GIDS Final Presentation

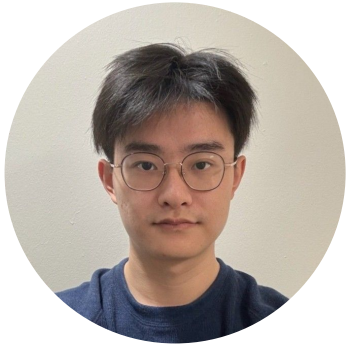
Francesco Colombo, Ajay Patel, Ruiyang Peng, Sen Liu, Yijie Bai

Project Members

Ajay Patel

Francesco Colombo

Ruiyang Peng



Sen Liu

Yijie Bai

Project Sponsor: Lisa Altman (GIDS)
Professor: Cantay Caliskan

Presentation Agenda

1. What is our project?
2. Timeline (Milestones)
3. Analysis Performed
    a. Exploratory Data Analysis
    b. Sentiment Analysis
    c. Predictive Modeling
4. Results
5. Challenges and Next Steps

Project Outline + Vision and Goals

Project Vision:

The Goergen Institute of Data Science is interested in extracting useful information from Statements of Purpose (SOP) to further understand admission decisions.

Project Goals:

Analyze the content of SOPs and determine whether there are correlations between the SOP's content and applicants' admission decision. We would like to be able to use the SOPs to tell a story about the person who wrote it. As a continuation create a predictive model to determine an applicant's admission decision based on the SOP. A further goal would be to achieve a model that can be applied to other departments.

Statements of Purpose

- Statements of Purpose (SOPs) are personal essays written by MIDS applicants to showcase their background, motivations, and interest in the program

As a millennial, I have constantly been fascinated by technology breakthroughs in computers, the internet and smartphones that literally reshape the world around us. In business, my long-term career goal was sparked by the fact that data has increasingly become a corporate asset and that the next generation's leaders are expected to have the expertise to turn data into business insights that will leverage corporate strategy and operation. With this in mind, in the long run I look forward to becoming a data-savvy decision-maker who drives "As-is" and "To be" business process at one of the global organizations.

To achieve this vision, upon graduation, I aim to become a data analyst in one of the leading consulting firms or financial institutions in the United States. In addition to building expertise in data analytics, this experience will also help me build my business sense and network on a global scale.

SOP Example Sentences

- “Challenging as it is, this project inspired my interest and passion in further exploring various applications of data analytics.” Demonstrate an **appreciation** for challenges and the **enthusiasm** found in the program.
- “Not satisfied with only dealing with numerical data, I started gaining interests in coping with text data.” Shows **motivation** to improve and explore more areas.
- “I asked myself whether I wanted to pursue a field that I enjoyed, one that would compensate well, or one that would allow me to make a difference in the world.” Shows deep **reflection** on future career choices.

Our Hypothesis

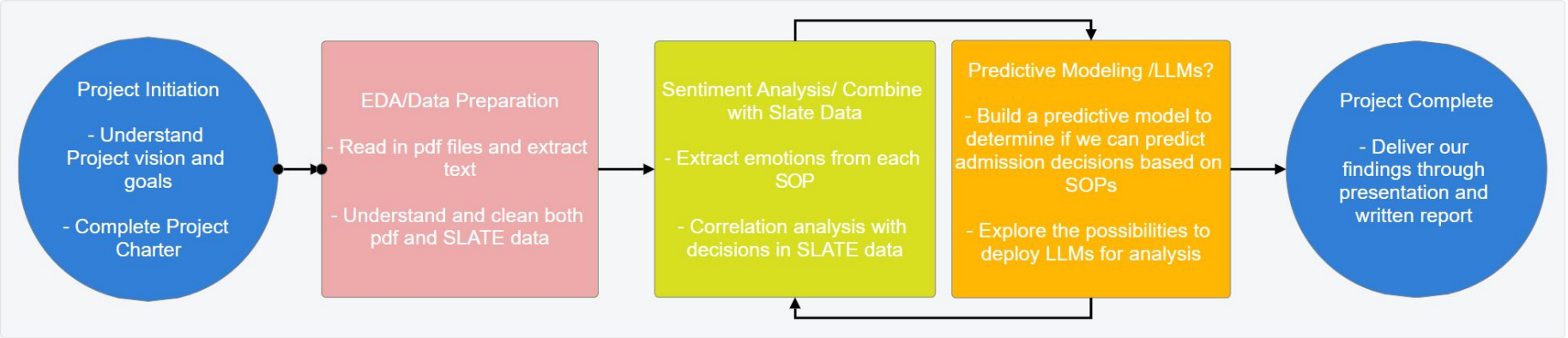
- Based on conversations within ourselves and our sponsor, we expected to see more desire and confidence among SOPs from admitted students
  - Denied students may not have as strong of a background, and that could lead to more anxious or dismissive emotions in their SOPs
- Prior research performed by Soleimani et al. found that while SOPs can be used to predict admission, no single feature will be valued highly and that the structure of the SOP (grammar, types of words used, etc.) mattered more than any extracted sentiment
  - This lowered our expectation for how much signal we'd be able to extract from the text

Milestones

Task	Date	Status
Create Project Charter	9/29/2024	Complete
Initial EDA on Slate	10/02/2024	Complete
Read SOP/ Extract test	10/06/2024	Complete
Sentiment Analysis	10/16/2024	Complete
Correlation Analysis of SOP and Slate	10/22/2024	Complete
Create Model to Determine Significance of SOP	11/06/2024	Complete
Finalize Analysis and Understand Results	11/18/2024	Complete
Final Presentation and Report	11/25/2024	In Progress

Table 1: Project Milestones

Project Pipeline



Technical Approach and Results

1. Data Introduction
2. Data Cleaning
3. Exploratory Data Analysis
4. Text Cleaning
5. Sentiment Analysis
6. Predictive Modeling
7. Results

