Analyzing Student Sentiments Through Reddit: Insights for University Well-Being

Team : CAT GPT

Name	Email	JHED
Jiadi(Jady) Li	jli517@jh.edu	jli517
Meishu Zhao	mzhao55@jh.edu	mzhao55
Zhenlong Zhang	zzhan352@jh.edu	zzhan352
An Roujin	ran8@jhmi.edu	ran8

1. Introduction

Our research focuses on the question: What are the major student concerns for each of the US News Top 50 U.S. universities, and how can sentiment analysis of Reddit posts provide actionable insights into student well-being and campus life? This question is particularly significant because universities often struggle to understand the real-time concerns of their students. Traditional methods like surveys and focus groups are limited by their structured nature and inability to capture the spontaneity of unfiltered student discussions. Reddit, as a platform widely used by university students, provides a rich source of organic, real-time data that can reveal trends in student sentiment over time. Additionally, this research helps the prospective students to choose schools that fit them well.

Understanding these trends is crucial for improving student well-being and addressing issues such as academic pressure, mental health, housing, and social integration. By analyzing sentiment patterns, universities can proactively respond to student needs and enhance the overall campus experience. This research also holds broader implications for public health, as it offers insights into the mental health challenges faced by young adults, a group often at risk of stress and anxiety.

Existing work in sentiment analysis has largely focused on platforms like X (Twitter), leveraging lexicon-based approaches and machine-learning models to understand public opinion. However, Reddit remains underexplored, despite its active university-specific communities. Our project builds on existing sentiment analysis tools by tailoring them to the unique structure and content of Reddit data. Additionally, by leveraging advanced techniques, such as combining other natural language pre-trained models for nuanced sentiment understanding and Generative Adversarial Networks (GANs) for addressing missing sentiment scores, we demonstrate the potential for improving student well-being through data-driven insights. Although GAN-predicted sentiment scores were excluded from the final dashboard due to consistency concerns, they highlight a promising direction for enhancing the completeness and reliability of sentiment analysis.

2. Objectives

The goal of this project is to analyze sentiment trends from Reddit posts for the top 50 U.S. universities. We wanted to better understand student concerns and provide actionable insights through an interactive dashboard so that school policymakers could enact better operations from the students' perspective. Another important goal is to help prospective students to find their dream school by letting them not only be able to compare traditional metrics such as basic information about universities, like tuition costs, acceptance rates, graduation rates, crime rates, and weather data, but also compare more insightful information such as the sentiment experience of existing students. With this, we hope to make it easier for students to pick the schools that best fit their needs and preferences.

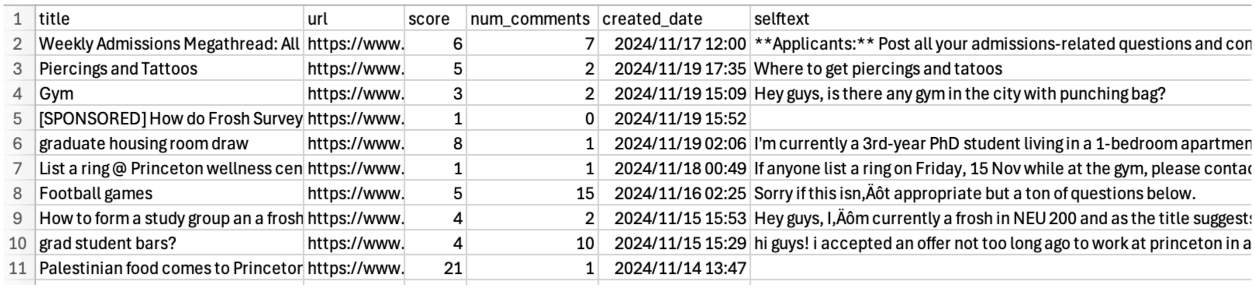
3. Data Collection

Data Source and Access:

The primary source of data for this project was Reddit, where we extracted posts from the subreddits of the top 50 U.S. universities. Using Python's "praw" library, we accessed the Reddit API to retrieve up to 1000 of the most recent posts for each university. Each post included metadata such as the post title, creation date, content, and the number of comments. These posts were exported as CSV files for further preprocessing and sentiment analysis in R.

In addition to Reddit posts, we collected supplementary data using R's "rvest" package. University-specific attributes, such as SAT scores, ACT scores, undergraduate tuition, graduate tuition, acceptance rates, and rankings, were scraped from the US News Rankings website. State-level violent crime rates (averaged from 2018 to 2022) were retrieved from the FBI database, while average annual temperatures were sourced from Current Results. These datasets were combined to create a comprehensive profile for each university, integrating external factors into the analysis.

Data Description:



1	title	url	score	num_comments	created_date	selftext
2	Weekly Admissions Megathread: All	https://www.	6	7	2024/11/17 12:00	**Applicants:** Post all your admissions-related questions and con
3	Piercings and Tattoos	https://www.	5	2	2024/11/19 17:35	Where to get piercings and tatoos
4	Gym	https://www.	3	2	2024/11/19 15:09	Hey guys, is there any gym in the city with punching bag?
5	[SPONSORED] How do Frosh Survey	https://www.	1	0	2024/11/19 15:52	
6	graduate housing room draw	https://www.	8	1	2024/11/19 02:06	I'm currently a 3rd-year PhD student living in a 1-bedroom apartmen
7	List a ring @ Princeton wellness cen	https://www.	1	1	2024/11/18 00:49	If anyone list a ring on Friday, 15 Nov while at the gym, please contac
8	Football games	https://www.	5	15	2024/11/16 02:25	Sorry if this isn,Äôt appropriate but a ton of questions below.
9	How to form a study group an a frosh	https://www.	4	2	2024/11/15 15:53	Hey guys, I,Äôm currently a frosh in NEU 200 and as the title suggests
10	grad student bars?	https://www.	4	10	2024/11/15 15:29	hi guys! i accepted an offer not too long ago to work at princeton in a
11	Palestinian food comes to Princetor	https://www.	21	1	2024/11/14 13:47	

Data Challenges:

Data collection posed several challenges, particularly with Reddit's API, which limits access to 1,000 posts per subreddit. This restriction resulted in varying periods and data volumes, especially for universities with less active subreddit communities. Additionally, Reddit's dynamically loaded content complicated scraping efforts in R, where only the first few posts could be retrieved for each subreddit. To address this limitation, we utilized Python's "praw" library, which provided more robust access to the data.

Preprocessing the Reddit data required significant effort, including separating date variables into day, month, and year, and handling missing titles or content. Cleaning the university-specific data also presented challenges, as several columns contained special characters and inconsistent formats. The final step involved joining multiple datasets using common variables, such as state or university names, and filling in any missing values to ensure a unified structure for analysis.

