What are the key factors driving negative sentiment among students in U.S. college subreddits, and do these factors vary across different colleges?

Yella Diekmann
Emory College of Arts & Sciences
2487775
yella.diekmann@emory.edu

Doron Czarny
Emory College of Arts & Sciences
2500127
dczarny@emory.edu

Claudia Schmidt
Emory College of Arts & Sciences
2483561
claudia.schmidt@emory.edu

***Abstract*—This project explores the key factors driving negative sentiment among students in U.S. college subreddits through a guided theme classification framework. Posts with strong negative sentiment were filtered and classified into five predefined themes: academic pressure, financial concerns, anxiety about the future, health and wellness, and social/personal issues. Using TF-IDF vectorization and cosine similarity, posts were assigned to the most relevant theme, enabling a nuanced understanding of the challenges faced by students. Our findings highlight financial concerns as the most prevalent theme, with significant variations in negative sentiment drivers across colleges. This analysis provides actionable insights for universities to develop targeted interventions addressing specific student concerns. Limitations include data constraints, platform bias, and the lack of temporal and demographic insights.**

I. Introduction

A. Motivation

The mental well-being of college students is increasingly at risk due to factors like academic pressure, isolation, and financial struggles. While students often express these concerns on platforms like Reddit, traditional sentiment analysis methods fail to provide detailed insights into the specific causes of negativity. This gap makes it difficult for universities to offer targeted support to address these issues effectively.

Our project seeks to bridge this gap by utilizing advanced text preocessing and clustering techniques to identify themes that contribute most to negative sentiment in college subreddit posts. We aim to reveal patterns that can inform student support strategies. Understanding the drivers of negative sentiment will allow institutions to better address the unique challenges their students face, improving overall well-being.

B. Problem Statement

This project aims to address the lack of insight into specific causes driving negative sentiment in U.S. college subreddits by text processing and clustering techniques to identify the key drivers behind it.

C. Related Work

Previous research in sentiment analysis has explored different data mining and machine learning techniques. A 2020 literature review analyzed current methods used for for detecting depression through social media posts. They found that a BERT-based model produced the most accuracy but emphasized the need for further optimization and better data handling [1].

Yan and Liu analyzed sentiment trends in US college Reddit spaces and compared these trends across pre- and post-pandemic times periods. They used a RoBERTa model and Graph Attention Networks (GAT) for classification and then employed a linear mixed-effects model to examine how sentiment changed over time and how other school factors affected the data [2].

However, NLP processing is sometimes ineffective on social media data because it is noisy and fragmented. Instead, TWEETEVAL is a benchmark that includes seven different classification tasks for Twitter data. It provides a unified framework to evaluate models like RoBERTa on classification problems [3]. There is still more work to be done on the best way to train these models and how to leverage them more effectively.

Another research paper focused on sentiment analysis has explored diverse methodologies and applications across domains such as politics, health, and marketing. Rodríguez-Ibánez et al. (2023) provided a comprehensive review of sentiment analysis on social media platforms, emphasizing the growing interest in advanced techniques like Transformer-based models, while noting the continued relevance of traditional approaches such as lexicons and support vector machines (SVM) [4]. This work highlights the critical gaps in applying state-of-the-art models, including their computational expense and limited integration into real-world tools, despite their potential for analysis.

For instance, the study pointed out that methods such as RoBERTa and BERT outperform traditional techniques in accuracy for sentiment analysis tasks but require substantial computational resources, making them less accessible for smaller-scale studies. These insights are relevant to our project as we employ the cardiffnlp/twitter-roberta-base-sentiment-latest model, chosen for its balance between state-of-the-art performance and computational feasibility.

Additionally, the paper underscores the importance of exploring temporal sentiment dynamics and causality, especially in domains like education and health. This aligns closely with our goal to analyze not only the sentiment trends within college subreddit posts but also the underlying drivers, such as seasonal differences. By focusing on temporal and contextual patterns, our research builds on Rodríguez-Ibánez et al.'s call for deeper analysis of sentiment trends, particularly in underexplored domains like student well-being.

