Group Project: Multimodal Network and Textual Analysis of Reddit Engagement during Exam Seasons

Report: Focus on Concordia & McGill universities’ Reddits

Introduction and Problem Definition

Increased awareness of student mental health challenges has underscored the need to better understand how students cope during high-stress academic periods. Online forums, particularly Reddit, have become critical platforms where students seek advice, vent frustrations, and share resources. Our project focuses on Reddit activity during final exam months (April and December) in 2021 and 2022 from two Canadian university communities, *r/mcgill* and *r/concordia.*

We define the problem as follows: **What factors drive high engagement (measured by post scores and reply activity) among students during exam periods, and how can we predict which users will generate high-impact content?** This problem is important because understanding engagement drivers can support targeted content moderation, early detection of distress, and enhanced digital well-being strategies for academic institutions.

Previous research has highlighted the utility of social media platforms such as Reddit for understanding mental health discourse and engagement dynamics among university students. For example, Turcan and McKeown (2019) introduced the Dreaddit dataset to analyze stress signals in online posts, providing a foundation for stress detection using textual features alone. Subsequent studies extended this work by investigating sentiment variation and help-seeking behavior in academic subreddits (Oryngozha et al., 2023), as well as correlating online discussions with real-world campus mental health consultations (PMC, 2022). While valuable, these efforts primarily treated textual content in isolation or focused on binary classification of mental health indicators. In contrast, our project combines graph-based social network analysis with natural language processing and deep learning, applying Graph Neural Networks (GNNs) to predict user-level engagement – defined by reply volume and post score – based on both structural and semantic features. This multimodal integration offers a novel perspective by modeling Reddit as a dynamic ecosystem where community structure, influence, sentiment, and topic specialization jointly shape engagement outcomes. Our findings not only improve predictive accuracy but also reveal how support-seeking behaviors and discussion themes differ across institutional contexts. These insights are socially relevant for improving mental health interventions, academically relevant for understanding digital peer support systems, and practically useful for platform moderators and student service teams aiming to prioritize impactful conversations.

Analytical Approach

Data Collection and Preparation

We extracted 5,375 comments and 1,967 submissions from r/mcgill, and 5,699 comments and 2,355 submissions from r/concordia, targeting the high-stress exam months of April and December (2021–2022)[1]. After filtering by date and normalizing the text, we constructed directed reply graphs using NetworkX, where each node represents a unique user and edges indicate reply interactions; see Appendix A.

Network Construction and Centrality Analysis

Each subreddit produced a dense user interaction graph with no isolated nodes. We calculated multiple centrality measures, including degree, closeness, betweenness, eigenvector, and Katz centrality, to understand user visibility, information flow efficiency, and network influence; see Appendix B. These metrics enabled the identification of users who were not only active but also structurally important in shaping discourse and engagement patterns (e.g. *Thermidorien* at McGill and *PurKush* at Concordia).

Community Detection and Topic Modeling

To uncover latent group structures within the network, we applied the Louvain algorithm for community detection, revealing 20 distinct user communities in McGill and 19 in Concordia; see Appendix C. We further explored content themes using Latent Dirichlet Allocation (LDA), which surfaced dominant topics such as final exams, study strategies, mental health, housing, and COVID-related anxieties; demonstrating the thematic clustering of user conversations within these subgroups; see Appendix D.

[1] Link for the data source: https://the-eye.eu/redarcs/

Sentiment and Engagement Analysis

Sentiment analysis was conducted using VADER, a lexicon-based model that captures tone intensity in short, informal text like Reddit posts. Interestingly, a positive sentiment trend emerged even during exam periods, particularly around help-seeking language; see Appendix E. Furthermore, smaller, more niche communities (e.g. McGill C12 and Concordia C8; see Appendix F) often produced highly engaging posts, suggesting that relevance and context may be stronger predictors of engagement than network size alone.

Predictive Modeling with Graph Neural Networks

To model engagement outcomes, we developed a **binary classifier** predicting whether a user's post would achieve above-median engagement. The feature set combined graph-derived metrics (centralities), semantic attributes (topic distributions), sentiment scores, and user historical performance (i.e., post scores). We tested three graph neural network architectures:

- **GCN**: because it aggregates neighborhood features equally, effective for homogenous networks.
- **GAT**: as it applies attention mechanisms to weight neighbor importance dynamically.
- **GraphSAGE**: since it samples and learns from neighborhoods inductively, allowing better scalability.