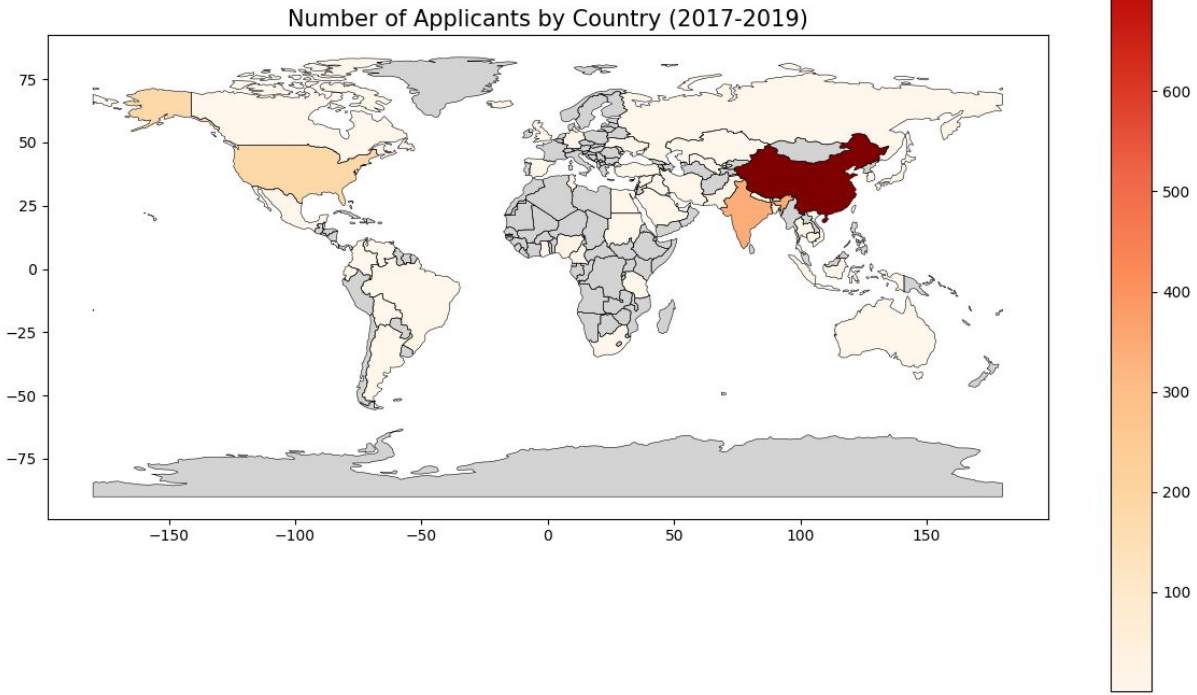
Introduction of our Data

- SOP Data (.pdf)
  - 1925 pdf files in a folder each containing an SOP
  - Each SOP has a reference number
- SLATE Dataset (.xlsx)
  - This dataset contains lots of information about the students applying to the Master's program (country, sex, major, grades, etc.)
  - It also contains information about admission decisions
  - A few unknown admission values (applicants who did not finish their application)

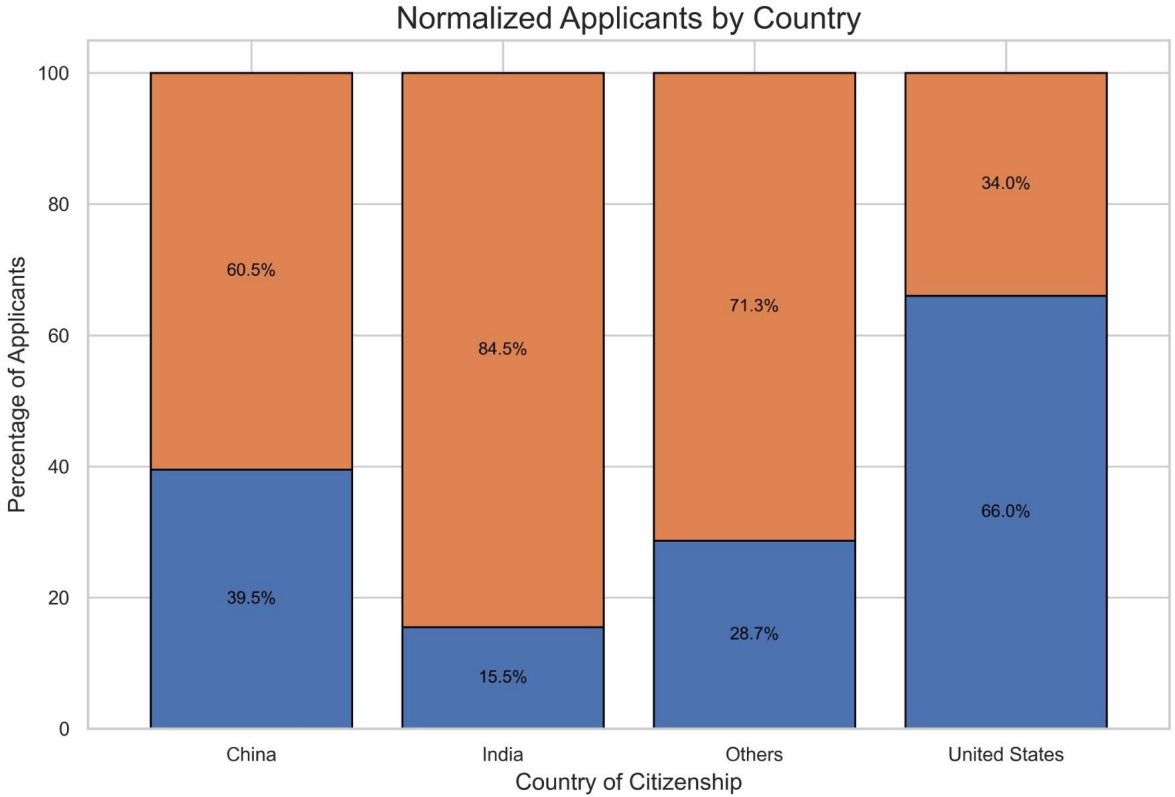
Exploratory Data Analysis

- Created visualizations on SLATE dataset and SOP text data to better understand the data.
- More than 90% of applicants are from China, U.S. and India.
- Most common major is Engineering, around 25%.
- Male applicants dominate
- Typical applicant is 20-30 years old with a strong academic background (GPA 3.0-4.0).
- Number of applicants increased over the course of 2017 to 2019.

