GCD-TM: Graph-Driven Community Detection for Topic Modelling in Psychiatry Texts

Anusuya Krishnan
Department of Mathematics
College of Science
United Arab Emirates University, UAE
anusuyababy@uaeu.ac.ae

Isaias Mehari Ghebrehiwet
Department of Computer Science
College of IT
United Arab Emirates University, UAE
ighebrehiwet@uaeu.ac.ae

Abstract

Psychiatry texts provide critical insights into patient mental states and therapeutic interactions. These texts are essential for understanding psychiatric conditions, treatment dynamics, and patient responses. However, the complex and diverse nature of psychiatric communications poses significant challenges for traditional topic modeling methods. The intricate language, subtle psychological nuances, and varying lengths of text segments make it difficult to extract coherent and meaningful topics. Conventional approaches often struggle to capture the depth and overlap of themes present in these texts. In this study, we present a novel approach to topic modeling that addresses these limitations by reformulating the problem as a community detection task within a graph constructed from the text corpus. Our methodology includes lemmatization for data standardization, TF-IDF vectorization to create a term-document matrix, and cosine similarity computation to produce a similarity matrix. This matrix is then binarized to form a graph, on which community detection is performed using the Louvain method. The detected communities are subsequently analyzed with Latent Dirichlet Allocation (LDA) to extract topics. Our approach outperforms traditional topic modeling methods, offering more accurate and interpretable topic extraction with improved coherence and lower perplexity.

1 Introduction

Psychiatric disorders like depression, bipolar disorder, anxiety, schizophrenia, and substance abuse are major contributors to disability, significantly impacting individuals' quality of life due to their prevalence and duration (James et al., 2018; Figueroa et al., 2020; Cuijpers et al., 2012). Traditionally, psychiatrists diagnose these conditions through detailed consultations, with linguistic research highlighting key speech patterns in various disorders (Cohen et al., 2008; Patra et al., 2020). Mental health interventions, including psychosocial, behavioral, pharmacological, and telemedicine methods, are vital for improving well-being but face systemic obstacles (DeRubeis et al., 2008; Miranda et al., 2008). The absence of objective diagnostic tools, variability in treatment quality, clinician shortages, especially in rural areas, and the high costs of training reduce the effectiveness of psychological therapies, underscoring the need for improved diagnostic and treatment tools (Firth et al., 2017; Wang et al., 2007).

Recently, there has been a surge in research aimed at diagnosing psychiatric disorders through Natural Language Processing (NLP). Studies have targeted a range of disorders, including schizophrenia, depression, bipolar disorder, obsessive-compulsive disorder, autism spectrum disorders, and dementia (Malgaroli et al., 2023; Rumshisky et al., 2016). Topic modeling has emerged as a key NLP technique for extracting meaningful patterns and themes from psychiatric texts (Nikolenko et al., 2017). Topic modeling algorithms are designed to uncover latent topics within large corpora of text by analyzing word co-occurrence patterns (Tong and Zhang, 2016). This approach can reveal underlying themes in patient narratives, therapy session transcripts, and clinical notes, providing valuable insights into the content and dynamics of psychiatric disorders. By applying topic modeling to mental health data, researchers can gain a deeper understanding of prevalent issues, treatment efficacy, and patient experiences, ultimately contributing to the improvement of diagnostic and therapeutic processes (Nikolenko et al., 2017).

Latent Dirichlet Allocation (LDA) is a widely used topic modeling technique based on the premise that documents are mixtures of topics and topics are distributions over words (Blei et al., 2003). Latent Semantic Analysis (LSA) is another topic modeling technique that employs singular value decomposition to reduce dimensionality

Proceedings of the 1st Workshop on NLP for Science (NLP4Science), pages 47–57
November 16, 2024 ©2024 Association for Computational Linguistics

and uncover underlying semantic structures (Dumais, 2004). Non-Negative Matrix Factorization (NMF) is also one of the topic modeling techniques, which factorizes term-document matrices into non-negative components, focusing on additive topic combinations (Lee and Seung, 2000). BERTopic enhances topic modeling by utilizing contextual embeddings from models like BERT and MPNet, combined with advanced term weighting, which improves topic identification, particularly in specialized domains such as legal documents (Grootendorst, 2022). Spectral clustering further refines topic coherence by analyzing document similarity structures, though it requires careful parameter tuning and can be sensitive to noise (Ng et al., 2001). K-means clustering is used to extract precise topics from unstructured data, such as biomedical texts, with improvements in accuracy and efficiency (Sinaga and Yang, 2020).

