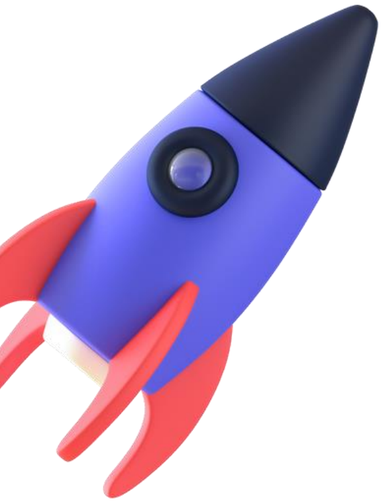
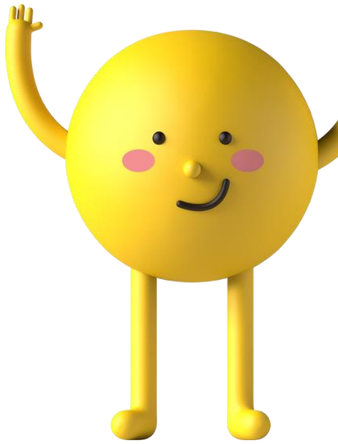
STATISTICAL PROGRAMMING

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



CAT GPT



PRESENTER

Jiadi(Jady) Li, Meishu Zhao, Roujin An, Zhenlong Zhang

Outline

- Motivation
- Data Collection
- Data Analysis
  - Machine learning
  - Objected-Oriented Paradigm
  - Functional Programming
- Dashboard Demo

Motivation

- Traditional resources focus on basic university metrics (e.g., location, tuition, ranking) Missing Student Experience (eg. academic & social support)
- Capture **authentic**, **student-centered insights** by analyzing sentiment from university Reddit posts

Target Audience

1. *Prospective Students*: Gain a clearer picture of campus life to choose the best-fit university
2. *College Policymakers*: Understand students' satisfactions and challenges, make informed adjustments

Data Collection

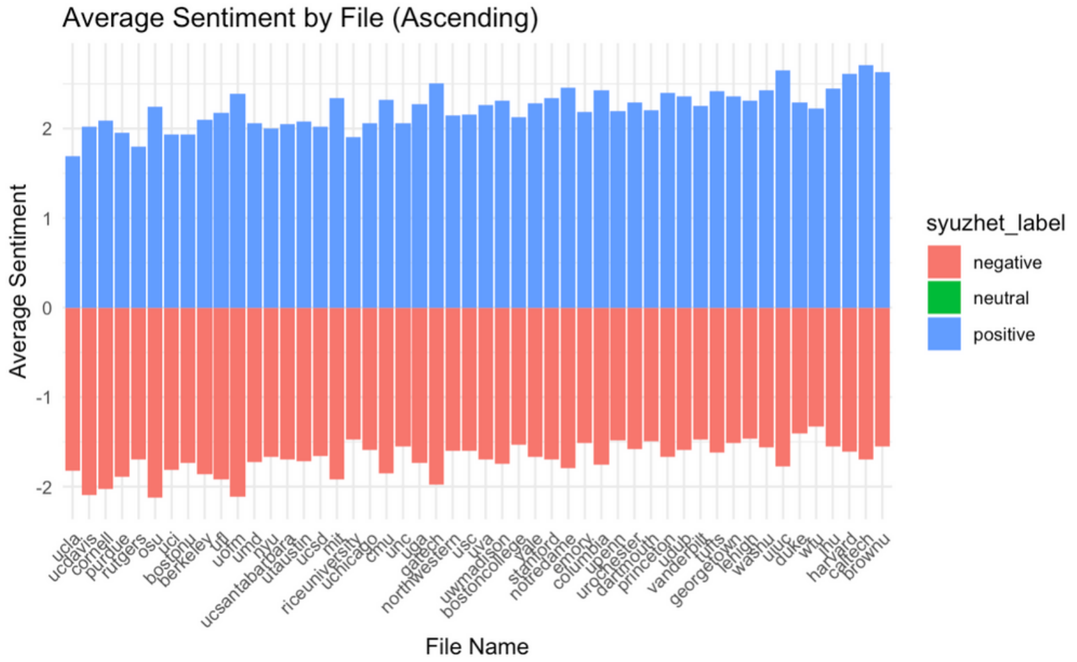
- **Data Sources:**
  - Reddit API
  - Supplemental Data:
    - University stats from US News.
    - Crime rates from FBI.
    - Weather data from Current Results.
- **Challenges:**
  - Reddit API limitations (700–1,000 posts per school)
  - Dynamic content scraping required Python `praw` & R `rvest`
  - Preprocessing: Handling missing values and aligning formats across datasets

