Tracking Student Mental Health through Reddit Sentiment Analysis

Aditya Mahesha Ballaki (ab3237@cornell.edu)
Maria DeCaro (msd249@cornell.edu)

December 19, 2024

Abstract

As mental health awareness continues to grow, universities are focusing on improving support for student well-being. In this project, we created a dataset from Reddit posts across 16 universities in the U.S., representing regions from the Northeast, South, Midwest, and California. By analyzing sentiment fluctuations during the fall semester, particularly around high-stress times like midterms and finals, we uncover patterns that help identify periods of emotional distress. The findings show regional differences, with California universities experiencing more frequent negative sentiment compared to schools in the South, Midwest, and Northeast. Through sentiment analysis and machine learning, we present a framework that universities can use to better allocate mental health resources and provide proactive support for students during critical times. This work highlights how universities can use these insights to improve mental health care and better support their students.

1 Introduction

Mental health awareness is on the rise, prompting universities to enhance their support for student wellbeing. This project seeks to gain deeper insights into the wellbeing of students across different regions in the United States by analyzing sentiment trends throughout the academic year.

We selected 16 universities from the Northeast, South, Midwest, and California to capture a diverse regional perspective. Using the Reddit API, we collected relevant discussions and sentiments expressed by students, filtered the data to remove irrelevant information, and converted it into a structured CSV format for analysis.

By examining these sentiment trends, universities can better allocate mental health resources during periods when student wellbeing may be low. Additionally, our machine learning model can predict future dips in sentiment, allowing institutions to proactively prepare and support their students in subsequent semesters.

This project provides a framework for universities to monitor and respond to student mental health needs effectively, fostering a healthier and more supportive academic environment.

2 Related Work

Sentiment Analysis of Social Media Content for Mental Health Monitoring

- Benrouba and Rachid (2023) examined sentiment analysis techniques to assess emotional states expressed in social media content.
- Highlighted the potential for early detection of mental health issues through textual data analysis.

Machine Learning-Based Prediction of Mental Well-Being Using Survey Data

- Lim et al. (2023) applied machine learning algorithms to predict negative psychological well-being states.
- Leveraged data from a large, multi-site cross-sectional survey of university students to demonstrate the effectiveness of predictive models in mental health assessment.

Predicting Mental Health Using Social Media Sentiment Analysis

- Seo et al. (2023) investigated the prediction of mental health status through sentiment analysis, text, and opinion mining of social media posts.
- Focused on South Korean adolescents, showcasing the relevance of sentiment analysis in mental health prediction.

3 Methods

3.1 Building the Dataset

We built this dataset by collecting data from 16 university-focused subreddits, including Harvard, MIT, Columbia, UChicago, Berkeley, Princeton, UCLA, Cornell, Stanford, UCSD, UIUC, UofM, Duke, Georgia Tech, Rice, and Vanderbilt. We made requests to the Reddit API to retrieve data from the response body in JSON format, filtered out unnecessary information to focus on essential fields like the title, body, and creation date, specifically targeting data from the beginning of August to December (capturing the time of the fall term). The processed data from all subreddits was converted into CSV files, labeled with their respective university names, and subsequently combined into a single pandas DataFrame for analysis.

3.2 Preprocessing the Dataset

For preprocessing, we focused on transforming raw text data into a structured and standardized format suitable for analysis. We started by using NLTK tools to remove stopwords—frequently occurring but non-informative words like "and" or "the"—to reduce noise in the dataset. Stemming was then applied using the Porter Stemmer, which reduces words to their root forms, such as converting "running" to "run," ensuring that variations of the same word are treated consistently throughout the analysis.

We also implemented a custom clean and process text function to refine the text further. This involved converting all text to lowercase for uniformity, normalizing spaces to remove unnecessary gaps, and eliminating URLs and special characters, leaving only alphanumeric content. Stopwords were removed, and stemming was applied to tokenized words to enhance data consistency. Together, these preprocessing steps produced standardized and concise text data, ready for downstream tasks such as sentiment analysis or machine learning-based classification.