Countries Applicants Are Born In



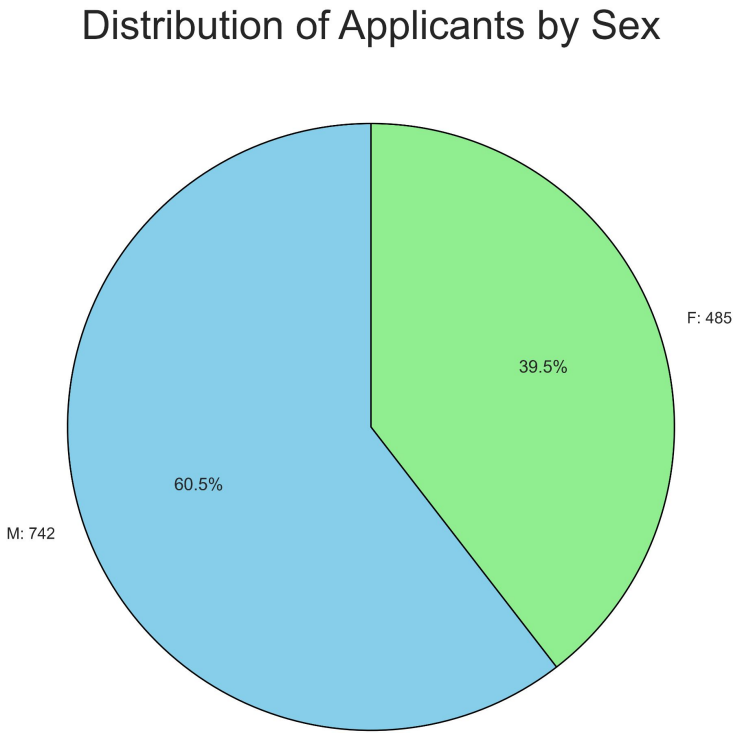
Admission Statistics



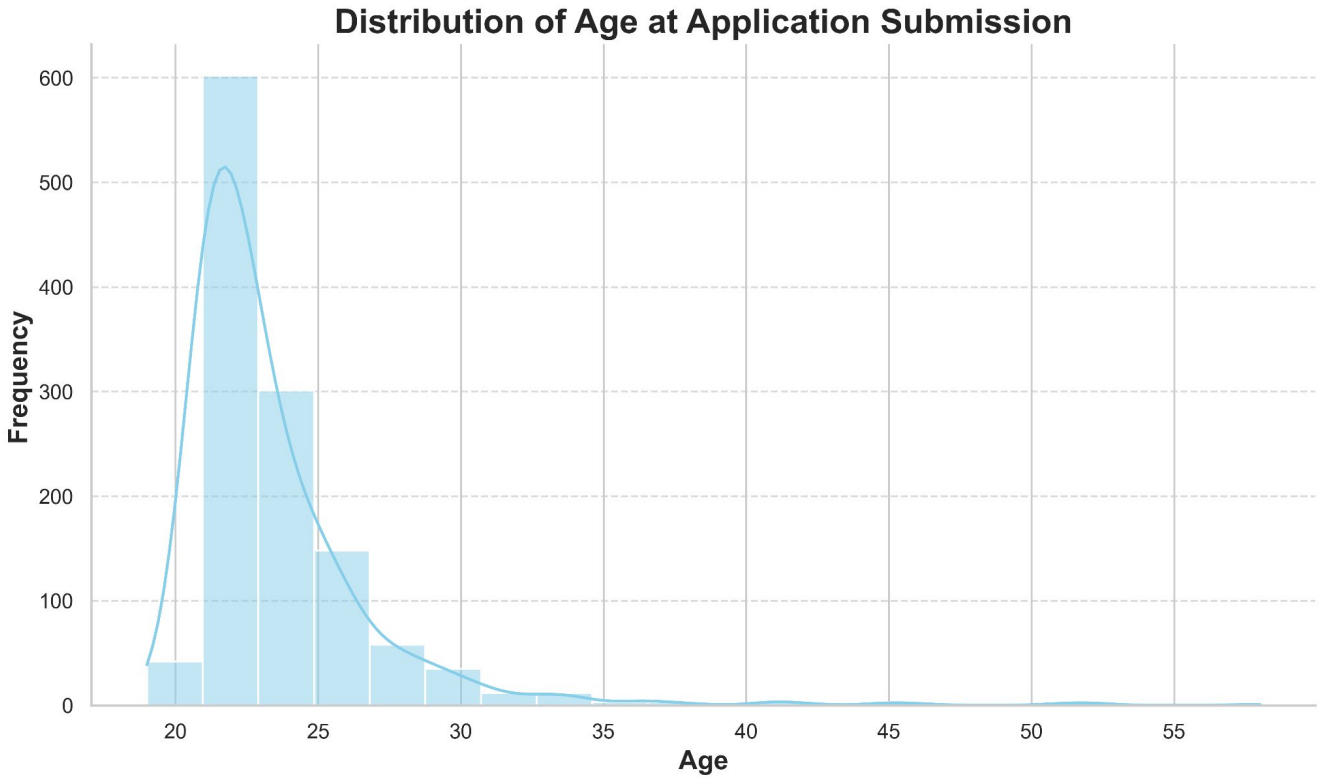
	China	India	Others	US
Admit	293	41	33	70
Deny	448	224	82	36
Total	741	265	115	106
%Admitted	39.5%	15.5%	28.7%	66.0%

