An Analysis of DSMs 1-5 Using Natural Language Processing

Analise Bottinger, Ryan Costa, Justin Guthrie, Tyler Nguyen, Joshua Yu
Northeastern University, Boston, MA, USA

Abstract

Our research makes use of natural language processing (NLP) to study how the Diagnostic and Statistical Manual of Mental Disorders (DSM volume 1–5) has evolved over time. The goal of this project is to track changes in the language surrounding mental disorders and their diagnoses and glean where research is headed. The group analyzed the DSM texts as a whole and specific disorders within DSM through Sankey diagrams, word clouds, topic modeling, and network graphs. The study was fruitful in mapping historical changes to the DSM, as well as gaining insight into future research. The group found more focused language around the disorders studied and greater emphasis on the effects of substance use, sleep, and hormonal changes on mental disorders as the DSM has evolved.

Introduction

As college students, the prevalence of mental health struggles are far from unfamiliar. In fact, The American College Health Association's 2017 study of 63,000 college students found that 60 percent reported feeling overwhelming anxiety, and 40 percent reported depression symptoms so severe that it was difficult for them to function. Overall, 50 percent of college students rate their mental health as poor [1]. As mental health concerns continue to grow, the diagnostic categories and criteria inevitably grow with it. One of the most popular mental illness diagnostic tools, known as the Diagnostic Statistics Manual (DSM), was first developed in the pre World War 2 era as a means of classification rather than clinical usefulness. Today, the DSM 5 is the standard for clinical classification and treatment of mental disorders. Even though the applications of the DSM have rapidly improved, The need for adequate mental health services, specifically for college-age adults is more prevalent now than it has ever been.

This project aims to dissect changes in the language surrounding mental disorders over the 5 installments of the DSM. Through mapping these changes, we hope to uncover trends across and within disorders through the progression of time, and potentially use these trends to predict the future of mental health discussions.

Methodology

* Python scripts to run start with "dsm", everything else are libraries to help with analysis

The group first collected PDFs of DSM 1-5 online. A function was then utilized to preprocess and parse the texts based on a user-specified page-range. The texts were preprocessed using "en_core_web_sm". From there, the texts could be passed to any visualization or analytical function (such as topic modeling).

The analysis was divided into two general sections: analysis of DSM texts as a whole and analysis of specific disorders. Visualizations for DSM texts as a whole were word clouds and a Sankey diagram that maps the most common words across each DSM edition to find commonalities and differences.

The specific disorders chosen by researchers included depressive, anxiety, and psychotic and schizophrenia disorders. These disorders were chosen due to the evolution of their understanding over time based on preliminary research as well as their relevance to the discussion on mental health today. Only DSMs 3-5 were used for this analysis as DSMs 1-2 were much less comprehensive in terms of the breadth and depth of discussion on mental disorders.

The texts for the specific disorders were fitted on a Latent Dirichlet Allocation model to find three topics per section containing 10 words. These topics were then plotted on bar charts based on relative word weight within each topic. This gave the group a general, quantifiable lens to see how relevant language for a specific disorder has evolved across the DSM.

Co-occurrence matrices were then made for the chosen disorders for each DSM which was then used to make a co-occurrence network graph. Edges and nodes in the network graph were filtered based on an arbitrary edge-weight threshold (k=0.00095) in order to make the networks less dense and more readable. The network graph allowed researchers to analyze commonly co-occurring words used to describe a disorder across DSM versions. The topic modeling was helpful here as researchers could use the topics as a starting point to find specific words in the network graphs and see how their use within the DSM has evolved over time.

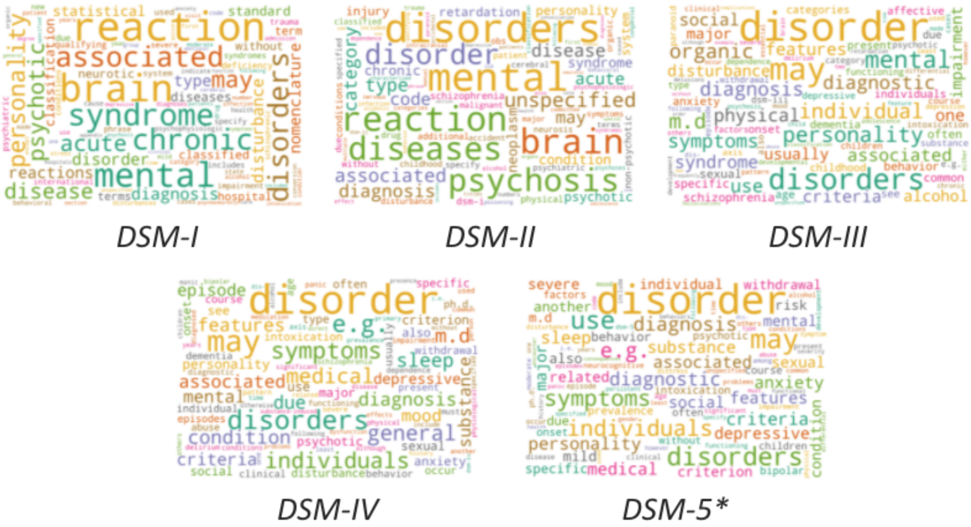
Analysis

General Word Cloud Analysis

A word cloud was created for each DSM edition to help visualize the change in language/topics between each DSM. Since the DSM's first published edition in 1952, its usage has evolved from a strictly classification piece of literature to an overarching text used to both diagnose disorders and help guide their treatments. In addition, new disorders have continued to be added while other behaviors have been amended or even removed – a testament to the progress made in the understanding of mental illnesses since the DSM's inception. Some words of note that have manifested in these five word clouds are "reaction" and "psychosis". Although prevalent in DSM I and II, the word "reaction" was dropped in subsequent editions. This is due to the evolution in how disorders as a whole were regarded. At the time of DSM I (and earlier), the predominant dictations on mental disorders followed the psychodynamic approach wherein mental abnormalities were treated as a reactionary manifestation to early life experiences.

Subsequent DSM editions later included biological understanding of mental illness. Similarly, the word "psychosis" began quite prevalent in early DSMs but has dwindled in recent editions. This is again due (in large part) to a greater understanding and finer classification of mental disorders. Many mental disorders including Bipolar Disorder and Schizophrenia were together bundled into the catch-all term "psychosis" in DSM I – a practice no longer reflective of present-day research and understanding which we will expound on later in the analysis of specific disorders.