Understanding the drivers of academic stress and emotional well-being among students is critical in the context of broader societal disruptions, such as the COVID-19 pandemic. Clabaugh et al. (2020) conducted a study assessing the academic perceptions and emotional well-being of college students during the early months of the pandemic [5]. Their findings revealed high levels of stress and uncertainty among students, driven by factors such as unfamiliarity with online learning, distractions in home environments, and a lack of access to academic resources. These stressors disproportionately affected students with higher neuroticism and those with an external locus of control, suggesting that individual personality traits play a significant role in coping with academic disruptions.

Additionally, Clabaugh et al. highlighted critical disparities in emotional well-being based on gender and ethnicity. Female students and students of color reported significantly higher stress levels and perceptions of academic risk compared to their counterparts. These disparities underscore the need for targeted interventions to address the unique challenges faced by marginalized groups. Interestingly, the study also found that students' emotional well-being correlated more strongly with academic stressors and immediate disruptions than with their general perceptions of COVID-19, suggesting that proximal stressors have a more significant impact on student well-being.

The insights from this study provide valuable context for our exploration of negative sentiment in college subreddit discussions. Many themes identified by Clabaugh et al., such as academic uncertainty, home distractions, and disparities in stress based on gender and ethnicity, are likely reflected in the content of student subreddit posts. Our research aims to identify the frequency of specific themes of negativity, such as those described by Clabaugh et al., and determine how they vary across colleges. Furthermore, the study's emphasis on individual differences and structural inequities aligns with our goal of uncovering actionable insights that can guide institutions in addressing student well-being effectively.

II. Methodology

A. Initial Methodology Exploration

Our prior approach relied heavily on sentiment analysis, attention mechanisms, and unsupervised clustering to uncover negative themes.

1) *Attention Weights:* We previously used attention weights extracted from the RoBERTa model to identify important thematic words or phrases. For example, in the post "I am so stressed because of exams", attention weights emphasized terms such as "stressed" and "exams". These words were then clustered into themes using unsupervised clustering algorithms, with the aim of uncovering themes behind negative posts.
   To improve clustering accuracy, Part-of-Speech (POS) tagging was applied to filter for nouns, verbs, and adjectives, as these parts of speech were deemed most meaningful for capturing thematic content. The top 5% of terms based on attention weights were retained, focusing the analysis on the most important words.
2) *Unsupervised vs. supervised learning:* In the initial methodology, unsupervised clustering was employed to group words and phrases extracted from attention weights into thematic clusters. This process relied on K-Means clustering to organize terms such as "stress", "exams", or "tuition" into categories representing potential themes like "academic pressure" or "financial concerns". We replaced unsupervised clustering with a guided classification framework.

The reliance on attention mechanisms and unsupervised clustering in the initial methodology introduced significant limitations. While attention weights highlighted individual words or phrases, they often failed to capture the broader context necessary for accurate interpretation. For example, a term like "pressure" could appear in both academic and financial contexts, resulting in ambiguous and noisy clusters.

As shown in Figure 1, the clustering process produced one overly large cluster containing the majority of posts, alongside several smaller, sparsely populated clusters. This uneven distribution hindered the extraction of statistically meaningful insights.

The root cause of this issue was likely the inherent thematic overlap in different stress causes. Many posts addressed multiple concerns simultaneously, yet the unsupervised clustering algorithm restricted each post to a single cluster. This rigid assignment failed to reflect the multifaceted nature of student concerns, reducing the interpretability and reliability of the results.

B. Revised Methodology

After exploring various analytical approaches including unsupervised clustering and different classification methods, we revised our methodology that balances automated processing with guided theme identification. The final methodology is outlined below.

1) Data Collection and Preprocessing: Our analysis utilizes sentiment-labeled social media posts from 128 university subreddits. Each post is pre-labeled with three sentiment scores: negative, positive, and neutral. To focus on significant negative experiences, we filter posts to include only those where the negative sentiment score exceeds both positive and neutral scores, with a minimum threshold of 0.6. Additional preprocessing steps include, amongst other, the removal of deleted posts and enforcement of minimum post-length requirements.