Existing topic modeling approaches, such as LDA and similar models, often face challenges in accurately capturing specific concepts of substantive interest within a corpus (Chemudugunta et al., 2008; Chang et al., 2009). While these models can explore the themes present in the data, they frequently produce multiple topics with overlapping content or merge distinct themes into a single topic. This can lead to difficulties in interpreting the topics and measuring key concepts accurately (Lei, 2012). Additionally, these models do not inherently incorporate information about the topics of interest, making it challenging for researchers to determine whether the generated topics align with the intended substantive concepts until after the model has been fitted (Brookes and McEnery, 2019). This limitation highlights the need for human validation to ensure the relevance and accuracy of the topics.

To address these issues, our graph-based method offers a more refined approach by constructing a network of terms, which better captures the relationships and nuances in the data, ultimately leading to clearer interpretations and more accurate measurements of the underlying thematic structures. In this study, we propose a novel approach to improve topic modeling in psychiatric text analysis by leveraging graph-driven community detection topic modelling (GCD-TM). This approach involves three key concepts:

- We build a graph based on cosine similarity computed from TF-IDF vectorized text data, where each document is represented as a node and edges denote the semantic similarity between nodes. This enables the identification of closely related groups of documents that share similar themes (Singh and Shashi, 2019).

- We apply the Louvain method for community detection on the constructed graph (Meo et al., 2011). This technique identifies densely connected subgroups within the graph, which correspond to communities of documents with closely related content. This step allows for more accurate grouping of related documents before topic modeling.

- Once communities are detected, Latent Dirichlet Allocation (LDA) is used to extract topics within each community. By focusing on these pre-grouped communities, the resulting topics are more coherent and accurately represent the underlying themes in the text, addressing the limitations of traditional topic models that often mix different themes or create overlapping topics.

The paper is structured as follows: Section 2 reviews related work, Section 3 describes the methodology, Section 4 discusses the results, section 5 gives the limitation, and Section 6 concludes with findings and future directions.

2 Related Works

This section explains topic modeling methodologies, including word-assisted, clustering-based, and sequence-based approaches.

Word-assisted topic modelling: Topic modeling is a technique used to identify themes and patterns in large text corpora by analyzing the co-occurrence of words and documents. Traditionally, fully automated models such as Latent Dirichlet Allocation (LDA) have been used to extract topics without requiring prior knowledge (Blei et al., 2003; Wood et al., 2017). However, these models often struggle with interpretability and can produce overlapping or ambiguous topics. To address these limitations, the keyATM approach has been proposed, which integrates human input by requiring researchers to specify a few keywords related to the topics of interest before fitting the model (Eshima et al., 2024; Lu et al., 2011). This enhancement has been shown to significantly improve both the interpretability and classification performance of the topics generated, providing more accurate and actionable insights from textual data. A model similar to the base keyATM assumes that each document has a single keyword topic, while other topics may lack keywords. In contrast, keyATM allows each document to belong to multiple keyword topics, providing a more flexible approach (Li et al., 2019).

Clustering based topic modelling: This study explores a hybrid topic modeling approach combining Bidirectional Encoder Representations (BERT) with Latent Dirichlet Allocation (LDA) and unsupervised clustering methods (George and Sumathy, 2023; Lim et al., 2017; Mu et al., 2022). Dimensionality reduction techniques such as PCA, t-SNE, and UMAP are employed to address computational inefficiencies in high-dimensional data. The approach, applied to the CORD19 dataset, integrates LDA's probabilistic topic assignments with BERT's sentence embeddings. Clustering is performed with k-means, and the Elbow Method identifies the optimal number of clusters (Subramani et al., 2018; Alharbi et al., 2021). The results indicate that this hybrid framework enhances topic coherence and effectiveness in topic modeling applications. STELLAR is an interactive tool for topic exploration, using BERT embeddings with UMAP and HDBSCAN to model topics. Human evaluation of the generated topics demonstrated their coherence and relevance (Eklund and Forsman, 2022).

3 Methodology

The proposed method, illustrated in Figure 1, involves several key stages: text preprocessing, text vectorization, graph construction, community detection, and topic modeling.

3.1 Text Preprocessing

Text preprocessing plays a crucial role in the quality of topic modeling results, and the pipeline for this study involved several key steps. First, text cleaning was applied to remove special characters, digits, non-informative tokens such as URLs, and excessive white spaces, while also converting all text to lowercase for consistency. Next, stopword removal was conducted using the NLTK library, eliminating common English stopwords (e.g., 'and', 'the', 'is') as well as domain-specific stopwords to prevent them from influencing topic formation. Finally, lemmatization was performed using the WordNet lemmatizer in NLTK, reducing words to their base or dictionary forms to standardize vocabulary and improve topic coherence by ensuring that variations like "running" and "ran" were treated as the same term, "run." This process helped reduce redundancy and enhance the overall clarity of the topics.

3.2 Text Vectorization

Text data is vectorized using TF-IDF with the term-document matrix X given by (Singh and Shashi, 2019):

$$X_{ij} = \text{TF-IDF}(t_i, d_j) \tag{1}$$

where TF-IDF(t_i, d_j) represents the TF-IDF score of term t_i in document d_j.