Figure 1. *General Word Clouds*



DSM-I DSM-II DSM-III

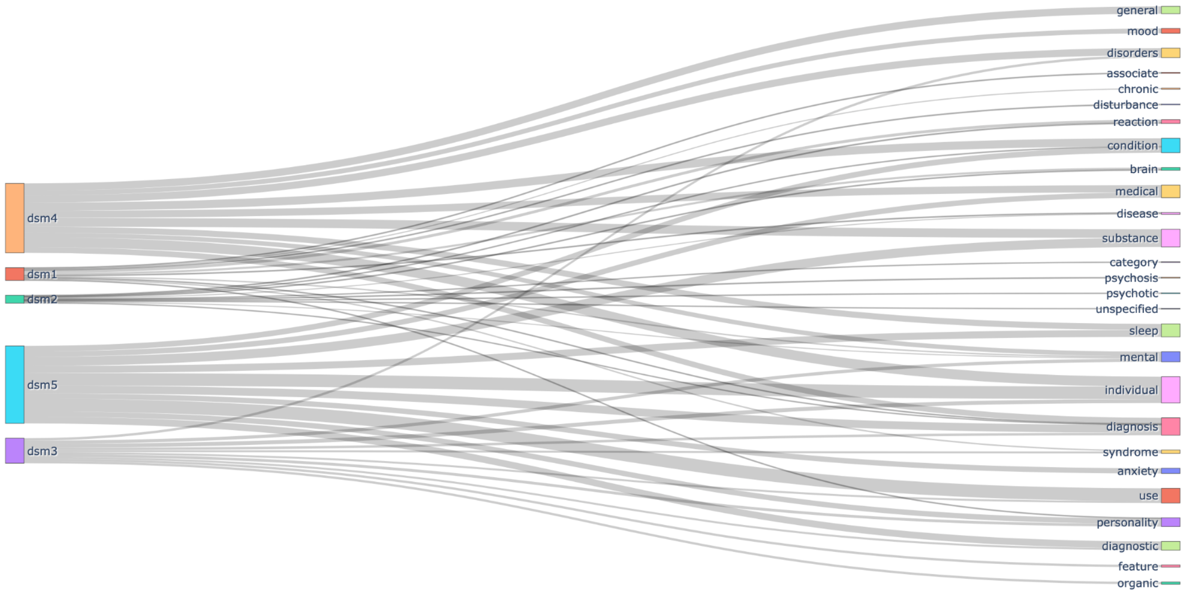
DSM-IV DSM-5*

General Sankey Diagram Analysis

A general Sankey diagram was created to show textual relationships between the different DSM versions as whole texts. The Sankey diagram that was created showed some interesting insights. Among all five versions of the DSM, among the most common words were “individual”, “diagnosis”, “substance”, and “sleep”. “Individual” is understandable as shared mental disorders are not common, and in each DSM “individual” is used to describe individual cases, especially since on a case-by-case basis, each diagnosis varies. For “diagnosis”, this also makes sense as the purpose of the DSM is to diagnose mental disorders, and it is a “diagnostic” manual. It is important to note, however, that in this Sankey diagram, this was the only word shared by all five versions. This is likely because even as diagnoses change over time, the purpose of the novel still remains the same, especially since this is a general look at each version of the novel. Another common word was “substance”. This word was most common in the later versions of the DSM (DSM-IV and DSM-5), which could reflect the increased use of pharmaceuticals to treat mental disorders, but could also shed light on increasing substance abuse.

Finally, “sleep” was also found most commonly in the later versions of the DSMs (DSM-IV and DSM-5). This could indicate the discovery of sleep-related disorders like narcolepsy and sleep apnea, but could also be indicative of a growing problem of people either not getting enough sleep or having irregular sleep patterns related to other diagnoses (like depression or anxiety).

Figure 2. *General Sankey Diagram*



DSM Depression Section Analysis

* Visualizations from here onwards will be included in an Appendix at the bottom due to their size

The topic modeling performed on DSM 3-5 depression sections (Appendix 1.1) highlights a few notable changes over time in how depression is discussed within the DSM. A few words which stand out in the latter DSM editions that are not in DSM 3 include “substance” (DSM IV & 5), “medication” (DSM IV & 5), “premenstrual” (DSM 5), and “suicide” (DSM 5). Conversely, DSM 3 topics seemed to include other disorders such as schizophrenia, dementia, and anxiety (also in DSM IV & 5 topics). “Sleep” and “hallucination” were also prevalent in DSM III topics, but not included in DSM IV & 5 topics. “Manic” and “hypomanic” were prevalent in DSMs III & IV, but not in DSM 5.

As far as commonalities across the DSM editions, “depressive” and related words (“depression”, “depressed”) were, as expected, ubiquitous throughout the topics, with “depressive” holding the highest relative weight across all topics.

These observations of words and their weights within topics will be used as a starting lens from which to analyze the network co-occurrence graphs in order to better understand how these words (and possibly others) and their interactions have evolved in the DSM as our understanding of depressive disorders has evolved.

From the network graphs (Appendix 1.2-1.4), the most notable evolution across the DSM editions has to do with the complexity of the networks. It would seem that from DSMs 3-5, there is a trend of decreasing graph complexity for depressive disorders. This might indicate more focused language surrounding depressive disorders, tying in to the earlier observation where other disorders such as schizophrenia, dementia, anxiety, etc. were present in the DSM III topics. It is also worth noting that in DSMs III & IV, depressive disorders are classified as "mood" disorders, and only in DSM 5 are they actually classified as depressive disorders with bipolar disorders getting their own section for the very first time. This coincides with a decrease in co-occurrence of the word "mood" with other words in the graph.

With bipolar disorders getting their own dedicated section, words relating to mania or manic disorders are no longer in the DSM 5 network graph. This reinforces the idea of an increasing use of focused and specific language to classify depressive disorders as our understanding of depression and related disorders increase.

DSM IV & 5 also include the word "substance" in increasing co-occurrence with other words. This could indicate a growing awareness of how substance use/abuse could affect the onset of depressive disorders. The word "premenstrual" also first appears in the DSM 5 graph co-occurring with "depressive" and "diagnosis", potentially indicating growing awareness of how women's reproductive cycles may affect the diagnosis of depressive disorders.

"Suicide" also first appears in the DSM 5 network graph, indicating a growing focus/awareness on the potential risks of depression as it relates to suicidality. This could be a part of a growing effort to highlight the severity of depressive disorders, in tandem with the increased use of focused language for depressive disorders mentioned above.

DSM Anxiety Analysis

Topic modeling was also performed on DSM anxiety diagnostic criteria sections for DSM III through DSM 5 (Appendix 2.1). Network graphs were then created (Appendix 2.2-2.4).

Across the three DSM topic modeling graphs, the association and the relative word weight of "anxiety" is one of the largest changes among the DSMs. In DSM III, "anxiety" is associated with words such as "common", "trauma", and "schizophrenia." in the DSM IV, anxiety's relative word weight increases from 0.045 in DSM III to 0.07 in DSM IV, becoming also more closely associated with words such as "social", "individual", and "situation". In the DSM 5, anxiety's relative word weight spikes to ~0.1, and its associations shift towards such as "separation", "association", and "medication". Evidently, the associations of words with anxiety surprisingly become more generalized as the DSM volumes progress, associating less with specific disorders and more with generalized terminology. Additionally, the word "medication" first appears in association to anxiety in the DSM 5, and first appears within the topics overall in the DSM IV. In the DSM IV, however, medication is more strongly associated with specific phobias, traumas, and obsessions. This coincides with an increase in the use of medication to treat generalized anxiety disorders outside of more severe, panic-type disorders. Based upon these results, a series of co-occurrence network graphs were created to further dissect these trends (Appendix 2.2-2.4).

Across the DSM III, there is a very large co-occurrence network compared to what is seen for DSMs IV and 5, similar to what we saw for depression as well. This could suggest that there is a broadening of language from DSMs I-II, but still little to no focused language to describe what anxiety is. For instance, there is a strong co-occurrence of disorder-specific words to anxiety, such as specific phobias and obsessions. In the DSM IV, there is still a larger co-occurrence network, however, still smaller than that of the DSM III. The co-occurence begins to trend towards more general anxiety terminology, with strong occurrences between anxiety and generalized diagnostic terminology as seen in the topic modeling graphs. This trend is still seen in the DSM 5, where the co-occurrence network has become more refined. However, the terms that co-occur to each other are arguably more general than those seen in previous DSMs, with strong co-occurrence between words such as "phobia", "diagnosis", "specific", and "condition". While this could look like a regression in focused language surrounding anxiety, it could also show that as the DSMs have progressed, the understanding of how anxiety connects to other disorders has also increased, and show that each of those disorders have different language for what anxiety looks like.

Across the three co-occurrence network graphs from DSMs 3-5, an interesting trend to note is the decrease overall in the co-occurrence frequency between words. Once again, this furthers the idea that as the DSM has progressed, more focused, specific language is being used to describe anxiety disorders, as well as characterize the overlap of anxiety with other disorders in the DSM. The more focused the language has become, the less likely it is to show up with strong co-occurrence, or even amongst a topic within the DSM topic modeling graph.

DSM Psychotic Disorder Analysis

Topic modeling for psychotic disorders saw a large focus around sensibly ubiquitous terms like "Schizophrenia" and "Psychotic", as schizophrenia leads as the most prominent psychotic disorder. One difference that develops primarily between DSM volume III and the subsequent two installations is the separation of schizophrenia from the topic basket associated with "Psychotic." This suggests that come the fourth and fifth DSMs there was some additional stress placed on schizophrenia as an individual disorder. A likely explanation for this separation in topics would be either the increased understanding of schizophrenia as an individual disorder, or the recognition of a high relative frequency of schizophrenia cases amongst all psychotic disorders. Either of the aforementioned would result in more specific, isolated use of the word "schizophrenia" apart from "psychotic".

