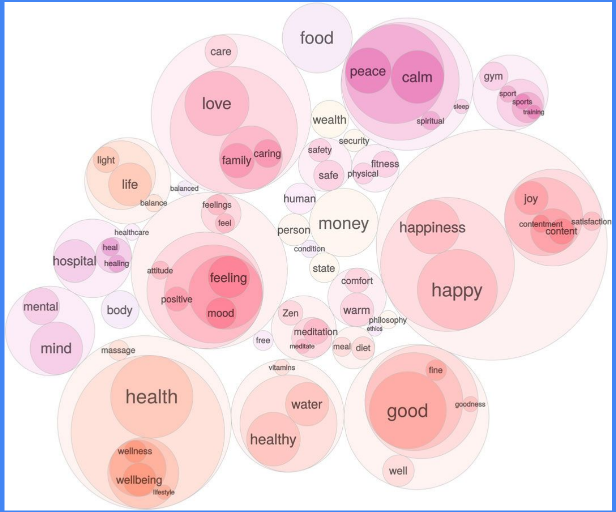
Analyzing Reddit Mental Health Discussions to Identify Key Themes and Influential Contributors



DS5230: Unsupervised Machine Learning (Fall-24)

Submitted By:
Yash Khare
Manjusha Motamarry
Karthikeyan Sugavanan

Under the guidance of:
Prof. David Brady

Table Of Contents

- Problem Statement
- Objectives
- Dataset Description
- Methodology
- Data Tidying
- Initial EDA
- Clustering Methods
- Topic Modeling
- Link Analysis
- Social Network Analysis
- Temporal Shifts in Discussion Themes
- Conclusions/Recommendations

Problem Statement & Objectives

Problem Statement:
Mental health issues are escalating globally, affecting millions and imposing substantial social and economic burdens. Online communities like Reddit provide crucial platforms for individuals to discuss their mental health experiences, generating extensive and complex data. Our project utilizes the Reddit Mental Health Dataset to uncover key themes and influential contributors by applying advanced analytical techniques, aiming to identify patterns that can inform effective support strategies.

Objectives:

1. Discover themes in mental health discourse through clustering and topic modeling.
2. Detect influential contributors by analyzing user interactions.
3. Examine temporal shifts to identify how community concerns change over time.

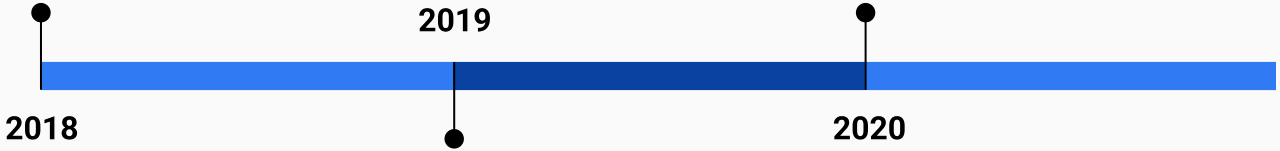
Reddit Mental Health Dataset Overview

Jan 1 to April 20, 2018. A control for seasonal fluctuations to match post data.

Unique users: 177,089

post: Jan 1 to April 20, 2020 (called "mid-pandemic" in manuscript; r/COVID19_support appears).

Unique users: 320,364



2018

2019

2020

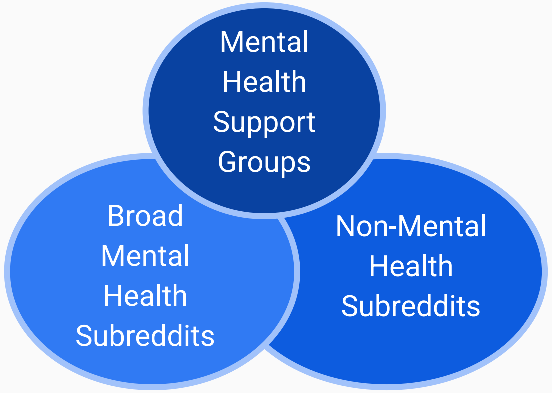
pre: Dec 2018 to Dec 2019. A full year which provides more data for a baseline of Reddit posts.

Unique users: 327,289.

Jan 1 to April 20, 2019

(r/EDAnonymous appears). A control for seasonal fluctuations to match post data.

Unique users: 282,560.



Mental Health Support Groups

Broad Mental Health Subreddits

Non-Mental Health Subreddits

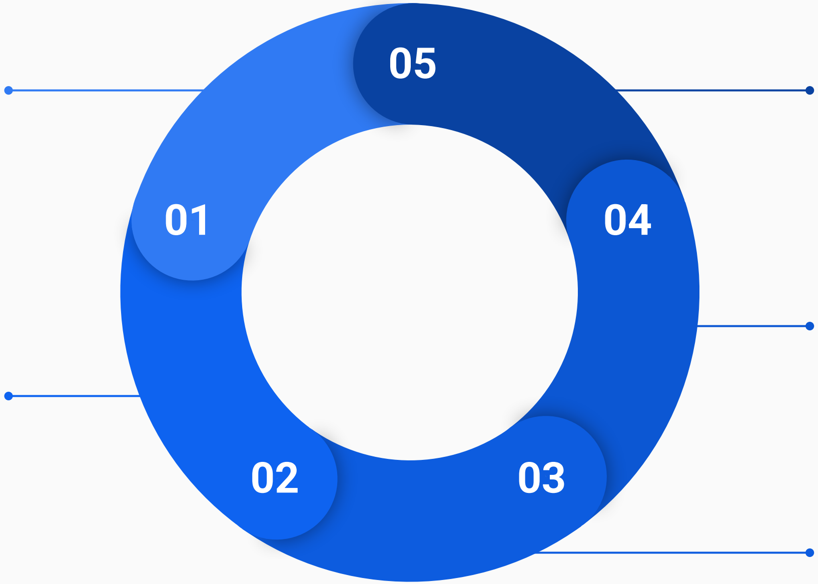
Methodology

Data Preprocessing

Data cleaning and normalization, TF-IDF vectorization of textual data, Metadata extraction (subreddit, timeframe)

Exploratory Data Analysis (EDA)

Initial examination of data distributions and trends, Identification of key patterns and anomalies.



Temporal Analysis

Tracking changes in discussion topics over time, Correlation with external events (e.g., COVID-19 pandemic).

Social Network Analysis

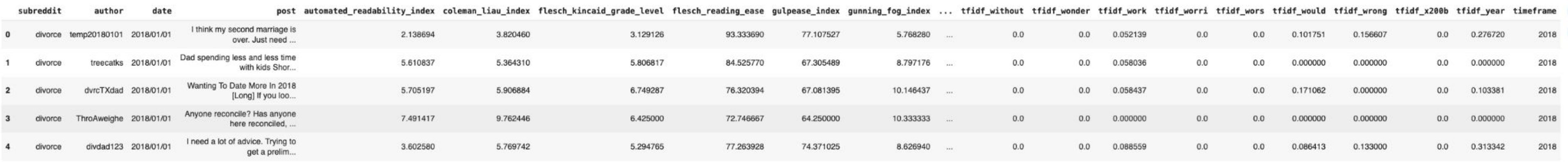
Analysis of user interactions to identify influential contributors, Metrics used: PageRank, degree centrality.