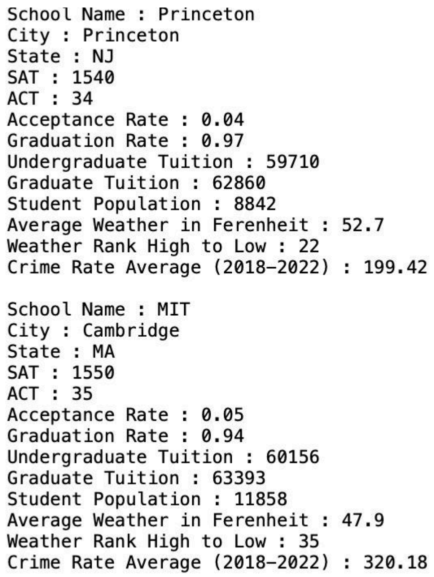
4. Programming Paradigms:

Object Oriented Paradigm

The Object-Oriented Programming (OOP) paradigm played a crucial role in managing and processing university-related data efficiently in our project. We implemented a School class to encapsulate key attributes and methods for each university, ensuring modularity and scalability. Each School object included attributes such as university name(character), city(character), state(character), SAT scores(double) and ACT scores(double), acceptance rates(double), graduation rates(double), tuition fees(double), student populations(double), average weather(double), and crime rates(double) over multiple years. The class methods facilitated data processing, including initializing objects with attributes, calculating average crime rates across years, and converting object data into a list for easy integration with the dashboard.

To create these objects, we loaded and preprocessed the school_info dataset, ensuring numeric fields were appropriately formatted and missing values were addressed. Using the School$new method, we instantiated objects for each university, encapsulating relevant data and calculations. The calculate_crime_rate method was applied to compute multi-year crime rate averages using functional programming (map2). These objects were stored in a lookup table (school_lookup), allowing seamless access to attributes for dashboard features like dropdown menus and comparison tables.

The adoption of OOP brought several advantages, including modularity, which made the code organized and easier to debug, and reusability, allowing straightforward integration of new universities or attributes. Additionally, the to_list method ensured smooth integration of object data into the Shiny dashboard, supporting interactive elements like tables and maps. Despite initial challenges with inconsistent data formats and integrating multi-year crime data, preprocessing steps and robust method design resolved these issues effectively. Overall, OOP provided a structured framework that streamlined development, enhanced the scalability of our system, and supported future extensions.

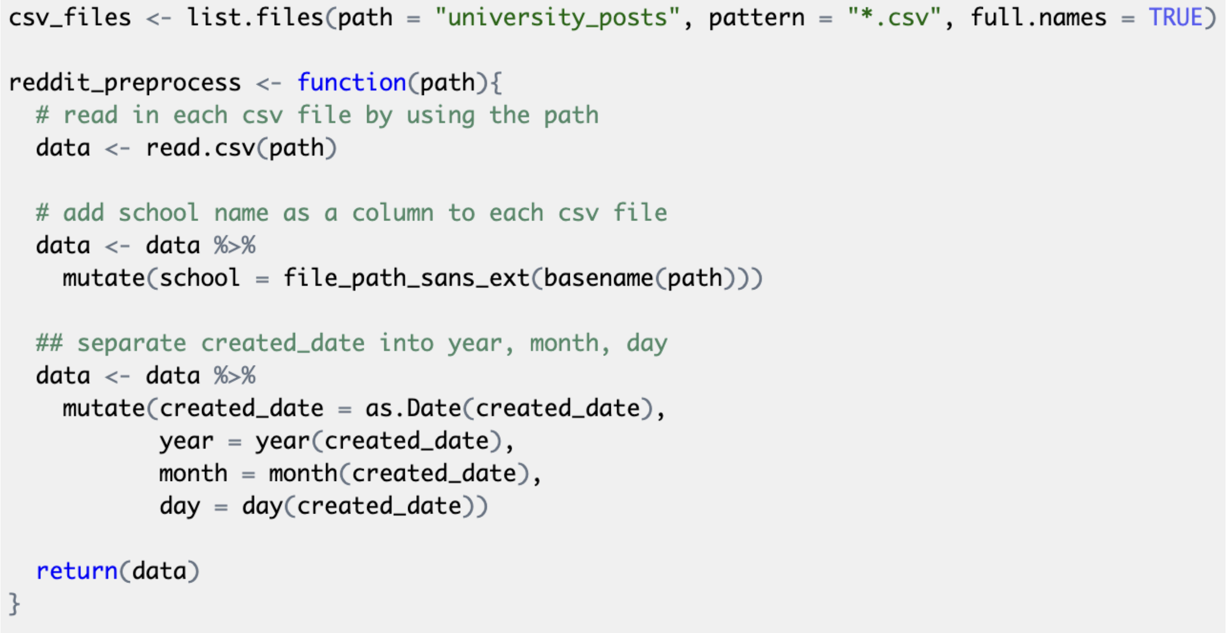


```
School Name : Princeton
City : Princeton
State : NJ
SAT : 1540
ACT : 34
Acceptance Rate : 0.04
Graduation Rate : 0.97
Undergraduate Tuition : 59710
Graduate Tuition : 62860
Student Population : 8842
Average Weather in Ferenheit : 52.7
Weather Rank High to Low : 22
Crime Rate Average (2018–2022) : 199.42

School Name : MIT
City : Cambridge
State : MA
SAT : 1550
ACT : 35
Acceptance Rate : 0.05
Graduation Rate : 0.94
Undergraduate Tuition : 60156
Graduate Tuition : 63393
Student Population : 11858
Average Weather in Ferenheit : 47.9
Weather Rank High to Low : 35
Crime Rate Average (2018–2022) : 320.18
```

Functional Programming Paradigm

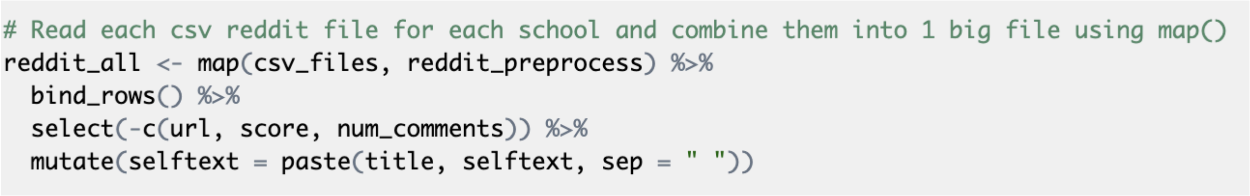
Functional Programming Paradigm is first used in data preprocessing, preparing data for the shiny dashboard. Since we have a total of 50 schools, and each school has its own csv file for Reddit posts, thus we have a total of 50 csv files. We want to first read in each file, add the school name as a column to each csv file, and separate the date of the Reddit post into year, month, day. And then, we want to combine the transformed 50 csv files into 1 big csv file.