Development of "niche" words that then get adopted in future installations appears to occur somewhat linearly with DSMs 1-5. One such example of this development would be the incorporation of the word "substance" in DSM IV, that would then carry onto DSM 5. Mention of words of the same ilk appear throughout the evolution of DSM and demonstrate evolving considerations in the manual.

Further exploration of co-occurrence graphs demonstrates similar patterns to prior specific disorder analyses (network graph). The web of DSM 5 shrank significantly from its predecessor, and DSM IV follows suit. As with the anxiety and depression analysis, it is possible that a greater focus in language streamlined the web such that the co-occurrence of low frequency connections were no longer present.

DSM 5 displays a large focus on the idea of the “substance”, which coincides with prior exploration of topic analyses in the same realm. A main connection for substance is the word “induced” possibly relating to an inclusion and further explanation of substance induced psychosis. Such an expansion upon substances could correlate to a creeping drug epidemic, whereby substances with the capacity to produce psychotic indications increase in variety and sheer quantity as time passes.

The appearance of “spectrum” in the network potentially indicates movement away from definitions of disorders that follow a binary. Instead, it is possible that the DSM understanding of psychosis has evolved to include a wide range of symptoms (for those which are appropriate).

Conclusions

The work delves into how the DSMs have changed over time between versions one through five, in both a general and a specific focus. First, the library parsed through all versions of the DSM, taking out certain stopwords like “the”, “and”, and “is”, which would generate no significant insights. Next, a general Sankey diagram and set of word clouds were made to show textual relationships between the different versions, showing which words are more prevalent and how they have changed between versions. The Sankey diagram explored all the versions together, while a word cloud was made for each DSM version. Delving further, depression, anxiety, and psychotic disorders were explored. Visualizing patterns using topic modeling and network graphs, a few insights were made on the prevalence of certain words within each disorder as well as connections between different words and phrases within each disorder.

However, even though many accomplishments were made, there are some limitations to recognize. The runtime of the program is lengthy and generally computationally expensive as the DSMs are such long texts. Adding future DSMs would complicate this further. Additionally, the stopwords were manually inputted, which could create confusion on what is technically considered a “stopword” between organizations/people, if adopting the library. Overall, this project has shed light on relationships between words in each of the DSMs, how they differ between versions, and how they connect within specific disorders.

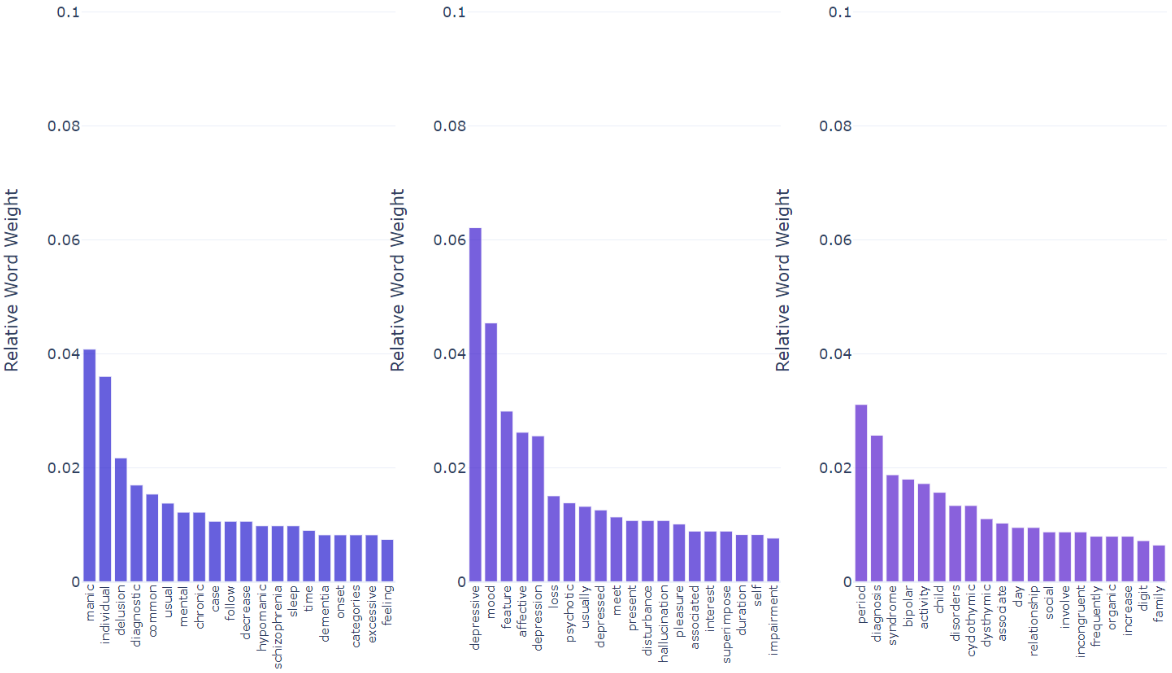
Author Contributions

Analise Bottinger: Analyzed anxiety disorders
Ryan Costa: creation/visualization/analysis of word cloud diagrams
Justin Guthrie: Created, visualized, and analyzed the general Sankey diagram
Tyler Nguyen: Analyzed psychotic disorders
Joshua Yu: Analyzed depressive disorders

