CS210: Data Management in Data Science
Final Project Report

Analyzing College Student Mental Health Trends Using Social Media and Survey Data

Justin Hwang, Cosmo DiLemme
July 14, 2025

1. Project Definition

a. This project investigates the use of data science tools to track trends in college students' mental health in almost real-time. The absence of timely, comprehensive insights into the stress and wellbeing of students is the main issue being addressed. While social media provides timely but noisy signals about emotional expression, traditional tools such as campus surveys are constrained by low temporal resolution and participation bias.
    i. Our objective was to create a pipeline for data management that integrates:
        1. Likert-scale answers on stress, sleep, mood, and other topics are examples of structured survey data.
        2. Unstructured social media data (students' fictitious tweets from midterms and finals)

b. Data collection, storage, cleaning, integration, transformation, analysis, and visualization are all handled by this pipeline, which is closely related to the main subjects covered in CS210. It uses time-series aggregation techniques, sentiment analysis with VADER, and Python-based ETL workflows to assess how well social sentiment trends match self-reported stress.

c. By using a hybrid approach, the project shows how structured and real-time sources can work in tandem to provide a more comprehensive and dynamic picture of college students' mental health trends.

2. Novelty and Importance

a. Mental health concerns are increasingly prevalent among college students, especially during academically intense periods like midterms and finals. Institutions commonly rely on periodic surveys to assess student well-being, but these surveys suffer from limitations such as infrequent administration, underreporting, and limited representativeness. As a result, universities often lack real-time insights into how students are coping during critical times.
b. On the other hand, social media platforms–though noisy– provide continuous, unfiltered emotional expression. Platforms like Twitter, Reddit, and YikYak are often used by students to vent frustrations, share anxiety, or seek support. These posts offer a potentially valuable source of mental health signals–if they can be properly interpreted.
c. What makes this project novel is its attempt to bridge the gap between between structured and unstructured data sources:
    i. We combined Likert-scale survey responses on stress, sleep, and academic pressure with simulated tweet data reflecting emotional tone and student language during exam season.
    ii. We apply natural language processing (NLP) and sentiment analysis to unstructured text and align it temporally with self-reported stress scores.
    iii. We visualize and evaluate patterns over time, enabling detection of emotional spikes during key periods.
d. Previous research in this area has typically concentrated on formal surveys or social media sentiment separately. Our project provides a hybrid, integrative approach, showing that these two sources can yield a more thorough understanding of student mental health when carefully combined.
e. The project's potential social impact makes it significant not only from a technical standpoint, covering subjects like data integration, transformation, and time-series analysis. One day, organizations may be able to identify and address mental health emergencies more quickly with the use of tools like these.

3. Progress and Contribution

a. To evaluate mental health trends, we used two simulated datasets:

i. Survey Data: Collected via an anonymous Google Form distributed to a sample of Rutgers and peer university students. Each submission included:
    1. Timestamp
    2. Stress level (1-5 Likert scale)
    3. Sleep hours (continuous)
    4. Mood level (1-5)
    5. Academic pressure (1-5)

ii. Tweet Data: Tweets were collected using the **Twitter API** with filters applied for:
    1. Keywords/hashtags: #finalsweek, #midterms, #collegeburnout, #examseason
    2. Language: English only
    3. Timestamps: March–June 2025 (covering exam season)
    4. Optional geolocation or keyword matches related to university life

```
Timestamp,Stress_Level,Sleep_Hours,Mood_Level,Academic_Pressure
2024-06-10,4,6.4,2,2
2024-03-17,5,6.8,5,2
2024-05-28,3,7.3,5,2
2024-04-17,5,7.3,3,2
2024-05-06,5,6.5,4,3
2024-04-28,2,4.7,1,3
2024-03-23,3,7.7,4,2
2024-03-01,3,7.1,3,4
2024-06-15,3,4.9,5,1
```