Clustering and Topic Modeling

Clustering: **Gaussian Mixture Models (GMM)** to group similar posts.

Topic Modeling: **Latent Dirichlet Allocation (LDA)** to extract underlying themes.

Data Tidying



	subreddit	author	date	post	automated_readability_index	coleman_liau_index	flesch_kincaid_grade_level	flesch_reading_ease	gulpease_index	gunning_fog_index	...	tfidf_without	tfidf_wonder	tfidf_work	tfidf_worri	tfidf_wors	tfidf_would	tfidf_wrong	tfidf_x200b	tfidf_year	timeframe
0	divorce	temp20180101	2018/01/01	I think my second marriage is over. Just need ...	2.138694	3.820460	3.129126	93.333690	77.107527	5.768280	...	0.0	0.0	0.052139	0.0	0.0	0.101751	0.156607	0.0	0.276720	2018
1	divorce	treecatks	2018/01/01	Dad spending less and less time with kids Shor...	5.610837	5.364310	5.806817	84.525770	67.305489	8.797176	...	0.0	0.0	0.058036	0.0	0.0	0.000000	0.000000	0.0	0.000000	2018
2	divorce	dvrcTXdad	2018/01/01	Wanting To Date More In 2018 [Long] If you loo...	5.705197	5.906884	6.749287	76.320394	67.081395	10.146437	...	0.0	0.0	0.058437	0.0	0.0	0.171062	0.000000	0.0	0.103381	2018
3	divorce	ThroAweighe	2018/01/01	Anyone reconcile? Has anyone here reconciled, ...	7.491417	9.762446	6.425000	72.746667	64.250000	10.333333	...	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.000000	2018
4	divorce	divdad123	2018/01/01	I need a lot of advice. Trying to get a prelim...	3.602580	5.769742	5.294765	77.263928	74.371025	8.626940	...	0.0	0.0	0.088559	0.0	0.0	0.086413	0.133000	0.0	0.313342	2018

Initial Data Handling:

- Loaded large-scale dataset using ***RAPIDS cuDF*** for GPU efficiency.
- Concatenated multiple CSV files into a unified DataFrame.

Selecting Relevant Features:

- Identified and extracted TF-IDF columns (**tfidf_ prefix**).
- Retained metadata columns: subreddit and timeframe.

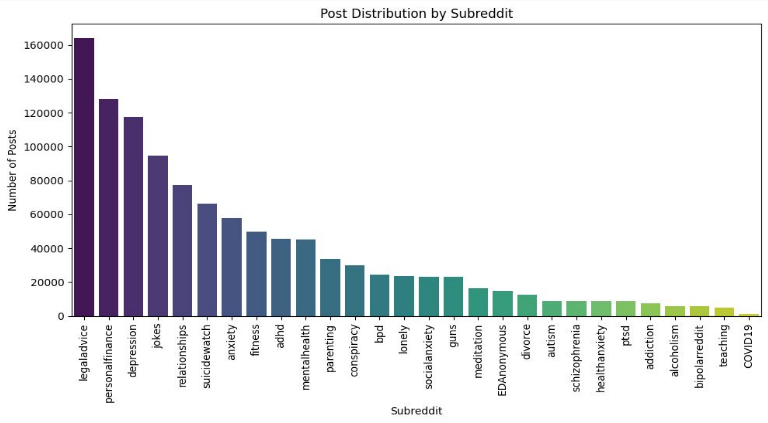
Final Dataset Structure:

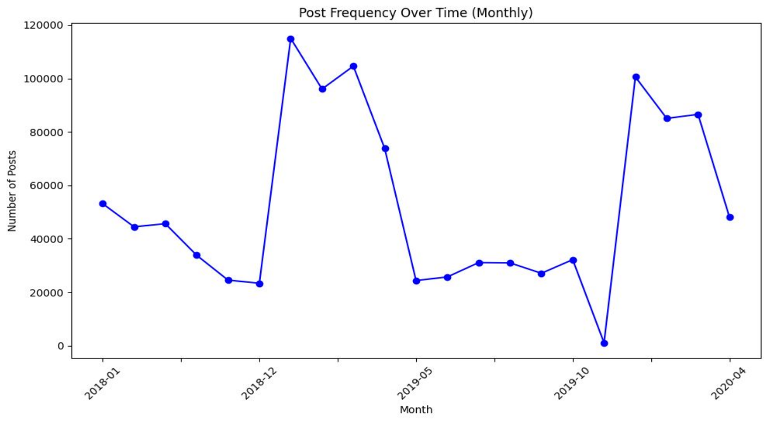
- Shape: (1**,107,302 rows, 351 columns**)
  - Components:
    - 256-dimensional TF-IDF features.
    - Metadata: subreddit, timeframe.

Initial EDA

1. **Subreddit Activity Distribution:**
- The highest post contributions came from r/legaladvice, r/personalfinance, and r/depression, as shown in the bar chart.
- Mental health-specific subreddits like r/suicidewatch and r/anxiety also had significant activity, highlighting their importance in community discussions.

2. **Post Frequency Over Time:**
- Monthly post frequency indicates fluctuations, with notable peaks during early 2019 & 2020, correlating with the onset of COVID-19.
- This trend reflects increased community engagement during critical global events.





Initial EDA

3. Top Contributors:

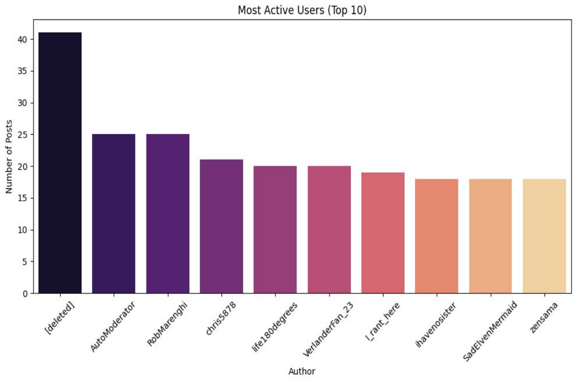
Analysis of user contributions revealed that [deleted], AutoModerator, and other frequent contributors play pivotal roles in shaping subreddit content.