We first create a folder called “university posts” that stores all 50 Reddit post csv files inside. Then I created a list called “csv_files” to fetch all 50 paths for each csv file inside the “university posts” folder. In order to apply transformations to each csv file efficiently and concisely, I created a function called “reddit_preprocess” that takes in a csv file path as input, read in the csv file, adds school name as a column by using that csv file’s file name, and separate created_date into year, month, day.



```r
csv_files <- list.files(path = "university_posts", pattern = "*.csv", full.names = TRUE)

reddit_preprocess <- function(path){
  # read in each csv file by using the path
  data <- read.csv(path)

  # add school name as a column to each csv file
  data <- data %>%
    mutate(school = file_path_sans_ext(basename(path)))

  ## separate created_date into year, month, day
  data <- data %>%
    mutate(created_date = as.Date(created_date),
           year = year(created_date),
           month = month(created_date),
           day = day(created_date))

  return(data)
}
```

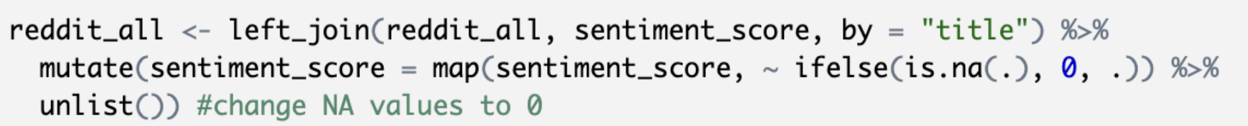
After setting up the “reddit_preprocess” function, we apply this function to the “csv_files” list by using the map function to make transformations for each of the 50 csv file, and then combine all of them into 1 big csv file.



```r
# Read each csv reddit file for each school and combine them into 1 big file using map()
reddit_all <- map(csv_files, reddit_preprocess) %>%
  bind_rows() %>%
  select(-c(url, score, num_comments)) %>%
  mutate(selftext = paste(title, selftext, sep = " "))
```

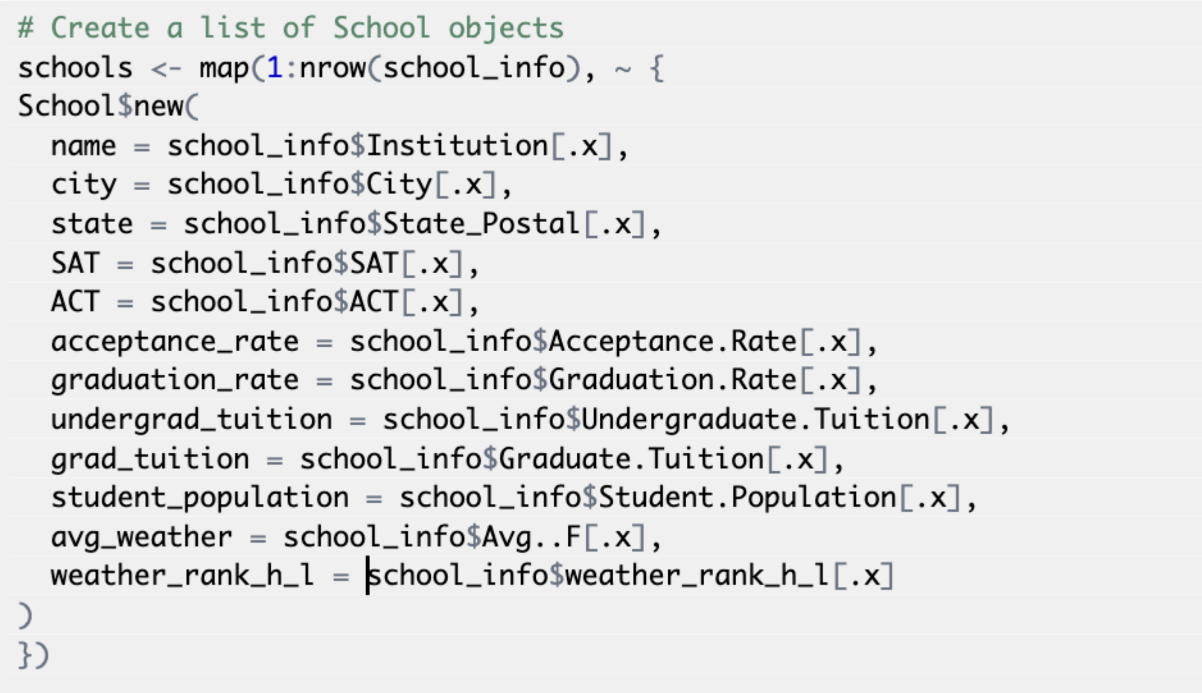
By applying the functional programming paradigm, we enhanced the conciseness and readability of code, reducing the probability of errors. In addition, it helps with parallel programming when the scale of the dataset is much larger. For instance, if we want to analyze more than 50 schools, like 500 schools, it will be very efficient, easy-to-understand, and easy-to-debug compared to using for loops.

Functional Programming Paradigm is also used to deal with missing values of sentiment scores for each Reddit post. After we created a new dataset called “sentiment_score” to store the sentiment score for each Reddit post by using the “afinn” package, we figured out that there are many missing values. In other words, many posts have an NA value for the sentiment_score column. To deal with the NA values for better sentiment analysis in later processes, we decided to change all NA values into 0. Thus, we used the map() function to take in the sentiment_score dataset, change all NA values to 0, and then left_join to the original dataset reddit_all. Since the result of the map() function is a list, we want to transform the list back into a data frame structure, so we applied to unlist() after using map().



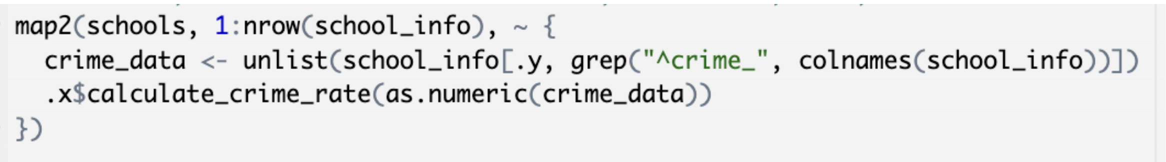
```
reddit_all <- left_join(reddit_all, sentiment_score, by = "title") %>%
  mutate(sentiment_score = map(sentiment_score, ~ ifelse(is.na(.), 0, .)) %>%
  unlist()) #change NA values to 0
```

We also applied functional programming to the creation and processing of OOP School class objects. By using the map function, we eliminated the need for verbose loops and manual initialization, significantly enhancing the code's conciseness and scalability. The map function iterated over each row of the school_info dataset to dynamically create a School object using the School$new() method. Each object encapsulated key attributes, including the university name, city, state, SAT scores, ACT scores, acceptance rates, graduation rates, tuition fees, student populations, average weather, and crime rates.



```
# Create a list of School objects
schools <- map(1:nrow(school_info), ~ {
School$new(
  name = school_info$Institution[.x],
  city = school_info$City[.x],
  state = school_info$State_Postal[.x],
  SAT = school_info$SAT[.x],
  ACT = school_info$ACT[.x],
  acceptance_rate = school_info$Acceptance.Rate[.x],
  graduation_rate = school_info$Graduation.Rate[.x],
  undergrad_tuition = school_info$Undergraduate.Tuition[.x],
  grad_tuition = school_info$Graduate.Tuition[.x],
  student_population = school_info$Student.Population[.x],
  avg_weather = school_info$Avg..F[.x],
  weather_rank_h_l = school_info$weather_rank_h_l[.x]
)
})
```

Additionally, we employed the map2 function to calculate the multi-year average crime rates for each university. This was achieved by pairing each School object with its corresponding row in the dataset and applying the calculate_crime_rate method. The method computed average crime rates using data from columns named crime_2018, crime_2019, ..., crime_2022. This functional approach streamlined the addition of calculated attributes to the objects and ensured consistency across all entries.



```
map2(schools, 1:nrow(school_info), ~ {
  crime_data <- unlist(school_info[.y, grep("^crime_", colnames(school_info))])
  .x$calculate_crime_rate(as.numeric(crime_data))
})
```