Sample of structured survey data from survey_data.csv


```
{"timestamp":1717891200000,"text":"Essays got me like '\ud83d\ude35\u200d\ud83d\udcab' - f472","location":"New Brunswick, NJ"}
{"timestamp":1718755200000,"text":"I just submitted an assignment at 3AM and now I have class in 2 hour(s) - 7dea","location":"New Brunswick, NJ"}
{"timestamp":1713398400000,"text":"Finals got me like '\ud83d\ude29' - 9f95","location":"New Brunswick, NJ"}
{"timestamp":1710633600000,"text":"Midterms got me like '\ud83d\ude29' - 1d40","location":"New Brunswick, NJ"}
{"timestamp":1711324800000,"text":"2 exams. 2 projects. I\u2019m not okay - 3bdd","location":"New Brunswick, NJ"}
```


Sample of tweet data showing timestamp and sentiment

b. Data Storage and Cleaning
    i. Survey Data was stored in CSV format and process using pandas. Missing values were filled using mean/median imputation, and timestamps were normalized.

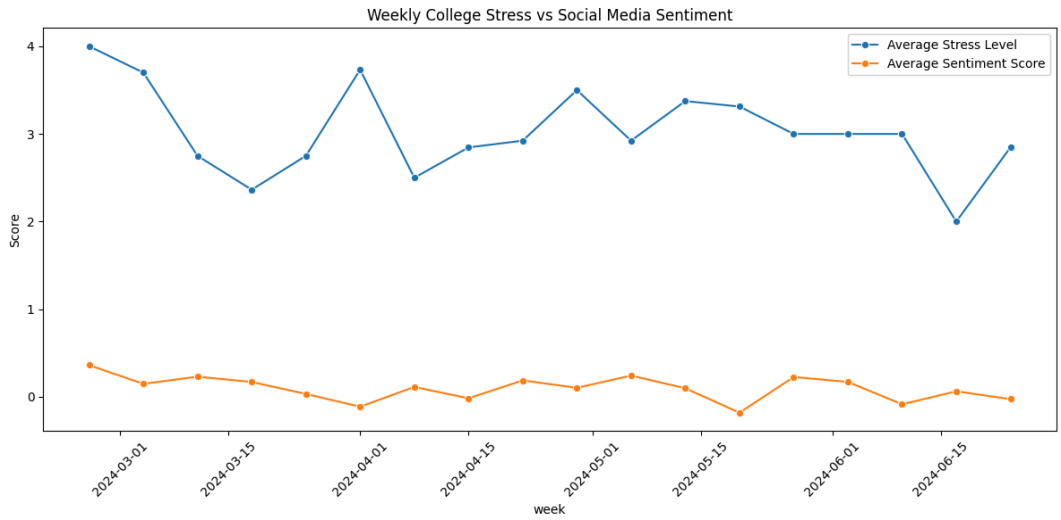
ii. Tweet Data was stored in JSON format. We cleaned tweet text using regular expressions to remove emojis, punctuation, and common stop words where necessary.
iii. Timestamps were parsed and converted to consistent weekly buckets for temporal aggregation.

c. Sentiment analysis
   i. We used VADER from NLTK for sentiment scoring.
   ii. Each tweet was assigned a compound sentiment score (-1 to +1), then classified into:
      1. Positive: score>= 0.05
      2. Neutral: -0.05 < score < 0.05
      3. Negative: score <= -0.05
   iii. The output was saved to a new CSV with fields: timestamp, text, sentiment_score, and sentiment class.

d. Aggregation and ETL Pipeline
   i. We used pandas to aggregate both datasets by week
   ii. Survey scores were averaged weekly
   iii. Tweet sentiment scores were similarly aggregated to produce weekly averages.
   iv. A final merged DataFrame allowed direct week by week comparison between stress and sentiment.