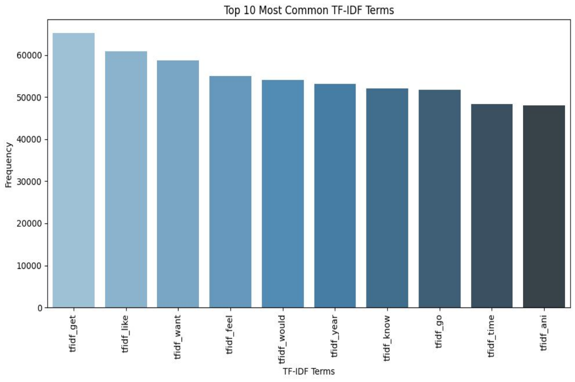
4. Posts per Subreddit Over Time:

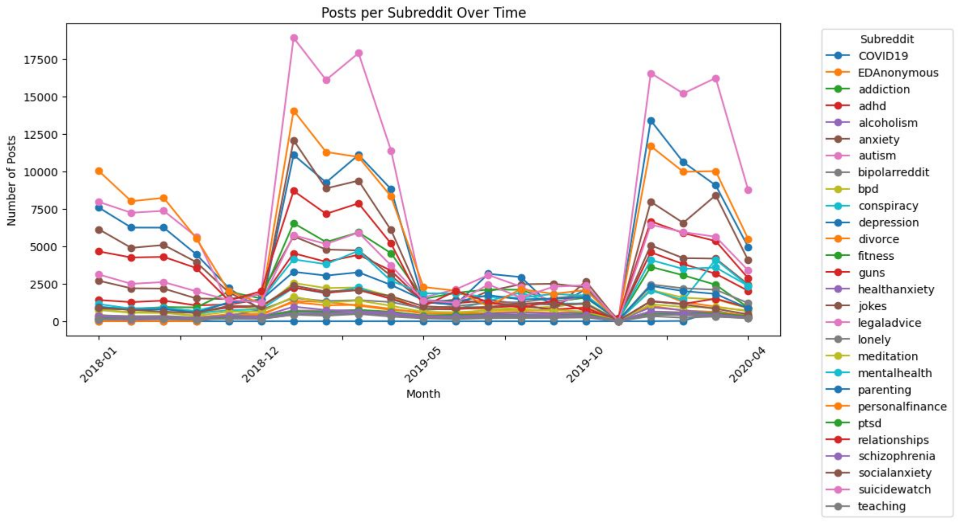
Time-series analysis shows activity surges in mental health-related subreddits like r/COVID19_support during pandemic peaks, emphasizing their role as support hubs.

5. Most Common TF-IDF Terms:

Key terms like "get," "like," and "feel" frequently appeared across posts, indicating the conversational and emotional nature of the discussions.





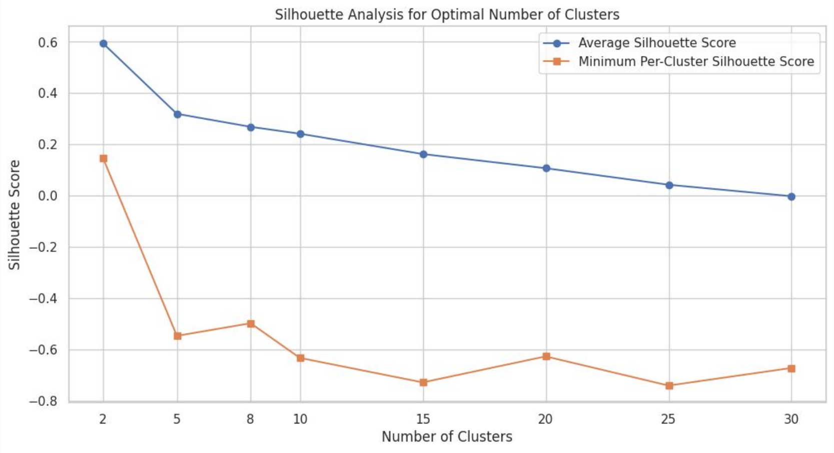


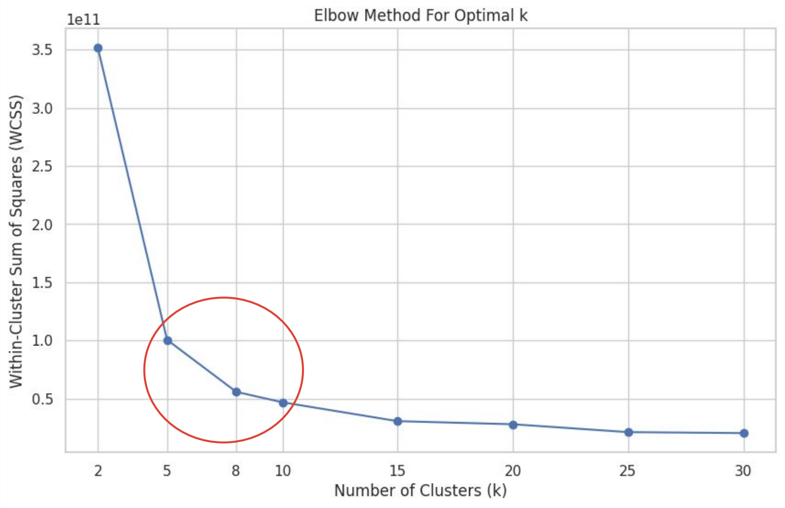
Clustering Analysis Using GMM

Key Insights:

1. **Optimal Number of Clusters:**
   - Using Silhouette Analysis and the Elbow Method, **10** clusters were identified as optimal.
   - Silhouette Score: **0.2406**, representing a trade-off between cohesion and separation.

2. **Cluster Evaluation:**
   - Calinski-Harabasz Score: **498420.2171**, indicating compact and well-separated clusters.
   - Heatmap analysis shows subreddit-specific clusters aligning with thematic content.



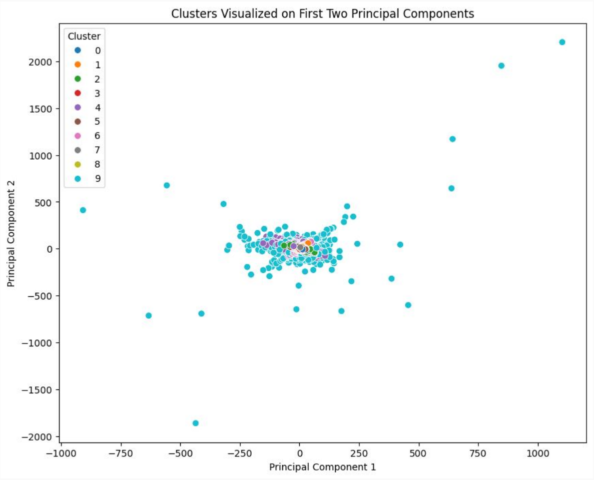


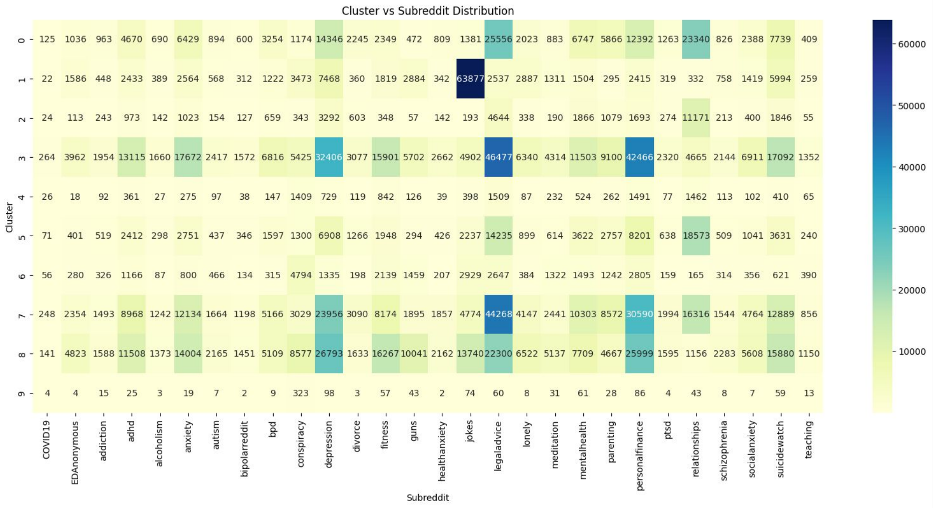
Cluster Visualization and Conclusion

1. **PCA Scatter Plot:**
- Clusters visualized on the first two principal components, highlighting distinct separations.
- Minimal overlap between clusters, ensuring meaningful grouping.