3.3 Graph Creation

Cosine similarity is computed for the TF-IDF matrix X to obtain a similarity matrix S: The cosine similarity between documents i and j is computed using:

$$S_{ij} = \frac{\sum_k X_{ki} \cdot X_{kj}}{\sqrt{\sum_k X_{ki}^2} \cdot \sqrt{\sum_k X_{kj}^2}} \tag{2}$$

where S_{ij} represents the cosine similarity between the term vectors of documents i and j. Here, X_{ki} denotes the TF-IDF score of term k in document i, and X_{kj} denotes the TF-IDF score of term k in document j. The numerator $\sum_k X_{ki} \cdot X_{kj}$ calculates the dot product of the term vectors for the two documents, which measures their similarity in terms of term distributions. The denominator $\sqrt{\sum_k X_{ki}^2} \cdot \sqrt{\sum_k X_{kj}^2}$ normalizes this dot product by the magnitudes of the term vectors for both documents, ensuring the similarity score lies between 0 and 1, where 1 indicates identical term distributions.

A binary distance matrix B is then created using a threshold τ:

$$\tau = k \times (\mu + 3\sigma) \tag{3}$$

where μ is the mean and σ is the standard deviation of the similarity values in the matrix S. The parameter k is optimized through hyperparameter tuning to determine the appropriate threshold for binarization. This binarization step, which converts the similarity matrix into a binary form based on the threshold, plays a crucial role in influencing the quality of the resulting topics.

To evaluate the sensitivity of topics to binarization, different threshold levels τ were tested by adjusting the value of k. If the threshold τ is set too

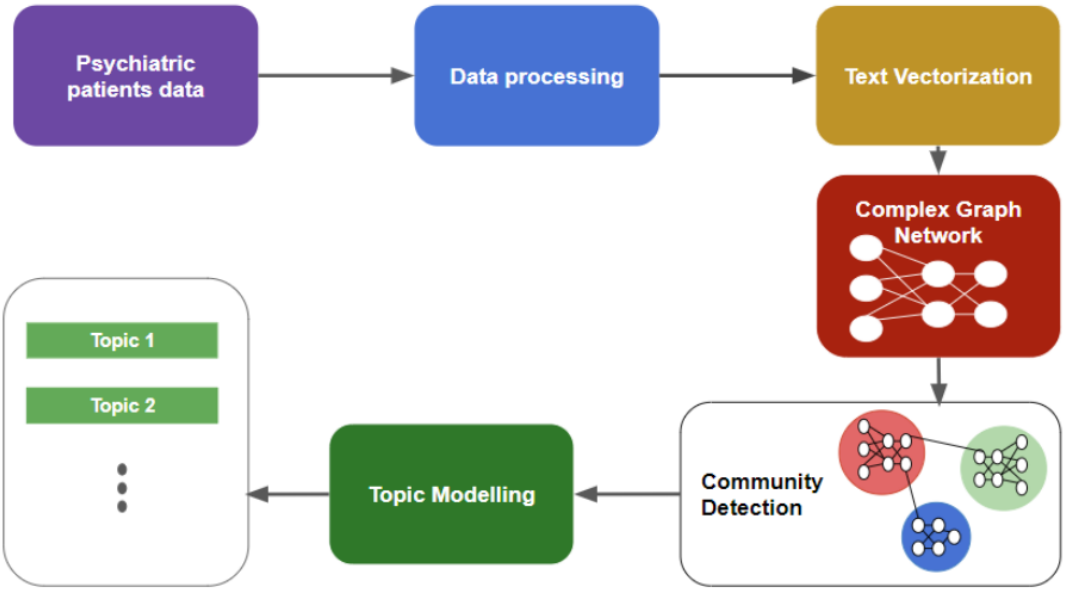

Figure 1: Proposed Architecture of Graph-Driven Community Detection for Topic Modelling.

low, many connections between documents are lost, leading to a sparse graph with fewer edges. This can result in over-fragmentation, where documents that should be grouped together are isolated, thus producing incoherent topics. Conversely, setting a higher threshold results in a densely connected graph, which diminishes the distinction between communities and blurs topic separation. Therefore, tuning k is critical to achieving the right balance between graph connectivity and effective community separation.

The binary distance matrix B is defined by:

$$B_{ij} = \begin{cases} 1, & \text{if } S_{ij} < \tau \\ 0, & \text{if } S_{ij} \geq \tau \end{cases} \quad (4)$$

In this matrix, B_{ij} represents the presence or absence of an edge between documents i and j. If the cosine similarity S_{ij} is less than the threshold τ, B_{ij} is set to 1, indicating a connection. If S_{ij} is greater than or equal to τ, B_{ij} is set to 0, indicating no connection. This binary matrix is used to construct a graph for subsequent community detection.