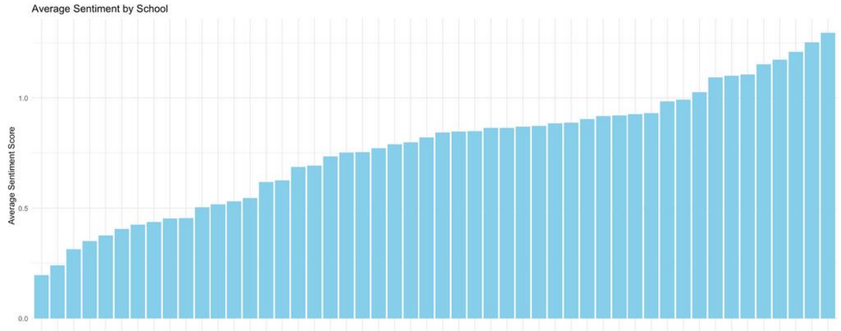
Data Analysis

Machine Learning

- To get more comprehensive data for ML model training, we used syuzhet package to get overall sentiment scores for each post of each university

- We are interested in the relationship between crime rates, weather etc. and overall university sentiments; And the possibility that the if the sentiments of universities are predictable with other factors such as SAT scores.



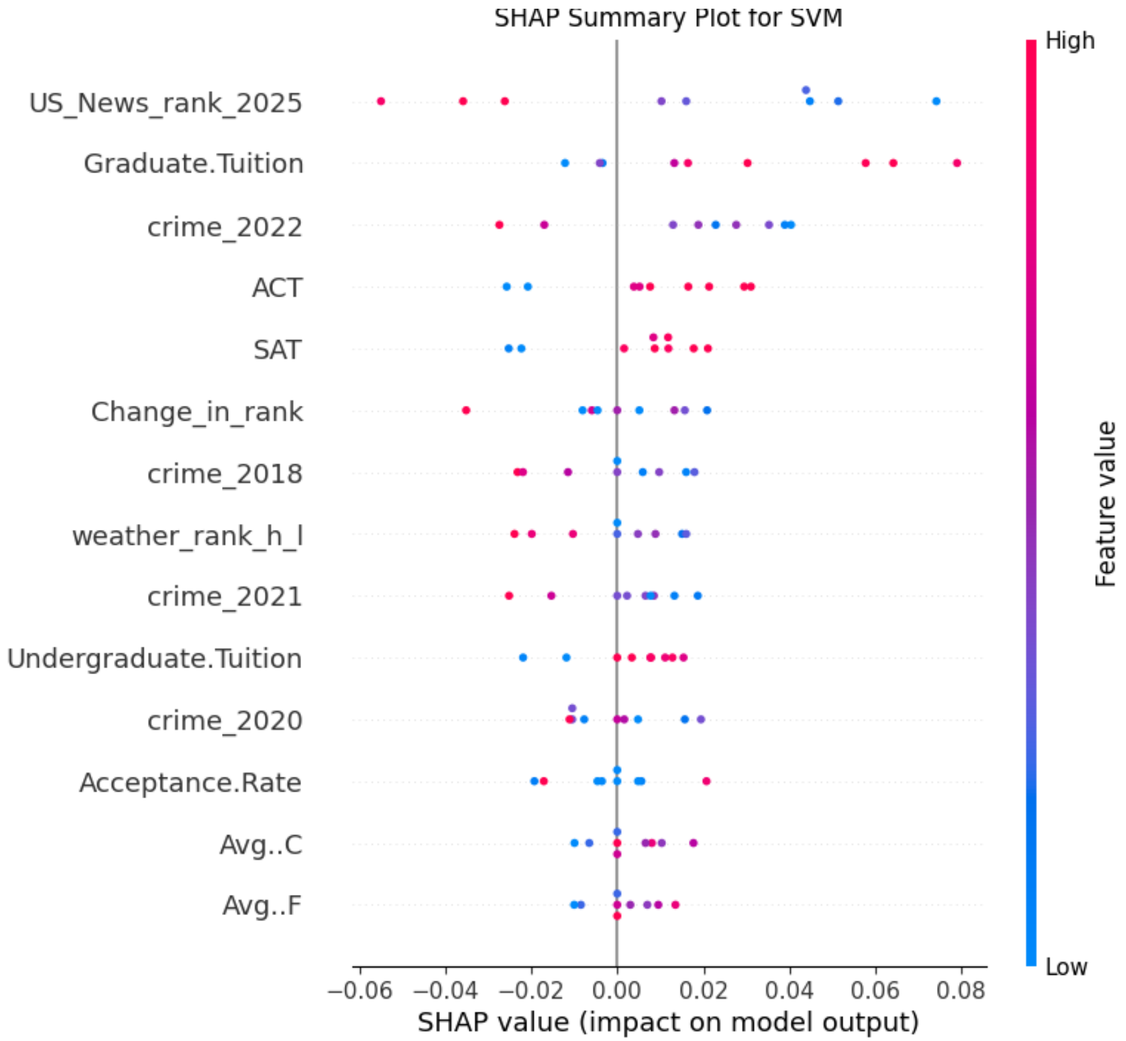


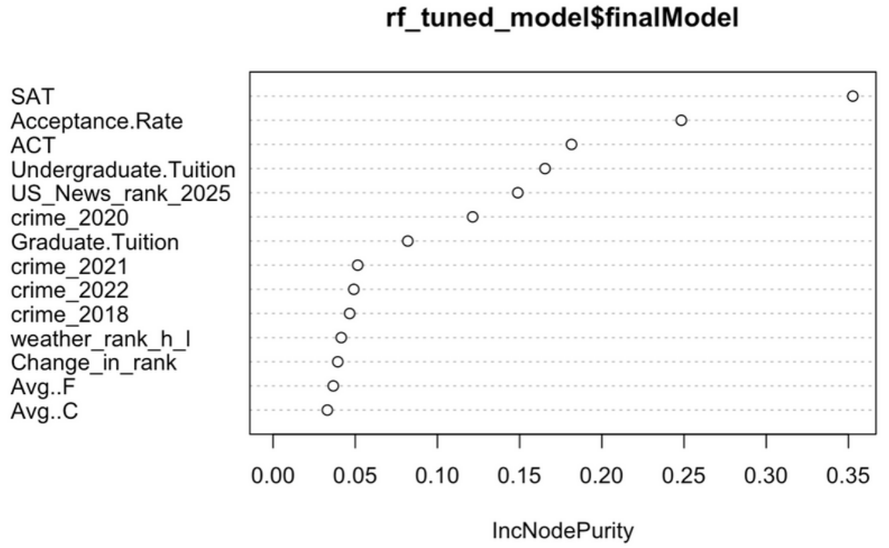

```python
param_grid = {
    'C': [0.1, 1, 10, 100],
    'gamma': ['scale', 'auto', 0.1, 0.01],
    'kernel': ['rbf', 'linear']
}
svm = SVR()
grid_search = GridSearchCV(svm, param_grid, cv=5, scoring='neg_mean_squared_error', n_jobs=-1)
grid_search.fit(X_train_scaled, y_train)

# Step 6: Get the best model
best_svm = grid_search.best_estimator_
print("Best Parameters:", grid_search.best_params_)
```

Model	Mean Squared Error (MSE)	R^2 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

Predictors included Undergraduate.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.

Machine Learning-Key Factors





SVM got best prediction results

Best Parameters: {'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}

SVM Mean Squared Error (MSE): 0.064

SVM R^2 Score: 0.15

Data Analysis

Object Oriented Programing -School Class

- **Attributes:**

A structured representation of each university with attributes: ***name, city, state, SAT, ACT, acceptance rate, graduation rate, undergraduate tuition, graduate tuition, student population, crime rate average, average weather, and weather rank***.

