Hidden in Plain Text: Social Media Mining of Neurodivergent Discourse at Scale on X and Reddit

Kartik Thakkar, Utah State University
Ludia Eka Feri, Utah State University
Autumn Cuellar, Utah State University

Autumn Cuellar is a Ph.D. student in Engineering Education. Her undergraduate and master's careers were both in Computer Science. She believes that everyone can achieve their goals, regardless of physical ability. This is why Autumn strives to make engineering accessible for everyone.

Vanessa Tran, Utah State University

Bachelor's Degree: Civil Engineering, the University of Architecture, Vietnam. Master's Degree: Global Production Engineering and Management, Vietnamese - German University, Vietnam. Ph.D. Student in Engineering Education, College of Engineering, Utah State University.

Dr. Catherine Mcgough Spence, Minnesota State University, Mankato

Catherine Spence is an Associate Professor at Iron Range Engineering in the Integrated Engineering Department at Minnesota State University, Mankato. She received her PhD in Engineering and Science Education in 2019 and a BS in Electrical Engineering in 2014 at Clemson University.

Dr. Cassandra McCall, Utah State University

Dr. Cassandra McCall is an Assistant Professor in the Engineering Education Department at Utah State University (USU). Her research focuses on the intersections of disability, identity formation, and culture and uses anti-ableist approaches to enhance universal access for students with disabilities in STEM, particularly in engineering. At USU, she serves as the Co-Director of the Institute for Interdisciplinary Transition Services. In 2024, Dr. McCall received a National Science Foundation CAREER grant to identify systemic opportunities for increasing the participation of people with disabilities in engineering. Her award-winning publications have been recognized by leading engineering education research journals at both national and international levels. Dr. McCall has led several workshops promoting the inclusion of people with disabilities and other minoritized groups in STEM. She holds B.S. and M.S. degrees in civil engineering with a structural engineering emphasis.

Dr. Hamid Karimi, Utah State University

I obtained my Ph.D. in Computer Science from Michigan State University (MSU) in 2021. My main research interest during my Ph.D. was artificial intelligence (AI) for social good and focused on several exciting directions, including AI in education, computational politics, and misinformation detection.

Hidden in Plain Text: Social Media Mining of Neurodivergent Discourse at Scale on X and Reddit

Abstract. This empirical study examines how neurodivergent individuals express lived experiences, challenges, and strengths on major social media platforms, focusing on X (formerly Twitter) and Reddit. These platforms function as emancipatory spaces where users construct self-defined identities, resist deficit-based narratives, and foreground strengths often undervalued in traditional educational and professional contexts. Grounded in inclusive pedagogy, the study links online discourse to educational inclusivity and awareness. We analyze over 90,000 posts using a hybrid natural language processing framework that combines text representation learning (TF-IDF, BERT, and OpenAI embeddings), clustering, topic modeling (LDA and Top2Vec), and GPT-based models for interpretability and thematic synthesis. Our results reveal clear platform-specific patterns: X posts are concise and advocacy-oriented, emphasizing diagnosis, stigma, and self-acceptance, while Reddit discussions are longer and more introspective, focusing on coping strategies, academic challenges, and mental health support. Across platforms, recurring themes include executive dysfunction, sensory regulation, burnout, and the pursuit of accessible education and employment. Overall, the findings show that neurodivergent users use social media not only for self-expression but also for collective knowledge-building and emotional support. This work demonstrates a scalable social media mining framework that integrates statistical, semantic, and generative AI methods to amplify neurodivergent voices and inform the design of more inclusive learning environments.

1 Introduction

Neurodivergence includes a variety of natural neurological differences, such as Autism, ADHD, Dyslexia, often bolstered by unique strengths like non-linear thinking, spatial awareness, and accompanied by challenges like sensory overload, executive dysfunction, etc. Social media platforms like Reddit, TikTok, and Twitter (X) have emerged as significant arenas for neurodivergent individuals to express their needs, discuss their struggles, celebrate their achievements, and seek help [1–4]. These digital experiences offer an organic, rich source of data reflecting genuine, first-hand experiences of neurodiverse minds and communities. In this work, we explore the content, context, and thematic structure of online conversations and discussions about neurodivergence using a combination of unsupervised machine learning, topic modeling, and large language models (LLMs). Beyond examining neurodivergent experiences, this work contributes a scalable method for analyzing large-scale discourse. This approach is relevant to engineering education research, enabling the study of student experiences and challenges using LLMs and data-driven topic discovery to extract patterns from unstructured data. In this context, we aim to address the following Research Questions (RQs)

- **RQ1.** How do neurodivergent individuals engage in online spaces, and what are the common discussions? What underlying topics and themes emerge within texts shared by or about the neurodivergent community?

- **RQ2.** What experiences, strengths, and challenges do individuals within the neurodivergent community face– both in academic contexts and in everyday life?

2 Background

Recent research on neurodiversity has increasingly moved away from deficit-based models toward inclusive and strength-oriented frameworks that recognize cognitive variation as a natural aspect of human diversity [5–7]. Studies have moved beyond purely medical perspectives to highlight the social and structural factors shaping how neurodivergence is defined and understood, framing neurodiversity not only as biological variation but also as a product of cognitive marginalization [8]. Within engineering education, this shift is particularly significant. A growing literature highlights that neurodiverse learners frequently demonstrate cognitive assets such as creative problem-solving, pattern detection, persistence, and systems thinking that are especially valued in engineering fields [9–13]. Yet neurodiverse students, including those with ADHD, autism, or dyslexia, remain underrepresented in engineering majors. Researchers argue that viewing neurodiversity predominantly through a deficit lens limits participation in engineering programs and reduces the breadth of perspectives and innovation within the engineering workforce [14, 15].

In educational contexts, including STEM and engineering programs, studies highlight persistent gaps between inclusive policy goals and the lived realities of neurodivergent learners, advocating for systemic accommodations, co-designed training, and pedagogical approaches such as Universal Design for Learning that reduce reliance on diagnosis and disclosure [16–24]. Strengths-based perspectives further call for transforming engineering education systems so that they move beyond accommodation toward empowering students to leverage their cognitive differences as assets in learning, design, and innovation. At the same time, social media platforms such as Reddit, X (Twitter), TikTok, and Instagram have emerged as important spaces for neurodivergent expression, advocacy, and community-building [3, 25–27]. Prior work has used social media analytics and natural language processing to examine public sentiment, emotional processing, inclusivity, and representation, while also critiquing how online spaces both challenge and reproduce dominant medicalized narratives [28–31]. Methodological studies further demonstrate how social media analysis can surface collective identities and community dynamics across discipline [32]. Building on this background, our study applies topic modeling and large language models to analyze neurodivergent discourse at scale, with the explicit goal of informing engineering education research. By examining how neurodiverse individuals discuss educational experiences, academic environments, strengths, and systemic barriers in online spaces, we seek to provide data-driven insights that can inform inclusive curriculum design, student support structures, and cultural change within engineering programs.

3 Dataset

We selected X (Twitter) and Reddit as the primary data sources to build a dataset that is both diverse and well-suited for analyzing neurodivergence-related discourse. These platforms are widely used for content creation, experience sharing, and community building, yet they differ in how users engage with them. X is largely public and encourages short-form posts, making it useful for capturing a broad range of perspectives, emerging ideas, and fast-moving conversations. Reddit supports longer, structured, community-driven discussions through subreddits, where users exchange context-rich narratives and detailed explanations that are especially valuable for qualitative thematic analysis. Using X's API, we accessed core entities such as users, posts, and search. Because of limited endpoint access, we primarily collected two types of data: timeline-based streams of posts and reshares, and non-timeline content. We manually identified influential neurodivergent community accounts, popular posts, and replies, and filtered records to include only content from August 2022 onward. For Reddit, we used a Python API wrapper to gather posts and comments from 36 curated neurodiversity-related subreddits (e.g., r/autism, r/adhd), where voting signals reflect popularity. In total, we collected 45,948 raw X records and 59,600 Reddit records.

To align with our goal of uncovering hidden themes in social media narratives, we focused on the main text fields: "tweet text" for X (Twitter) and "selftext" for Reddit. Data preprocessing included lowercasing, removing mentions, URLs, new lines, and accented characters, and normalizing contractions (e.g., "can't" $\rightarrow$ "cannot"). For better readability and consistency, we expanded common social media abbreviations, converted percentage expressions into words to preserve context, and removed single-character tokens. We removed stop words, commonly used words such as "the," "is," and "and" that typically carry limited semantic meaning using a combined list from the Natural Language Toolkit (NLTK) [33], a widely used open-source Python library for Natural Language Processing (NLP). This custom list was refined iteratively through exploratory analysis, repeatedly updating it until irrelevant terms no longer appeared in word clouds for either platform. Overall, preprocessing was iterative, with steps added, removed, or reordered through trial and error to maximize coherence. Finally, we removed duplicate texts and excluded entries with fewer than four words.