A graph G is constructed from the binary distance matrix B.

3.4 Community Detection

Community detection is performed using the Louvain method on G. The Louvain method is an algorithm designed to optimize the modularity of a partition of the graph into communities. Modularity is a metric that measures the density of edges within communities compared to edges between different communities.

Modularity Q for a given partition of the graph into communities is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[B_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (5)$$

where B_{ij} is the adjacency matrix of G, where $B_{ij} = 1$ if there is an edge between nodes i and j, and $B_{ij} = 0$ otherwise. k_i and k_j represent the degrees of nodes i and j, respectively, and m is the total number of edges in the graph. The term $\delta(c_i, c_j)$ is the Kronecker delta, which equals 1 if nodes i and j are in the same community and 0 otherwise.

The Louvain algorithm maximizes modularity in two phases. In the Local Moving Phase, each node starts in its own community and is iteratively moved to increase modularity (Meo et al., 2011). Once no further improvement is possible, the Aggregation Phase begins, where communities are combined into single nodes, and the process repeats. These steps are iterated until no more modularity gains can be achieved. The result is a partition of the graph into non-overlapping communities with denser internal connections compared to external ones.

3.5 Topic Modelling

The Latent Dirichlet Allocation (LDA) model is employed to extract latent topics from the communities detected in the previous steps. For this, a corpus and a dictionary are first prepared using the words from each community. The corpus consists of a bag-of-words (BoW) representation of the text, where each document (community) is represented by a list of tuples. Each tuple contains a word and its corresponding frequency in that community. The dictionary maps each unique word to an integer ID. The LDA model is then trained on this corpus, with the dictionary providing the mapping between words and their IDs. The model is configured to extract a predefined number of topics, denoted as K. The training involves iterating over the corpus multiple times, controlled by the parameter 'passes', to refine the topic distribution for each document. In this study, the model is trained with $K = 5$ topics and 15 passes to ensure convergence and optimal topic extraction. The result is a set of topics, each represented by a distribution over the words in the dictionary, which characterizes the underlying themes within the communities.

4 Experimental Discussion

4.1 Experimental setup

Dataset: The datasets utilized in this study, sourced from Kaggle, include the Suicidal Mental Health Dataset, Reddit Mental Health Data, and Predicting Anxiety in Mental Health Data. Each dataset comprises three key features: patient IDs, textual statements, and corresponding mental health status labels.

The Suicidal Mental Health Dataset(SMH) encompasses a wide range of textual data related to suicide, capturing personal experiences, mental health struggles, and appeals for help. Reddit Mental Health Data (RMH) comprises posts and comments from mental health-focused subreddits, offering candid insights into everyday experiences with conditions like depression and anxiety. The Predicting Anxiety in Mental Health Data (AMH) focuses on anxiety-related content, including forum posts and social media comments, detailing symptoms, triggers, and coping mechanisms. Together, these datasets provide a rich foundation for analyzing mental health themes and language patterns in written communication. Table 1 summarizes the stastistics of the three datasets.

Dataset	Total Number of samples
SMH	5000
RMH	10000
AMH	3500

Table 1: A summary of the datasets used in this work. Dataset statistics including total number, majority samples, and minority samples.

Evaluation Measures: To evaluate the effectiveness of topic modeling, we use two key measures in this study: coherence score and perplexity. The coherence score assesses how coherent the topics are by evaluating the degree to which the top words of a topic frequently appear together in the text. A higher coherence score indicates that the topics are more semantically consistent and meaningful. On the other hand, perplexity measures the model's ability to predict a set of words within the text. It provides an indication of how well the model captures the underlying structure of the data, with lower perplexity values suggesting better predictive performance and a more accurate representation of the text (Newman et al., 2011).

Baseline Models and Implementation Details: We performed a comprehensive comparison between our proposed method and four baseline approaches: Latent Dirichlet Allocation (LDA), Spectral Clustering (SC), BERTopic, and K-means Clustering (KMC). This evaluation encompasses a range of traditional topic modeling techniques to ensure thorough benchmarking. Each topic modeling method was tested with multiple configurations, and hyperparameter tuning was employed to optimize their performance.

Our proposed method was developed using Python version 3.11. All experiments were conducted on a desktop computer equipped with a Ryzen 9 5950X processor, 128GB of RAM, and an NVIDIA GeForce RTX 3090 graphics card with 24GB of memory.

4.2 Comparative Analysis on SMH Dataset

The table 2 presents a performance comparison of different topic modeling techniques on the Suicide Mental Health (SMH) dataset. The results highlight the effectiveness of each method based on two metrics: Coherence and Perplexity.