- **Methods:**

1. **initialization**: Populate attributes for each university.
2. **calculate_crime_rate**: Compute the average crime rate from 2018 to 2022
3. **to_list**: Convert class attributes into a list format for dashboard integration.

Data Analysis

Functional Programming

Prepare Reddit post data from all schools for sentiment analysis

- Approach:
  - Create a function to:
    - Read each file
    - Add school name
    - Split time into year, month, and day.

- Outcome: Apply the function to all 50 files with map() and combine results into one dataset

Handling Missing Values

- Approach:
  - Compute sentiment scores
  - Replace NA values with 0 using map()
  - Convert results back to a dataframe

- Outcome：Ensure a clean dataset ready for deeper exploration

Dashboard Demo

CampusLens Dashboard Link

Future Direction

Refine Sentiment Analysis

- Use advanced pre-trained natural language models for better accuracy.
- Improve handling of complex or nuanced sentiments.

Expand Data Sources

- Include other data sources to strengthen external validity.

Enhance Features

- Add demographic filters and predictive analytics.
- Use tools like heatmaps or time-series graphs for better visualization.

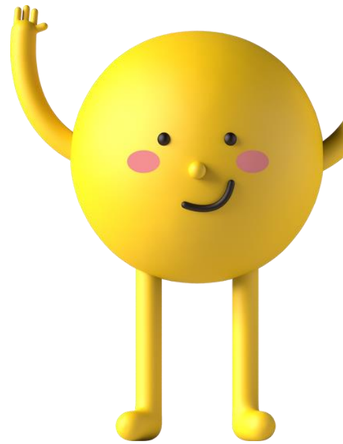
Upgrade to Real-Time Analytics

- Allow universities to respond to issues quickly.

GAN Improvements

- GAN showed promising results with high-quality outputs.
- Continue fine-tuning and validation to address missing data.
- Focus on integrating GAN-predicted scores into the dashboard.

Thank You !



GROUP: CAT GPT

Q&A Session

Please raise your hand and provide your question.