3.3 Clustering

Sentiment analysis was conducted using the TextBlob library to calculate the polarity and subjectivity of user-generated posts. Posts falling within the lowest 15% of sentiment polarity values were flagged for deeper exploration. This focused analysis allowed us to isolate discussions that were predominantly negative.

We utilized TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert textual data into numerical feature sets. These features were then input into a K-Means clustering algorithm, which grouped posts based on shared characteristics. Two separate clustering exercises were performed:

1. On the entire dataset.

2. Exclusively on the subset of low-sentiment posts.

3.4 Topic Modeling with Temporal and Regional Analysis

To uncover underlying themes within the data, Latent Dirichlet Allocation (LDA) was applied to the low-sentiment posts using two methods. The first method identified the lowest 15% of posts based on sentiment. The second method focused on isolating the lowest sentiment weeks, where posts with sentiment scores below 0.05 were selected for each school. Posts from these low-sentiment weeks were then analyzed. These approaches revealed four key topics, each characterized by distinct sets of recurring terms. To visualize the results, t-SNE (t-distributed Stochastic Neighbor Embedding) was used to create a two-dimensional scatter plot of topic clusters.

Temporal patterns in sentiment were further explored by aggregating weekly average sentiment polarity scores for posts grouped by region during the lowest sentiment weeks. Line graphs highlighted fluctuations over time, with notable sentiment drops during midterms and finals.

3.5 Sentiment Analysis

The sentiment analysis was conducted using TextBlob, a natural language processing library that calculates sentiment scores based on text inputs. For this analysis, sentiment scores were derived from the combined titles and content of Reddit posts related to various schools during the fall semester. Data was collected and aggregated across the academic period, spanning from early August to mid-December. A time series approach was used to map average sentiment trends, allowing for a detailed exploration of temporal and geographical variations. Sentiment scores were averaged over specific time intervals to identify key patterns and fluctuations. Additionally, a school-specific analysis was performed to assess differences in sentiment between individual institutions.

3.6 Sentiment Prediction

Sentiment labels were initially populated using TextBlob and manually corrected to address discrepancies. This resulted in a high-quality dataset for model training and evaluation.

Deep Learning Approach

A deep learning model was implemented using PyTorch. The text data was tokenized with the BERT tokenizer, and a bidirectional LSTM model was built with an embedding layer, LSTM, and a fully connected classification layer. The model was trained over 10 epochs using cross-entropy loss and the Adam optimizer, achieving competitive accuracy on the test set.

Random Forest Classifier

The text data was vectorized using TF-IDF, and a Random Forest classifier with 100 estimators was trained. Model performance was evaluated using accuracy and F1-score, demonstrating its effectiveness on the dataset.

Logistic Regression

A Logistic Regression model was also trained on TF-IDF vectorized data. It was evaluated using accuracy and a classification report, providing a benchmark for sentiment prediction.

4 Results

4.1 Clustering Results

Full Dataset Clustering

When clustering the entire dataset, Cluster 3 contained the majority of posts (55%), indicating its broad thematic scope. Upon manual review:

- **Cluster 0:** Posts related to class-based queries, such as registration and course information.
- **Cluster 1:** No clear patterns identified.
- **Cluster 2:** Posts about campus living.
- **Cluster 3:** A diverse mix of topics without a specific thematic focus.

Low-Sentiment Posts Clustering

For posts with low sentiment polarity, K-Means clustering revealed distinct patterns. Similar to the first clustering, Cluster 3 contained a majority of the posts (about 75%):

- **Cluster 0:** Questions about classes, including difficulty and exam-related inquiries.
- **Cluster 2:** Academic queries, similar to Cluster 0.
- **Cluster 3:** A mix of topics, again showing no clear pattern.

These clusters highlight that admissions concerns and academic challenges are predominant themes among low-sentiment posts. Visualizations reinforced these findings. Bar charts revealed the dominance of Cluster 3 in both datasets, while manual review highlighted academic and logistical topics (Clusters 0 and 2) as key areas of concern.