4 Experiments

To address our research questions, we conducted a series of experiments to perform a comprehensive thematic analysis of the emancipatory language used by individuals in the neurodivergent community, along with the thoughts, opinions, and views shared by people around the world about neurodiversity. Further, we discuss in detail the methods we used in our experiments, such as unsupervised and hybrid learning methods, to identify both the overarching themes explicitly discussed in the data and the latent topics that require additional modeling. As a result, we develop a framework of methodologies for modeling textual data from social networking websites for exhaustive analysis. We also use a proprietary Large Language Model (GPT-3.5 Turbo) to enhance the interpretability of topic modeling output. To improve transparency and reduce hallucinations, we employ Retrieval-Augmented Generation (RAG) [34], a framework that combines information retrieval with text generation. In RAG-based prompting, relevant documents or extracted topic outputs are first retrieved from the dataset and then provided to the LLM as contextual input. This process enables the model to generate explanations that are directly informed by the underlying data, thereby increasing explainability and reliability.

4.1 Cluster Analysis X

We converted the text into numeric vectors using OpenAI's context-based embeddings, where vector direction captures semantic relatedness. Using the text-embedding-3-large model, we generated 3,072-dimensional vectors for each tweet. We chose agglomerative clustering as the clustering method since it is more robust to high-dimensional embeddings, as it hierarchically merges the closest points without centroid assumptions, unlike K-Means clustering, which depends on randomly initialized centroids that can converge to the mean of the data, or Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [35], which is sensitive to parameter tuning at this scale.

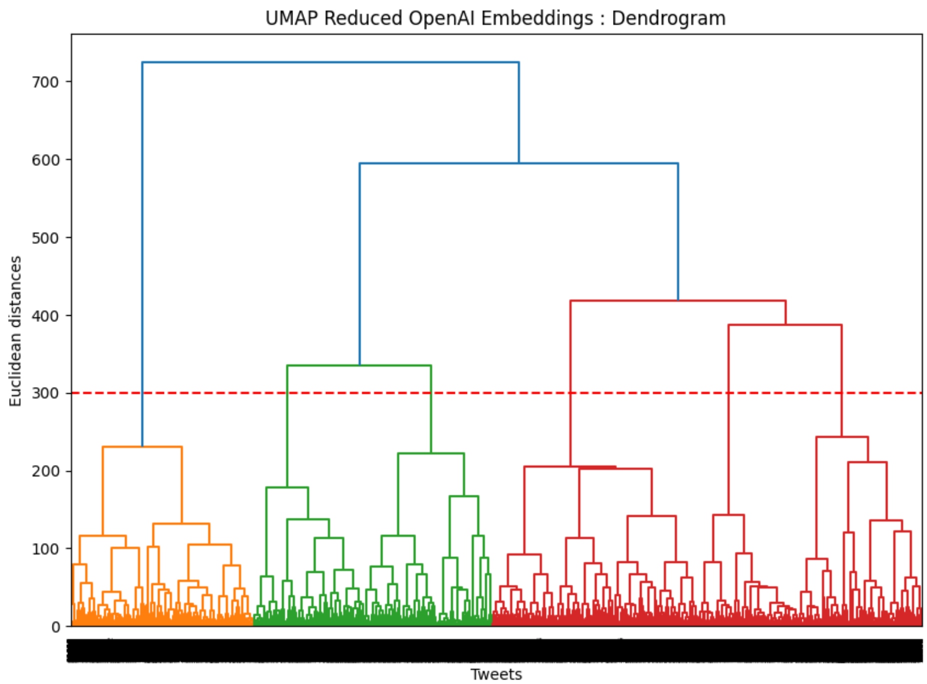

Figure 1: Dendrogram for agglomerative clustering of UMAP-reduced OpenAI embeddings, showing cluster merge points and the horizontal cut resulting in six partitions.

To mitigate the curse of dimensionality, we reduced the embeddings from 3,072 to 50 dimensions with UMAP [36], which preserves both local neighborhoods and global structure. Because unequal variances across the reduced dimensions can distort Euclidean distances, we standardized the UMAP features using a StandardScaler before fitting the hierarchical model. Figure 1 shows the hierarchical merges and a horizontal cut suggesting a reasonable six-cluster partition.

4.2 Topic Modeling: Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a popular unsupervised topic modeling method used in natural language processing (NLP) [37]. Training an LDA model requires two key inputs: a predetermined number of topics (n) to identify and a corpus (i.e., a collection of documents). Additionally, other parameters like α and β are set to auto by default unless specified otherwise. Higher values of α and β make the documents and topics less distinct and more similar to one another. Therefore, in our study, we chose lower values ($\alpha = 0.01$ and $\beta = 0.01$) to capture finer nuances in our corpus related to neurodivergent-related topics.

To determine the optimal number of topics for the LDA models, we trained models on X and Reddit data separately over different values of n and calculated their corresponding coherence scores. The coherence score measures the likelihood of words belonging to the same topic being semantically related, based on word pairs and their probabilities. The optimal number of topics can be identified by plotting coherence scores and selecting the point of inflection.

The optimal number of topics was selected based on coherence analysis (see Appendix B.1), with an inflection point observed at n=5. Figure 4 in Appendix B.1 shows the PyLDAvis [38] visualization of the 5 topics in the X data learned by the LDA model. Specifically, the figure displays the relevant terms for the 4th topic in the right panel. The relevance metric λ is set to 0.0 to show only the terms relevant to the current topic. In the left panel, the inter-topic distance map shows the five topics on a two-dimensional plane, revealing a slight overlap among topics 1, 2, and 3, as indicated by their relative positions.

Reddit texts tend to be longer and more detailed, and with nearly 1.5 times as many records as X, we anticipated a broader array of topics. Consequently, we explored a wider range of topic values (*n=5* to *n=49*) to determine the optimal number of topics. Based on the coherence curve in Appendix B.2 (Figure 5), *n=6* yielded the highest coherence score, but the PyLDAvis visualization showed substantial topic overlap, with 4 of 6 topics not well separated. Although *n=9* achieved the second highest coherence, its visualization indicated that topics formed tight clusters, with separation occurring primarily between clusters rather than among individual topics. We therefore examined higher-coherence settings (*n=17* and *n=24*) and found that *n=24* (visualized in Figure 6 in Appendix B.1) provided the best trade-off, producing topics that were comparatively well separated and balanced in size; hence, we use the *n=24* LDA model for subsequent analysis.

4.3 Topic Modeling: Top2Vec (Topic to Vector)

Top2Vec is an algorithm for topic modeling and semantic search. It automatically detects topics in text and jointly embeds topic, document, and word vectors [39]. Top2Vec provides nuanced insights through its useful functionalities, including topic strengths and sizes, topic words and scores, and document-topic distributions. The most essential inputs to the model include the text data and the embedding model to be used. The cleaned and lemmatized text was used for training the Top2Vec model. For selecting the embedding model, Top2Vec provides a list of options to choose from. For our study, we selected the Universal Sentence Encoder (USE) [40] because it is optimized for performance on short-form text such as sentences and phrases. USE has been trained on a diverse range of web data, including question-answering pages and forums, making it robust in understanding informal language commonly found on social media. For both X and Reddit, Top2Vec models were trained using the input parameters described above. 120 and 121 topics were discovered from X and Reddit data, respectively.

The variation in results between LDA and Top2Vec is attributable to their differing methodological assumptions rather than to inconsistencies in the data. LDA requires a predefined number of topics, producing broader themes, whereas Top2Vec automatically identifies semantically coherent clusters, resulting in more fine-grained topics. The subsequent LLM-assisted labeling and theme aggregation steps were therefore designed to structure these micro-topics into coherent higher-level themes. This hierarchical abstraction enables us to retain semantic detail while presenting analytically meaningful patterns. This multi-level analysis advances engineering education research by identifying systemic patterns and nuanced experiences that inform inclusive curriculum design, student support strategies, and equity-focused policy within engineering programs.

4.4 LLM assisted insight enhancement for Top2Vec topics

Because Top2Vec forms topics by clustering documents, we describe in Results and Discussion how these topic groupings were used to derive insights. However, the topics vary significantly in size (i.e., the number of documents per topic) and, without explicit labels and descriptions, are difficult to interpret and analyze in depth. To address this limitation, we adopted a systematic few-shot prompting strategy to generate a concise label and short description for each topic from both models. Specifically, we leveraged OpenAI's *gpt-3.5-turbo* model to improve topic interpretability by generating a relevant title and a summary for each topic. Listing 1 in Appendix A presents the Python code used to prompt the model. We carefully crafted our prompt by specifying a role for the AI in the *system_message* variable in line 4 as a "helpful assistant" and instructing it to generate a label and description for a topic based on the provided documents. Later, in the *user_message* variable, we provide clear instructions for the task by explicitly stating its structure ("analyze these documents and assign a meaningful label and description"), specifying an expected output format, and providing contextual information by including the top 10 documents for each topic. We use a delimiter-based context control approach to help the model distinguish between documents, wrapping them in '####' to ensure they are treated as distinct units.