We evaluated model performance using accuracy, comparing predicted and true labels for nodes held out during training. **GraphSAGE** achieved the highest accuracy on the **McGill** graph at **71.4%**, which had more nodes and denser interaction patterns (5,323 users and 21,095 interactions). On the **Concordia** graph, **GCN** performed best with **59.9%** accuracy, likely due to the graph's smaller and more uniform structure (2,634 users and 8,099 interactions). All models were trained using standard hyperparameters (e.g., learning rate = 0.01, epochs = 100). Our results confirm that engagement on Reddit can be predicted with reasonable accuracy using a combination of network structure and textual features.

Additionally, we used a **node-level transductive split** for training and evaluation, meaning that all nodes and edges were visible to the model during training, but only a subset of nodes had labels provided. The model could learn from the full network structure (including unlabeled nodes for message passing), but it was only evaluated on its ability to predict the labels of nodes it had not seen during training. This approach was chosen to preserve the graph's structural integrity and allow the model to fully leverage connectivity patterns and neighborhood features during learning which are important strengths of GNN architectures when applied to social media engagement networks; see Appendix G for full performance metrics details.

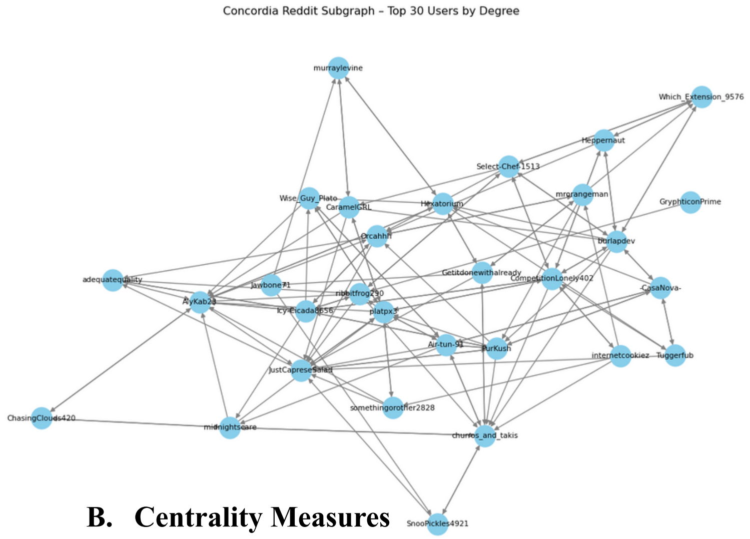
Expected Impact

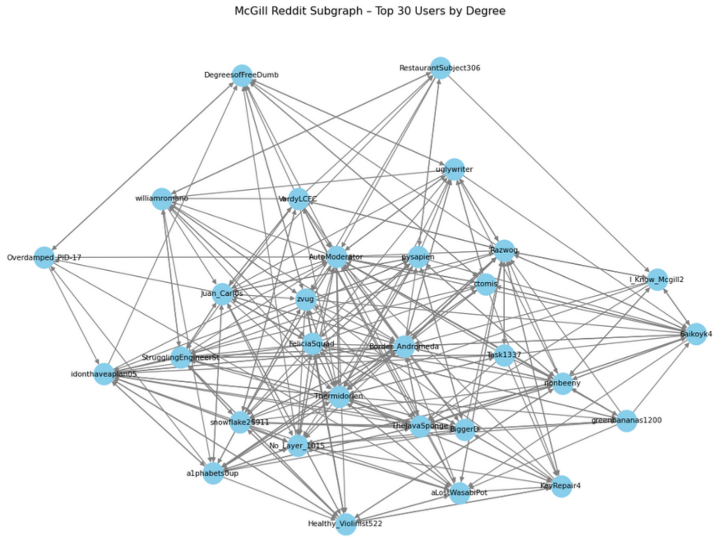
This project provides both explanatory and predictive insights into online student engagement during high-stress academic periods. By combining social network analysis with textual modeling, we identified important drivers of engagement: **users in structurally central positions** (high betweenness or eigenvector centrality), posts with **positive** and **help-seeking sentiment**, and discussions within **small**, **thematically** focused **communities**. These findings suggest that engagement is shaped by more than activity level: it's influenced by community placement, emotional tone, and topical relevance. On the predictive side, we developed a multimodal GNN classifier that achieved up to 71.4% accuracy in identifying high-impact users based on their network position, post history, sentiment, and content. These results demonstrate that Reddit engagement patterns during exam periods are both interpretable and predictable. This framework can inform future work on digital well-being, targeted support strategies for academic institutions to help their students; specifically in areas such as 'financial aid' topics for McGill where they can share their resources with students or 'tutoring' topics for Concordia where the university can offer out-of-class tutor sessions; see Appendix D for topic lists identified.