Table 1: Admission Statistics by Country

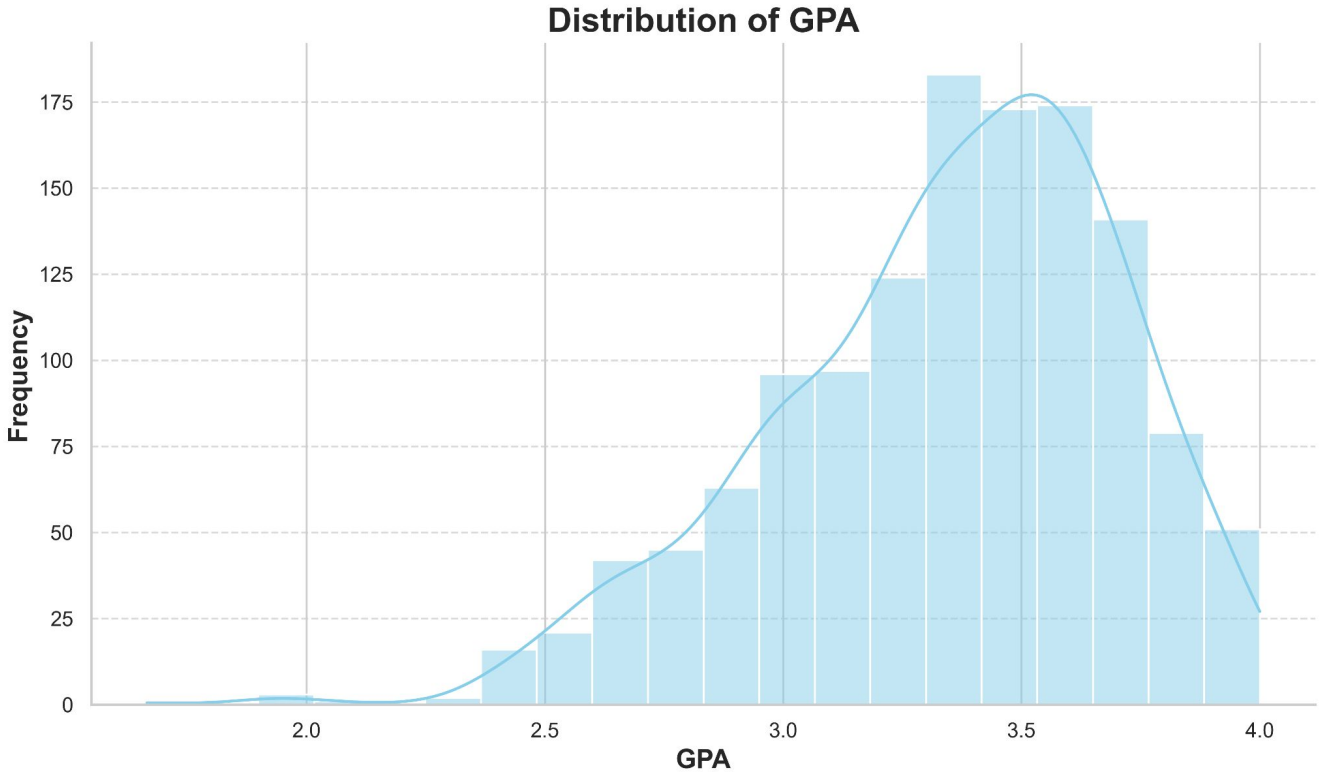
Applicants by Sex



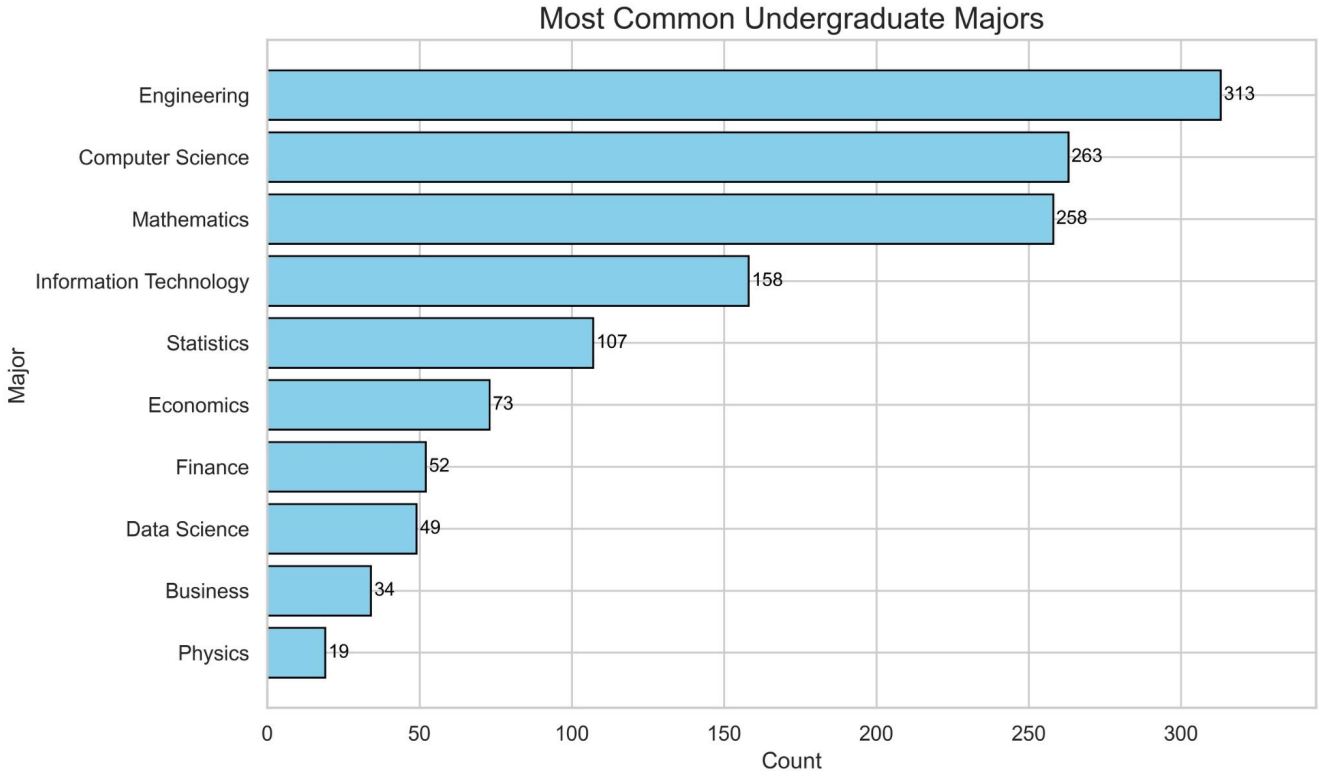
Applicants By Age



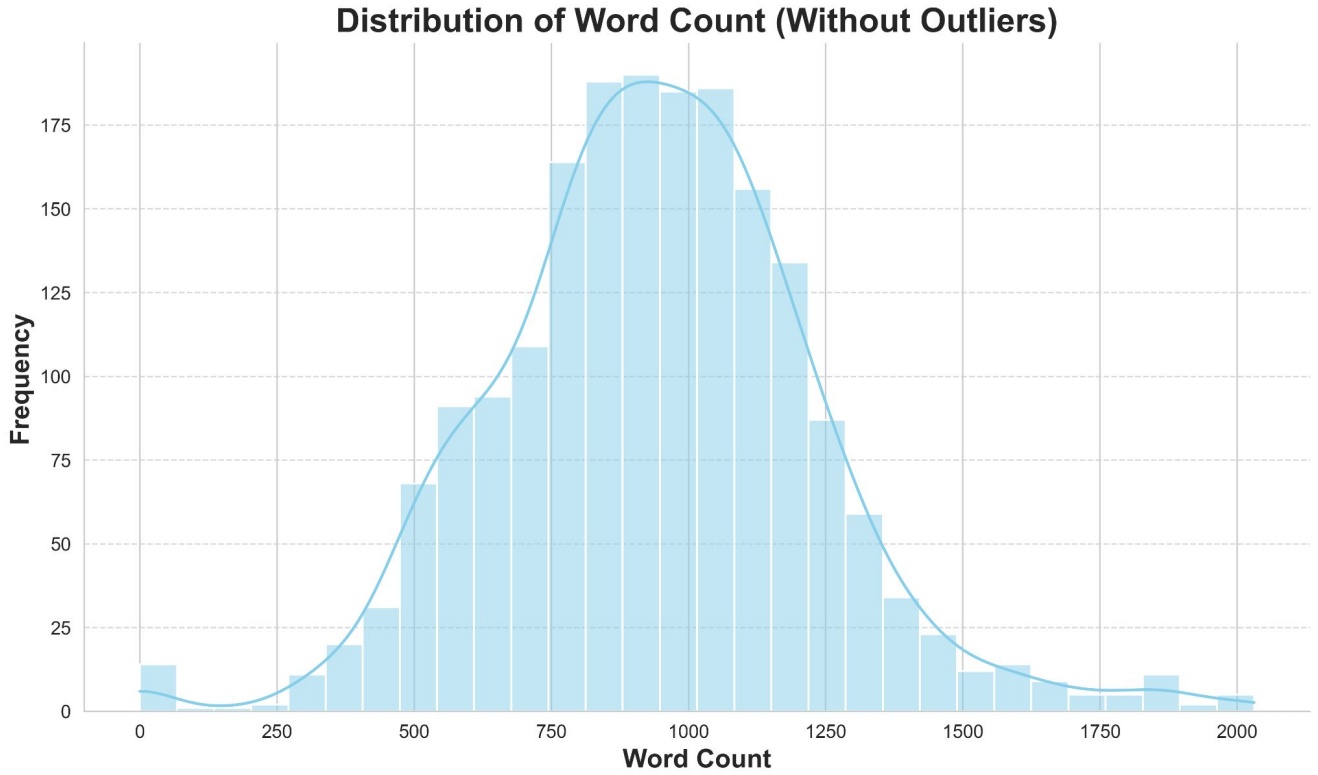
Applicants By GPA



Applicants by Major



Word Count in SOPs



Grammar and Spelling

The model we are using is the language tool in python with the language set to be US english