2. **Cluster vs Subreddit Heatmap:** Displays subreddit distribution across clusters, revealing strong associations for subreddits like **r/depression** and **r/anxiety**.

3. **Conclusion:**
- 10 Clusters Selected: Maintains balance between interpretability and statistical rigor.
- Aligns well with subreddit-specific content, capturing thematic distinctions.




Cluster vs Subreddit Distribution

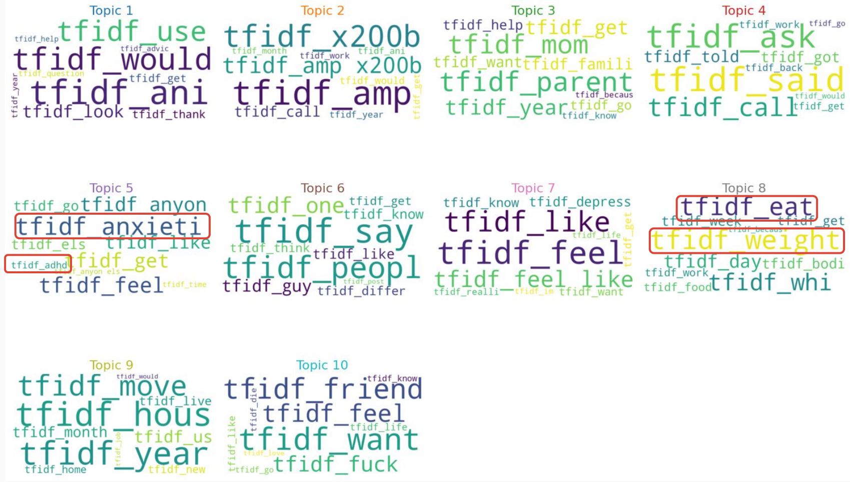
Cluster	COVID19	EDAnonymous	addiction	adhd	alcoholism	anxiety	autism	bipolarreddit	bpd	conspiracy	depression	divorce	fitness	guns	healthanxiety	jokes	legaladvice	lonely	meditation	mentalhealth	parenting	personalfinance	ptsd	relationships	schizophrenia	socialanxiety	suicidewatch	teaching
0	125	1036	963	4670	690	6429	894	600	3254	1174	14346	2245	2349	472	809	1381	25556	2023	883	6747	5866	12392	1263	23340	826	2388	7739	409
1	22	1586	448	2433	389	2564	568	312	1222	3473	7468	360	1819	2884	342	63877	2537	2887	1311	1504	295	2415	319	332	758	1419	5994	259
2	24	113	243	973	142	1023	154	127	659	343	3292	603	348	57	142	193	4644	338	190	1866	1079	1693	274	11171	213	400	1846	55
3	264	3962	1954	13115	1660	17672	2417	1572	6816	5425	32406	3077	15901	5702	2662	4902	46477	6340	4314	11503	9100	42466	2320	4665	2144	6911	17092	1352
4	26	18	92	361	27	275	97	38	147	1409	729	119	842	126	39	398	1509	87	232	524	262	1491	77	1462	113	102	410	65
5	71	401	519	2412	298	2751	437	346	1597	1300	6908	1266	1948	294	426	2237	14235	899	614	3622	2757	8201	638	18573	509	1041	3631	240
6	56	280	326	1166	87	800	466	134	315	4794	1335	198	2139	1459	207	2929	2647	384	1322	1493	1242	2805	159	165	314	356	621	390
7	248	2354	1493	8968	1242	12134	1664	1198	5166	3029	23956	3090	8174	1895	1857	4774	44268	4147	2441	10303	8572	30590	1994	16316	1544	4764	12889	856
8	141	4823	1588	11508	1373	14004	2165	1451	5109	8577	26793	1633	16267	10041	2162	13740	22300	6522	5137	7709	4667	25999	1595	1156	2283	5608	15880	1150
9	4	4	15	25	3	19	7	2	9	323	98	3	57	43	2	74	60	8	31	61	28	86	4	43	8	7	59	13

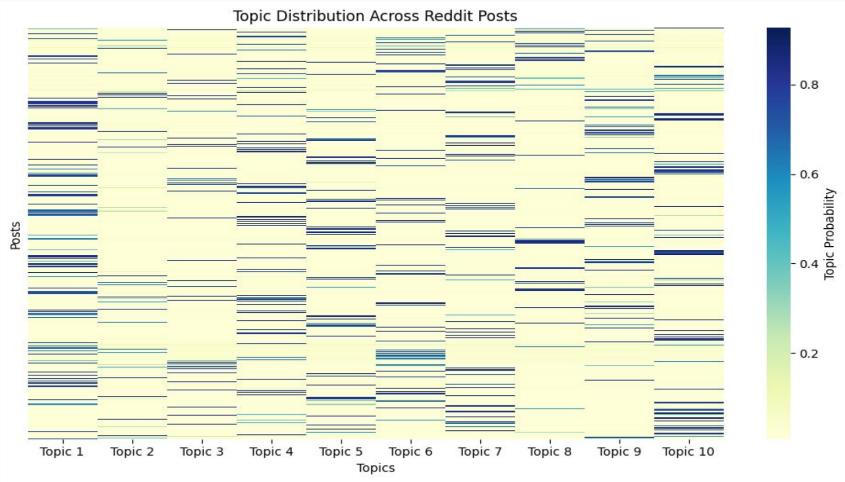
Topic Modeling Using LDA

1. **Purpose of Topic Modeling:**
   - Identify and interpret key themes in Reddit discussions using Latent Dirichlet Allocation (LDA).
   - Analyze how different subreddits contribute to specific discussion topics.

2. **Topics Extracted:**
   - Derived 10 topics, aligned with the optimal number of clusters suggested by Gaussian Mixture Modeling (GMM).