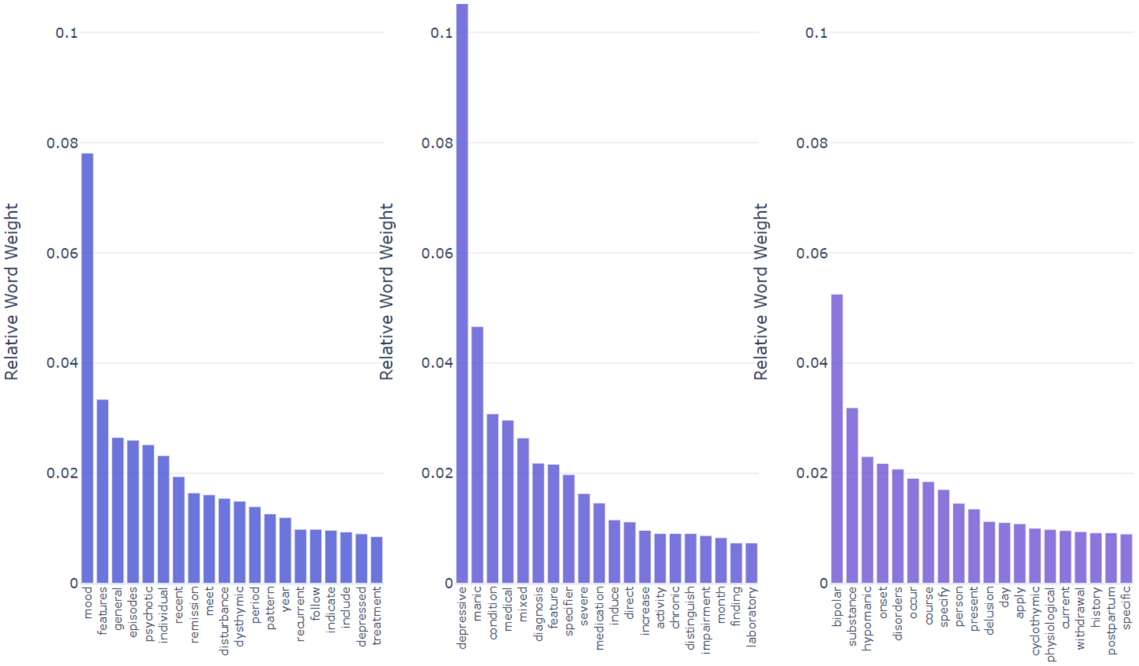
Appendix

1.1 Depression Topics

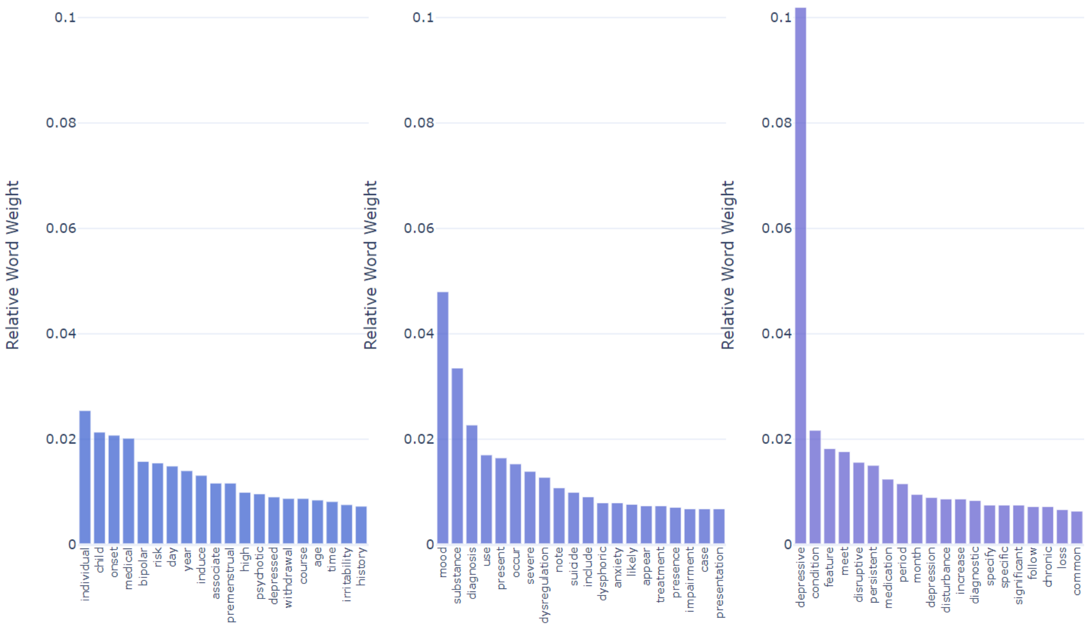
DSM 3 Topics



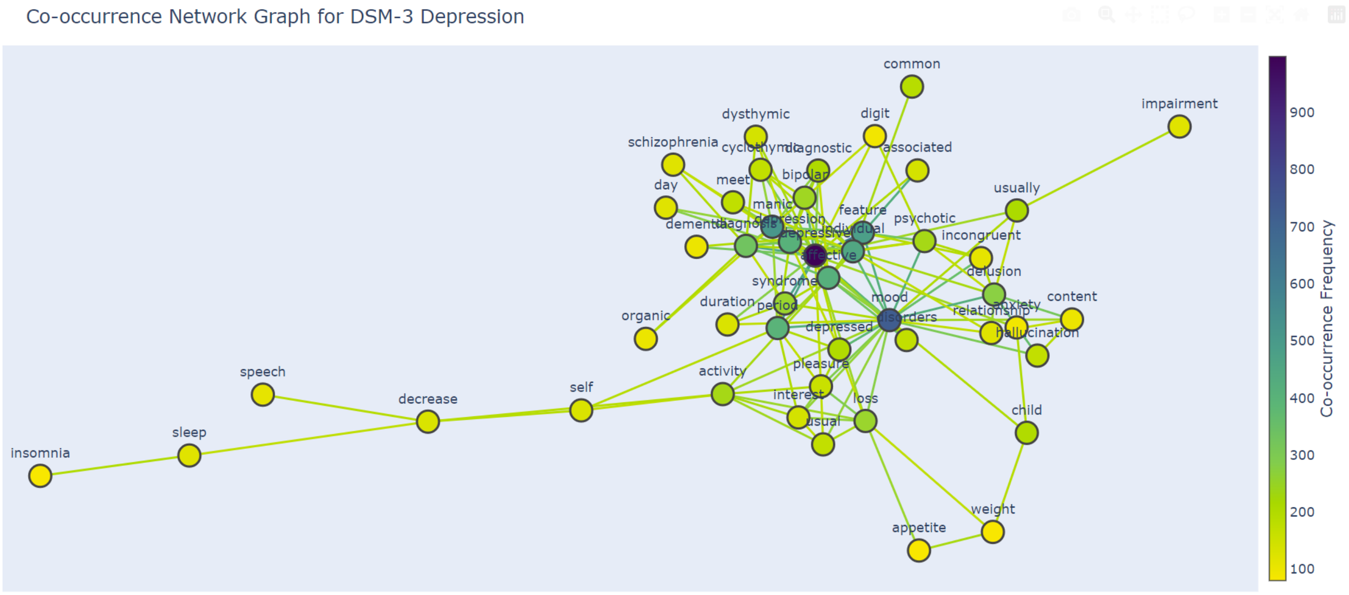
DSM 4 Topics



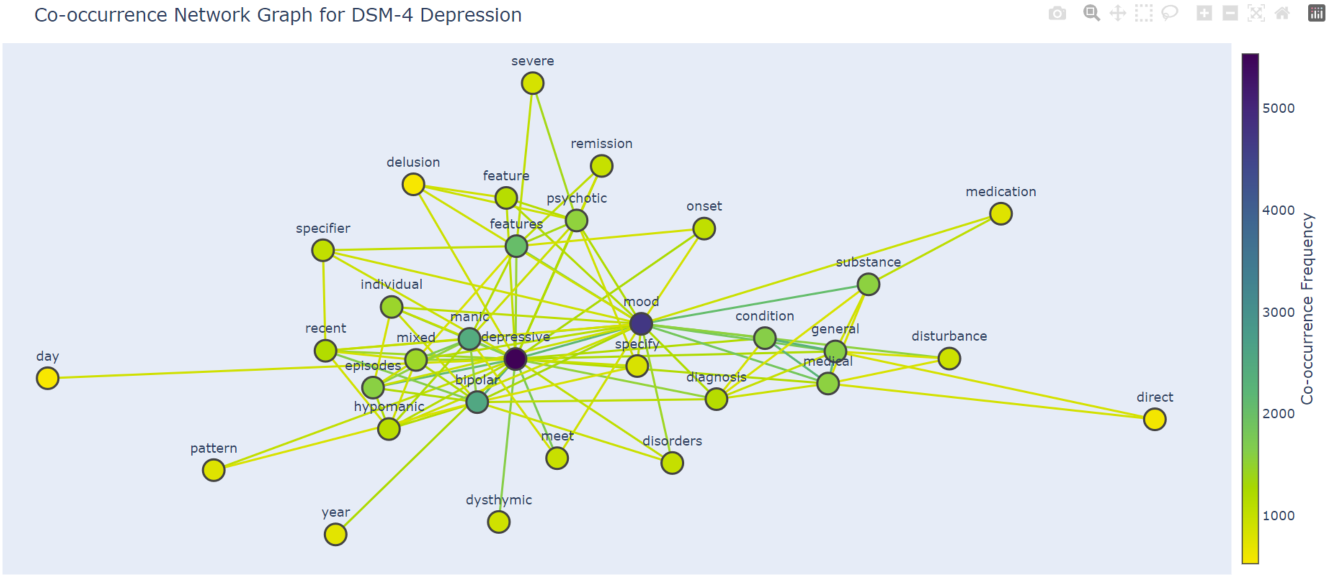
DSM 5 Topics



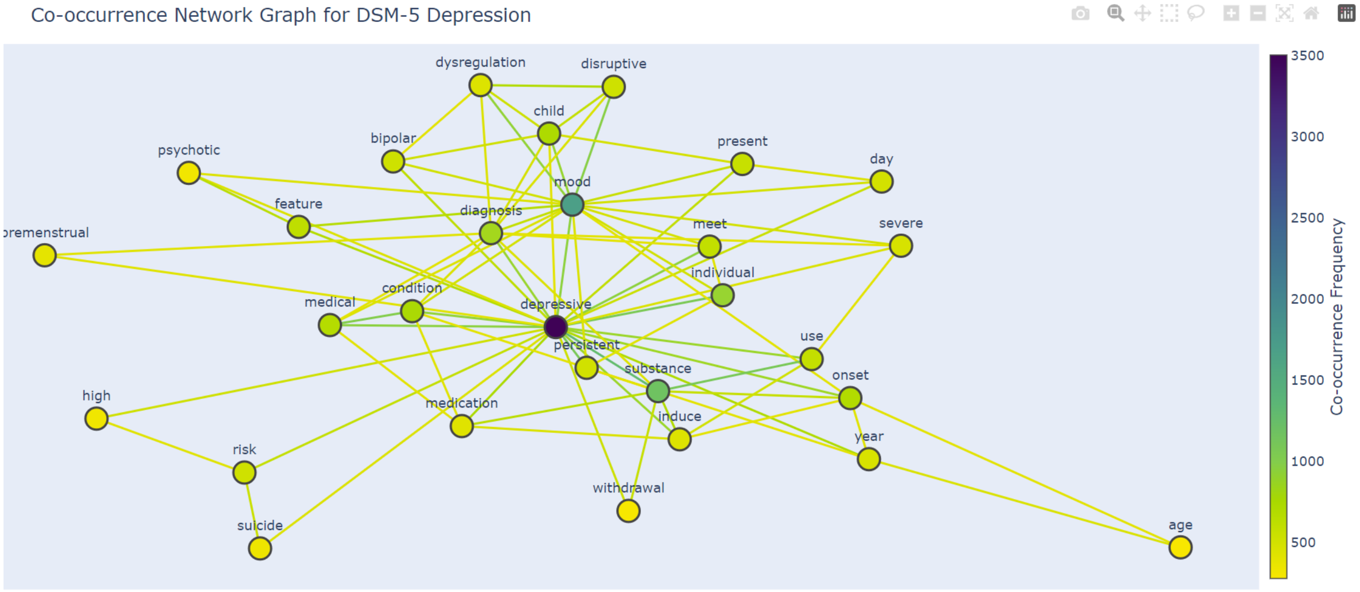