- To check spelling it compares each word to a large dictionary to check if it's different.
- For grammar checking it is rule based, where it checks whether a sentence matches some predefined rules in the model.

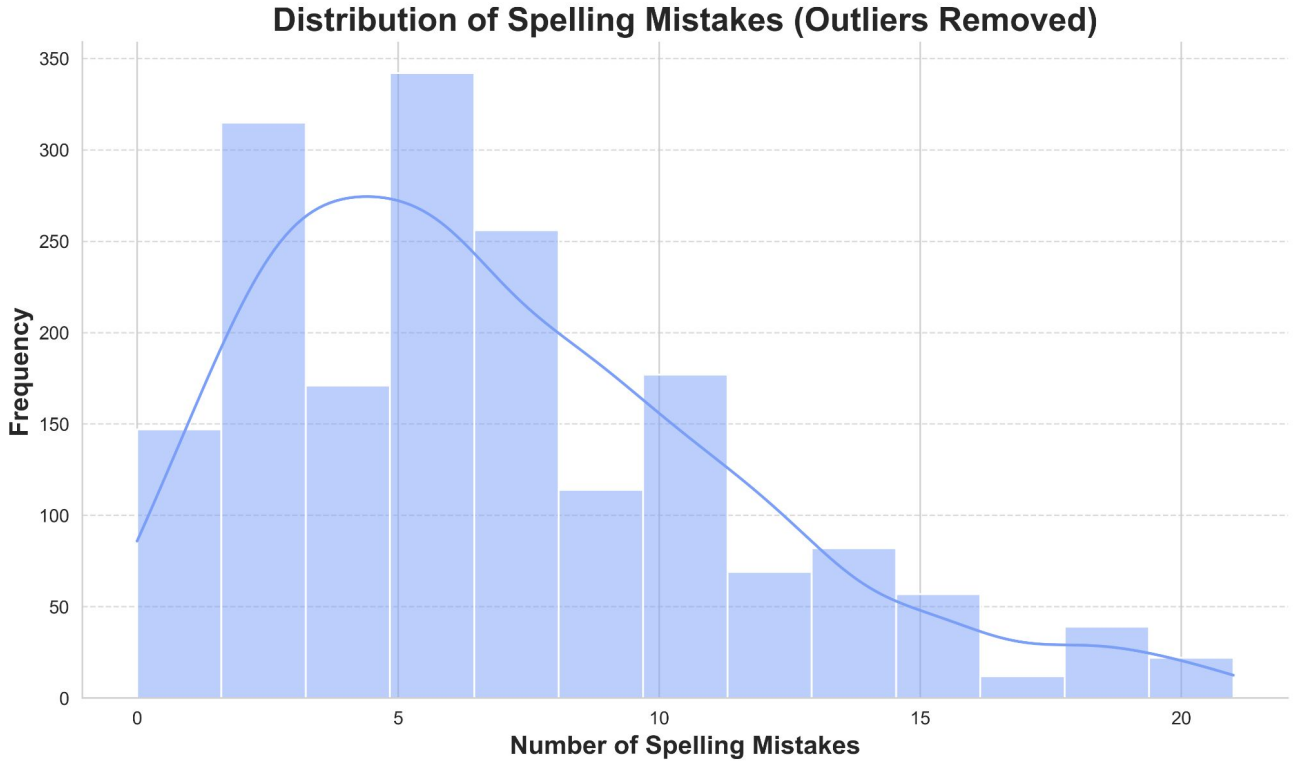
- Example #1: "This is a smple text with errrors."