Fig. 1. Prior Clustering Results

2) Theme Classification Framework: Rather than using unsupervised clustering, we employ a guided theme classification approach based on five predefined categories of student concerns. These categories are developed through literature review and are designed to be distinctive while representing common student stress factors:

1) Academic Pressure: Course workload, grades, learning difficulties
2) Anxiety about Future: Career prospects, job interviews, post-graduation concerns
3) Social/Personal: Interpersonal relationships, belonging, community integration
4) Financial: Costs, fees, housing, living expenses
5) Health and Wellness: Physical health, mental health, stress management

For each theme, we develop a comprehensive set of 20-25 seed words that characterize the category. These seed words are generated through prompting ChatGPT and validated through manual review to ensure distinctiveness and relevance to each theme.

3) Text Processing and Theme Assignment: Our text processing pipeline utilizes the spaCy natural language processing library to extract meaningful information from posts. The process begins with basic preprocessing steps including tokenization, special character removal, and lemmatization. The main part of our phrase extraction process focuses on three key elements:

- Verb phrases with their objects (capturing actions and their targets)
- Noun phrases with sentiment-bearing modifiers (identifying key concerns and their characterizations)
- Pattern-matched expressions that capture common ways students express concerns

The theme assignment process utilizes TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert processed text into numerical representations. Our vectorization parameters are tuned to balance between common and rare terms (minimum document frequency of 2, maximum of 0.9) and to capture both individual words and phrases up to three words long. We calculate cosine similarity between each post's vector and our theme vectors, assigning posts to the theme with the highest similarity score, provided it exceeds our minimum threshold of 0.1.

4) Quality Control: To ensure robust analysis, we implement several quality control measures. At the data level, we perform strict filtering following certain criteria including minimum post length. For theme assignment quality, we conduct manual reviews of random assignments. Additionally, we monitor inter-theme similarity and analyze similarity score distributions to validate the meaningfulness of our classifications.

5) Analysis Framework: Our analysis framework consists of two main components:

1) Theme Distribution Analysis: We analyze the distribution of posts across themes and calculate similarity scores to assess the strength of theme assignments and verify their quality.
2) Academic Rigor Correlation: We examine the relationship between university rankings and the prevalence

of certain stress factors in higher-ranked versus lower-ranked schools.

3) Institution Type Correlation: We investigate a potential correlation between the occurrence of stress factors in private versus public institutions.

III. Results

1) Data Processing Results: Our initial analysis examined 147,580 posts from 128 university subreddits. Through our filtering process focusing on posts with strong negative sentiment (negative score > 0.6 and exceeding both positive and neutral scores), and after applying our quality control measures, 638 posts (0.43% of total) were successfully classified into our five themes. This strict filtering ensures high-quality theme assignments, though at the cost of reduced data volume. The classified posts came from 76 different universities, with varying levels of representation. The most represented institutions included UC Santa Cruz (29 posts), UIUC (27 posts), and Stony Brook University (25 posts). Major universities such as Georgia Tech, Cornell, and Ohio State each contributed 23 posts to the final dataset. Notably, some prestigious institutions like Harvard (2 posts), Stanford (2 posts), and Northwestern (1 post) had fewer posts meeting our filtering criteria.

2) Theme Distribution: The 638 classified posts are distributed across the five identified themes, as shown in Figure 2. Posts are generally well distributed across the five categories, with "Financial" having the highest amount of posts assigned at 25.4% and "Social/Personal" having the lower amount of posts with 13.5%.

3) Detailed Theme Analysis: In this next section, we will go into detail on each category for further analysis.

As previously mentioned, financial concerns is the most common theme (25.39%), centered primarily on tuition costs, housing expenses, and financial aid issues. An example post showcases this apparent financial pressures students face:

> "Just checked and I messed up. The $18k was tuition AND housing..."