GCD-TM outperforms all other methods with a coherence score of 0.67, indicating that it generates the most semantically meaningful and internally consistent topics. This suggests that GCD-TM is particularly effective at identifying coherent patterns within the text data. Additionally, GCD-TM has the lowest perplexity score (-8.48), demonstrating its strong ability to generalize to new data. This

Figure 2: Wordcloud for SMH Dataset.

low perplexity value means that the topics generated by GCD-TM are not only coherent but also accurate in representing the underlying structure of the dataset.

In comparison, BERTopic shows a decent performance with a coherence score of 0.45, which is better than LDA, SC, and KMC, but still falls short of GCD-TM. Its perplexity score of -7.32, while better than some methods, is also not as low as GCD-TM, indicating room for improvement in predictive accuracy. SC and KMC perform moderately, with coherence scores of 0.41 and 0.39, respectively, and perplexity scores of -7.35 and -8.01. These results suggest that while these methods can produce somewhat coherent topics, they do not perform as well in terms of generalization. Finally, LDA has the lowest coherence score (0.38) and a relatively high perplexity score (-6.15), indicating that it struggles the most with producing coherent topics and accurately modeling the dataset compared to the other techniques.

Methods	**Coherence**	**Perplexity**
LDA	0.38	-6.15
SC	0.41	-7.35
KMC	0.39	-8.01
BERTopic	0.45	-7.32
GCD-TM	0.67	-8.48

Table 2: Performance Comparison of different topic modelling techniques on Suicide Mental Health dataset (SMH).

Figure 2 shows the word cloud for the SMH dataset, highlighting prominent terms such as "life", "tired", "hate", "suicide," "kill," and "dead." Table 3 lists the top five words for each topic generated by five different topic modeling methods (LDA, SC, KMC, BERTopic, and GCD-TM). This table demonstrates how GCD-TM uniquely segments the dataset into distinct topics using the most relevant keywords, showcasing its capability to organize the text data into meaningful clusters, distinguishing itself from the other methods.

Topics	**Top five words by LDA**
1	Time, worry, stress, sadness, solitude
2	Event, joyful, bad, pressure, decision
3	Workplace, chat, help, colleague, hear
4	Prepare, harm, daybreak, close, shadow
5	Love, people, nurture, optimism, remain
Topics	**Top five words by SC**
1	Right, worry, nothing, sadness, isolation
2	Life, individuals, conflict, better, dead
3	Support, idea, advice, partner, attention
4	Passing, back, hope, someone, void
5	Every, bonds, empathy, always, endure
Topics	**Top five words by KMC**
1	Years, fear, anxiety, depression, alone
2	Thing, happy, worst, tension, choice
3	Office, talk, support, friend, listen
4	Ready, kill, morning, end, dark
5	Love, family, care, hope, stay
Topics	**Top five words by BERTopic**
1	Exhaustion, dread, nervous, sad, solitude
2	Living, society, struggle, urge, kill
3	Comfort, chat, courage, buddy, care
4	End, suffer, gloom, silence, empty
5	Warm, connection, kind, belief, family
Topics	**Top five words by GCD-TM**
1	Suicide, fear, anxiety, depression, alone
2	Life, people, kill, tension, hate
3	Help, talk, support, friend, listen
4	Death, pain, hopeless, end, dark
5	Love, family, care, hope, stay

Table 3: Sample topics generated by different topic modelling techniques (top 5 topics) from the Suicide Mental Health dataset (SMH).

4.3 Comparative Analysis on RMH Dataset

Table 4 compares different topic modeling techniques on the Reddit Mental Health Data (RMH) dataset using coherence and perplexity. GCD-TM is the top performer with a coherence score of 0.73

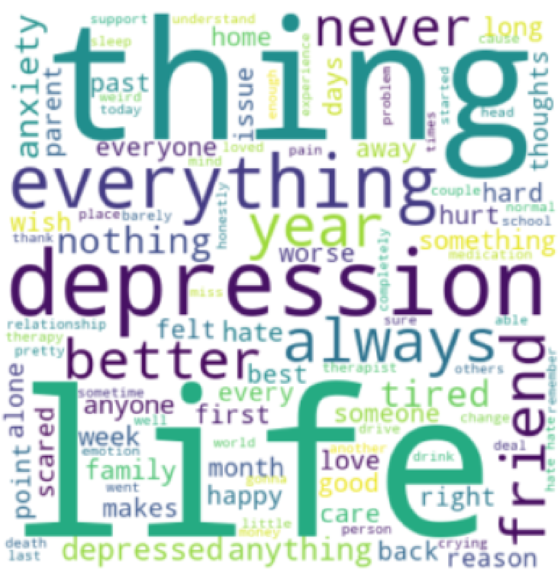

Figure 3: Wordcloud for RMH Dataset.

and a perplexity score of -6.52, indicating it produces the most coherent topics and excels in predicting new data. Spectral Clustering (SC) follows with a coherence score of 0.54 and a perplexity score of -6.08, demonstrating good topic coherence and prediction but not as effectively as GCD-TM.