  Model output: 2 spelling mistakes, 0 grammar mistakes.

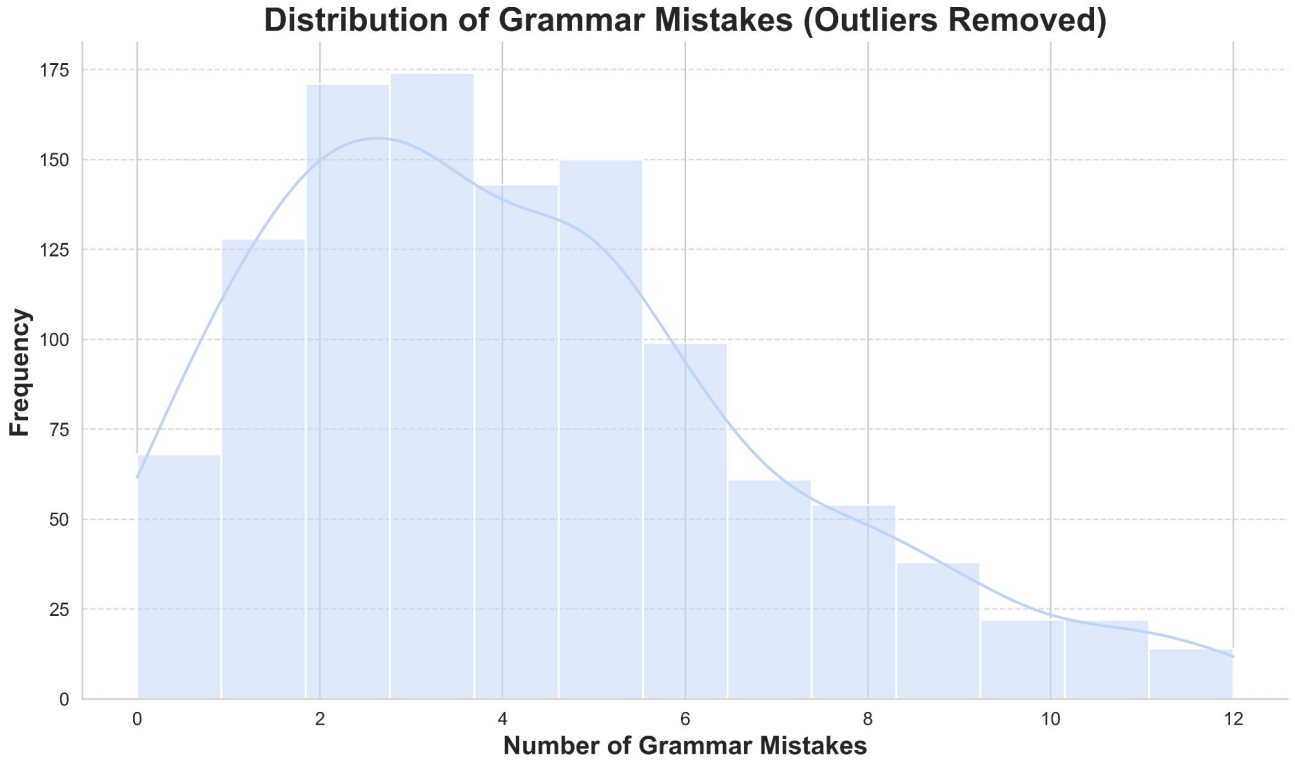
- Example #2: "He go to the store."

  Model output: 0 spelling mistakes, 1 grammar mistake.