Academic Pressure (23.20%) emerged as the second most common theme. These were posts primarily concerned with course workload, grade-related stress, and challenges with specific academic requirements, as exemplified by:

> "I retook 103, just felt I didn't really learn anything from the class the first time and ended up failing out the first time. Retook it again and still got [a low grade]..."

Next, anxiety about Future (18.97%) posts frequently discussed themes such as job market concerns, internship competition, and career preparation. An example post representing this cause of student stress is:

> "People who only know Python wouldn't be able to pass any real job interview. They wouldn't know what a virtual function, or a SIMD instruction, or a virtual processor [is]..."

Health & Wellness posts (18.97%) revealed a strong focus on mental health issues, particularly stress management and sleep difficulties. A representative post exemplifies these concerns:

> "There is no dignity, only danger, and health educators at Vaden Health Center..."

Social/Personal concerns (13.48%) was the least populated cluster and contained student concerns related to their personal life such as feelings of isolation and difficulties with peer relationships:

> "Nobody should be getting mad. Not like it concerns them on a personal level or something..."

A. Quality Control

Fig. 3. Distribution of Similarity Scores by Theme

The similarity scores, calculated using cosine similarity, represent the alignment between posts and their assigned themes, provide valuable insights into the robustness of the classification process. Moreover, they are a validation of the assignment correctness and quality. As shown in Figure 2, the scores generally align well with the predefined thresholds and reflect the distinctiveness of the identified themes. This is indirectly a result of already filtering out any posts that were assigned to a cluster, however, did not meet the minimum cosine similarity threshold.

While the similarity scores appear numerically low (ranging from 0.121 to 0.130), several factors support the validity of our classification approach. First, manual verification of the clustered posts demonstrates strong thematic coherence, with posts consistently aligning with their assigned themes in meaningful ways, as evidenced by the examples discussed in the previous section. Second, the relative stability of scores across themes, despite their low absolute values, suggests systematic rather than random classification.

The relatively low similarity scores can be attributed to the inherent complexity of natural language, particularly in informal student posts where themes may be expressed through diverse vocabularies and contextual nuances that challenge traditional similarity metrics. However, the consistency between our quantitative clustering and qualitative validation supports the reliability of these classifications for practical analysis. Future work could potentially enhance these scores through more sophisticated text processing techniques, advanced embedding

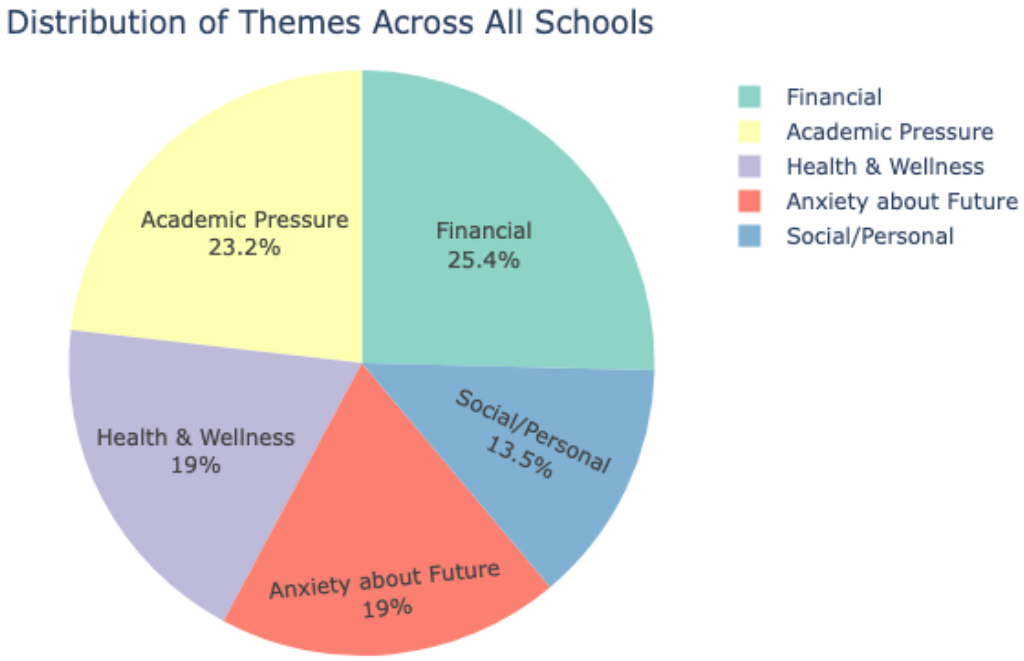

Fig. 2. Distribution of themes across the classified posts.

models, or refined similarity metrics, but the current framework provides a solid foundation for understanding student concerns.