4.5 Finding Global Themes with ChatGPT

After generating topic labels and descriptions using the LLM model, we used OpenAI's ChatGPT to aggregate the large set of Top2Vec topics into higher-level themes. This grouping step was intended to summarize the global structure of conversations, posts, and narratives in online content related to neurodiversity. We prompted the model to cluster topics based on semantic similarity among the topic labels and to output the set of topics assigned to each cluster. We interpret these clusters as themes identified in the data. Listing 2 in Appendix A presents the prompt used for the above.

5 Results and Discussion

5.1 Cluster Analysis

We clustered X (Twitter) posts using agglomerative clustering on OpenAI embeddings reduced with UMAP. The six clusters are visualized in a UMAP-reduced 2D plane in Figure 2. The clusters are moderately separated, and the overlaps suggest shared themes across them. Cluster 0 is widely dispersed, while Cluster 3 overlaps with nearby clusters, indicating less distinct boundaries. Additionally, differences in cluster sizes (notably Clusters 1 and 3) suggest unequal document counts across themes.

More exploration of these clusters highlighted prominent terms and topic tendencies. Overall, the clusters reflect a wide range of lived experiences and everyday conversations. *Cluster 0* contains frequent references to children or students "catching up," alongside "support," "help," and "pain," indicating academic challenges and associated stress. *Cluster 2* includes "bird," "game," "movie," and "laughing," showing casual routines and daily-life interests. *Cluster 3* emphasizes

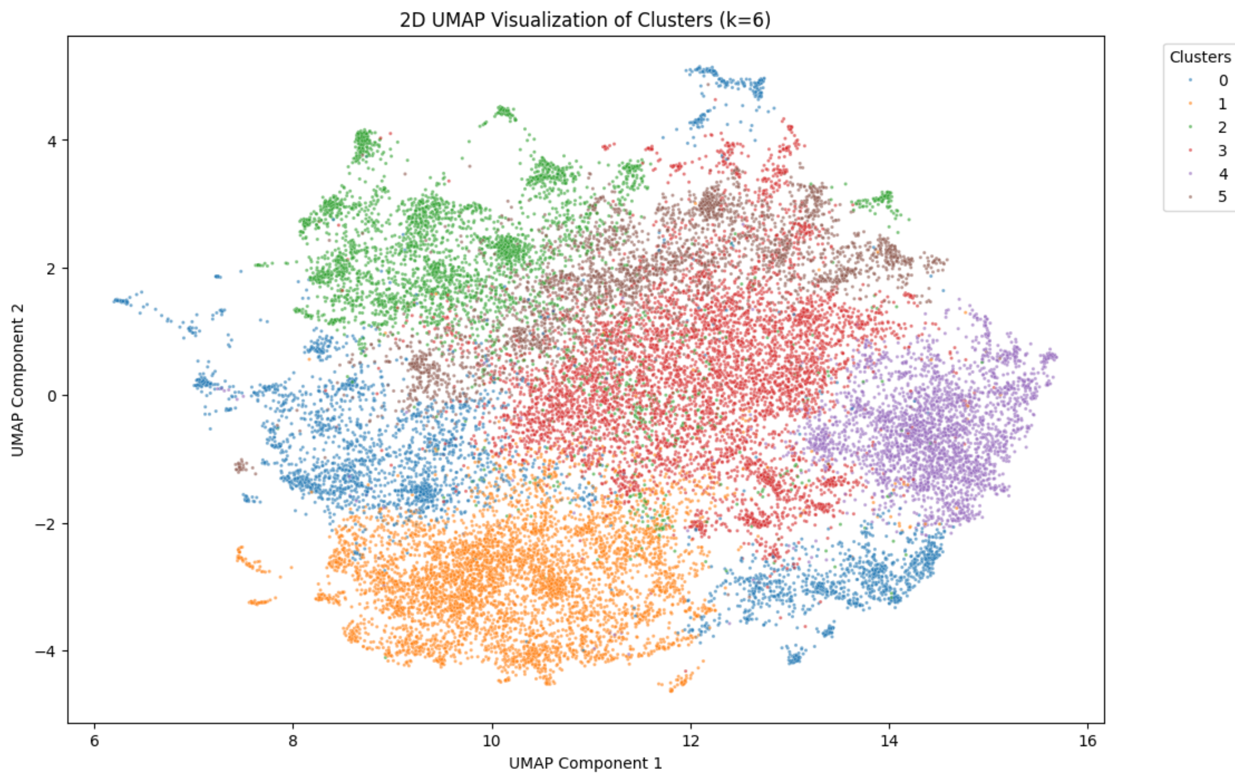

Figure 2: Tweet text cluster visualization (2D UMAP).

terms like "brain," "need," "struggle," "writing," and "working," reflecting overwhelm, overcommitment, and difficulties organizing tasks. *Cluster 5* features digital engagement through words such as "read," "love," "email," and "post," suggesting active online interaction and community participation.

5.2 Latent Dirichlet Allocation (LDA)

Figure 7 to 16 in Appendix B.2 present the top 20 words for topics (0-indexed) discovered by the X and Reddit LDA models. The left y-axis shows each word's raw count within documents where that topic is dominant, whereas the right y-axis shows the word's topic weight. In Figure 9, three patterns stood out. First, terms such as adhd, autism, want, need, and think appear across multiple topics, revealing shared vocabulary that cut across subthemes. Second, many words show higher frequency than weight. While such words are often treated as less discriminative, here they still address RQ1 by revealing the common linguistic anchors through which neurodivergent users describe experiences across contexts.

LDA model for Reddit produced 24 topics spanning themes from internet culture to ADHD medication. Due to space constraints, we manually selected five topics most closely aligned with our research questions (without excluding the rest). Topic 8 centers on anxiety and mental health, while Topic 13 highlights supportive community dynamics, with frequent terms such as autistic, help, thank, support, and community pointing to strong peer-to-peer support.

5.3 Topic to Vector (Top2Vec) and LLM assisted topic enhancement

Table 1 presents the LLM-generated titles and descriptions for randomly selected Top2Vec topics for X and Reddit. However, this does not imply that other topics discovered are not relevant to our study; we simply did this due to space limitations. We highlight the following to acknowledge RQ2.

Table 1: LLM-generated titles and descriptions for randomly selected Top2Vec topics

Source	Title	Description
X	Challenges of Task Initiation for Individuals with ADHD	The documents highlight common struggles faced by individuals with ADHD regarding initiating and completing tasks. Challenges include procrastination, feeling overwhelmed, lack of motivation, need for novelty, and breaking tasks into smaller steps.
X	Rejection Sensitivity and Emotional Reactivity in ADHD	Describes intense emotional sensitivity, anxiety, and severe emotional reactions associated with criticism or rejection, commonly linked to ADHD.
Reddit	Academic Pressure and Mental Health Struggles	Discusses issues related to academic stress, mental health challenges such as anxiety and depression, procrastination, medication effects, and feelings of overwhelm.
Reddit	Challenges with Teachers in School	Discusses parents' challenges with teachers, including behavior management issues, denial of accommodations, lack of support for special needs children, and related frustrations.

- For X data, Topic 4 highlights executive dysfunction, where individuals with ADHD describe difficulty initiating tasks and low motivation. Topics 61 and 57 emphasize repeated burnout, emotional overload, and the resulting sense of ineffectiveness, including intense sensitivity to criticism or rejection (Rejection Sensitive Dysphoria). Topics 21 and 41 capture professional challenges such as interviews, ghosting, and disclosure-related anxiety, alongside stress from maintaining productivity in environments not designed for neurodivergent minds.
- For Reddit data, academic stress is a major theme, with users describing struggles around studying, exams, assignments, and burnout when demands feel unmanageable (Topic 20). Parents also report frustration and conflict when teachers misunderstand needs or deny accommodations (Topic 58). Other discussions focus on fears of one-sided communication and difficulty sustaining conversations (Topic 12), as well as on sensory overstimulation, especially noise sensitivity (Topic 35). Despite these challenges, strengths are evident: many individuals demonstrate proactiveness and adaptability by seeking practical strategies to manage hyperfixation and overwhelm, and by building structured routines and supportive rituals (Topic 18).