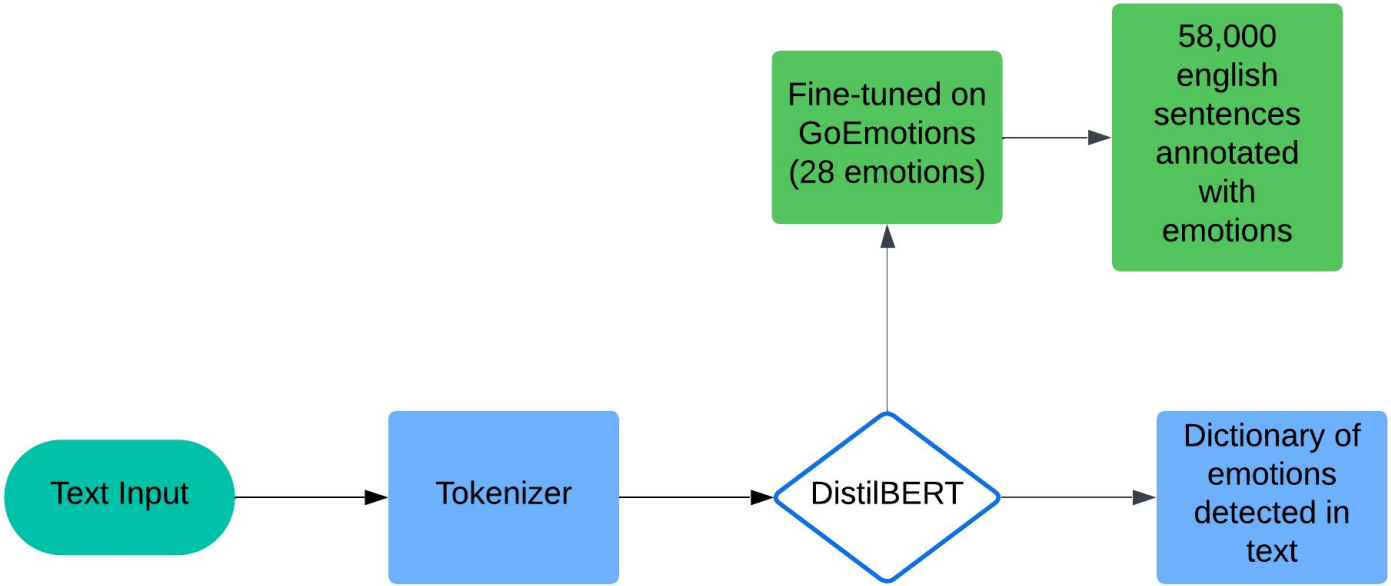
Spelling Mistakes

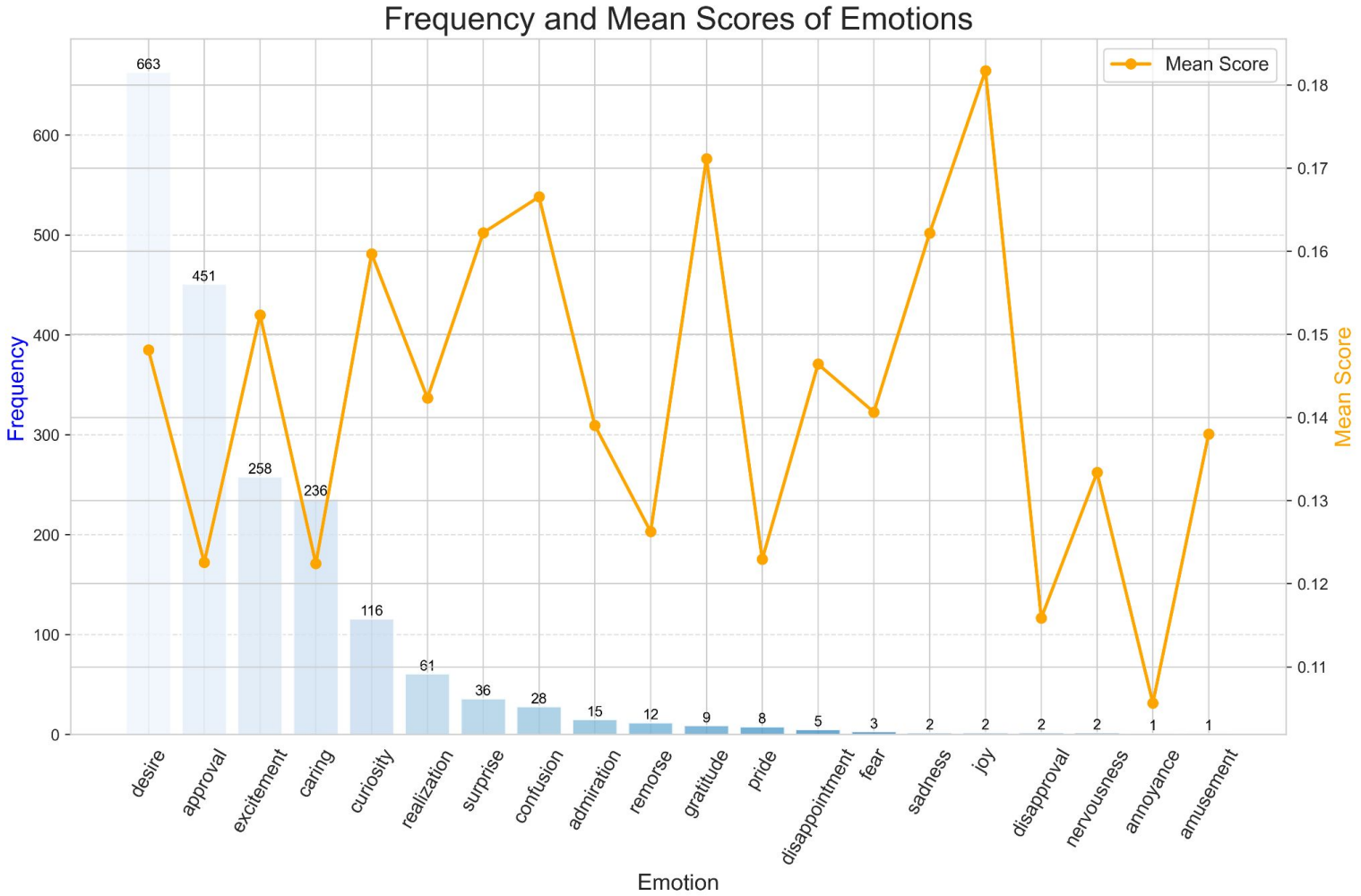


Grammar Mistakes



Emotion Analysis (DistilBERT)

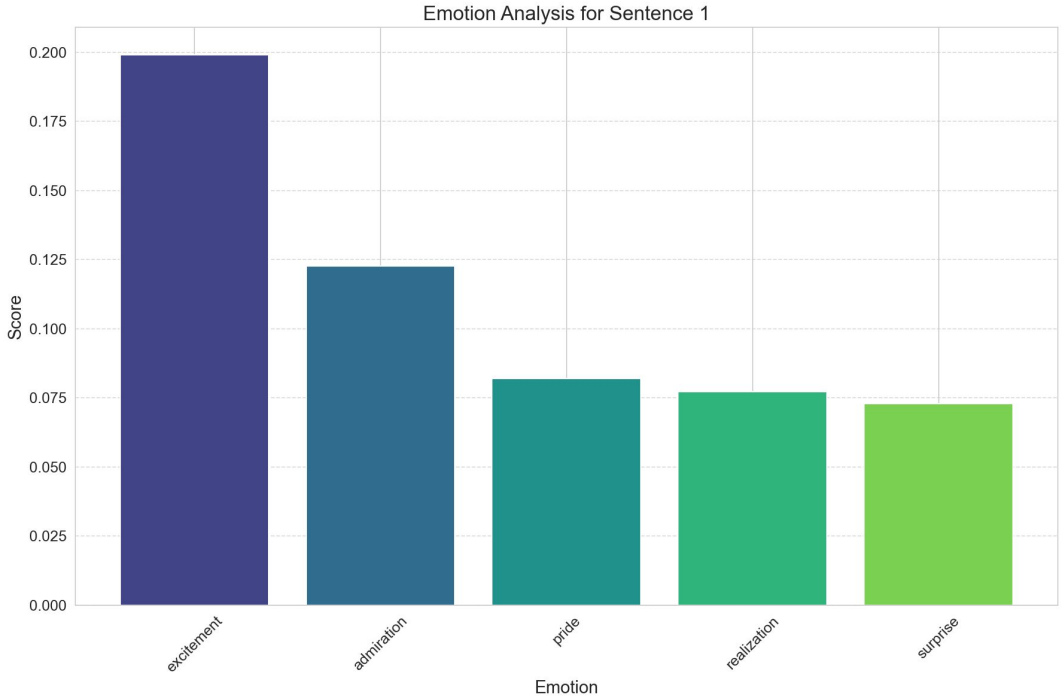



Frequency and Mean Scores of Emotions
Mean Score
Frequency
Mean Score
Emotion
663
451
258
236
116
61
36
28
15
12
9
8
5
3
2
2
2
2
1
1
desire
approval
excitement
caring
curiosity
realization
surprise
confusion
admiration
remorse
gratitude
pride
disappointment
fear
sadness
joy
disapproval
nervousness
annoyance
amusement

Emotion Analysis Example #1

Sentence #1: “The Goergen Institute of Data Science is the greatest place on earth!”