Machine Learning Paradigm

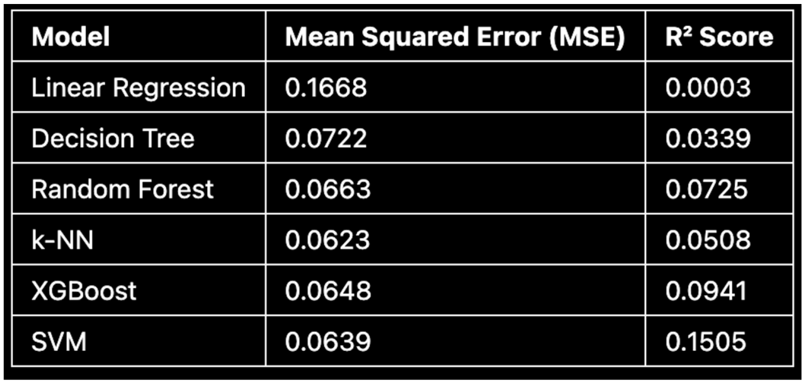
Initially we were interested in if the sentiment scores of each school is predictable with related factors including weather, crime rate, rankings, tuition and SAT scores. We made machine learning prediction models with linear regression, decision tree, random forest, k-NN and SVM, we fine-tuned the decision tree model and SVM model to get the best hyperparameters for prediction.

Data preprocessing: The dataset was first cleaned by dropping rows with missing values (na.omit). Predictors included Undergraduates. Tuition, Graduate.Tuition, Acceptance.Rate, SAT, ACT, Avg..F, Avg..C, weather_rank_h_l, crime_2018, crime_2020, crime_2021, crime_2022, US_News_rank_2025, and Change_in_rank. The target variable was average_sentiment.

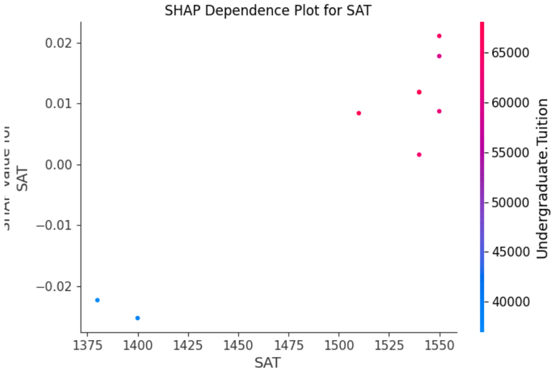
Model training and fine-tuning: The dataset was then split into training (80%) and testing (20%) subsets using stratified sampling to preserve the distribution of the target variable. Features were standardized using StandardScaler to ensure that all predictors were on a comparable scale. An SVR (Support Vector Regression) model was eventually used to predict the average sentiment. A grid search with 5-fold cross-validation was performed to fine-tune the hyperparameters of the SVR model. The hyperparameters tested included: C (regularization parameter): [0.1, 1, 10, 100]; gamma (kernel coefficient): ['scale', 'auto', 0.1, 0.01] and kernel (type of kernel): ['rbf', 'linear'].The best combination of hyperparameters was determined based on minimizing the mean squared error (MSE) during cross-validation.

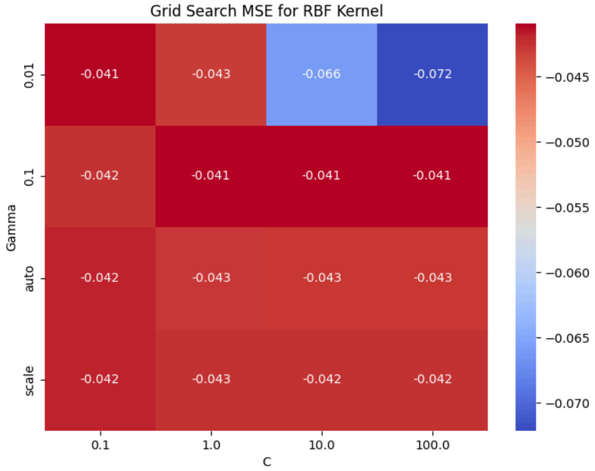
Model Evaluation: The model was evaluated on the test dataset using Mean Squared Error (MSE): To measure the average squared difference between predicted and actual values; and R² Score: To assess the proportion of variance explained by the model.
The contribution of each predictor variable was analyzed using SHAP (SHapley Additive exPlanations). SHAP values were computed for each test sample to understand how individual predictors influenced the model's predictions.SHAP Summary Plot was produced to visualize the average importance of predictors. And SHAP Dependence Plot on SAT and crime rates are also generated.

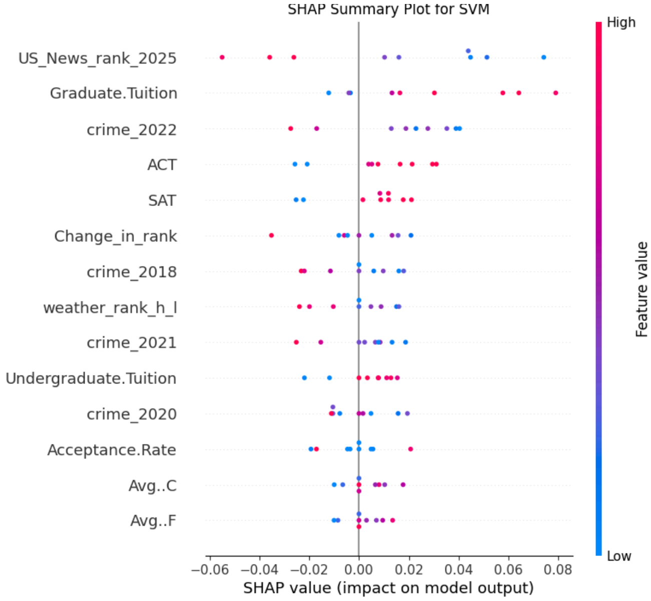
The implementation was performed in Python using libraries such as pandas, scikit-learn, seaborn, matplotlib, and shap. All plots were saved as images for further analysis and reporting.

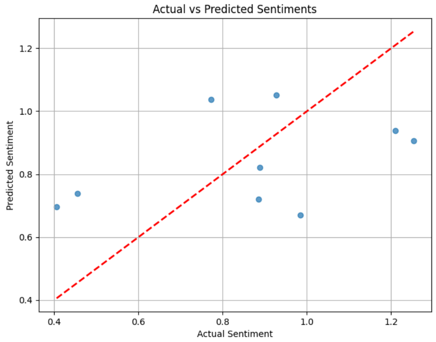


Model	Mean Squared Error (MSE)	R² Score
Linear Regression	0.1668	0.0003
Decision Tree	0.0722	0.0339
Random Forest	0.0663	0.0725
k-NN	0.0623	0.0508
XGBoost	0.0648	0.0941
SVM	0.0639	0.1505