5.4 Extracting Global Themes from Top2Vec topics using ChatGPT

For X, the response to the prompt yielded 12 thematic groupings, while for Reddit, 18 groups were created. While there were overlaps in some groups' definitions, the non-overlapping parts, especially the specific topics that construct those groupings, contributed significantly to understanding the narratives. To enable the use of these groupings for additional meaningful insights, further prompting was provided, explicitly asking for the information to be consolidated into a structured, downloadable format. Table 2 illustrates some of the thematic groups and their characteristics generated through contextually engineered ChatGPT prompts for the Top2Vec topics from the X and Reddit datasets, respectively.

Table 2: Randomly selected Top2Vec topics grouped into themes using ChatGPT prompts. 'Thematic Grouping' and 'Group Characteristics' are the generated responses.

Source	Thematic Grouping	Group Characteristics
X	Emotional Well-Being, Mental Health, and Trauma	Anger management, trauma responses, therapy experiences, mental health stigma.
X	Hobbies, Entertainment, and Leisure Interests	Personal passions, media interests, collecting, gaming, birdwatching, body modifications.
X	Everyday Routines, Organization, and Personal Productivity	Practical strategies for cleaning, decluttering, digital organization, cooking, scheduling, and productivity techniques.
Reddit	Family, Parenting, and Children's Education	Family dynamics, parenting, special-needs parenting, conflicts with teachers, parental concerns.
Reddit	Broader Neurodiversity, Personal Growth, and Community	Community building, stigma reduction, race/neurodivergence intersectionality, coping strategies.
Reddit	Anxiety (General and Specific Fears)	General anxiety, health anxiety, panic attacks, specific fears such as driving, nuclear war, insects.

5.5 Discussion

Our findings in align with prior research that moves beyond deficit-based narratives of neurodiversity and instead recognizes both challenges and strengths within educational contexts [5, 14, 15]. Consistent with studies documenting executive dysfunction, emotional regulation difficulties, and academic stress among neurodivergent learners [41–43], our topic modeling results reveal sustained discussions of burnout, task initiation challenges, disclosure anxiety, and accommodation barriers. These patterns reinforce existing evidence of gaps between inclusive policy and the current realities of neurodivergent students in STEM education [44–47]. At the same time, themes of peer support, strategy-sharing, adaptability, and systems thinking reflect strengths-based perspectives that highlight cognitive assets relevant to engineering contexts.

The identified themes underscore important implications for engineering education. Persistent discussions of burnout and accommodation barriers suggest that performance-intensive engineering environments may unintentionally marginalize neurodivergent students. At the same time, recurring narratives of adaptability, peer collaboration, and systems thinking highlight strengths that align with core engineering competencies. Together, these insights support a shift from accommodation-centered approaches toward strengths-informed instructional design and inclusive learning environments that foster belonging, retention, and innovation [15, 48].

This study has several limitations. The data were drawn from self-selected users on public social media platforms and may not represent the full diversity of neurodivergent individuals, particularly those outside online communities. Topic modeling and embedding-based clustering simplify complex narratives, and LLM-generated labels provide interpretive summaries rather than definitive classifications. Additionally, user identities and educational affiliations cannot be verified, limiting direct claims about engineering students. Future research integrating qualitative or institutional data could strengthen contextual validity within engineering education.

6 Conclusion

This work demonstrates the value of computational methods in surfacing the authentic voices of neurodivergent individuals across two major social media platforms: X and Reddit. By leveraging a rich blend of natural language processing techniques, including classical clustering, probabilistic topic modeling, and embedding-driven approaches, this study captures the multi-layered, community-driven discourse around neurodiversity in online spaces.

The work goes beyond content detection to explore how neurodivergent users employ these platforms not only for expression and advocacy but also for knowledge-sharing and emotional co-regulation. While X supports concise, often affective exchanges centered around personal experience, Reddit fosters longer, reflective, and advice-oriented narratives. Each platform reveals distinct dimensions of the neurodivergent experience, highlighting both common and platform-specific themes in diagnosis, education, workplace dynamics, sensory management, and emotional well-being.

For engineering education, these insights underscore the importance of recognizing both structural challenges such as burnout, disclosure anxiety, and accommodation gaps, also strengths such as persistence, systems thinking, and collaborative problem-solving. By leveraging Top2Vec with LLM-assisted summarization, this study also contributes a scalable methodological framework for analyzing large-scale discourse to inform inclusive curriculum design, student support systems, and policy development.

Ultimately, this research not only answers the posed research questions about usage patterns, recurring themes, and experiences in neurodivergent discourse but also contributes a transferable methodology for social-data-driven inquiry in engineering education and beyond. It affirms that social media, when ethically harnessed, can serve as a meaningful lens to the challenges, strengths, and evolving identity of marginalized groups, and that such insights can inform inclusive design, policy, and pedagogical strategies attuned to real-world needs.

Acknowledgments

This material is based upon work supported by the National Science Foundation under EEC Award #2306830 and #2306831. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] J. Logan, "Queer and neurodivergent identity production within the social media panopticon," *The Macksey Journal*, vol. 1, no. 1, 2020.

[2] A. Yeung, E. Ng, and E. Abi-Jaoude, "Tiktok and attention-deficit/hyperactivity disorder: a cross-sectional study of social media content quality," *The Canadian Journal of Psychiatry*, vol. 67, no. 12, pp. 899–906, 2022.

[3] T. Eagle, "Exploring collective medical knowledge and tensions in online adhd communities," in *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, 2022, pp. 245–250.

[4] J. Egner, "# actuallyautistic: Using twitter to construct individual and collective identity narratives," *Studies in Social Justice*, vol. 16, no. 2, pp. 349–369, 2022.

[5] L. Clouder, M. Karakus, A. Cinotti, M. V. Ferreyra, G. A. Fierros, and P. Rojo, "Neurodiversity in higher education: a narrative synthesis," *Higher Education*, vol. 80, no. 4, pp. 757–778, 2020.

[6] A. M. Alcorn, S. McGeown, W. Mandy, D. Aitken, F. Murray, L. J. Peacock, and S. Fletcher-Watson, "A mixed-methods evaluation of the learning about neurodiversity at school (leans) programme: teaching neurodiversity concepts in mainstream primary schools," *INSAR 2023*, 2023.

[7] L. Brown, C. Spence, A. Cuellar, K. Thakkar, H. Karimi, and M. Tsugawa, "Wip: Developing an interview protocol to unveil the stories of neurodivergent engineering students." [Online]. Available: https://par.nsf.gov/biblio/10626353

[8] M. Legault, J.-N. Bourdon, and P. Poirier, "From neurodiversity to neurodivergence: the role of epistemic and cognitive marginalization," *Synthese*, vol. 199, no. 5, pp. 12 843–12 868, 2021.

[9] C. von Karolyi, "Visual-spatial strength in dyslexia: Rapid discrimination of impossible figures," *Journal of learning disabilities*, vol. 34, no. 4, pp. 380–391, 2001.

[10] H. A. White and P. Shah, "Uninhibited imaginations: creativity in adults with attention-deficit/hyperactivity disorder," *Personality and individual differences*, vol. 40, no. 6, pp. 1121–1131, 2006.

[11] L. Mottron, "The power of autism," *Nature*, vol. 479, no. 7371, pp. 33–35, 2011.

[12] A.-A. S. Meilleur, P. Jelenic, and L. Mottron, "Prevalence of clinically and empirically defined talents and strengths in autism," *Journal of autism and developmental disorders*, vol. 45, no. 5, pp. 1354–1367, 2015.

[13] C. L. Taylor, A. E. Zaghi, J. C. Kaufman, S. M. Reis, and J. S. Renzulli, "Characteristics of adhd related to executive function: Differential predictions for creativity-related traits," *The Journal of Creative Behavior*, vol. 54, no. 2, pp. 350–362, 2020.

[14] M. Chrysochoou, A. E. Zaghi, and C. M. Syharat, "Reframing neurodiversity in engineering education," in *Frontiers in Education*, vol. 7. Frontiers Media SA, 2022, p. 995865.

[15] M. Chrysochoou, A. E. Zaghi, C. M. Syharat, S. Motaref, S. Jang, A. Bagtzoglou, and C. A. Wakeman, "Redesigning engineering education for neurodiversity: New standards for inclusive courses," in *2021 ASEE Virtual Annual Conference Content Access*, 2021.

[16] A. Alcorn, S. Mcgeown, W. Mandy, D. Aitken, and S. Fletcher-Watson, "Learning about neurodiversity at school: Evaluating a new classroom programme for mainstream primary schools," *Journal of Educational Research*, 2023.

[17] L. G. Hamilton and S. Petty, "Compassionate pedagogy for neurodiversity in higher education: A conceptual analysis," *Frontiers in psychology*, vol. 14, p. 1093290, 2023.