While our model performed well on the McGill dataset, results were lower for Concordia, suggesting room for improvement. Future work could include additional temporal features, test more advanced models like GAT with multi-head attention or integrate transformer-based text embeddings, and expand the dataset to other subreddits like r/ubc or r/uoft to improve generalizability.

Appendices

A. User Interaction Networks





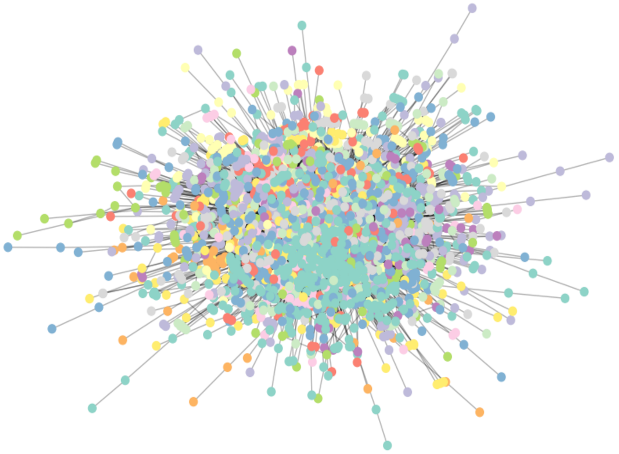
B. Centrality Measures

User	Degree	Closeness	Betwenness	Eigenvector	Katz	Page Rank
PurKush	0.097227	0.252425	0.091338	0.292360	0.027978	0.007893
CompetitionLonely402	0.065325	0.231656	0.049678	0.122849	0.024583	0.004482
JustCapreseSalad	0.057349	0.264271	0.058117	0.243568	0.029663	0.009610
Air-tun-91	0.049753	0.238794	0.048296	0.172353	0.024063	0.004788
burlapdev	0.039499	0.222563	0.028982	0.078937	0.022274	0.002007
Orcahhh	0.035321	0.243511	0.032683	0.143446	0.025407	0.005595
adequatequality	0.034182	0.226971	0.022256	0.098747	0.022876	0.002881
AlyKab23	0.030384	0.243111	0.023817	0.170118	0.025347	0.004561
Hexatorium	0.029624	0.227256	0.023830	0.089879	0.023585	0.003437
Heppernaut	0.026965	0.219029	0.014918	0.066306	0.022057	0.003597

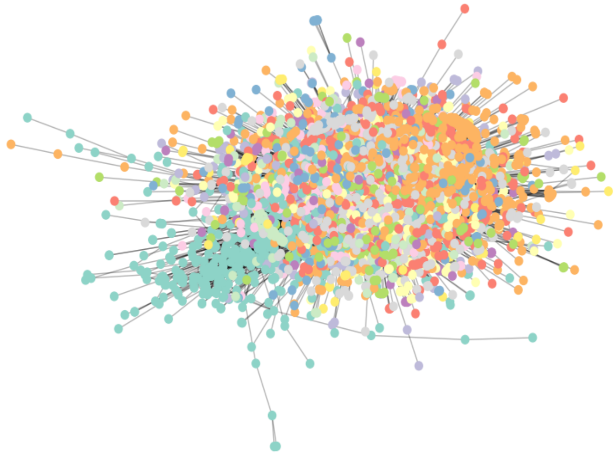
User	Degree	Closeness	Betweness	Eigenvector	Katz	Page Rank
Thermidorien	0.127208	0.310865	0.126049	0.126049	0.126049	0.126049
AutoModerator	0.111424	0.327943	0.083623	0.293152	0.043068	0.016299
snowflake25911	0.102217	0.291015	0.082139	0.204149	0.030639	0.012889
TheJavaSponge	0.060691	0.282348	0.038027	0.158340	0.022797	0.005012
Razwog	0.043593	0.279169	0.025008	0.148765	0.024079	0.004655
uglywriter	0.040962	0.275567	0.024412	0.119405	0.023358	0.004643
No_Layer_1015	0.039083	0.289991	0.019939	0.150079	0.023454	0.004044
BiggerD	0.037392	0.265339	0.018641	0.087028	0.018539	0.002362
zvug	0.036452	0.273505	0.020415	0.094027	0.018931	0.002396
nonbeeny	0.036452	0.276621	0.023678	0.112816	0.019911	0.002966