4.2 Topic Modeling with Temporal and Regional Analysis Results

Method 1:

This method produced the following topics:

- **Topic 0:** ”anyon”, ”els”, ”anyon”, ”know”, ”grad”, ”student”, ”hi”, ”everyon”, ”financi”, ”aid”, ”wonder”, ”anyon”, ”greatli”, ”appreci”, ”let”, ”know”, ”dont”, ”know”, ”feel”, ”like”
- **Topic 1:** ”subject”, ”remov”, ”admiss”, ”site”, ”undergrad”, ”grad”, ”admiss”, ”offic”, ”chanc”, ”admiss”, ”graduat”, ”admiss”, ”help”, ”resourc”, ”princeton”, ”resourc”, ”princeton”, ”princeton”, ”undergradu”, ”princeton”, ”undergradu”, ”admiss”
- **Topic 2:** ”anyon”, ”know”, ”feel”, ”like”, ”dont”, ”know”, ”anyon”, ”taken”, ”thank”, ”advanc”, ”hi”, ”im”, ”im”, ”look”, ”im”, ”wonder”, ”ive”, ”heard”, ”grad”, ”student”
- **Topic 3:** ”anyon”, ”know”, ”anyon”, ”els”, ”dont”, ”know”, ”incom”, ”freshman”, ”financi”, ”aid”, ”intern”, ”student”, ”feel”, ”like”, ”hi”, ”everyon”, ”let”, ”know”, ”im”, ”tri”
- **Topic 4:** ”grad”, ”student”, ”anyon”, ”know”, ”anyon”, ”els”, ”dont”, ”want”, ”dont”, ”know”, ”hi”, ”im”, ”footbal”, ”game”, ”im”, ”wonder”, ”let”, ”know”, ”placement”, ”test”

Method 2:

When analyzing unigrams, five topics emerged, along with their corresponding average sentiment polarities:

- **Topic 0:** ”im”, ”look”, ”year”, ”appli”, ”class”, ”major”, ”unit”, ”student”, ”spring”, ”quarter” (Polarity: 0.0514)
- **Topic 1:** ”anyon”, ”campu”, ”know”, ”pleas”, ”park”, ”want”, ”look”, ”im”, ”student”, ”today” (Polarity: 0.0630)

- **Topic 2:** ”anyon”, ”class”, ”need”, ”student”, ”cours”, ”ticket”, ”thank”, ”stat”, ”physic”, ”lectur” (Polarity: 0.0131)

- **Topic 3:** ”final”, ”class”, ”im”, ”anyon”, ”math”, ”grade”, ”like”, ”know”, ”major”, ”help” (Polarity: 0.0266)

- **Topic 4:** ”like”, ”im”, ”peopl”, ”make”, ”dont”, ”know”, ”time”, ”feel”, ”year”, ”berkeley” (Polarity: 0.0389)

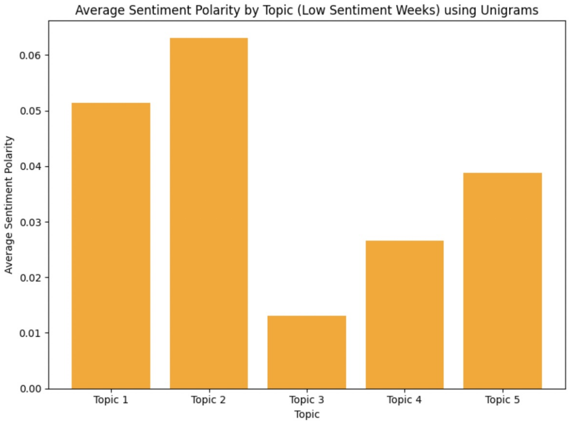

Figure 1: Average Sentiment Polarity by Topic using Unigrams

When analyzing bigrams and trigrams, the following topics and their average sentiment polarities were identified:

- **Topic 0:** ”anyon”, ”els”, ”feel”, ”like”, ”anyon”, ”taken”, ”wonder”, ”anyon”, ”dine”, ”hall”, ”im”, ”look”, ”look”, ”someon”, ”im”, ”wonder”, ”extra”, ”credit”, ”minut”, ”walk” (Polarity: 0.0265)