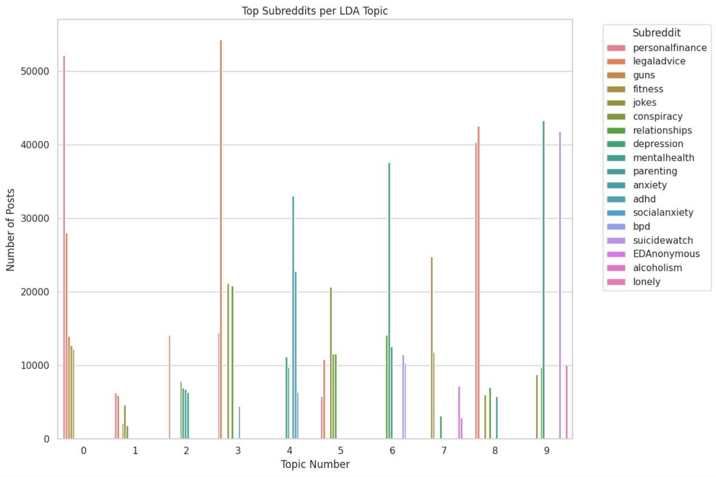
3. **Visualizations:**
   - Word Clouds: Highlight dominant words for each topic, providing intuitive insights into the content of discussions.
   - Topic Distribution Heatmap: Shows topic probabilities across posts, indicating overlaps and dominant themes.

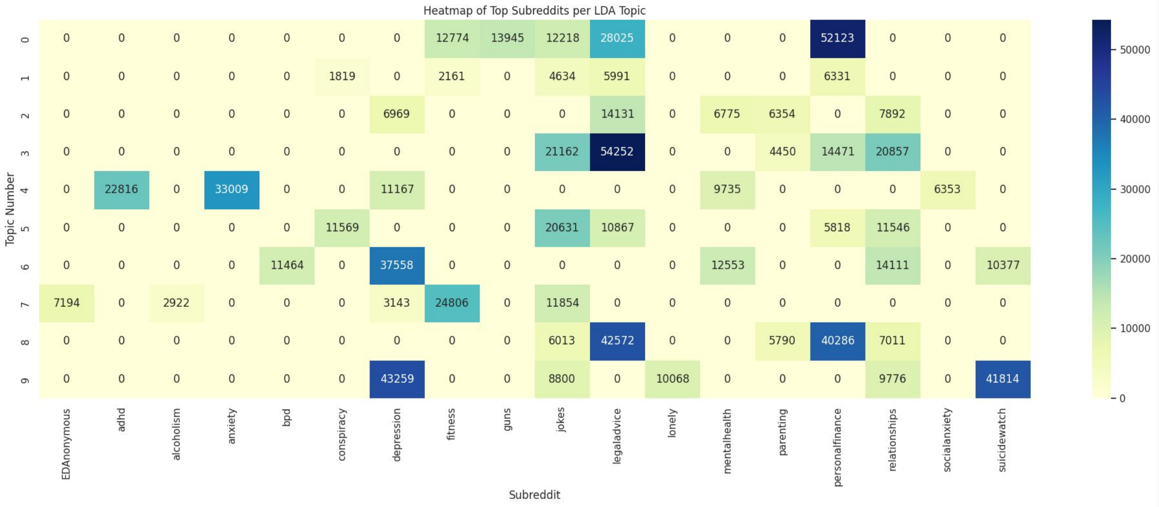




Topic Modeling Insights and Subreddit Contributions

1. **Connection Between Clustering and Topic Modeling:**
   - GMM clustering indicated 10 as the optimal number of clusters, which guided the selection of 10 topics for LDA.
2. **Subreddit Distribution Across Topics:**
   - Subreddits such as r/anxiety and r/depression align closely with mental health topics, while r/legaladvice dominates legal discussions.
3. **Visualizations:**
   - Bar Plot: Highlights top contributing subreddits for each topic, illustrating thematic alignment.
   - Subreddit-Topic Heatmap: Shows detailed subreddit contributions, helping link subreddits with topics.





	EDAnonymous	adhd	alcoholism	anxiety	bpd	conspiracy	depression	fitness	guns	jokes	legaladvice	lonely	mentalhealth	parenting	personalfinance	relationships	socialanxiety	suicidewatch
0	0	0	0	0	0	0	0	12774	13945	12218	28025	0	0	0	52123	0	0	0
1	0	0	0	0	0	1819	0	2161	0	4634	5991	0	0	0	6331	0	0	0
2	0	0	0	0	0	0	6969	0	0	0	14131	0	6775	6354	0	7892	0	0
3	0	0	0	0	0	0	0	0	0	21162	54252	0	0	4450	14471	20857	0	0
4	0	22816	0	33009	0	0	11167	0	0	0	0	0	9735	0	0	0	6353	0
5	0	0	0	0	0	11569	0	0	0	20631	10867	0	0	0	5818	11546	0	0
6	0	0	0	0	11464	0	37558	0	0	0	0	0	12553	0	0	14111	0	10377
7	7194	0	2922	0	0	0	3143	24806	0	11854	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	6013	42572	0	0	5790	40286	7011	0	0
9	0	0	0	0	0	0	43259	0	0	8800	0	10068	0	0	0	9776	0	41814

Connectivity, Influence, and Behavioral Insights of Clusters

Objective: Identify influential user clusters using degree centrality and PageRank and analyzing engagement and sentiment metrics to understand user behavior. Visualize the clusters' connectivity and rank them based on influence and activity.

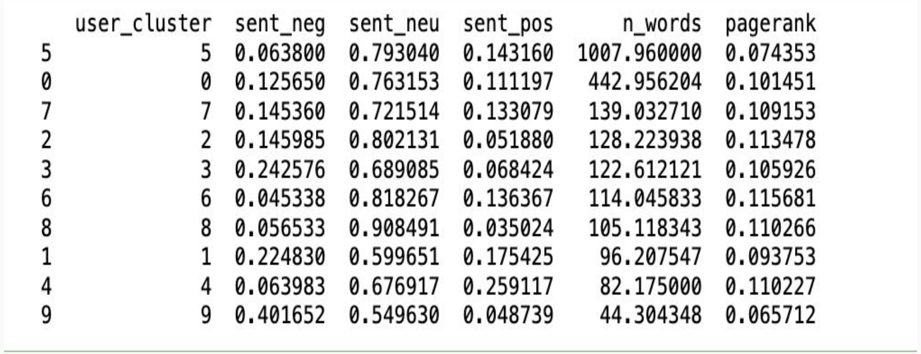
Method:

- Computed degree centrality and PageRank to measure connectivity and importance.
- Aggregated sentiment and engagement metrics, and applied KMeans clustering.
- Visualized connection between influential clusters and engagement insights.