K-Means Clustering (KMC) scores 0.49 in coherence and has the highest perplexity score of -4.28, showing that while it offers better coherence than some methods, it struggles with generalizing to new data. Latent Dirichlet Allocation (LDA) achieves a coherence score of 0.43 and a perplexity score of -5.51, suggesting moderate coherence and weaker predictive performance. BERTopic, with the lowest coherence score of 0.36 and a perplexity score of -5.37, exhibits the least interpretable topics and only moderate prediction capability.

Methods	**Coherence**	**Perplexity**
LDA	0.43	-5.51
SC	0.54	-6.08
KMC	0.49	-4.28
BERTopic	0.36	-5.37
GCD-TM	0.73	-6.52

Table 4: Performance Comparison of different topic modelling techniques on Reddit Mental Health Data (RMH).

Table 5 presents the topics generated by five topic modeling methods applied to the Reddit Mental Health (RMH) dataset, with each topic represented by its top five associated words. This table highlights the key themes identified, with the GCD-TM model outperforming the other methods. Figure 3 displays the word cloud for the RMH dataset.

Topics	**Top five words by LDA**
1	Hopes, first, better, best, calm
2	Year, stress, rooms, dawns, change
3	Love, thing, bliss, rough, worse
4	Goals, share, ties, smile, hurt
5	Family, odds, paths, trusts, days
Topics	**Top five words by SC**
1	Lives, weary, buddy, tough, right
2	Worry, pains, house, early, major
3	Weeks, close, relax, worst, anger
4	Hopes, trusts, cares, loves, happy
5	Hard, worse, plans, helps, times
Topics	**Top five words by KMC**
1	Life, better, tired, friend, hard
2	Patient, issue, home, long, first
3	Week, family, rest, best, hate
4	Wish, help, support, love, happy
5	Hard, worse, right, care, month
Topics	**Top five words by BERTopic**
1	Goals, heavy, buddy, tough, quiet
2	Tasks, worries, homes, early, shift
3	Break, folks, peace, rough, anger
4	Dreams, gives, bonds, laugh, cheer
5	Trials, risks, steps, trusts, weeks
Topics	**Top five words by GCD-TM**
1	Depression, sadness, tired, empty, life
2	People, struggle, daily, difficult, survive
3	Sleep, insomnia, restless, night, wake
4	Therapy, help, support, treatment, better
5	Pain, cry, tears, hurt, sorrow

Table 5: Sample topics generated by different topic modelling techniques (top 5 topics) from the Reddit Mental Health Data (RMH).

4.4 Comparative Analysis on AMH Dataset

Table 6 compares the performance of various topic modeling techniques on the Anxiety Mental Health Data (AMH) dataset, evaluating each method based on coherence and perplexity. GCD-TM demonstrates the highest coherence score of 0.69 and the lowest perplexity score of -7.92, indicating that it produces the most coherent topics and performs the best in terms of predictive accuracy among the methods evaluated.

In comparison, BERTopic has a coherence score of 0.46 and a perplexity score of -6.77, showing relatively strong performance but not as effective

Figure 4: Wordcloud for RMH Dataset.

as GCD-TM. K-Means Clustering (KMC) follows with a coherence score of 0.45 and a perplexity score of -6.54, indicating reasonable topic coherence but higher perplexity. Spectral Clustering (SC) and Latent Dirichlet Allocation (LDA) have coherence scores of 0.44 and 0.41, respectively, with perplexity scores of -6.08 and -6.18. Both methods show lower coherence and higher perplexity compared to GCD-TM and BERTopic, reflecting their less effective performance in generating and predicting topics.

Methods	Coherence	Perplexity
LDA	0.41	-6.18
SC	0.44	-6.08
KMC	0.45	-6.54
BERTopic	0.46	-6.77
GCD-TM	0.69	-7.92

Table 6: Performance Comparison of different topic modelling techniques on Anxiety Mental Health Data (AMH).

Table 7 presents the top five words associated with each topic generated by different topic modelling methods from the Anxiety Mental Health Data (AMH).Figure 4 illustrates the wordcloud of AMH dataset.

