach did alright in terms of accuracy when classifying OTCs, Substances, and Depression (although its precision and F1-scores were very low). The traditional machine learning approach performed the best overall using a support vector machine to classify each entity (although F1-Scores were still low). The overall poor performance of these models, while unfortunate, was helpful in motivating the identification of critical limitations which should be addressed in future investigations, including annotating more data, accounting for ambiguous language, balancing training data by entity classes, and removing truncation. Despite these less-than-ideal results, the future improvements identified by this investigation along with the overall framework provides a critical step in laying the groundwork to begin designing a system to effectively identify popular use cases around prescription stimulants to inform more targeted treatments and better patient outcomes.

References

1. Jetson, J., Hu, A., & Xiong, F. Trends in Central Nervous System (CNS) Stimulant Prescription Drugs, 2012-2022. *Prescription Monitoring Program; Washington State Department of Health*; 2023, July. https://doh.wa.gov/sites/default/files/2023-08/600-092-StimulantPMPReport.pdf?uid=64f8f14ca269f
2. Robitaille C. "This drug turned me into a robot": an actor-network analysis of a web-based ethnographic study of psychostimulant use. *Canadian journal of public health = Revue 8anadienne de sante publique*, *109*(5-6), 653–661; 2018. https://doi.org/10.17269/s41997-018-0149-z
3. Favrod-Coune, T., & Broers, B. The Health Effect of Psychostimulants: A Literature Review. *Pharmaceuticals (Basel, Switzerland)*, *3*(7), 2333–2361; 2010. https://doi.org/10.3390/ph3072333
4. DeSantis, A. D., & Hane, A. C. "Adderall is definitely not a drug": justifications for the illegal use of ADHD stimulants. *Substance use & misuse*, *45*(1-2), 31–46; 2010. https://doi.org/10.3109/10826080902858334
5. Abdelnour, E., Jansen, M. O., & Gold, J. A. ADHD Diagnostic Trends: Increased Recognition or Overdiagnosis?. *Missouri medicine*, *119*(5), 467–473; 2022.
6. Kim, M. L., Dalvi, N., Valerio, D. D., Strickler, G. K., & Young, L. D. Prescribed stimulant medications: Trends in the last decade, pre and post COVID-19 response. *Exploratory research in clinical and social pharmacy*, *11*, 100314; 2023.
7. Boe B. PRAW: The Python Reddit API Wrapper. https://github.com/praw-dev/praw/; 2012. [Online; accessed 2024-01-15].
8. Stenetorp B., Pyysalo S., Topić G., Ohta T., Ananiadou S., & Tsujii J. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics; 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation.
9. Wolf, Thomas, et al. "HuggingFace's Transformers: State-of-The-Art Natural Language Processing." *ArXiv:1910.03771 [Cs]*, 11 Feb. 2020, arxiv.org/abs/1910.03771.