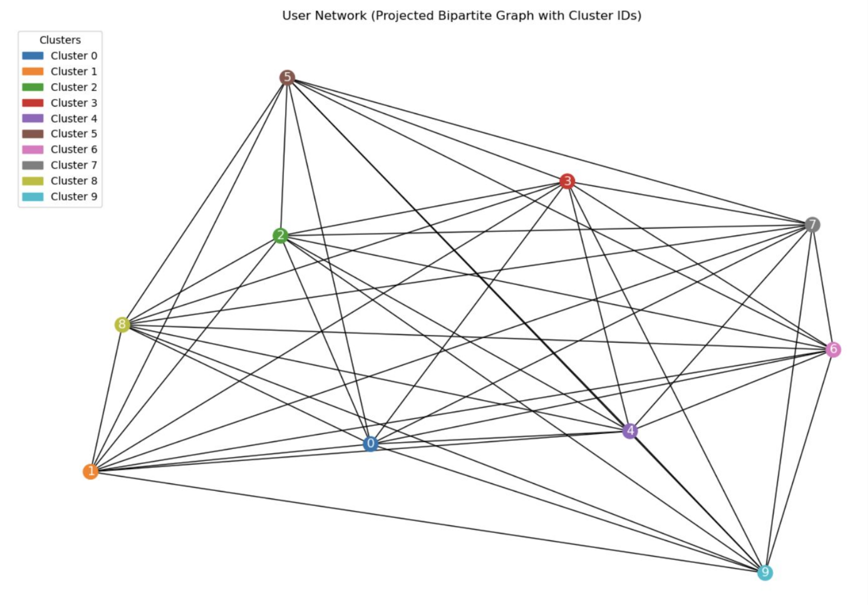
Techniques Used: NetworkX for graph algorithms (degree centrality, PageRank) and visualization, KMeans for clustering, and Pandas for sentiment and engagement aggregation. Matplotlib was used for graph and metric visualizations.

Outcome: Identified top user clusters ranked by connectivity and influence, with their engagement and sentiment profiles analyzed. Provided actionable insights through clear visualization of network structure and user behavior metrics.

Connectivity, Influence, and Behavioral Insights of Clusters



```
   user_cluster  sent_neg  sent_neu  sent_pos     n_words  pagerank
5             5  0.063800  0.793040  0.143160  1007.960000  0.074353
0             0  0.125650  0.763153  0.111197   442.956204  0.101451
7             7  0.145360  0.721514  0.133079   139.032710  0.109153
2             2  0.145985  0.802131  0.051880   128.223938  0.113478
3             3  0.242576  0.689085  0.068424   122.612121  0.105926
6             6  0.045338  0.818267  0.136367   114.045833  0.115681
8             8  0.056533  0.908491  0.035024   105.118343  0.110266
1             1  0.224830  0.599651  0.175425    96.207547  0.093753
4             4  0.063983  0.676917  0.259117    82.175000  0.110227
9             9  0.401652  0.549630  0.048739    44.304348  0.065712
```



User Network (Projected Bipartite Graph with Cluster IDs)

Influential Users in a Social Network

Objective: Identify influential users in a social network based on their importance and connectivity.

Method: Utilized PageRank and Degree Centrality algorithms to measure user influence and connectivity.

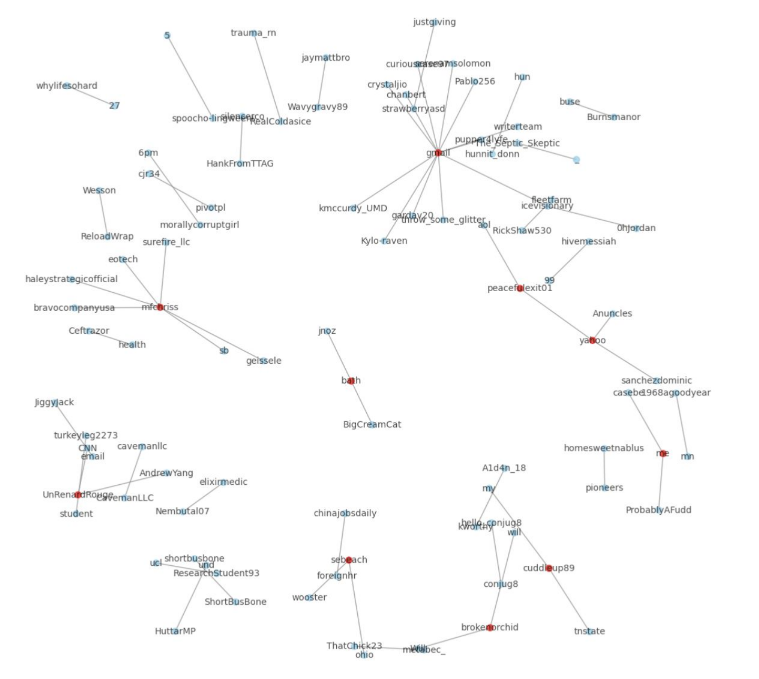
Techniques Used:

- **PageRank:**
  Looks at how important someone is by seeing how likely people are to visit their profile based on the links they have with others.
- **Degree Centrality:**
  Measures how well-connected someone is by counting how many friends or direct connections they have.

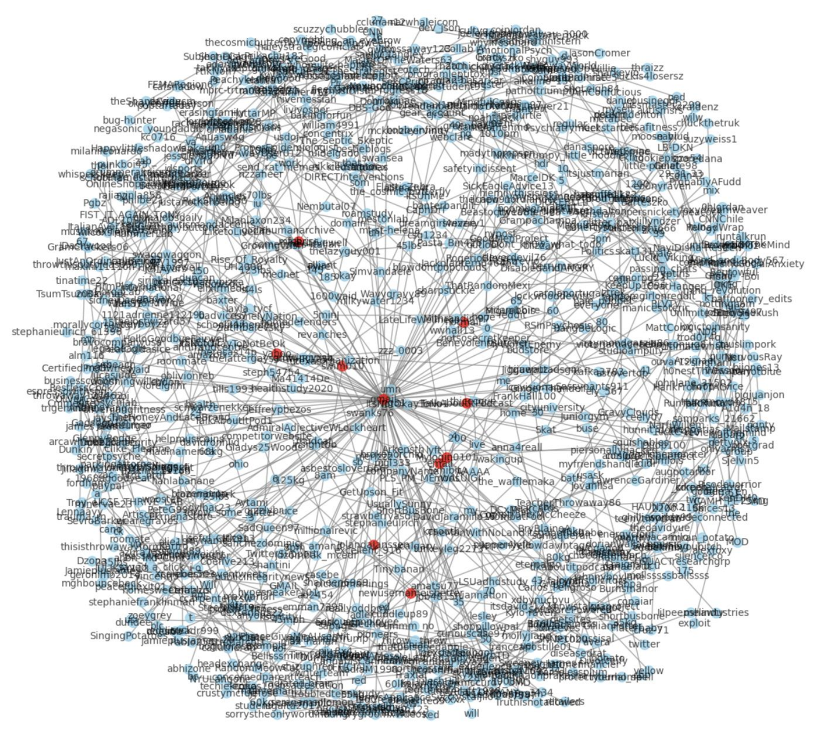
Visualization:

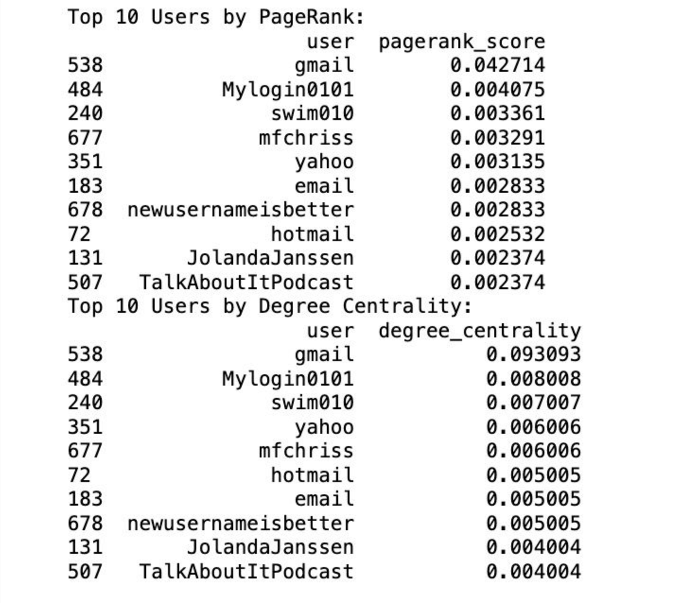
1. Highlighted top central users using nodes in the network graph.
2. Simplified the graph for only 100 nodes.