C. Community Detection

Louvain Communities in Concordia Reddit Network



Louvain Communities in McGill Reddit Network

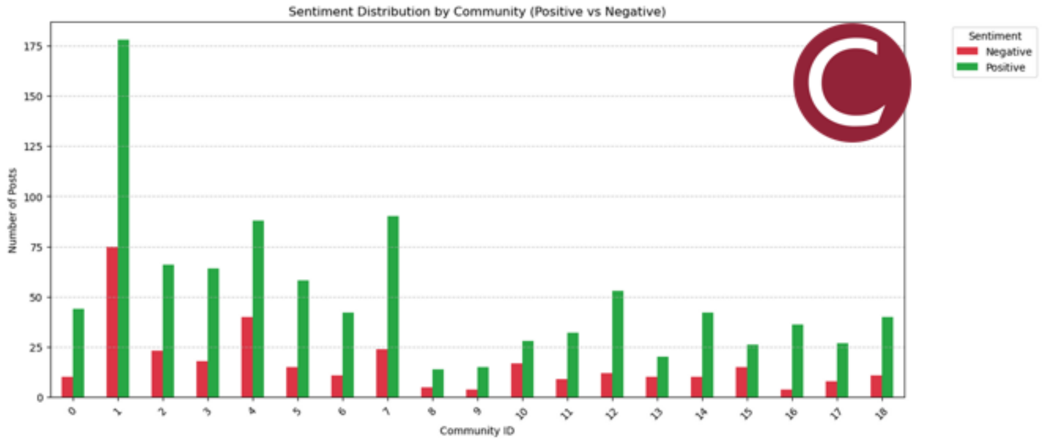


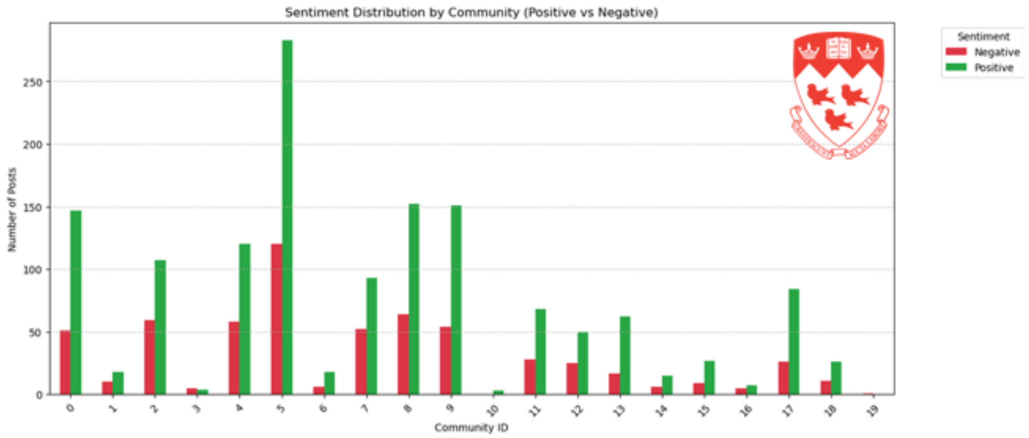
D. Topic Modeling

Topic 1	math	course	transfer	final	comp	class	econ	elective	jmsb	im
Topic 2	prof	moodle	rplace	mid	pm	free	covid	teacher	arent	final
Topic 3	gpa	grade	course	final	class	got	cgpa	semester	internal	retake
Topic 4	tutor	fee	insurance	tip	mark	survey	form	pay	private	asap
Topic 5	coen	le	grey	nun	guy	concordia	online	aware	final	note
Topic 6	poll	housing	view	association	application	campus	link	response	discord	dorm
Topic 7	course	summer	elective	engineering	program	student	im	concordia	coop	class
Topic 8	class	im	exam	know	student	course	like	concordia	semester	time
Topic 9	final	comm	exam	engr	grade	pas	phys	thought	midterm	acco
Topic 10	removed	permit	study	caq	student	disc	international	application	letter	university