1.2 Co-occurrence Network Graph for DSM-3 Depression



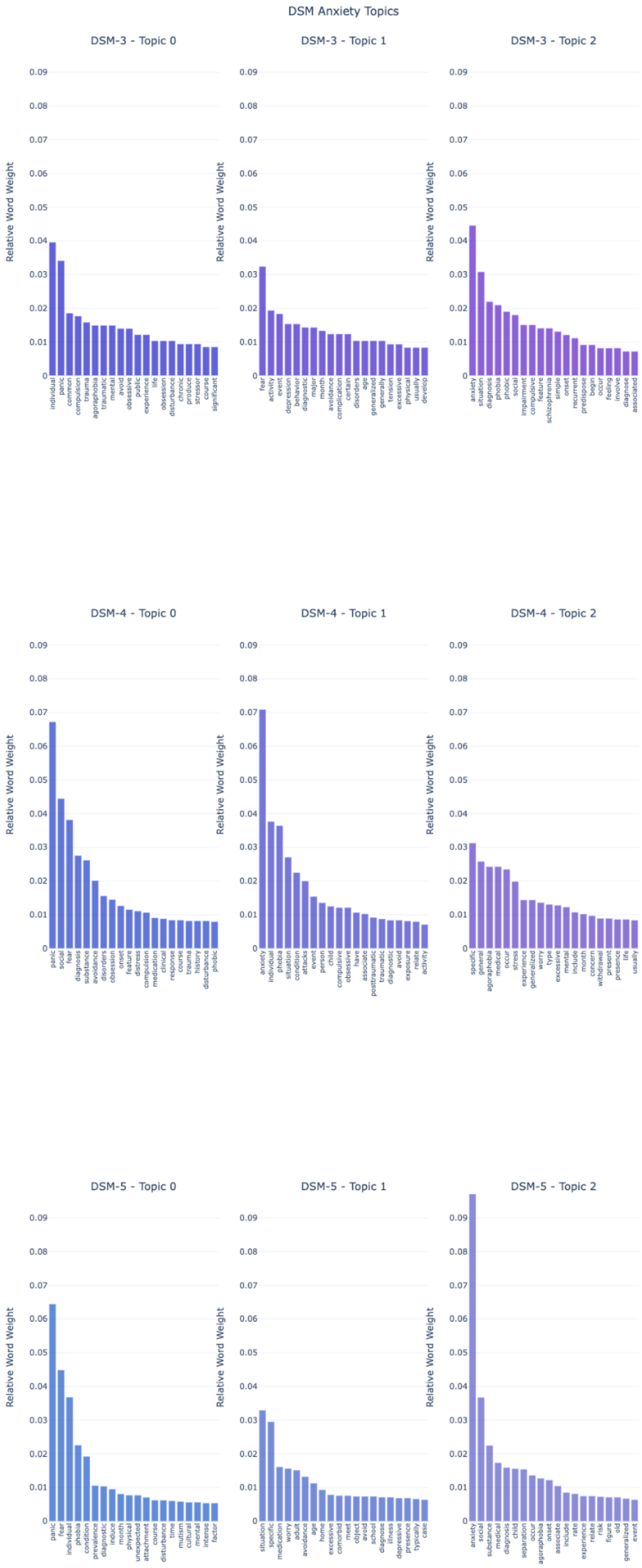
1.3 Co-occurrence Network Graph for DSM-4 Depression



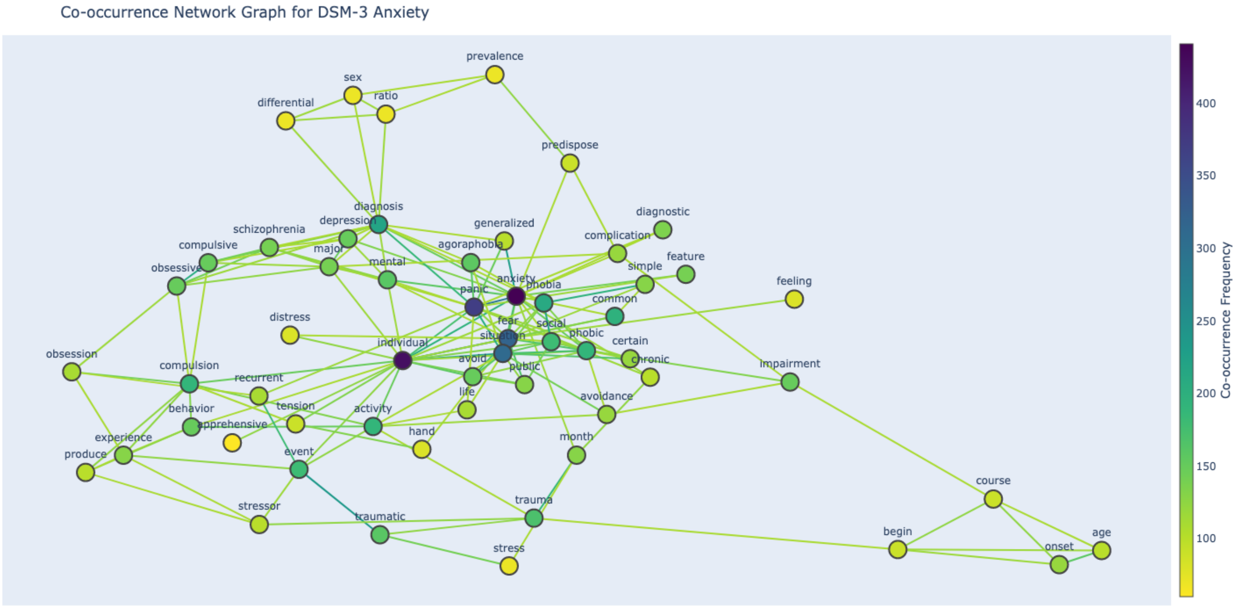
1.4 Co-occurrence Network Graph for DSM-5 Depression