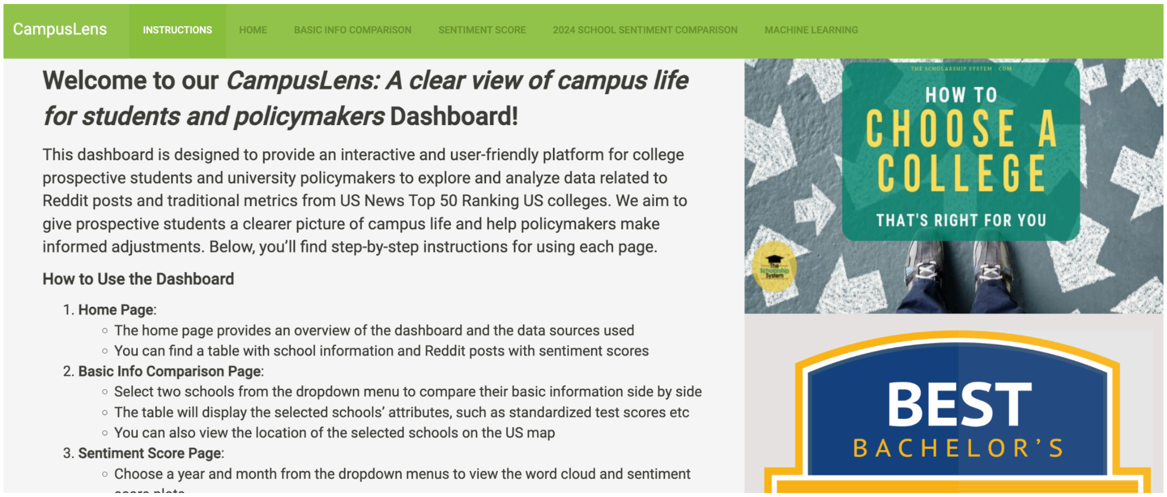


5. Data Analytic Product

The final data analytic product is a Shiny dashboard that provides a comprehensive and interactive platform to explore sentiment trends across Reddit posts for the top 50 U.S. universities. The dashboard is structured into several intuitive interfaces, each serving a specific purpose. Key features of the dashboard include sentiment trend analysis, sentiment distribution visualization, and a university comparison tool. These features are enhanced by interactive elements, such as word clouds, filtering options, and customizable reports. The dashboard is designed to be user-friendly, visually engaging, and adaptable to various analytical needs. Its intuitive design ensures accessibility for a wide range of users, from university administrators to researchers, while providing actionable insights into student sentiment trends.

Instruction Page

The instruction page provides a clear, step-by-step guide for using our Dashboard. It explains key features, including the Home Page for data overviews, the Basic Info Comparison Page for side-by-side school comparisons, and the Sentiment Score Page for analyzing trends through word clouds and sentiment scores. This helps users effectively navigate and analyze college-related data for informed decision-making.



Welcome to our *CampusLens: A clear view of campus life for students and policymakers* Dashboard!

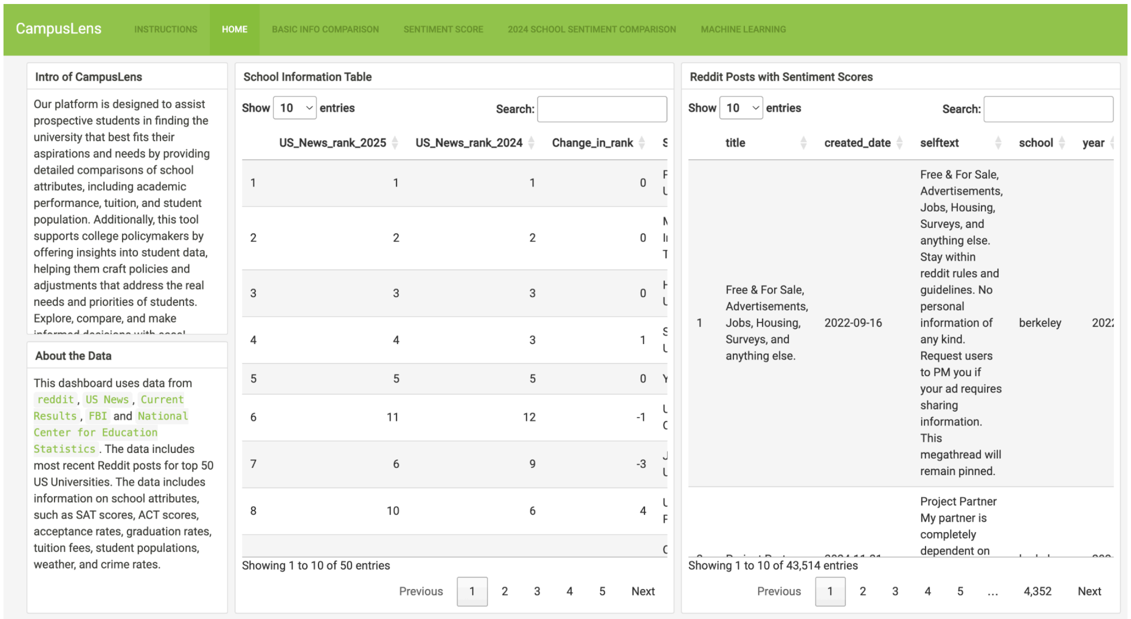
This dashboard is designed to provide an interactive and user-friendly platform for college prospective students and university policymakers to explore and analyze data related to Reddit posts and traditional metrics from US News Top 50 Ranking US colleges. We aim to give prospective students a clearer picture of campus life and help policymakers make informed adjustments. Below, you'll find step-by-step instructions for using each page.

How to Use the Dashboard

1. **Home Page**:
   - The home page provides an overview of the dashboard and the data sources used
   - You can find a table with school information and Reddit posts with sentiment scores
2. **Basic Info Comparison Page**:
   - Select two schools from the dropdown menu to compare their basic information side by side
   - The table will display the selected schools' attributes, such as standardized test scores etc
   - You can also view the location of the selected schools on the US map
3. **Sentiment Score Page**:
   - Choose a year and month from the dropdown menus to view the word cloud and sentiment score plots

Home Page

The Home Page introduces the dashboard and its purpose, providing users with an overview of the dataset and its structure. It includes a table showcasing school attributes like rankings, tuition, and acceptance rates, as well as another table displaying Reddit posts with sentiment scores, dates, and content. This page is designed to familiarize users with the data and outline the dashboard's goal: to help prospective students find universities that match their needs while offering policymakers insights into student well-being and campus trends.



Intro of CampusLens

Our platform is designed to assist prospective students in finding the university that best fits their aspirations and needs by providing detailed comparisons of school attributes, including academic performance, tuition, and student population. Additionally, this tool supports college policymakers by offering insights into student data, helping them craft policies and adjustments that address the real needs and priorities of students. Explore, compare, and make

About the Data

This dashboard uses data from `reddit`, `US News`, `Current Results`, `FBI` and `National Center for Education Statistics`. The data includes most recent Reddit posts for top 50 US Universities. The data includes information on school attributes, such as SAT scores, ACT scores, acceptance rates, graduation rates, tuition fees, student populations, weather, and crime rates.