4.5 Performance Analysis and Insights

The superior performance of the Graph-Driven Community Detection for Topic Modeling (GCD-TM) method over traditional techniques like LDA, Spectral Clustering (SC), K-Means Clustering

Topics	Top five words by LDA
1	Hate, anxious, reason, little, long
2	Times, haste, able, point, fear
3	Today, crowd, worry, stress, every
4	Sleep, noise, dreams, hours, rush
5	Quiet, calm, clear, focus, panic
Topics	**Top five words by SC**
1	Stress, worse, shock, creep, today
2	Month, speed, thank, every, always
3	Long, alone, past, shame, nerves
4	Cause, awake, noise, hours, rush
5	Quiet, ease, focus, peace, clear
Topics	**Top five words by KMC**
1	Worry, tensed, month, first, strain
2	Heart, fear, able, time, days
3	Friends, good, alone, avoid, good
4	Sleep, awake, night, dream, chest
5	Calm, heart, relax, focus, hour
Topics	**Top five words by BERTopic**
1	Worry, thing, head, doubt, strain
2	Pain, back, ache, gasp, dread
3	Groups, throng, always, year, crowd
4	Snooze, uneasy, dusk, alert, life
5	Peace, every, ease, days, steer
Topics	**Top five words by GCD-TM**
1	Anxiety, nervous, panic, worry, stress
2	Heart, race, chest, breath, fear
3	Social, crowd, alone, avoid, public
4	Sleep, restless, night, awake, thoughts
5	Calm, breathing, relax, focus, control

Table 7: Sample topics generated by different topic modelling techniques (top 5 topics) from the Anxiety Mental Health Data (AMH).

(KMC), and BERTopic is primarily due to its innovative integration of graph-based community detection with topic modeling. This approach enables GCD-TM to uncover more coherent and contextually meaningful topics, resulting in higher coherence scores and lower perplexity values.

One of the key reasons GCD-TM achieves better results is its ability to capture the intricate relationships between words by constructing a graph where nodes represent terms, and edges represent the similarities between them. Unlike traditional methods that rely solely on statistical distributions, GCD-TM's use of graph theory allows it to identify clusters (or communities) of related words that are more likely to appear together in meaningful contexts. This graph-based approach is particularly effective in detecting nuanced and contextually relevant word associations, which often go unnoticed in methods like LDA, SC, or KMC, where the focus is more on word frequency and document-term distributions. The Louvain method, used for community detection within GCD-TM, further enhances the model's performance by optimizing modularity, which measures the strength of the division of a network into communities. By maximizing modularity, the Louvain method ensures that the communities (or clusters of words) identified are densely connected internally but sparsely connected with other communities. This results in well-defined groups of related words that contribute to more coherent topic extraction during the subsequent LDA phase.

Furthermore, GCD-TM's approach to binarizing the similarity matrix, based on a carefully tuned threshold, allows it to effectively filter out noise and focus on the most significant word associations. This selective process leads to the construction of a more accurate and representative graph, which is crucial for the success of community detection and, ultimately, topic modeling. By integrating these techniques, GCD-TM is able to provide a more detailed and accurate representation of the underlying thematic structure in the data. This holistic approach not only captures the global context within the text but also reveals the subtle, localized patterns that are often missed by other models. As a result, GCD-TM produces topics that are not only more coherent but also more reflective of the actual content and structure of the dataset, leading to its superior performance compared to traditional topic modeling methods.

5 Limitations

Despite its strong performance, GCD-TM has certain limitations. First, the method's reliance on graph construction and community detection makes it computationally intensive, especially with large datasets, which can lead to increased processing times. Second, the model's effectiveness is sensitive to the choice of threshold for graph binarization and the number of topics specified, requiring careful hyperparameter tuning to achieve optimal results. Additionally, while GCD-TM excels at capturing well-defined themes, it may struggle with topics that are highly interrelated or overlap significantly, potentially leading to less distinct topic separation. Finally, the complexity of the method might make it less accessible for users who are not familiar with graph-based approaches or community detection techniques, limiting its broader applicability in different research contexts.

6 Conclusion

In this study, we introduced the Graph-Driven Community Detection for Topic Modeling (GCD-TM) methodology and demonstrated its effectiveness in uncovering thematic structures within mental health datasets. By leveraging a combination of text preprocessing, TF-IDF vectorization, graph-based community detection, and topic modeling with LDA, GCD-TM offers a robust approach to identifying and analyzing latent topics. Our method outperforms traditional techniques in terms of coherence and perplexity, highlighting its capacity to deliver a more nuanced and accurate representation of underlying themes.

The comparative evaluation reveals that GCD-TM not only achieves superior results but also provides a more detailed understanding of thematic structures compared to methods like LDA, SC, KMC, and BERTopic. The combination of community detection and topic modeling enhances the model's ability to capture complex relationships within the data, leading to more meaningful and interpretable topics. To further enhance the GCD-TM methodology, it would be beneficial to explore the integration of additional data features and modalities, experiment with alternative algorithms, and assess the model's performance across more diverse datasets to improve robustness and adaptability. Additionally, investigating methods to address computational efficiency and scalability will be crucial for handling larger and more complex datasets. Finally, applying GCD-TM to other domains and comparing its performance with emerging techniques could provide valuable insights and drive innovation in topic modeling.

References

Adel R. Alharbi, Mohammad Hijji, and Amer Aljaedi. 2021. Enhancing topic clustering for arabic security news based on k-means and topic modelling. *IET Networks*, 10(6):278–294.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Gavin Brookes and Tony McEnery. 2019. The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1):3–21.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22:288–296.