Across themes, "Health Wellness" achieves the highest mean score. The narrower ranges and lower standard deviations observed in "Anxiety about Future" and "Academic Pressure" suggest a high degree of thematic consistency, reflecting clear boundaries for these categories.

In contrast, "Social/Personal" exhibited the highest variability in scores, with a standard deviation of 0.033. This is expected, as this theme captures a broader spectrum of subjective concerns, such as interpersonal relationships and feelings of isolation which might also not be fully covered by the seed words used in our clustering technique. Similarly, "Financial," despite its relatively higher mean score, showed a moderate range in similarity scores, reflecting the diversity of financial stressors discussed by students, from tuition costs to housing expenses.

Overall, the similarity score distributions confirm the effectiveness of the guided theme classification framework in capturing nuanced patterns of student concerns. While some variability is inherent to themes with more subjective or diverse content, the high mean similarity scores and narrow standard deviations for most categories underscore the robustness of the methodology. These results align with the quality control measures implemented during theme assignment, supporting the validity of the classifications.

B. Case Studies

In the following sections, we performed a series of case studies investigating different aspects of our results. Sources for school rankings and other data include U.S. News & World Report, QS World University Rankings, and Times Higher Education. These platforms provide comprehensive and up-to-date evaluations of university performance, making them ideal for benchmarking against themes such as academic pressure and student concerns.

1) Top 30 vs Other Schools: This case study examines the differences in theme distributions between schools ranked in the top 30 and those outside this category, as shown in Figure 4. In top-ranked schools, financial concerns and health and wellness issues are most prevalent, appearing in 30.9% and 23.5% of posts, respectively. Social and personal concerns are the least mentioned at 11.8%. This trend might demonstrate that the competitive environments of top-ranked institutions may cause increases in financial stress and concerns about maintaining physical and mental well-being.

In contrast, schools outside the top 30 exhibit a more balanced distribution across themes, with academic, anxiety about the future, financial, and health and wellness concerns each appearing in between 19.3% and 27.1% of posts. The more even distribution of themes may reflect the more diverse number of challenges faced by students in less competitive environments where the focus may not be as singularly tied to academic performance or prestige. For these schools, anxiety about the future emerges as the most prevalent theme at 27.1% while social and personal concerns remain the least discussed

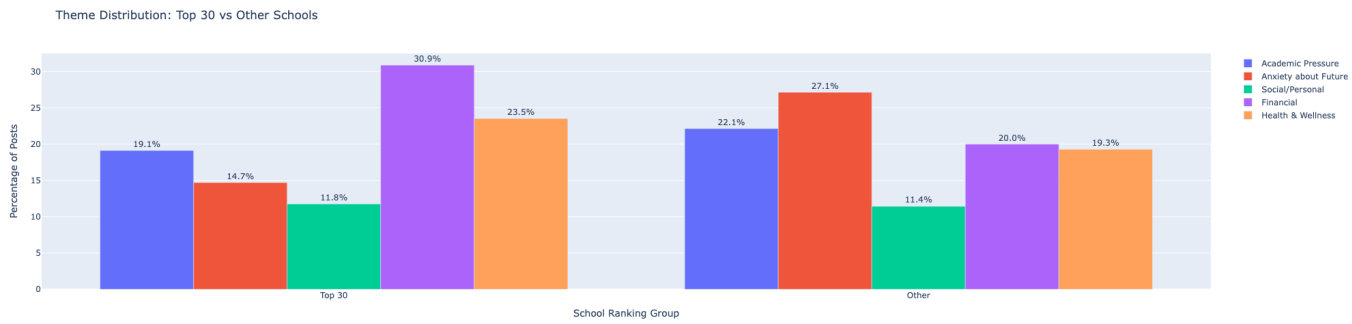

Fig. 4. Theme Distribution of Top 30 vs Other Schools

at 11.4%.