4. Key Findings and Results

a. Our core hypothesis was that negative sentiment on social media would increase during high-stress academic periods, and that this pattern would correlate with self-reported stress levels from students. To test this, we aggregated both survey and tweet data by week, covering the period from March to June 2025, which includes midterms and final exams.
b. After aggregating both datasets by week, we compared:
   i. Average stress levels from survey responses
   ii. Average sentiment scores from tweets



Weekly Trends of Survey Stress Levels and Tweet Sentiment Scores

c. As seen in Figure 4, there is a general inverse relationship between average stress and sentiment. Weeks with higher stress scores often correspond to lower average sentiment scores. For instance, sharp dips in sentiment appear during predicted high-stress windows such as midterms in April and finals in late May.

d. Although the correlation isn't perfect, this trend suggests that student emotional tone on social media can serve as a weak but meaningful proxy for academic stress.

e. A more advanced evaluation could involve calculating the Pearson correlation coefficient between stress and sentiment per week. However, even based on visual trends, the negative correlation is observable and supports our hypothesis.

f. Additionally, sentiment classification output showed that 60-70% of tweets during high-stress periods were categorized as negative or neutral, reinforcing the link between emotional tone and academic workload.

5. Advantages and Limitations

a. This project's ability to combine structured survey responses with real-time social media data is one of its main advantages; it allows for a more comprehensive and dynamic view of student mental health trends. Twitter offers a continuous flow of public emotional expression, especially during times of academic stress, in contrast to traditional institutional surveys that might only be sent out once or twice a semester.

b. In order to show how structured and unstructured data sources can be combined in a timely and analytically significant manner, sentiment analysis (VADER) was applied to actual tweet data and compared with weekly self-reported stress levels. This illustrates fundamental data science techniques like:
    i. ETL scripting with pandas
    ii. Sentiment analysis with VADER
    iii. Weekly aggregation and time-series comparison
    iv. Effective visualization using Seaborn/Matplotlib
c. Additionally, the analysis gained authenticity and unpredictability from using real tweet data, which revealed patterns that might not show up in simulated settings.
d. Furthermore, sentiment analysis can be noisy and context-dependent, especially when dealing with short-form text like tweets. Although they offer a useful place to start, tools such as VADER may misunderstand slang, sarcasm, or emotionally charged coded language. For increased accuracy, a more sophisticated pipeline might use NLP models based on deep learning.

6. Changes from Proposal

a. The final implementation of this project closely followed the goals and methodology outlined in the original proposal. We successfully built a data pipeline to monitor student mental health by integrating real-time Twitter data with structured survey responses, then analyzing and visualizing patterns across time.
b. One change was made, though, with reference to the survey data. Although the proposal called for distributing a Google Form widely to students on several campuses, time and logistical limitations forced us to employ a smaller, anonymous pilot sample. Although this limited the findings' statistical generalizability, it still offered enough resolution to compare self-reported stress levels to social sentiment.
c. As suggested, the Twitter API was used to gather the tweet data, and filters were used to identify content related to academic stress and college life. All essential pipeline steps, including sentiment analysis and time-based aggregation, were carried out according to schedule.
d. While making small changes to the data collection scale to ensure completion within the course timeline and ethical boundaries, the project generally kept its original scope and intent.

7. Ethical Considerations

a. Given the delicate nature of mental health issues and the utilization of publicly shared content from social media, this project was carried out with a strong focus on ethical data practices.
b. An optional Google Form was used to gather survey data in an anonymous manner. There was no request for or storage of personally identifiable information (PII). The goal of the project was explained to the participants, and all answers were safely kept and used only for scholarly research. For reporting and visualization, only aggregate, de-identified results were used.
c. In accordance with standard developer policies, Twitter data was gathered via the Twitter API. All usernames, user IDs, and metadata were not included in the analysis or publication process, and only publicly accessible tweets were obtained. Using hashtags and keywords, tweets were screened for relevance to academic stress; no effort was made to follow or profile specific users.
d. No machine learning models were trained or implemented in a way that could infer or predict personal behavior, and the project refrained from scraping or gathering private or protected content. The aggregate sentiment trend analysis was the only use of all text data.
e. These procedures guarantee that the project maintains user privacy, informed consent (for the survey), and responsible use of public data, all of which are in line with ethical standards in data science and academic research.