- **Topic 1:** ”big”, ”game”, ”meal”, ”swipe”, ”late”, ”drop”, ”hi”, ”guy”, ”studi”, ”abroad”, ”data”, ”scienc”, ”final”, ”week”, ”undi”, ”run”, ”financi”, ”aid”, ”im”, ”look” (Polarity: 0.0519)

- **Topic 2:** ”let”, ”know”, ”pleas”, ”let”, ”pleas”, ”let”, ”know”, ”dont”, ”want”, ”winter”, ”quarter”, ”hi”, ”everyon”, ”look”, ”like”, ”hi”, ”im”, ”rooter”, ”bu”, ”anyon”, ”taken” (Polarity: 0.0480)

- **Topic 3:** ”high”, ”school”, ”hey”, ”guy”, ”winter”, ”break”, ”thank”, ”advanc”, ”feel”, ”free”, ”dont”, ”know”, ”im”, ”hope”, ”im”, ”appli”, ”hi”, ”anyon”, ”studi”, ”room” (Polarity: 0.0605)

- **Topic 4:** ”anyon”, ”know”, ”dont”, ”know”, ”good”, ”luck”, ”doubl”, ”major”, ”intern”, ”student”, ”im”, ”really”, ”final”, ”exam”, ”final”, ”anyon”, ”greatli”, ”appreci”, ”im”, ”sure” (Polarity: 0.0525)

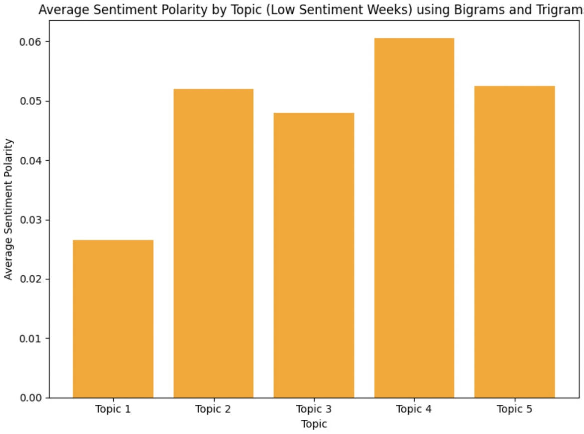

Figure 2: Average Sentiment Polarity by Topic using Bigrams and Trigrams

Analysis of Method 1 and Method 2

Method 1 focused on the lowest 15% of posts based on sentiment. By targeting this subset, the analysis naturally captured more specific, practical concerns of students, such as graduate studies, financial aid, and placement tests. These topics often involve clear, straightforward queries or issues that lead to repetitive terms, resulting in topics with highly specific terms. This method likely identified more targeted student concerns because it filtered posts that may have been directly asking for help or guidance on certain academic or logistical matters.

In contrast, Method 2 took a different approach by analyzing posts from weeks with sentiment scores below 0.05. This method produced more general topics that reflected the overall student experience. The use of unigrams (single words) in this method created more broad topics focused on common experiences, like course registration, class needs, and social interactions. These topics are more abstract and less detailed because unigrams represent isolated words without context, which limits the specificity of the topics identified.

4.3 Sentiment Analysis

The analysis revealed intriguing trends in sentiment across various schools during the fall semester. A time series analysis showed that average sentiment peaked during the first two weeks of the semester, a period often characterized by optimism and excitement as students adjusted to their routines. Sentiment declined significantly during midterm weeks, reaching a low point likely due to academic stress. Following midterms, sentiment rebounded during fall break and the subsequent two weeks, reflecting a period of relief and recovery. However, sentiment reached its lowest point toward the end of the semester during finals season, highlighting a strong correlation between negative sentiment and exam-related stress.