[18] E. Spaeth and A. Pearson, "A reflective analysis on neurodiversity and student wellbeing: Conceptualising practical strategies for inclusive practice," *Journal of Perspectives in Applied Academic Practice*, vol. 11, no. 2, pp. 109–120, 2023.

[19] D. Taneja, P. Viswanathan, and S. V. Rajan, "Navigating neurodiversity in higher education in india," *eLife*, vol. 13, p. e104071, 2024.

[20] C. A. Hinton and S. F. Hinton, "Neurodiversity in education and the workplace: Understanding and celebrating differences," in *Diversity and Inclusion in Global Business and Education*. IGI Global Scientific Publishing, 2025, pp. 471–492.

[21] J.-A. Johnson and S. Ahluwalia, "Neurodiversity in the healthcare profession," *Postgraduate medical journal*, vol. 101, no. 1192, pp. 167–171, 2025.

[22] N. G. Azar and A. A. Taha, "Embracing neurodiversity in the nursing workforce," *AJN The American Journal of Nursing*, vol. 125, no. 1, pp. 18–19, 2025.

[23] D. Gottardello, T. Calvard, and J.-W. Song, "When neurodiversity and ethnicity combine: Intersectional stereotyping and workplace experiences of neurodivergent ethnic minority employees," *Human Resource Management*, 2025.

[24] M. H. Khan, M. K. Grabarski, M. Ali, and S. Buckmaster, "Insights into creating and managing an inclusive neurodiverse workplace for positive outcomes: A multistaged theoretical framework," *Group & Organization Management*, vol. 48, no. 5, pp. 1339–1386, 2023.

[25] V. McDermott, ""tell me something you didn't know was neurodivergence-related until recently. i'll start": Tiktok as a public sphere for destigmatizing neurodivergence," in *Rethinking perception and centering the voices of unique individuals: Reframing autism inclusion in praxis*. IGI Global, 2022, pp. 127–147.

[26] A. P. A. Mair, M. Gonzalez-Figueroa, D. McConachie, K. Goodall, and K. Gillespie-Smith, "36 autism & grief: a social media study of experiences of grief and loss in the context of neurodivergence," 2024.

[27] S. Solanki, K. Kheiri, M. A. Tsugawa, and H. Karimi, "Leveraging social media analytics in engineering education research," in *2023 ASEE Annual Conference & Exposition*, no. 10.18260/1-2–43472. Baltimore , Maryland: ASEE Conferences, June 2023, https://peer.asee.org/43472.

[28] D. Z. Zhang *et al.*, "Brainbow: A real-time inclusivity index for neurodiversity using sentiment analysis of news and social media," in *CS & IT Conference Proceedings*, vol. 14, no. 19. CS & IT Conference Proceedings, 2024.

[29] M. Krazinski, "Celebrating neurodivergence amid social injustice," *Hypatia*, vol. 38, no. 4, pp. 726–745, 2023.

[30] J. P. Barnett, "Neuroqueer frontiers: Neurodiversity, gender, and the (a) social self," *Sociology Compass*, vol. 18, no. 6, p. e13234, 2024.

[31] P. Baillargeon, J. Yoon, and A. Zhang, "Who puts the" social" in" social computing"?: Using a neurodiversity framing to review social computing research," *arXiv preprint arXiv:2410.15527*, 2024.

[32] A. Standlee, "Social media analysis," *The Blackwell Encyclopedia of Sociology*, pp. 1–3, 2007.

[33] E. Loper and S. Bird, "Nltk: The natural language toolkit," 2002. [Online]. Available: https://arxiv.org/abs/cs/0205028

[34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[35] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[36] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[38] C. Sievert and K. Shirley, "Ldavis: A method for visualizing and interpreting topics," in *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014, pp. 63–70.

[39] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv preprint arXiv:2008.09470*, 2020.

[40] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. L. U. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, Y. hsuan Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," in *In submission to: EMNLP demonstration*, Brussels, Belgium, 2018, in submission. [Online]. Available: https://arxiv.org/abs/1803.11175

[41] N. A. Tcherdakoff, P. Marshall, A. Dowthwaite, J. Bird, and A. L. Cox, "Burnout by design: How digital systems overburden neurodivergent students in higher education," in *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, 2025, pp. 1–18.

[42] P. Ezeamii and K. Shinohara, "Navigating stem doctoral programs with adhd: Barriers, workflow challenges, and adaptive strategies," in *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*, 2025, pp. 1–13.

[43] A. Vincent, K. Shnaikat, E. Marasco, and A. Barcomb, "Categorizing interactive digital learning tools to support neurodivergent students in engineering education," *Proceedings of the Canadian Engineering Education Association (CEEA)*, 2025.

[44] C. M. Syharat, A. Hain, A. E. Zaghi, R. Gabriel, and C. G. Berdanier, "Experiences of neurodivergent students in graduate stem programs," *Frontiers in Psychology*, vol. 14, p. 1149068, 2023.

[45] S. Salvatore, C. White, and S. Podowitz-Thomas, ""not a cookie cutter situation": how neurodivergent students experience group work in their stem courses," *International Journal of STEM Education*, vol. 11, no. 1, p. 47, 2024.

[46] J. L. Hsu and G. R. Goldsmith, "Instructor strategies to alleviate stress and anxiety among college and university stem students," *CBE—Life Sciences Education*, vol. 20, no. 1, p. es1, 2021.

[47] K. D. Wang, J. McCool, and C. Wieman, "Exploring the learning experiences of neurodivergent college students in stem courses," *Journal of Research in Special Educational Needs*, vol. 24, no. 3, pp. 505–518, 2024.

[48] M. L. Wehmeyer and J. A. Kurth, "Inclusive education in a strengths-based era: changing attitudes and practices," *Cz lowiek-Niepe lnosprawność-Spo leczeństwo*, no. 54 (4, pp. 5–28, 2021.

APPENDIX A

Python Function and Contextual Prompt for LLM

This appendix provides provides Python code and contextual prompt that were omitted from the main paper due to space limitations.

A.1 Python function

```python
def analyze_topics(topic_number, representative_docs, model):

    delimiter = '####'
    system_message = "You're a helpful assistant. Your task is to provide a label and a description for a predefined topic based on the documents provided."
    user_message = f'''
    Below is a set of documents, corresponding to a predefined topic number. The documents are delimited with {delimiter}.
    Please analyze these documents and assign a meaningful label and a brief description to the topic number.

    Expected output format:
    {{
        "topic_label": "<label>",
        "topic_description": "A brief description of the topic based on the posts."
    }}

    The documents are organized as follows:
    {delimiter}
    {delimiter.join(representative_docs)}
    {delimiter}
    '''
    messages = [
        {'role': 'system', 'content': system_message},
        {'role': 'user', 'content': user_message},
    ]
    api_response = get_model_response(messages, model=model, temperature=0, max_tokens=1000)
    api_response_dict = json.loads(api_response.choices[0].message.content)
    return api_response_dict
```

Listing 1: Python function to prompt and generate labels and descriptions using GPT

A.2 Finding Global Themes with ChatGPT

```
The text below the "TEXT DATA" heading contains a list of 120 topic labels and
topic descriptions separated by the ":" delimiter.
You are an expert in neurodivergence and social sciences. I want you to group
these topic labels based on how similar they are. You can use the topic
description to analyze the semantic similarity of the topic labels.

For each group of topics you create, give its description, reasoning behind why
these topic labels form a group and a list of all the topics that form each of
these groups.

"TEXT DATA"
Topic 1: Topic 1 description
Topic 2: Topic 2 description
.
.
.
Topic 120: Topic 120 description
```

Listing 2: Contextual prompt for ChatGPT

APPENDIX B

Plots and Figures

B.1 LDA Model Selection

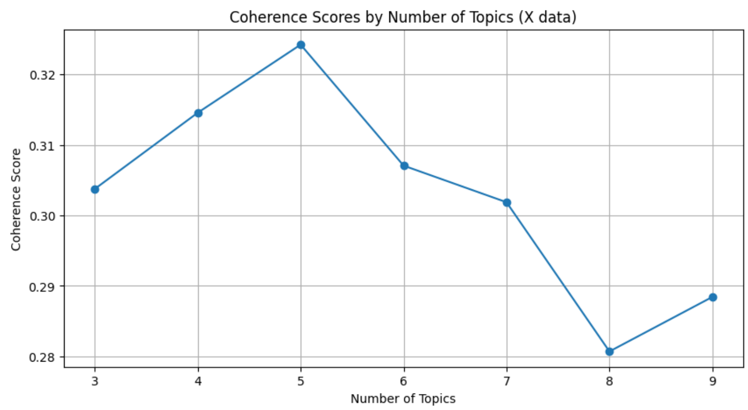

Figure 3: Coherence scores for LDA models trained on X data across varying numbers of topics. The point of inflection occurs at *n*=5.