These differences reflect the differing priorities and challenges experienced by students at various levels of institutional ranking.

2) Private vs Public Institutions: The second case study explores the differences in theme distributions between private and public institutions, as depicted in Figure 5. In private institutions, themes were distributed relatively evenly with social and personal concerns being the least mentioned at 9.7%. All other themes, including academic, financial, health and wellness, and anxiety about the future, appeared in 21% to 22.9% of posts. This even distribution suggests that private institutions may foster a more balanced set of stressors, potentially reflecting their emphasis on personalized support and smaller student populations.

In contrast, public institutions revealed a more imbalanced distribution of theme prevalence. Financial concerns were significantly higher than all other themes, appearing in 33.9% of posts. This highlights the economic pressures faced by students in public institutions which could be a result of factors such as higher student-to-faculty ratios, resource constraints, and reliance on state funding. As such, this may lead to increased tuition or fewer financial aid opportunities.

The differences demonstrate that stress factors are sometimes dependent on institutional differences

IV. Limitations

A. Data Constraints

Out of 147,580 initial posts, only 638 (0.43%) met the filtering criteria, highlighting the strict selection process. However, this was necessary to ensure the quality of the final clusters. In future work, it might be valuable to explore the use of additional datasets in hopes that more posts qualify for the selection criteria or, alternatively, to examine alternative data processing techniques that could improve the quality of more posts, allowing them to be included in the final analysis.

Additionally, there was a significant imbalance in university representation. Some institutions were disproportionately represented. For example, UC Santa Cruz had 29 posts, while Harvard had only 2 posts. Similarly, this issue could be addressed by including more posts overall, ensuring that even if the selection percentage remains the same, a larger number of posts are included. At the current rate, it was not feasible to ensure an even representation of all schools, as some schools simply had too few posts. Limiting the contributions of other schools to match these numbers would have left too little data. Alternatively, schools that do not have enough posts after filtering could be eliminated entirely; however, this would further reduce the size of the final dataset.

B. Methodology Limitations

The five predefined themes may not have fully captured all student concerns, leaving certain issues unaddressed. This is relevant to: 1) the overall availability of themes—potentially, there are more than five themes, which may also vary by school—and 2) the description of themes. As evident in the theme similarity analysis, some themes, such as the personal cluster, exhibited lower similarity. This could be because such a cluster might encompass broader topics, and the keywords generated by ChatGPT may not fully represent these, thus affecting cluster quality. Although our attempts were not successful, a future step would be to explore unsupervised clustering methods that could result in insightful clusters based on themes specific to each school.

Additionally, the TF-IDF similarity threshold of 0.1 may have excluded valid but nuanced posts, limiting the scope of the analysis. Furthermore, while posts can belong to multiple categories, they are currently sorted into only one, which could oversimplify their classification.

C. Platform Bias

There was self-selection bias in determining who posts negative content, which may affect the overall representativeness of the analysis. Specifically, there is bias both in who chooses to post and in the reasons they post. It is possible that a small group of individuals posts disproportionately compared to the overall student body, thus not accurately representing the causes of student stress at their schools.

Moreover, even if a diverse range of students contributes, they may post selectively about certain topics while omitting other stressors they experience, leading to an incomplete picture of student concerns.