A school-specific analysis indicated that the University of Chicago exhibited the highest levels of positive sentiment throughout the semester, identifying it as the ”happiest” school in the dataset. Conversely, UC San Diego consistently showed the lowest levels of positive sentiment, with the disparity between these two schools widening significantly toward the end of the semester. Geographically, schools in California displayed the most pronounced sentiment fluctuations and experienced an overall negative sentiment throughout the year, particularly during midterms. In contrast, schools in the South, Midwest, and Northeast exhibited more stable sentiment levels, with comparable averages that were generally less negative. Although California experienced the most frequent bad weeks, it was interesting to find that the Midwest and South, while having fewer bad weeks, showed significantly lower average sentiment polarity during those weeks compared to California and the Northeast.

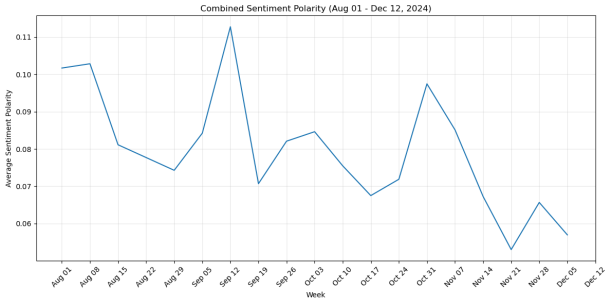

Figure 3: Sentiment through the weeks

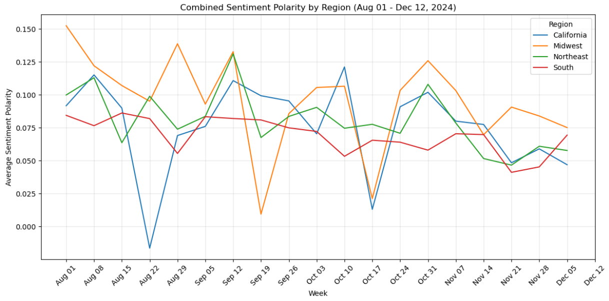

Figure 4: Sentiment through the weeks by region

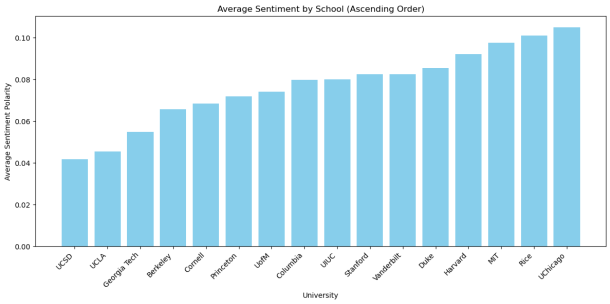

Figure 5: Average Sentiment by school

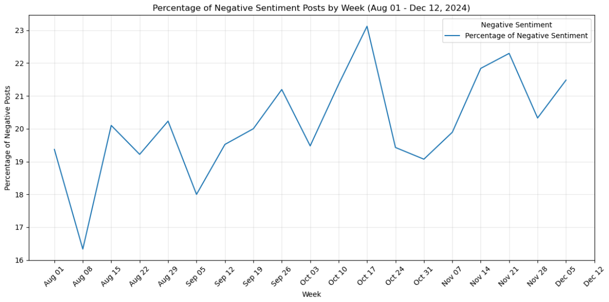

Figure 6: Proportion of posts with negative sentiment

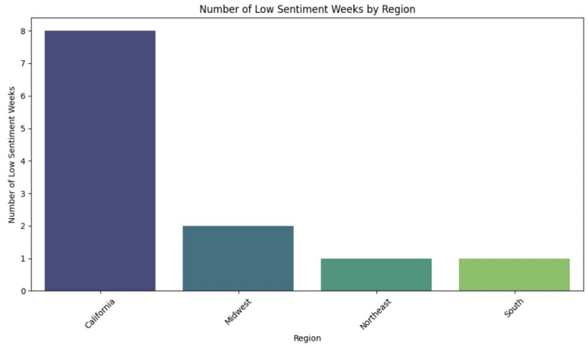

Figure 7: Number of Weeks with Polarity <0.05 by Region

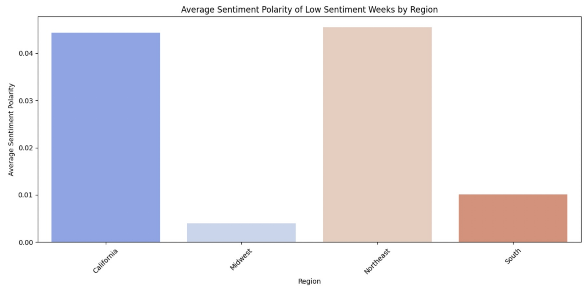

Figure 8: Average Polarity of Low Sentiment Weeks (<0.05) by Region

Model	**Accuracy**	**F1 Score**
Logistic Regression	0.7449	0.7304
Random Forest	0.7946	0.7890
Deep Learning	0.7639	0.7522

Table 1: Performance Metrics for Different Models

4.4 Sentiment Prediction Performance

The **Random Forest Classifier** achieved the best performance with an accuracy of 0.7946 and an F1 Score of 0.7890. We recommend using this model in the university system to effectively track students' mental wellbeing.

5 Discussion

The findings of our project highlight promising strategies for enhancing mental health support for university students. By leveraging sentiment analysis and predictive modeling, universities can identify and address critical periods of emotional distress among students, fostering a proactive approach to mental health care.

Analyzing temporal sentiment trends revealed predictable drops during high-stress periods like midterms and finals. These insights emphasize the importance of tailoring mental health resources to peak stress times, such as scheduling additional counseling services, hosting stress-relief workshops, or providing free mental health screenings during these weeks. Institutions can also use regional insights to customize interventions, as California schools exhibited the most frequent negative sentiment periods compared to other regions like the South and Midwest, where sentiment was more stable.