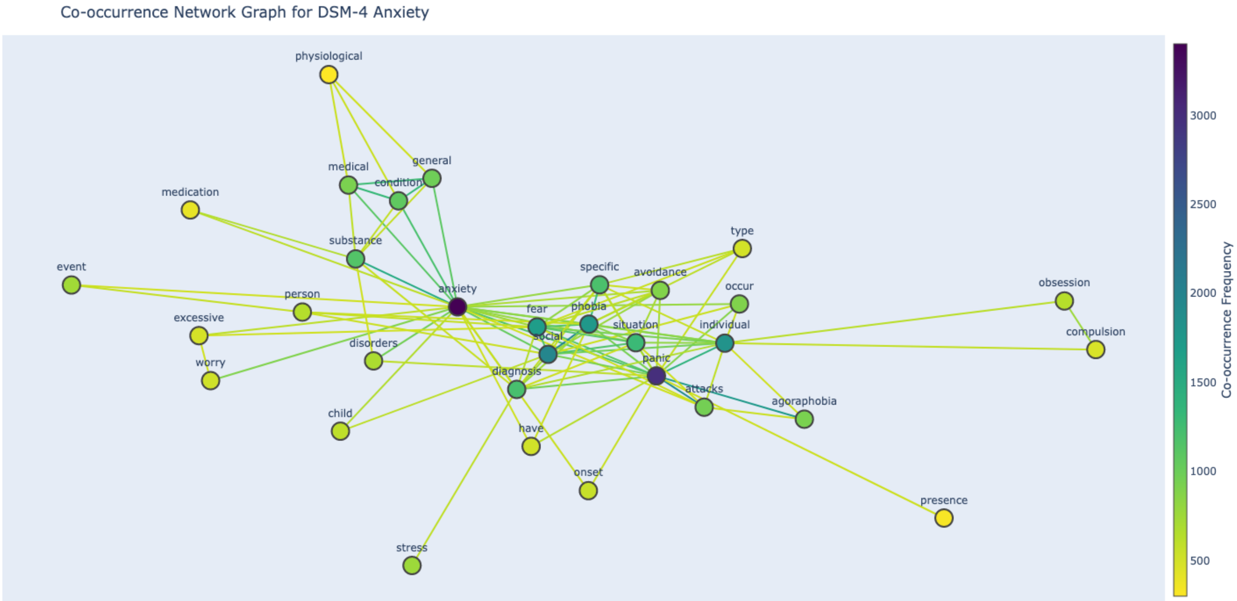
2.1 Anxiety Topics



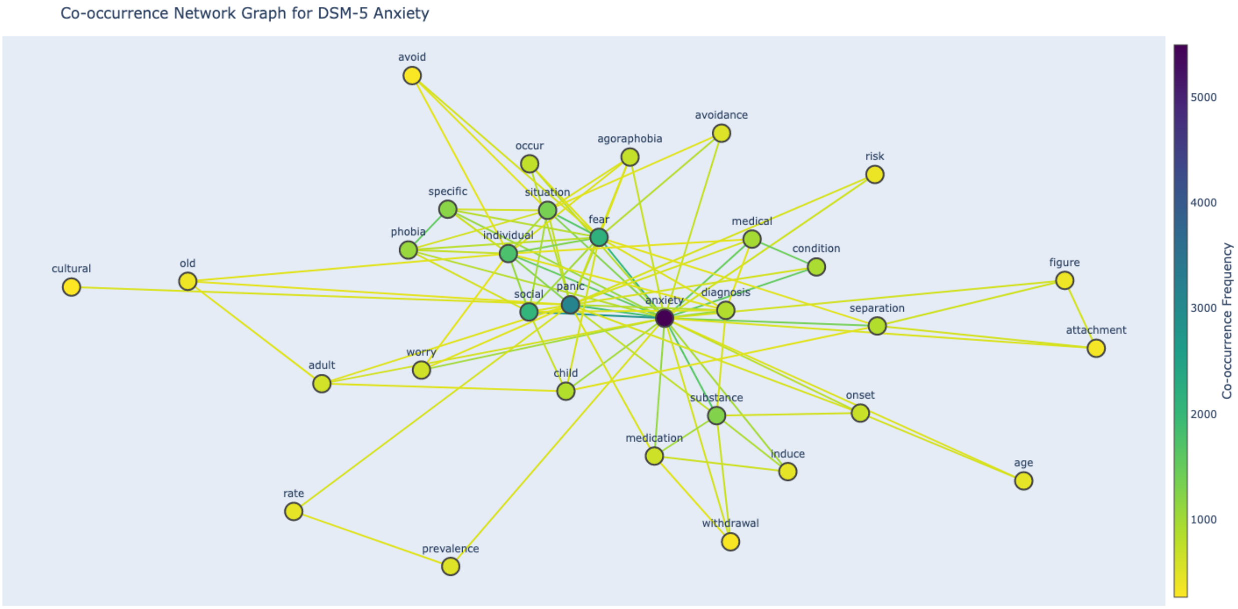
2.2 Co-occurrence Network Graph for DSM 3 Anxiety



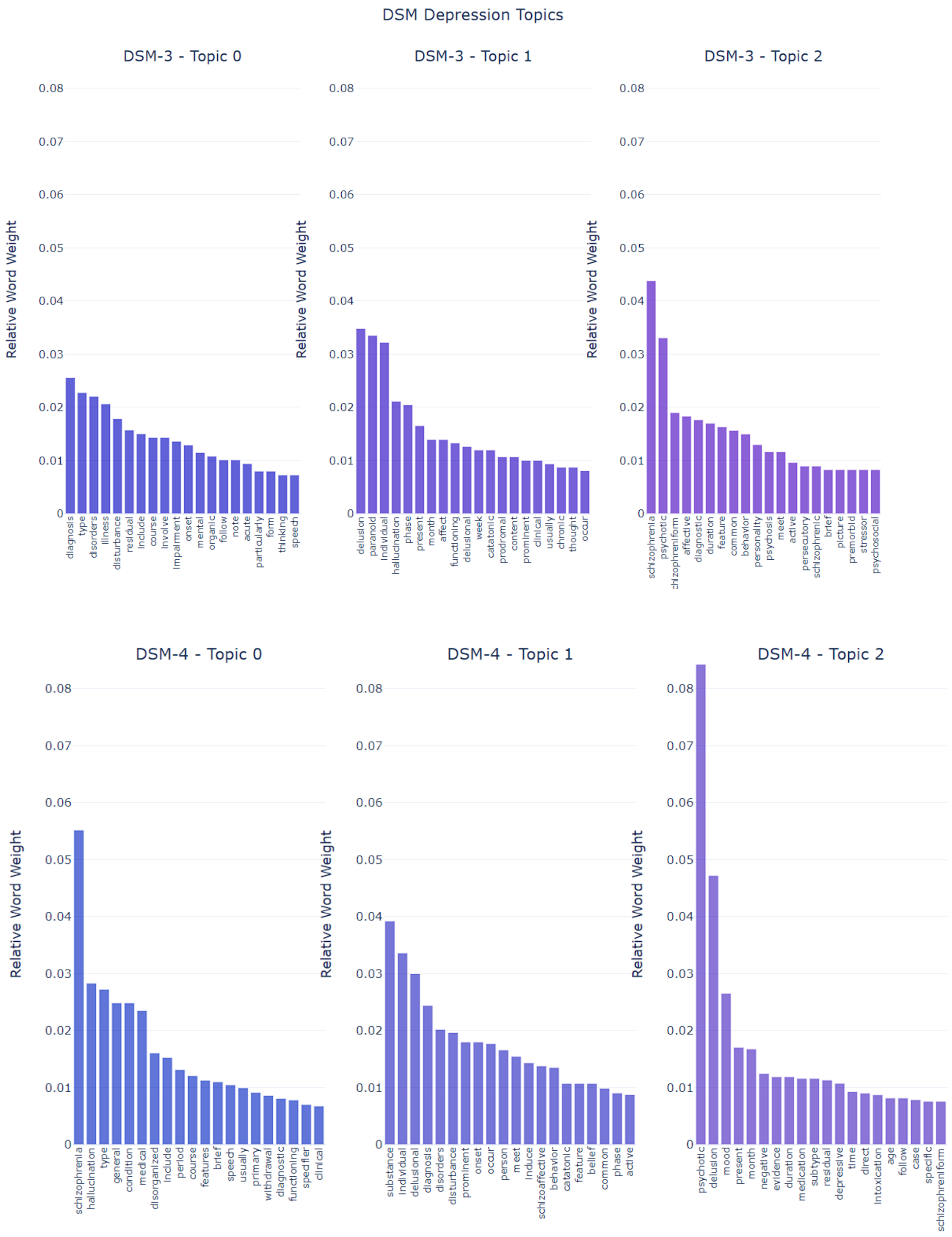
2.3 Co-occurrence Network Graph for DSM 4 Anxiety