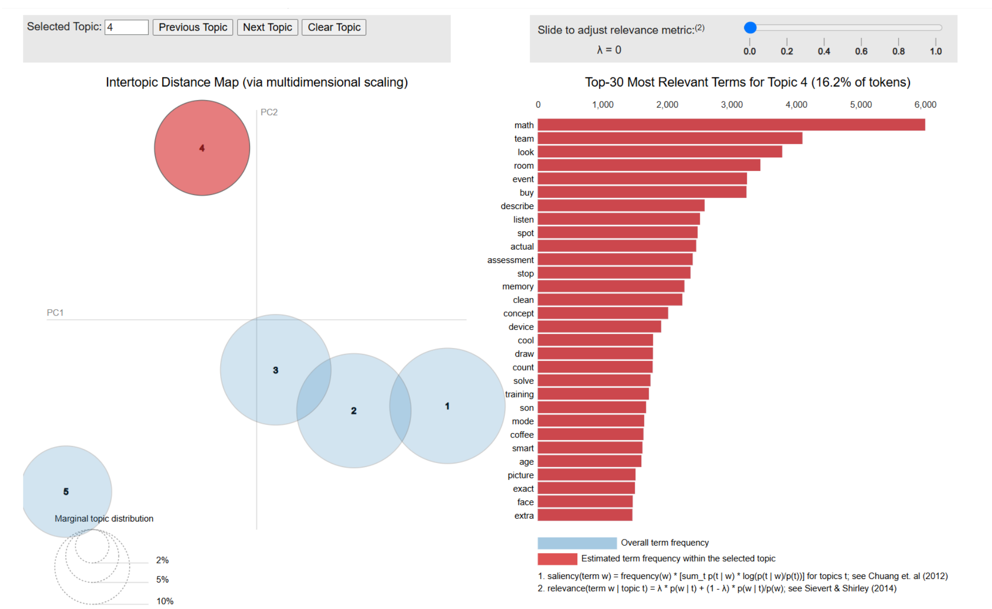

Figure 4: PyLDAvis visualization for the LDA model trained on X data (*n*=5). The left panel shows inter-topic distances, and the right panel shows the most relevant terms for each topic.

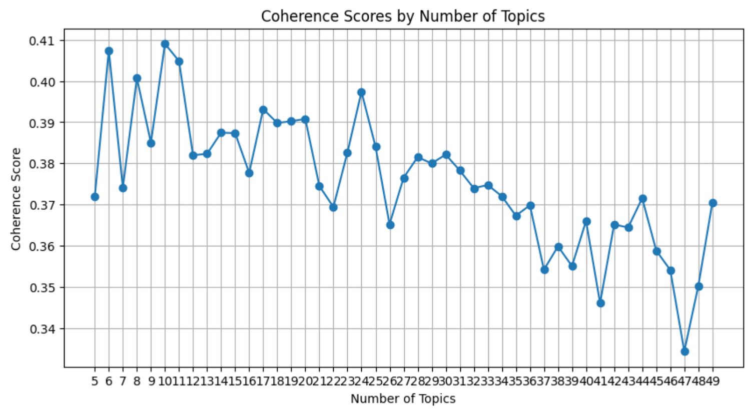

Figure 5: Coherence scores for LDA models trained on Reddit data across varying numbers of topics, showing multiple points of inflection.

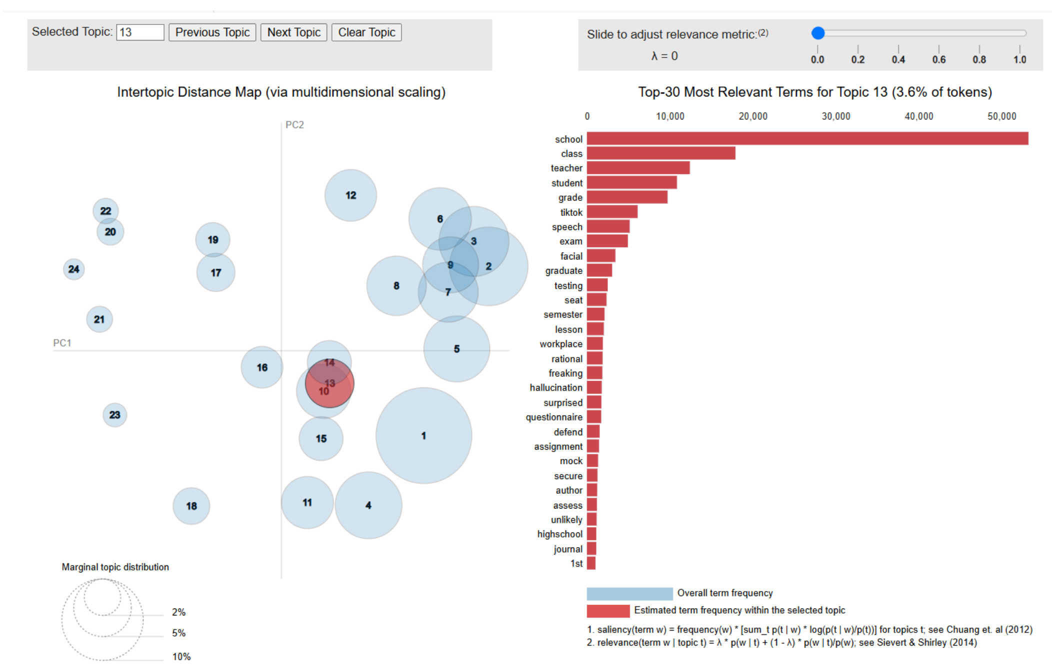

Figure 6: PyLDAvis visualization for the LDA model trained on Reddit data (n=24). The left panel shows intertopic distances, and the right panel displays the most relevant terms for each topic.

B.2 Top 20 words per LDA topic (X data)