School Information Table

Show 10 entries   Search:

	US_News_rank_2025	US_News_rank_2024	Change_in_rank
1	1	1	0
2	2	2	0
3	3	3	0
4	4	3	1
5	5	5	0
6	11	12	-1
7	6	9	-3
8	10	6	4

Showing 1 to 10 of 50 entries

Previous 1 2 3 4 5 Next

Reddit Posts with Sentiment Scores

Show 10 entries   Search:

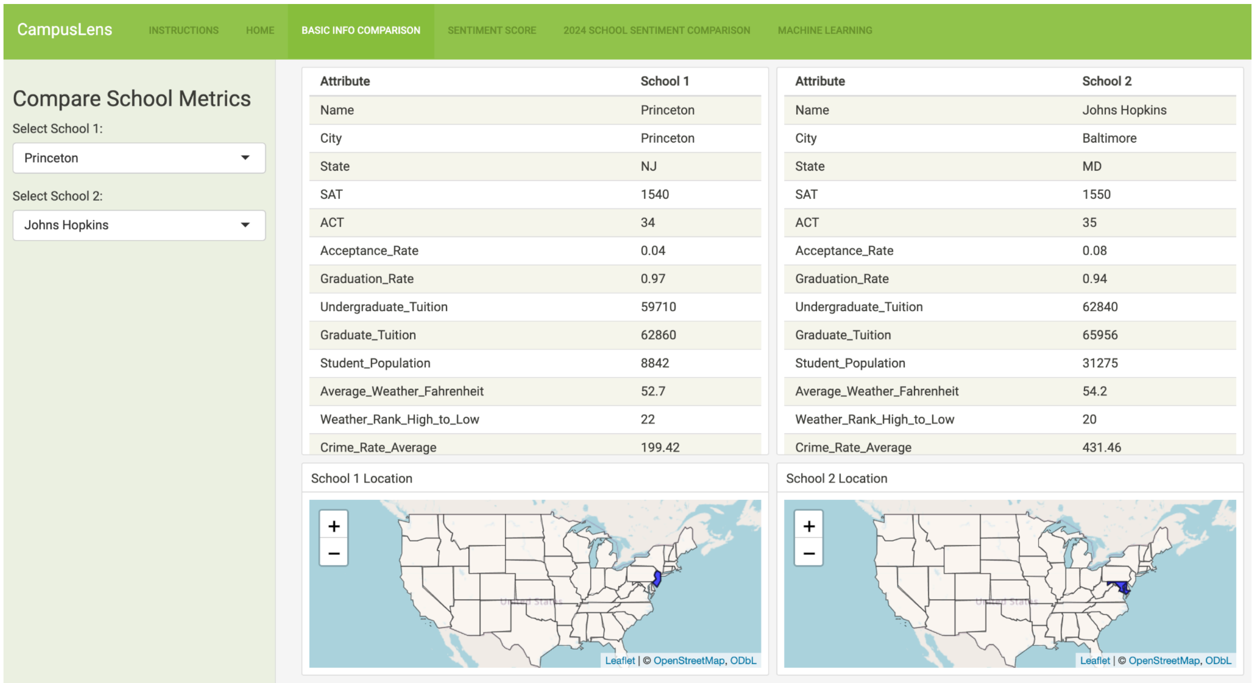
	title	created_date	selftext	school	year
1	Free & For Sale, Advertisements, Jobs, Housing, Surveys, and anything else.	2022-09-16	Free & For Sale, Advertisements, Jobs, Housing, Surveys, and anything else. Stay within reddit rules and guidelines. No personal information of any kind. Request users to PM you if your ad requires sharing information. This megathread will remain pinned.	berkeley	202
			Project Partner My partner is completely dependent on		

Showing 1 to 10 of 43,514 entries

Previous 1 2 3 4 5 … 4,352 Next

Basic Info Comparison page

This Basic Info Comparison page allows users to compare two universities side-by-side by selecting them from dropdown menus. Key attributes such as standardized test scores, tuition costs, graduation rates, student population, and average weather are displayed in a comparison table. Interactive maps highlight the geographical locations of the selected schools, offering a clear and concise view of how these institutions differ in various metrics.



CampusLens
INSTRUCTIONS
HOME
BASIC INFO COMPARISON
SENTIMENT SCORE
2024 SCHOOL SENTIMENT COMPARISON
MACHINE LEARNING
+
−
Leaflet | © OpenStreetMap, ODbL
+
−
Leaflet | © OpenStreetMap, ODbL

Compare School Metrics

Select School 1: Princeton

Select School 2: Johns Hopkins

Attribute	School 1
Name	Princeton
City	Princeton
State	NJ
SAT	1540
ACT	34
Acceptance_Rate	0.04
Graduation_Rate	0.97
Undergraduate_Tuition	59710
Graduate_Tuition	62860
Student_Population	8842
Average_Weather_Fahrenheit	52.7
Weather_Rank_High_to_Low	22
Crime_Rate_Average	199.42

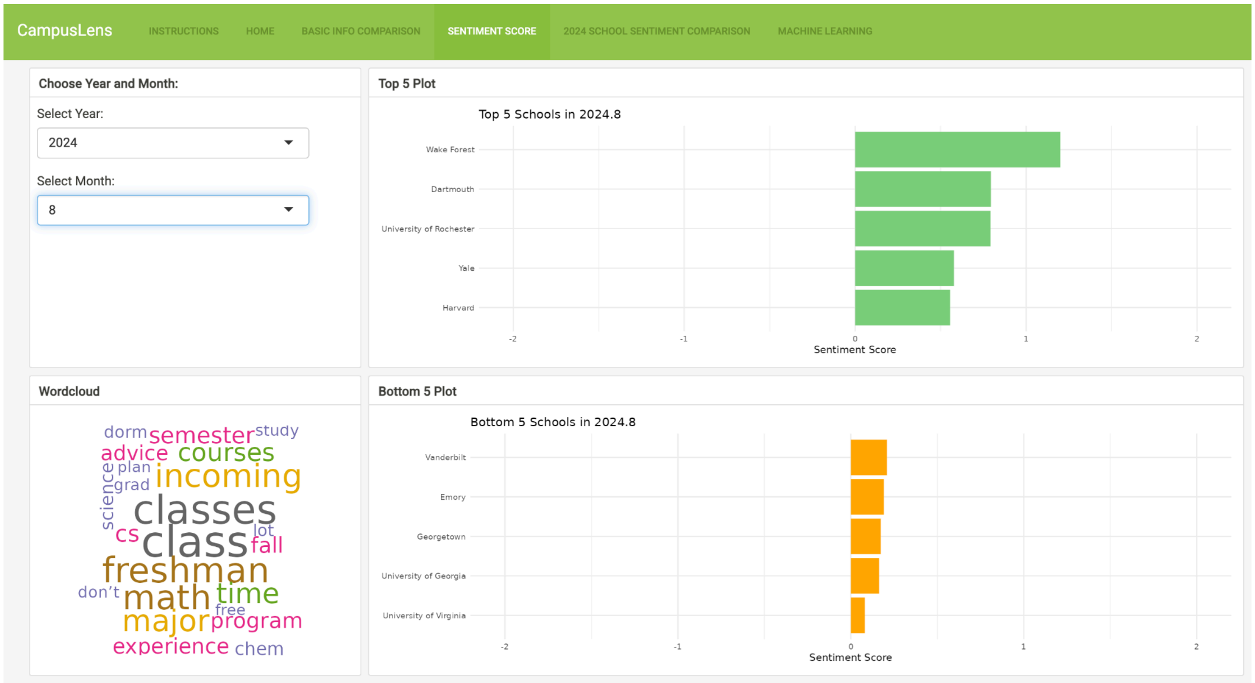
Attribute	School 2
Name	Johns Hopkins
City	Baltimore
State	MD
SAT	1550
ACT	35
Acceptance_Rate	0.08
Graduation_Rate	0.94
Undergraduate_Tuition	62840
Graduate_Tuition	65956
Student_Population	31275
Average_Weather_Fahrenheit	54.2
Weather_Rank_High_to_Low	20
Crime_Rate_Average	431.46

School 1 Location

School 2 Location

Sentiment Score page

The Sentiment Score page focuses on sentiment trends, providing bar charts to display the top 5 and bottom 5 schools based on average sentiment scores for a selected month and year. A word cloud highlights commonly discussed terms in Reddit posts during that period, giving users additional context into the prevalent topics driving sentiment. This page offers an intuitive way to explore student sentiment patterns and the key issues affecting each university.