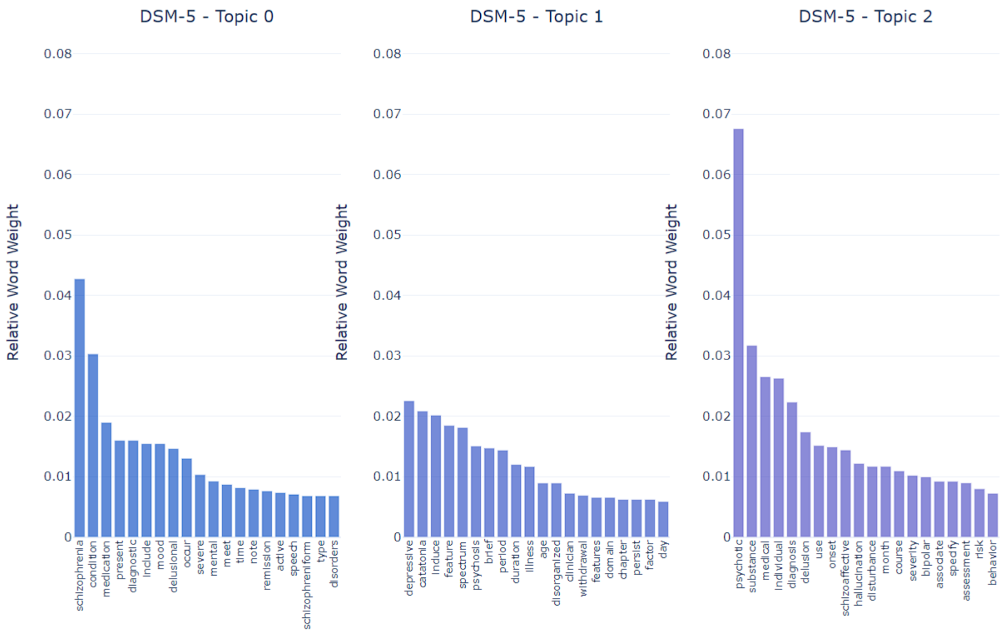


2.4 Co-occurrence Network Graph for DSM-5 Anxiety

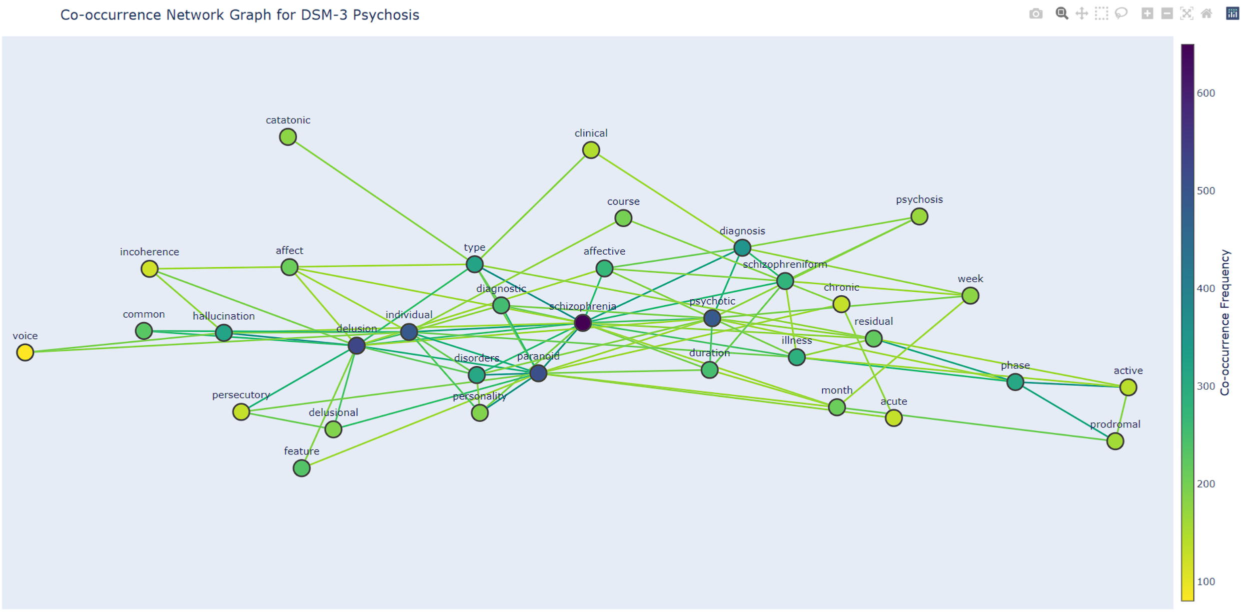


3.1 Psychosis/Schizophrenia Topics

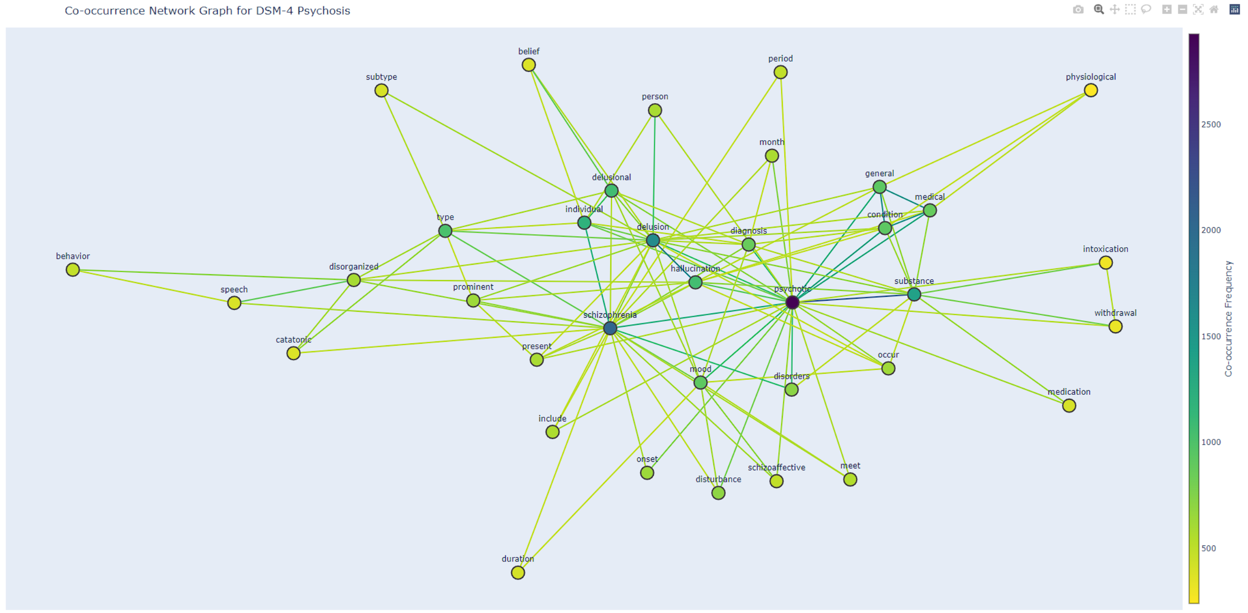




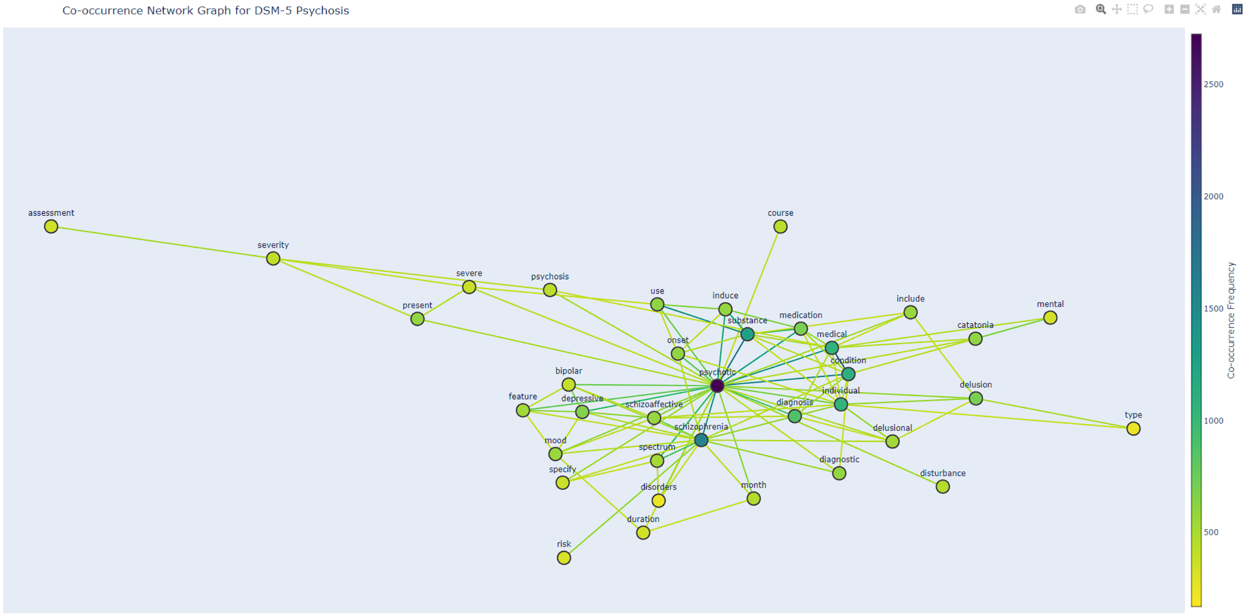
3.2 Co-occurrence Network Graph for DSM 3-Psychosis/Schizophrenia



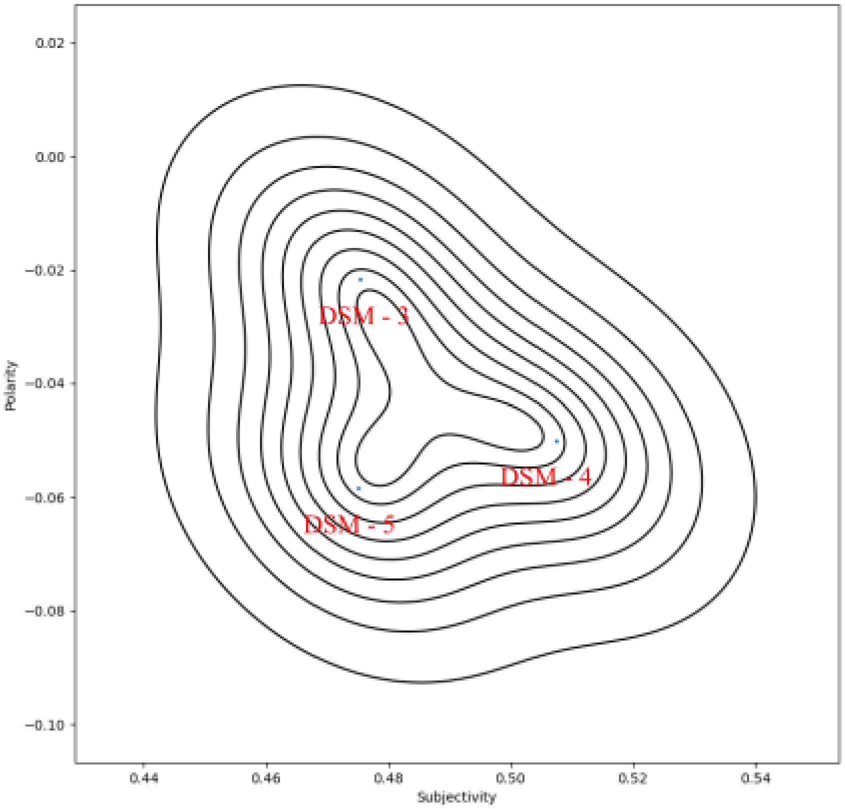
3.3 Co-occurrence Network Graph for DSM 4-Psychosis/Schizophrenia



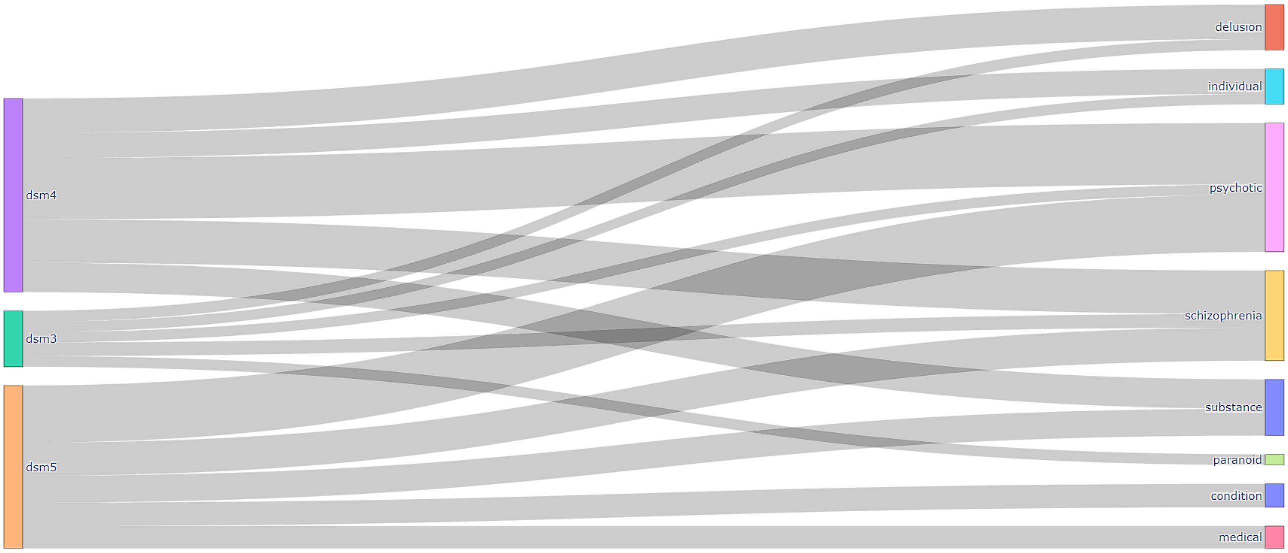
3.4 Co-occurrence Network Graph for DSM 5-Psychosis/Schizophrenia



3.5 DSM 1-5 Psychosis Sentiment Analysis



3.6 DSM 1-5 Psychosis Sankey



References

1. Chai, Michaela. “College Students Face Transitional Mental Health Challenges Entering Workplaces.” *Mindsharepartners*, Mindsharepartners, 9 July 2019, https://www.mindsharepartners.org/post/mental-health-and-college-students-entering-workplaces?utm_source=GOOGLE&utm_medium=CPC&gclid=CjwKCAjw_MqgBhAGEiwAnYOAen_pYX4MVKgbYgbmy189DrMS2_-4gDqcOusgfyo-zrsxRE1EEiB8qhoCXyUQAvD_BwE.
2. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, Mental Hospital Service, 1961.
3. *Diagnostic and Statistical Manual of Mental Disorders: DSM-3*. American Psychiatric Association, 1980.
4. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Association, 2017.
5. *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. American Psychiatric Association, 1994.
6. *DSM-II: Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 1968.
7. “DSM History.” *Psychiatry.org - DSM History*, https://www.psychiatry.org/psychiatrists/practice/dsm/about-dsm/history-of-the-dsm.