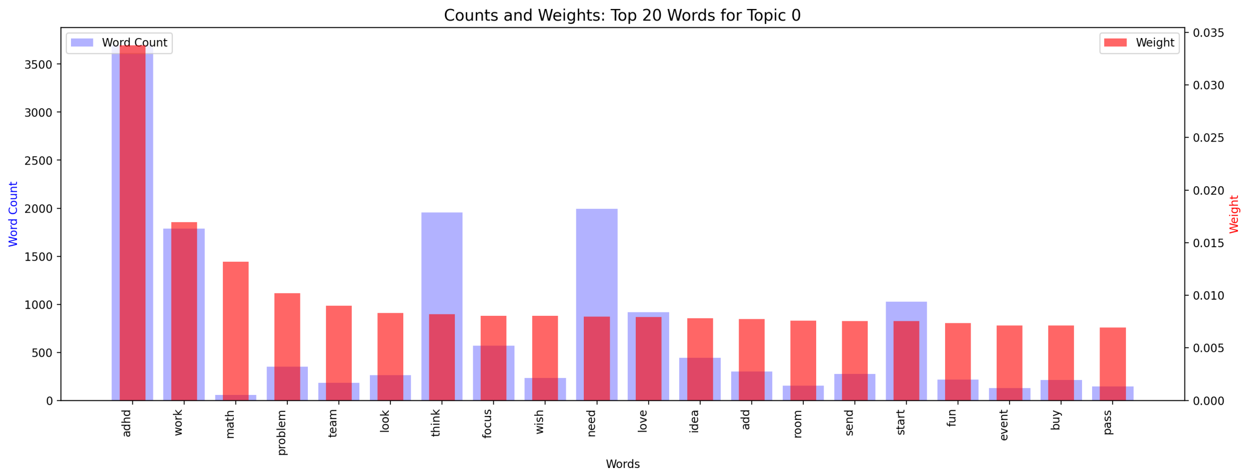

Figure 7: Topic 0

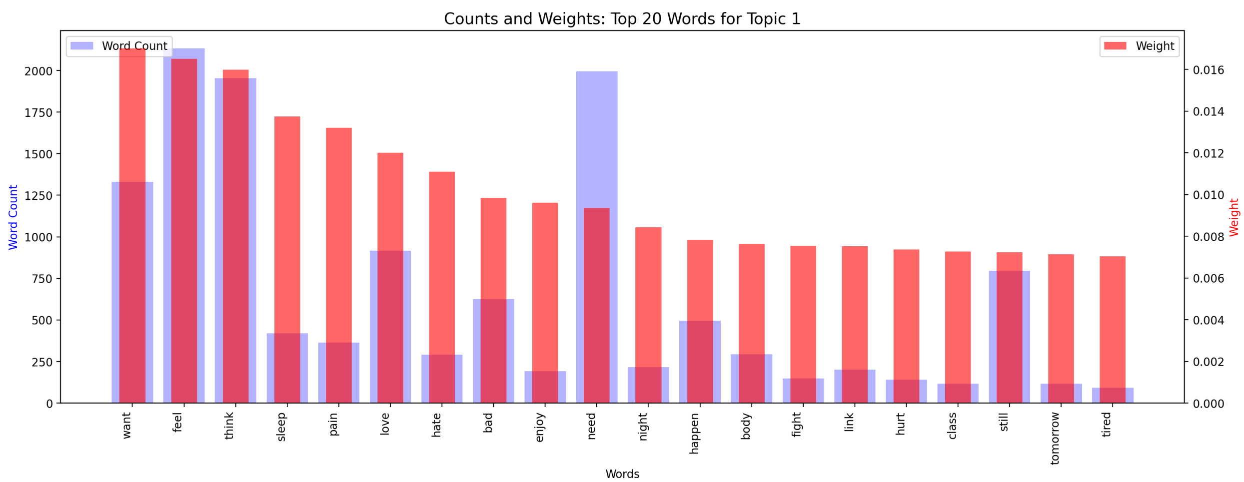

Figure 8: Topic 1

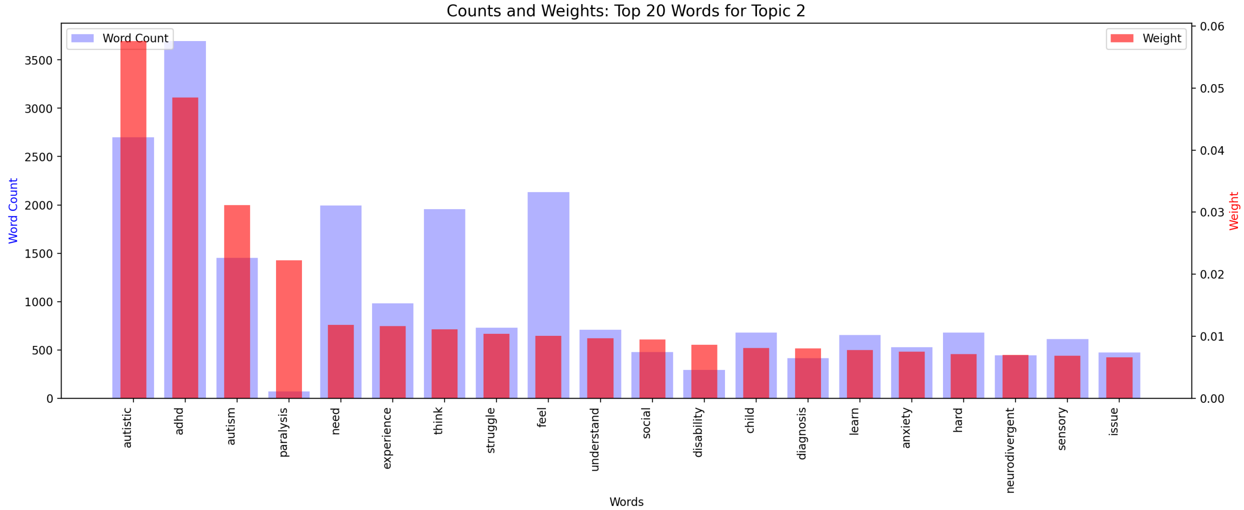

Figure 9: Topic 2

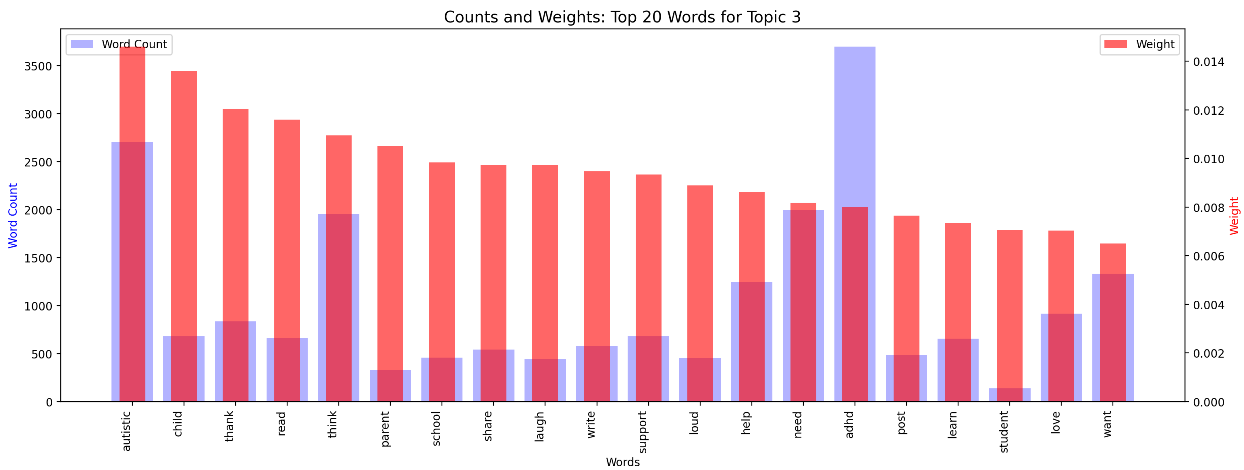

Figure 10: Topic 3

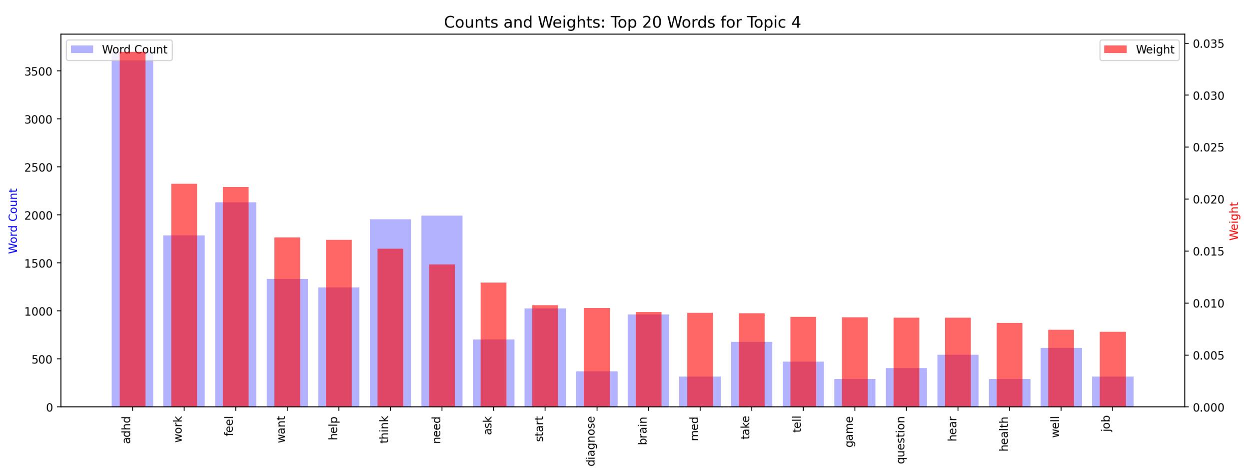

Figure 11: Topic 4

B.3 Top 20 words per LDA topic (Reddit data, five randomly selected topics)