Outcome: The most influential and connected users were identified and visualized.

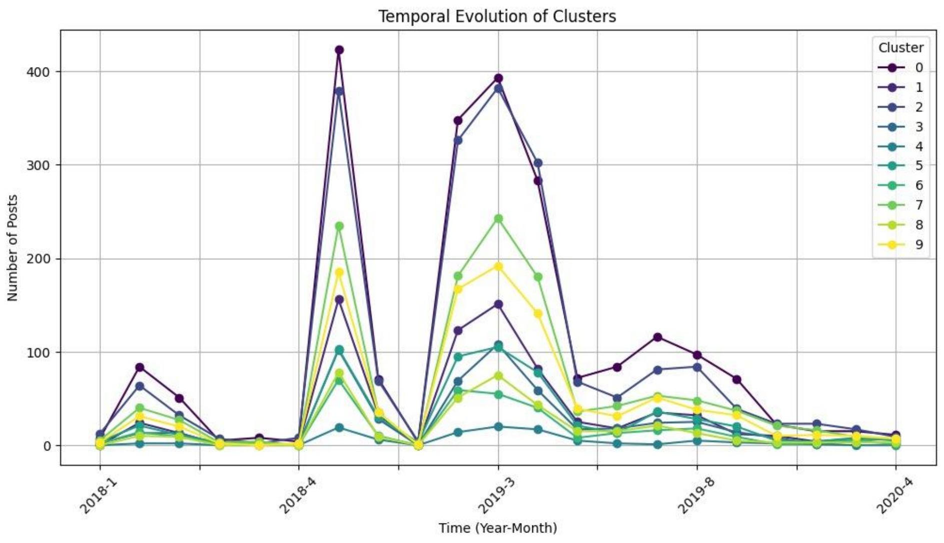


Influential Users in a Social Network





```
Top 10 Users by PageRank:
                        user  pagerank_score
538                    gmail        0.042714
484              Mylogin0101        0.004075
240                  swim010        0.003361
677                  mfchriss        0.003291
351                    yahoo        0.003135
183                    email        0.002833
678     newusernameisbetter        0.002833
72                   hotmail        0.002532
131            JolandaJanssen        0.002374
507         TalkAboutItPodcast        0.002374
Top 10 Users by Degree Centrality:
                        user  degree_centrality
538                    gmail           0.093093
484              Mylogin0101           0.008008
240                  swim010           0.007007
351                    yahoo           0.006006
677                  mfchriss           0.006006
72                   hotmail           0.005005
183                    email           0.005005
678     newusernameisbetter           0.005005
131            JolandaJanssen           0.004004
507         TalkAboutItPodcast           0.004004
```



Temporal Shifts in Discussion Themes

Objective:

Analyze shifts in discussion topics and temporal trends to understand evolving patterns.

Method:

- **PCA**: Reduced data dimensions to highlight theme changes.
- **KMeans Clustering**: Categorized discussions into 10 clusters.
- **Temporal Grouping**: Tracked cluster activity by year-month.

Outcome:

Identified key topic shifts and high-activity periods, enabling timely responses to emerging mental health issues.

Heatmap of Cluster Activity

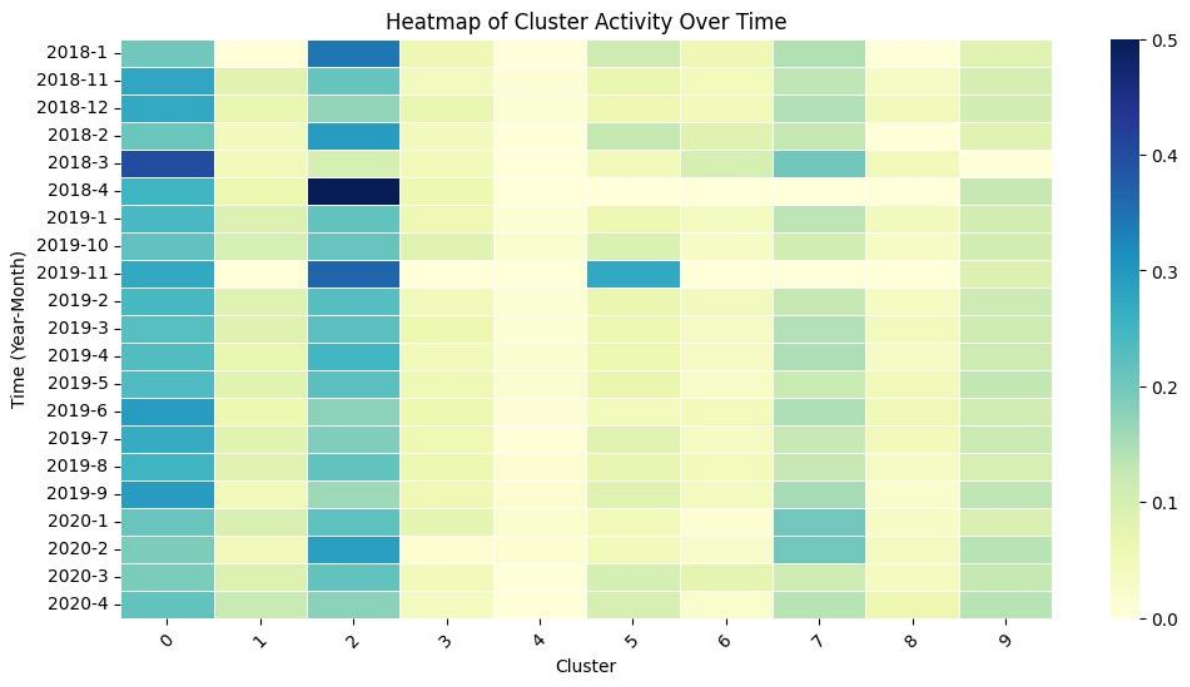
Outcome:

- Identified periods of dominance for specific clusters.
- Provided insights into the ebb and flow of key discussion topics.

Temporal Focus: The heatmap highlights the intensity of cluster activity over time, with darker shades indicating periods of dominance.

Key Insights:

- Cluster 2 shows significant activity spikes in specific months, likely indicating heightened discussions around specific topics.
- Cluster dominance varies significantly, reflecting the shifting focus of discussions across time.