Clustering and topic modeling results revealed recurring student concerns, including academic stress, campus living, and admissions. These themes provide actionable directions for improving university support systems. For instance, creating peer mentorship programs to help students navigate course registration, enhancing access to affordable housing, or implementing stronger resources to help students during the process of finding student living can alleviate common stressors identified in the analysis. Social and academic support networks could also be strengthened by implementing community-building events that encourage connections among students.

It is essential to note that the data for our project was collected exclusively from Reddit posts, which, while valuable, presents certain limitations. The sentiments analyzed are self-reported and reflect the perspectives of users active on social media platforms, who may not fully represent the broader student population. Additionally, online discussions often skew toward extremes, potentially amplifying negative sentiment or omitting nuances present in in-person communication. Future research could incorporate diverse data sources, such as anonymized campus survey responses or focus groups, to provide a more comprehensive understanding of student well-being.

The machine learning models demonstrated the potential for sentiment prediction. The Random Forest Classifier emerged as the most effective tool, achieving an accuracy of 79.46% and an F1 Score of 78.90%. This model can help universities anticipate future sentiment trends, enabling proactive mental health interventions before issues escalate. Coupled with real-time monitoring of sentiment data, these models could be embedded into campus wellness strategies to identify at-risk groups and deploy targeted resources effectively.

Future initiatives should also involve mental health professionals to refine predictive tools, ensuring interventions are not only timely but also clinically informed. By addressing the inherent limitations of relying solely on social media data and expanding the scope of sentiment analysis, universities can foster a more inclusive, responsive, and supportive academic environment for their student populations.

References

[1] Lim, W., Gunasekara, A., Pallant, J., Pallant, J., Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21, 100790. https://doi.org/10.1016/j.ijme.2023.100790

[2] Benrouba, F., Rachid, B. (2023). Emotional sentiment analysis of social media content for mental health safety. *Social Network Analysis and Mining*, 13. https://doi.org/10.1007/s13278-022-01000-9

[3] Song, J., Song, T.-M., Lee, S., Seo, D.-C. (2023). Depression in South Korean Adolescents Captured by Text and Opinion Mining of Social Big Data. *International Journal of Environmental Research and Public Health*, 20(17), 6665. https://doi.org/10.3390/ijerph20176665