Chaitanya Chemudugunta, America Holloway, Padhraic Smyth, and Mark Steyvers. 2008. Modeling documents by combining semantic concepts with unsupervised statistical learning. In *Proceedings of the 7th International Conference on The Semantic Web (ISWC '08)*, pages 229–244, Berlin, Heidelberg. Springer-Verlag.

Trevor Cohen, Brett Blatter, and Vimla Patel. 2008. Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. *Journal of Biomedical Informatics*, 41(6):1070–1087.

Pim Cuijpers, Ellen Driessen, Steven D Hollon, Patricia van Oppen, Jürgen Barth, and Gerhard Andersson. 2012. The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. *Clinical Psychology Review*, 32:280–291.

Robert J DeRubeis, Greg J Siegle, and Steven D Hollon. 2008. Cognitive therapy versus medication for depression: treatment outcomes and neural mechanisms. *Nature Reviews Neuroscience*, 9:788–796.

Susan T. Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST)*, 38:189–230.

Anton Eklund and Mona Forsman. 2022. Topic modeling by clustering language model embeddings: Human validation on an industry dataset.

Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. 2024. Keyword-assisted topic models. *American Journal of Political Science*, 68(2):730–750.

Jose F Figueroa, Jessica Phelan, E John Orav, Vikram Patel, and Ashish K Jha. 2020. Association of mental health disorders with health care spending in the medicare population. *JAMA Netw Open*, 3:e201210.

Joseph Firth, John Torous, Jennifer Nicholas, Rebekah Carney, Abhishek Pratap, Simon Rosenbaum, and Jerome Sarris. 2017. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry*, 16:287–298.

Lijimol George and P. Sumathy. 2023. An integrated clustering and bert framework for improved topic modeling. *International Journal of Information Technology*, 15(4):2187–2195.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet*, 392:1789–1858.

Daniel Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13.

David M. Lei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Chenliang Li, Shiqian Chen, Jian Xing, Aixin Sun, and Zongyang Ma. 2019. Seed-guided topic model for document filtering and classification. *ACM Transactions on Information Systems*, 37(1):1–37.

Kwan Hui Lim, Shanika Karunasekera, and Aaron Harwood. 2017. Clustop: A clustering-based topic modelling algorithm for twitter using word networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 2009–2018. IEEE.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, pages 81–88, Vancouver, BC, Canada. IEEE.

Matteo Malgaroli, Thomas D Hull, James M Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: A systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. 2011. Generalized louvain method for community detection in large networks. In *Proceedings of the 11th International Conference on Intelligent Systems Design and Applications*, pages 88–93. IEEE.

Jeanne Miranda, Thomas G McGuire, David R Williams, and Philip Wang. 2008. Mental health in the context of health disparities. *American Journal of Psychiatry*, 165:1102–1108.

Wenchuan Mu, Kwan Hui Lim, Junhua Liu, Shanika Karunasekera, Lucia Falzon, and Aaron Harwood. 2022. A clustering-based topic model using word networks and word embeddings. *Journal of Big Data*, 9(1):38.

David Newman, Edwin V. Bonilla, and Wray Buntine. 2011. Improving topic coherence with regularized topic models. *Advances in Neural Information Processing Systems*, 24.

Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14.

Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102.

Braja Gopal Patra, Reshma Kar, Kirk Roberts, and Hulin Wu. 2020. Mental health severity detection from psychological forum data using domain-specific unlabelled data. *AMIA Summits on Translational Science Proceedings*, 2020:487.

Anna Rumshisky, Marzyeh Ghassemi, Tristan Naumann, Peter Szolovits, VM Castro, TH McCoy, and RH Perlis. 2016. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, 6(10):e921–e921.

Kristina P. Sinaga and Miin-Shen Yang. 2020. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727.

Anita Kumari Singh and Mogalla Shashi. 2019. Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*, 10(7).

Sandhya Subramani, Vaishnavi Sridhar, and Kaushal Shetty. 2018. A novel approach of neural topic modelling for document clustering. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2169–2173. IEEE.

Zhou Tong and Haiyi Zhang. 2016. A text mining research based on lda topic modelling. In *Proceedings of the International Conference on Computer Science, Engineering and Information Technology*, pages 201–210.

Philip S Wang, Sergio Aguilar-Gaxiola, Jordi Alonso, Matthias C Angermeyer, Guilherme Borges, Evelyn J Bromet, Ronny Bruffaerts, Giovanni De Girolamo, Ron De Graaf, Oye Gureje, et al. 2007. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the who world mental health surveys. *Lancet*, 370:841–850.

J. Wood, P. Tan, W. Wang, and C. Arnold. 2017. Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In *Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 411–422, San Diego, CA, USA. IEEE.