Stacked Area Plot of Cluster Trends

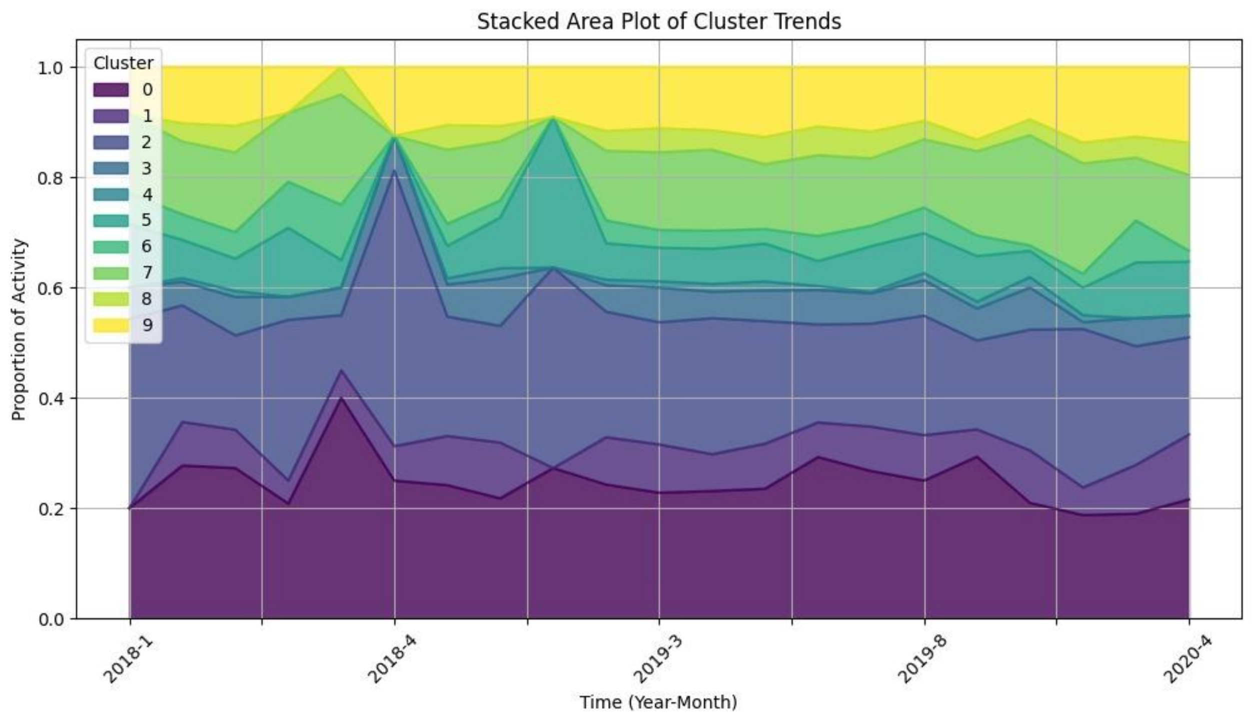
Purpose: The stacked area plot provides a proportional view of how cluster contributions evolve over time.

Outcome:

- Showed how discussions shifted between clusters over time.
- Highlighted rising and declining trends in specific clusters.

Insights:

- Clearly shows the rise and fall of specific clusters' influence in discussions.
- Cluster 0 consistently contributes a large proportion of discussions, while others fluctuate.



PCA-Based Scatter Plot with Temporal Coloring

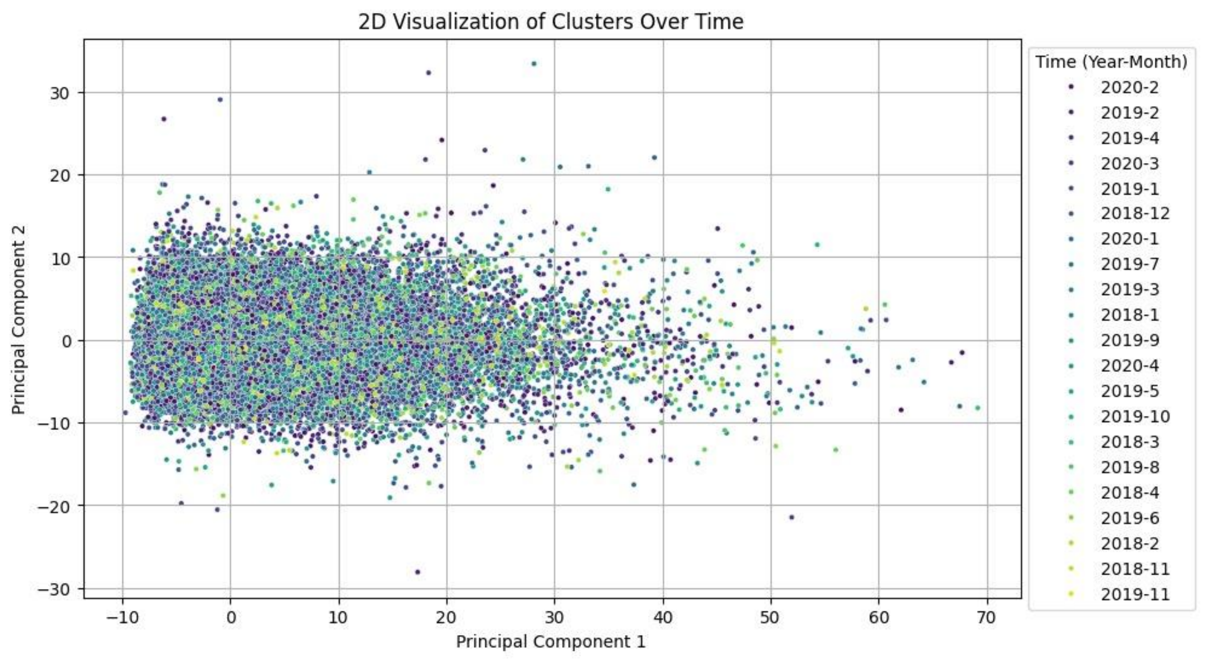
Purpose: Visualize temporal evolution in reduced dimensions to observe how cluster structures shift over time.

Outcome:

- Visualized temporal shifts in cluster structure.
- Enabled the identification of evolving patterns in discussions.

Key Highlights:

- Data points from different time periods show overlapping but distinct patterns, indicating gradual changes in discussion themes.
- Temporal color coding reveals the flow of clusters, enabling the identification of emerging trends and diminishing topics



Conclusions and Recommendations

1. **Successful Methods:**

- *GMM* clustering and *LDA* topic modeling effectively identified key themes and aligned with subreddit-specific topics.
- *Social network analysis* highlighted influential contributors driving discussions.

2. **Challenges:**

- *Temporal analysis* lacked precision in tracking detailed discussion shifts.
- *PCA* scatter plots showed overlap, limiting distinct trend identification.

3. **Recommendations:**

- Leverage clustering insights for targeted mental health interventions.
- Engage identified influencers to enhance community participation.
- Incorporate additional metadata for more refined temporal analysis.

Thank you!