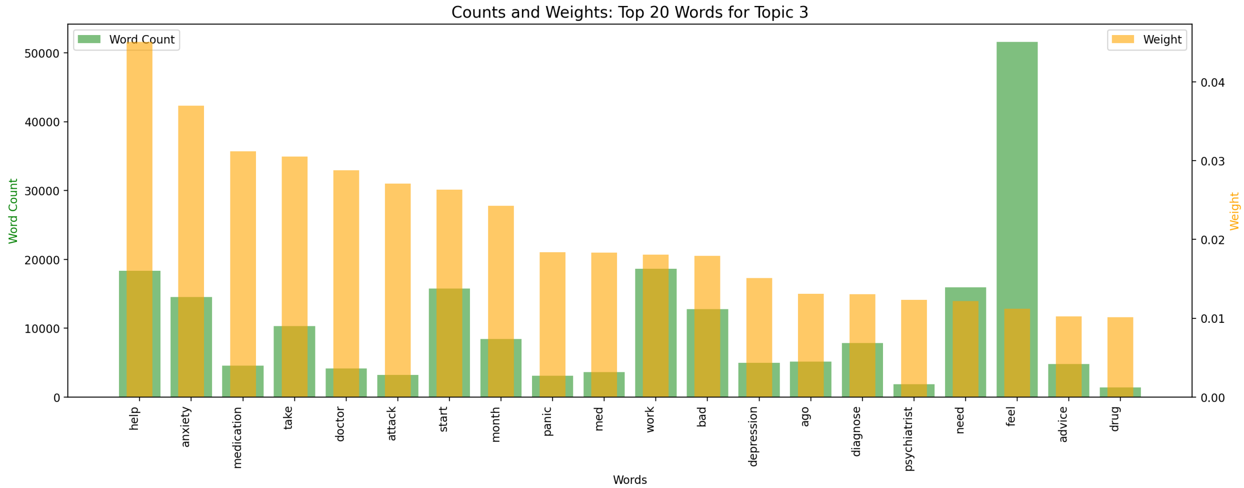

Figure 12: Topic 3

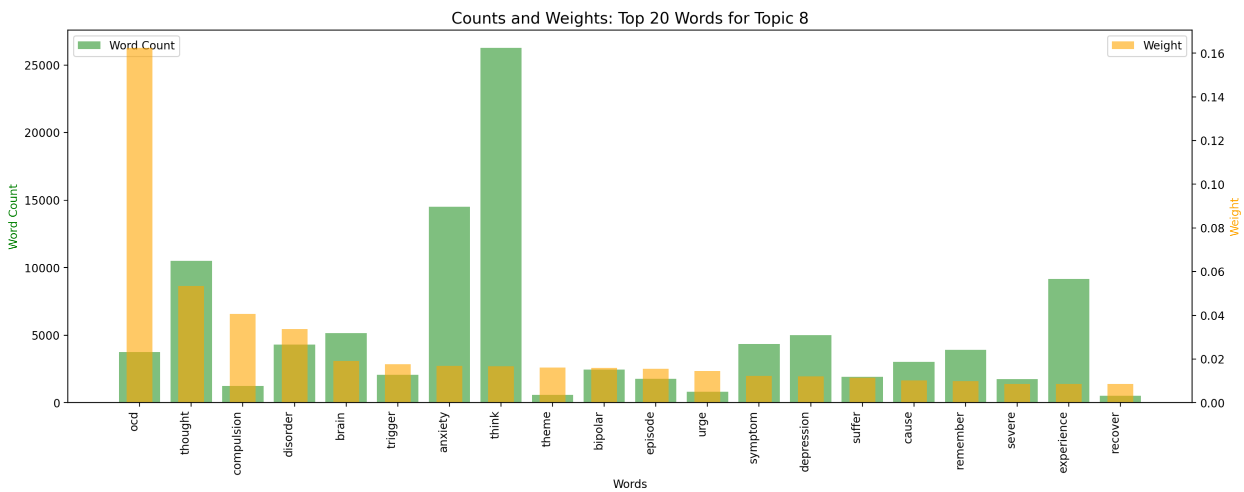

Figure 13: Topic 8

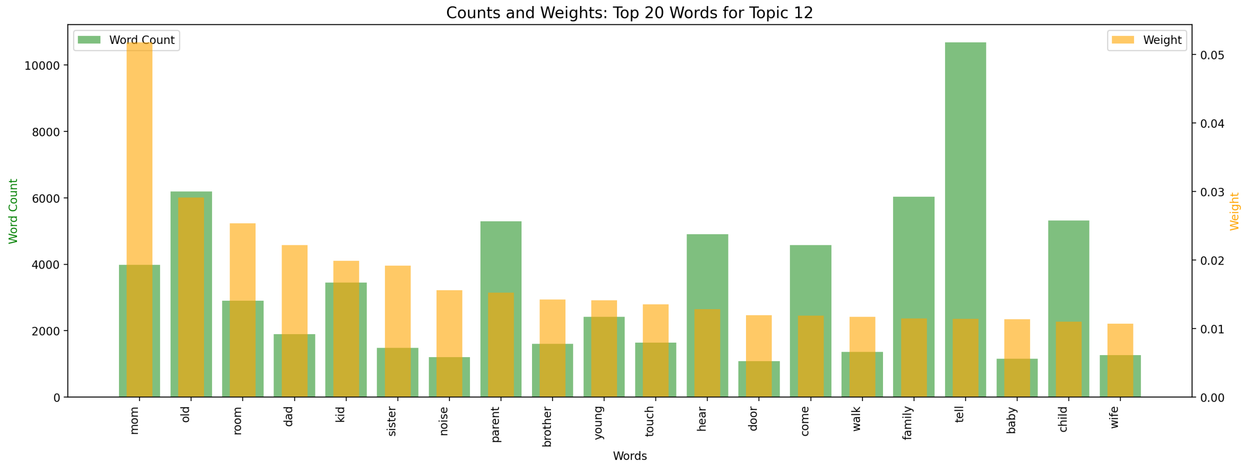

Figure 14: Topic 12

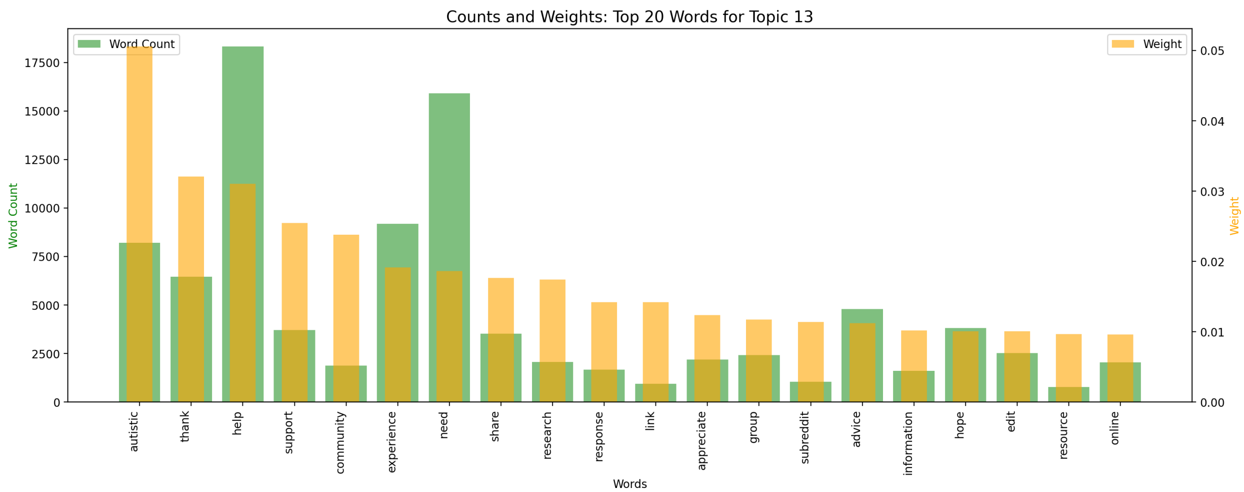

Figure 15: Topic 13

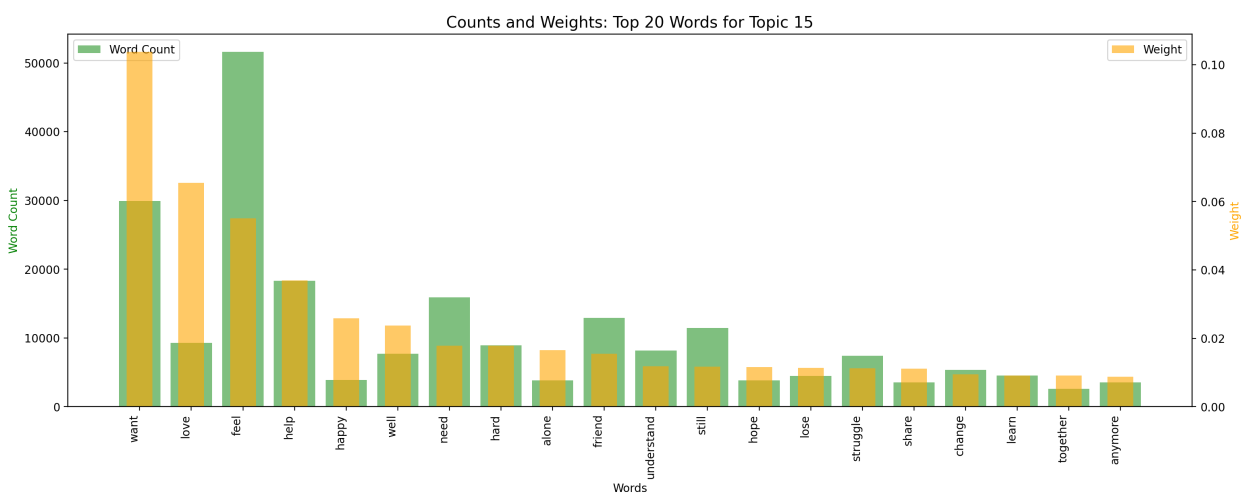

Figure 16: Topic 15