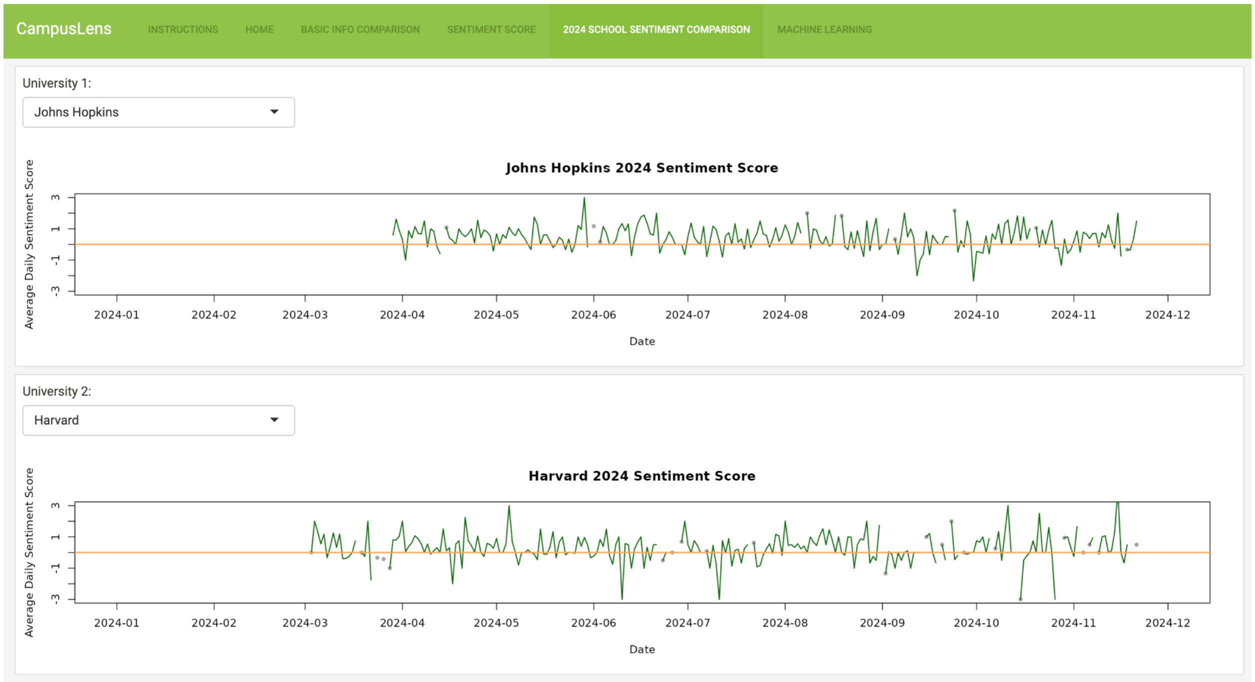
Choose Year and Month:

Select Year: 2024

Select Month: 8

2024 School Sentiment Comparison Page

This page enables users to compare sentiment scores across multiple universities in 2024 using bar charts and scatter plots. The bar charts rank universities by average sentiment scores, while the scatter plots map sentiment scores against post-activity levels, helping users identify patterns, outliers, and potential correlations. This comparative analysis helps prospective students and policymakers make data-driven decisions.



University 1: Johns Hopkins

University 2: Harvard

Machine Learning page

This Machine learning page outlines the machine learning models and techniques used to predict sentiment scores, leveraging features like weather, crime rates, tuition costs, and rankings. Multiple models were tested, with Support Vector Machines (SVM) achieving the highest R2R^2R2 score of 0.15, though performance remains unsatisfactory. Hyperparameter tuning for SVM identified C=100C=100C=100 and γ=0.01\gamma=0.01γ=0.01 as optimal parameters. SHAP analysis highlighted the importance of factors such as U.S. News rankings, tuition, and crime rates in sentiment prediction. The residual plot reveals significant deviations between actual and predicted sentiment, indicating room for model improvement in future iterations.

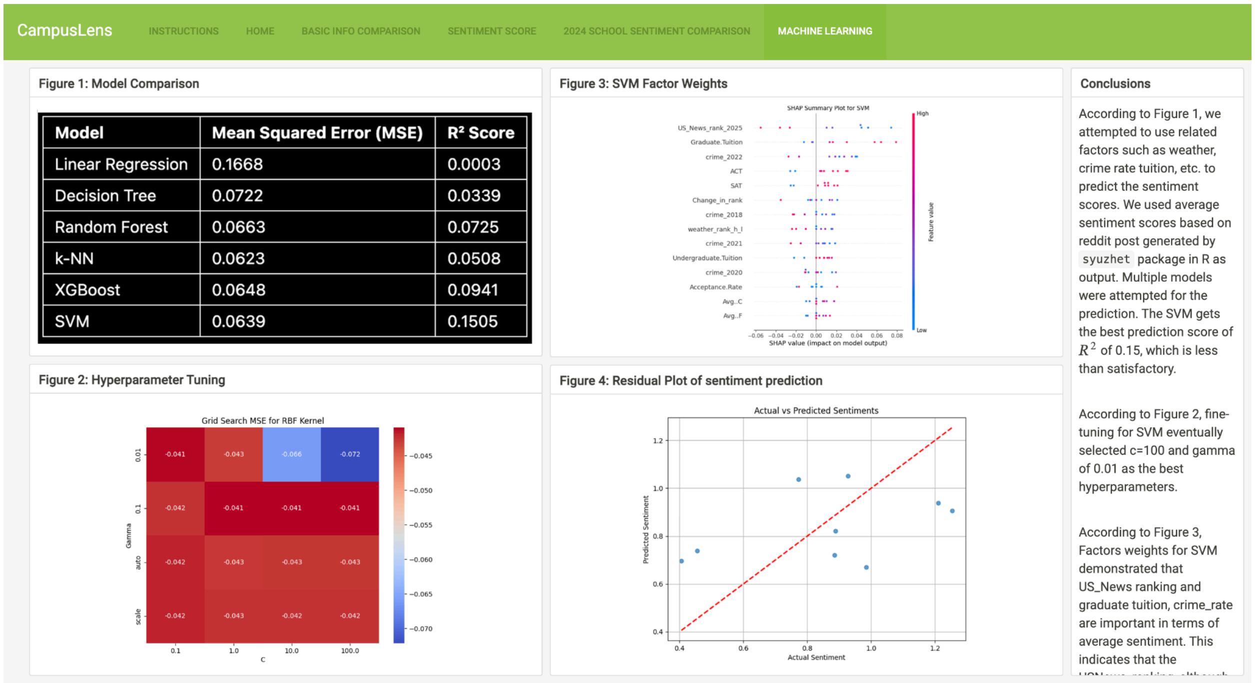


Figure 1: Model Comparison

Model	Mean Squared Error (MSE)	R² Score
Linear Regression	0.1668	0.0003
Decision Tree	0.0722	0.0339
Random Forest	0.0663	0.0725
k-NN	0.0623	0.0508
XGBoost	0.0648	0.0941
SVM	0.0639	0.1505