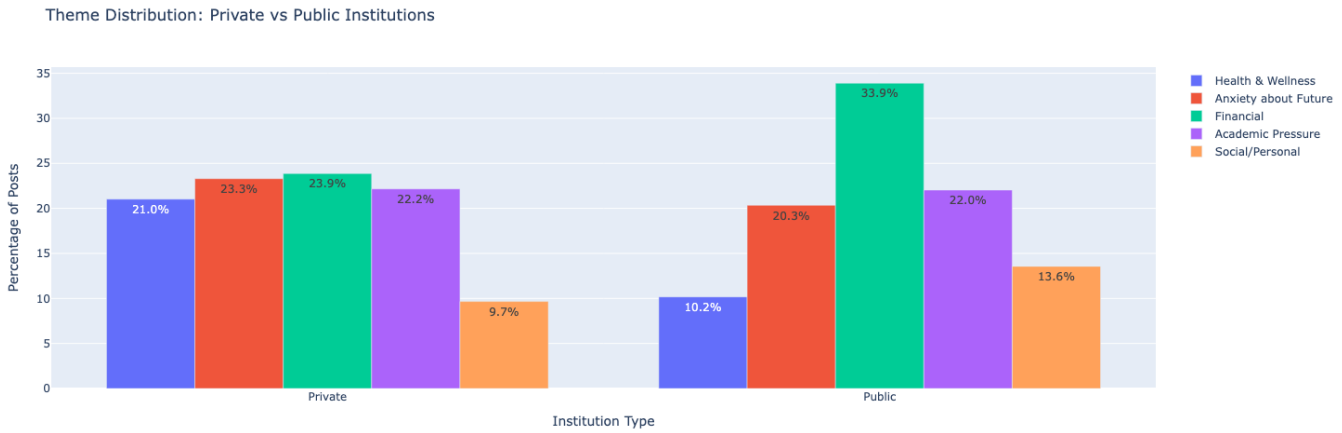

Fig. 5. Theme Distribution of Private vs Public Institutions

In future work, analyzing post metadata, such as user posting frequency, demographics, and other factors, could provide better insight into the quality of the data and the meaningfulness of the results.

D. Analysis Gaps

The analysis lacks a temporal component, as it does not account for variations across the academic calendar. The data is limited to posts from 2019 to avoid bias introduced by the COVID-19 pandemic. However, there is no consideration of when during the year the posts were made (e.g., students posting during finals week may exhibit significantly more stress indicators than those posting during summer break).

Although the dataset includes posts from throughout the year, applying a weighting function to place greater emphasis on posts made during the school year could improve the quality and relevance of the results presented here.

E. Relative Frequency Analysis

The higher frequency of negative posts in certain themes may reflect a greater overall discussion volume rather than a proportionally higher negative sentiment. For example, some schools may generally favor discussions on their Reddit forums about financial topics compared to academics. As a result, a higher frequency of financial-related posts appearing after filtering for negative sentiment, and the subsequent analysis of financial stress as a key factor, might be misleading, as the school may simply discuss this topic more frequently overall.

The analysis lacks a baseline comparison between the distribution of negative sentiment and the overall topic distribution across each university's subreddit, as well as an analysis of the frequency of negative posts proportional to overall occurrences.

This is an area that could be explored in future work.

V. Conclusion

Rising stress levels among college students underscore the need for qualitative counseling services tailored to their specific needs. By understanding the root causes of stress, institutions can adopt targeted strategies to address these challenges effectively. Our approach aims to help colleges identify the primary stressors affecting their students, allowing for more customized interventions. Our findings and analysis support this thesis, revealing that different institutions often experience distinct themes of stress more frequently, emphasizing the importance of individualized solutions.

References

[1] D. William and D. Suhartono, "Text-based depression detection on social media posts: A systematic literature review," Feb. 2021, accessed: 2024-10-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921000491

[2] T. Yan and F. Liu, "Has sentiment returned to the pre-pandemic level? a sentiment analysis using u.s. college subreddit data from 2019 to 2022," Mar. 2024, accessed: 2024-10-8. [Online]. Available: https://arxiv.org/pdf/2309.08845

[3] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval:unified benchmark and comparative evaluation for tweet classification," Oct. 2020, accessed: 2024-10-8. [Online]. Available: https://arxiv.org/pdf/2010.12421

[4] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Aug. 2023.

[5] A. Clabaugh, J. Duque, and L. Fields, "Academic stress and emotional well-being in united states college students following onset of the covid-19 pandemic," Mar. 2021.