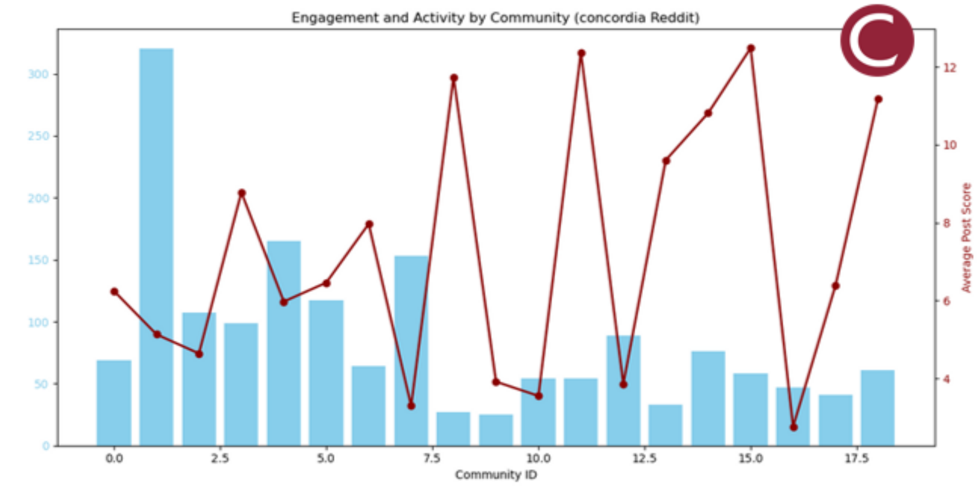
Topic 1	summer	course	im	mcgill	student	research	job	experience	minor	psyc
Topic 2	comp	removed	monday	transfer	news	course	allowed	student	good	class
Topic 3	removed	gym	biol	mcgill	final	hour	library	engineering	fieldhouse	room
Topic 4	removed	final	mcgill	mgcr	admission	econ	application	psyc	anat	exam
Topic 5	math	removed	course	science	major	art	final	chem	computer	faculty
Topic 6	final	removed	grade	thought	course	class	exam	semester	taken	thanks
Topic 7	concordia	removed	mcgill	protest	history	logo	rplace	positive	english	dm
Topic 8	removed	phgy	mcgill	account	aid	bursary	textbook	financial	money	lease
Topic 9	removed	cloudberry	mycourses	calculator	housing	membership	cat	aria	mcgill	dose
Topic 10	exam	im	final	like	dont	class	time	course	know	feel

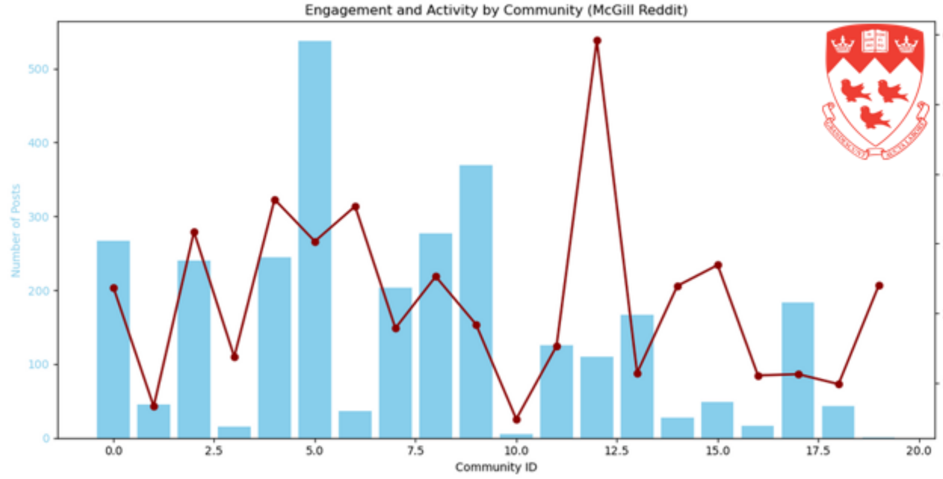
E. Sentiment Analysis





F. Niche Communities





G. Performance Metrics

Model	Concordia Accuracy	McGill Accuracy
GCN	0.5994	0.6885
GAT	0.5577	0.6758
GraphSAGE	0.5321	0.7140

References

Turcan, E., & McKeown, K. (2019). Dreaddit: A Reddit dataset for stress analysis in social media. arXiv. https://arxiv.org/abs/1911.00133

Oryngozha, N., Shamoi, P., & Igali, A. (2024). Detection and analysis of stress-related posts in Reddit academic communities. https://arxiv.org/html/2312.01050v2

Saha, K., Yousuf, A., Boyd, R. L., Pennebaker, J. W., & De Choudhury, M. (2022). Social media discussions predict mental health consultations on college campuses. Scientific Reports, 12, Article 123. https://doi.org/10.1038/s41598-021-03423-4