Figure 2: Hyperparameter Tuning

Figure 3: SVM Factor Weights

Figure 4: Residual Plot of sentiment prediction

Conclusions

According to Figure 1, we attempted to use related factors such as weather, crime rate tuition, etc. to predict the sentiment scores. We used average sentiment scores based on reddit post generated by `syuzhet` package in R as output. Multiple models were attempted for the prediction. The SVM gets the best prediction score of R^2 of 0.15, which is less than satisfactory.

According to Figure 2, fine-tuning for SVM eventually selected c=100 and gamma of 0.01 as the best hyperparameters.

According to Figure 3, Factors weights for SVM demonstrated that US_News ranking and graduate tuition, crime_rate are important in terms of average sentiment. This indicates that the

6. Conclusion:

Results

Our analysis provided valuable insights into student sentiments and factors influencing their experiences across the top 50 U.S. universities. By analyzing sentiment trends and supplementary data, we identified patterns that reveal the dynamic nature of student well-being.

One significant finding was the variation in sentiment trends over time. Word clouds captured shifts in student discussions, with keywords reflecting seasonal concerns. For instance, in April, words like "intern," "summer," and "grad" were prominent, indicating a focus on internships and graduation. In August, keywords such as "GPA," "freshman," and "housing" highlighted the excitement and adjustments of new students starting the academic year. Sentiment scores aligned with these trends, showing overall positivity in August 2024 across both top and bottom-performing schools. However, by September, sentiment scores for the bottom 5 schools turned negative, likely reflecting the impact of academic pressures and mid-term exams.

Crime rates have added one more context to our analysis. Universities in California, for example, exhibited higher state-level crime rates compared to Johns Hopkins University (JHU). Interestingly, sentiment scores for California schools did not directly correlate with these crime rates, suggesting that campus-specific factors may mitigate the influence of state-level statistics. JHU, on the other hand, maintained positive sentiment throughout 2024, with minor dips in September and October, likely due to midterm stress. Comparatively, Harvard University experienced a sharper sentiment decline in mid-October, potentially linked to academic challenges, and also showed negative sentiment during June and July, possibly tied to stress from internships and summer research.

These patterns highlight disparities in student experiences and emphasize the importance of ongoing support. Schools that started the academic year with positive sentiment often faced challenges maintaining this positivity as the semester progressed. Our visualizations, including word clouds, sentiment trends, and interactive maps, made these findings accessible and actionable. For instance, sentiment score trends showed clear shifts across schools and months, while word clouds vividly illustrated evolving student priorities.

The insights gained from this project can be meaningful for both prospective students and university policymakers. For students, sentiment trends offer a deeper understanding of campus life, complementing traditional metrics like tuition and acceptance rates. For policymakers, these data points provide a real-time lens into student well-being, enabling proactive responses to issues such as mental health, housing, and academic stress during high-pressure periods.

In conclusion, our analysis underscores the potential use of Social Media sentiment data as a powerful tool for understanding student experiences. It not only captures authentic, real-time insights but also supports data-driven decisions to enhance campus life. Future work could refine the analysis by incorporating city-level crime data, expanding sentiment models for greater accuracy, and integrating additional datasets to provide a more comprehensive perspective.

Limitation & Bias

Although there are many good findings, there are also limitations within this project. In terms of the scope of data, we only scraped university Reddit posts from 50 schools, which is a relatively small sample size, and could lead to bias since they are all Top 50 US News Ranking US schools and might share some common features. Also, for each university's Reddit website, we are only able to get between 700 to 900 posts due to limitations from API. This leads to the variance in the time range of posts. For instance, for public schools like UWM and UC school, the posts cover a very short period of only three months. However, for private schools like Wake Forest University, we have data for over ten years. This large variability in the time range of posts makes it harder to compare sentiments. Together, our sentiment comparison page's line plots in the shiny dashboard could only be limited to cover data in the year 2024. Also, we realized that when scraping the crime rate data, it is only at the state level, which misses city-specific details; and weather data uses state averages, making it less precise to more specific locations.

In terms of the methods, the tools that we used to do sentiment analysis, *afinn* package in R, is struggling with complex or unclear posts, resulting in many zero scores—only 28,366 scores from 43,514 posts. Additionally, the current machine learning models we use might miss important factors beyond weather, tuition, graduation rate, and standardized scores, affecting performance. Word Clouds also need to be tuned better to improve relevance. Addressing these issues with better data and tools can improve future studies.

Future Work

This project establishes a foundation for analyzing sentiment trends across universities, but several extensions can enhance its impact. Refining sentiment analysis with advanced models like BERT or RoBERTa can improve accuracy, especially for complex sentiments. Expanding data sources to other platforms and integrating university surveys or mental health reports can provide a broader context and validate insights.

Additional features, such as demographic filters and predictive analytics, could make the dashboard more targeted and actionable. Enhancing visualizations with tools like heatmaps or time-series graphs would improve accessibility. Upgrading to real-time analytics could allow universities to address emerging issues promptly. Collaborating with university stakeholders could ensure the findings translate into meaningful interventions for student well-being.

Additionally, in order to better deal with missing data, we think about other ways to impute NA values instead of just transforming them into 0. We tried Generative Adversarial Networks (GANs) to address the challenge of incomplete sentiment data by predicting the missing sentiment scores. In our dataset, the GAN model demonstrated clear improvements over 20 epochs. The discriminator's ability to identify real data (D Real Scores) remained consistently high, while its ability to detect fake data (D Fake Scores) improved steadily, indicating the generator was producing more realistic outputs that were increasingly difficult to distinguish. The generator's quality (G Generated Scores) also improved significantly, as evidenced by closer alignment with real scores in later epochs. Furthermore, the loss values for both the generator (G Loss) and discriminator (D Loss) decreased steadily, showcasing effective gradient descent and stable training. Despite these promising results, we decided not to use GAN-predicted sentiment scores in our final analysis due to remaining concerns about consistency and reliability. This remains a key area for future work, where additional fine-tuning and validation could enhance GAN performance and ensure its integration into sentiment analysis pipelines.

7. Data Sources:

- https://www.reddit.com/
- https://www.collegetuitioncompare.com/best-schools/us-top-100/
- https://www.scouting.org/resources/los/states/
- https://www.currentresults.com/Weather/US/average-annual-state-temperatures.php
- https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_violent_crime_rate
- https://www.collegekickstart.com/blog/item/u-s-news-world-report-posts-2025-college-rankings

8. Appendix

Dashboard link: https://jiadili.shinyapps.io/FlexDashboard/