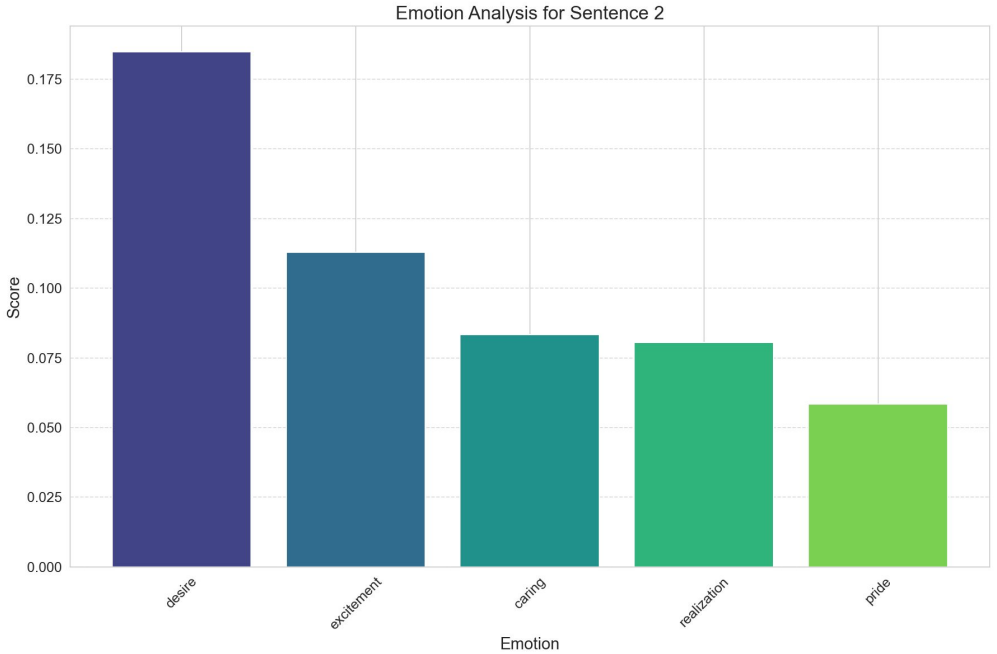
Emotion	Score
Excitement	0.1990
Admiration	0.1227
Pride	0.0820
Realization	0.0771
Surprise	0.0730



Emotion Analysis Example #2

Sentence #2: “According to my passion for statistical modeling and calibration, I am applying for a Master’s degree in Data Science so that I can pursue my life-mission discovering better methods in data modeling and forecasting to make decisions, performing better explanations on given statistical data.”

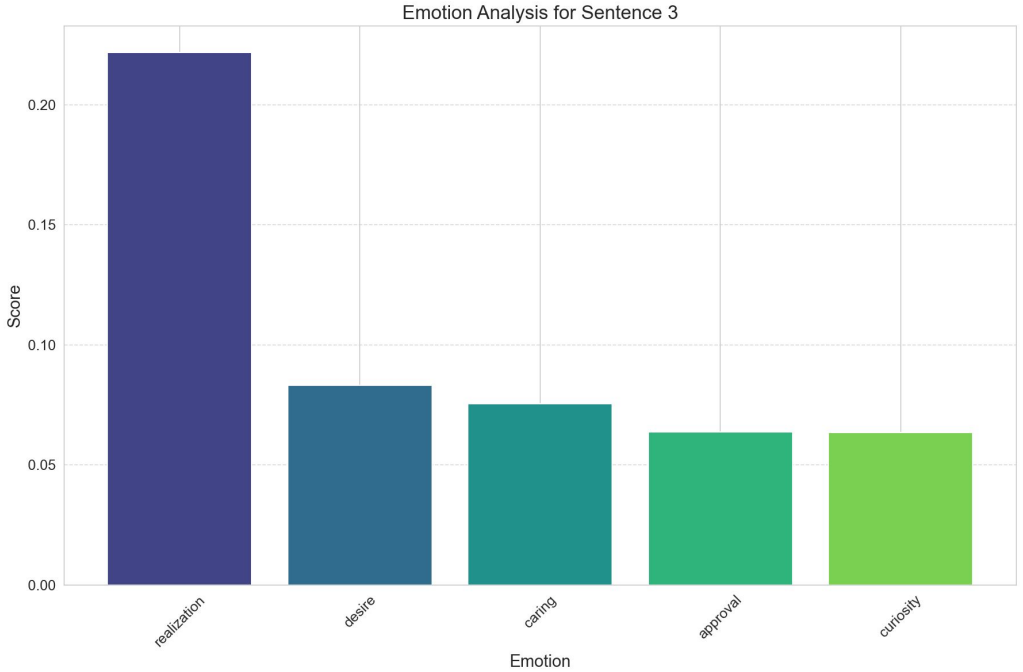
Emotion	Score
Desire	0.1848
Excitement	0.1130
Caring	0.0834
Realization	0.0806
Pride	0.0585



Emotion Analysis Example #3

Sentence #3: “During my undergraduate formation, I recognized data science, and in particular, machine learning, as a promising tool to tackle challenging and complex issues while being able to apply its results to the real world.”

Emotion	Score
Realization	0.2217
Desire	0.0833
Caring	0.0756
Approval	0.0637
Curiosity	0.0635



Correlation Analysis

- Analyzed correlation between emotions and key demographic variables to see if there were any existing relationships
- Correlation values were not too strong on the whole
- Older applicants tend to show more remorse and nervousness in their SOPs

Variable 1	Variable 2	Correlation
Remorse	Age At Submission	0.1
Nervousness	Age At Submission	0.1
Gratitude	Sub-Category: Health and Biomedical Sciences	0.1

Predictive Models

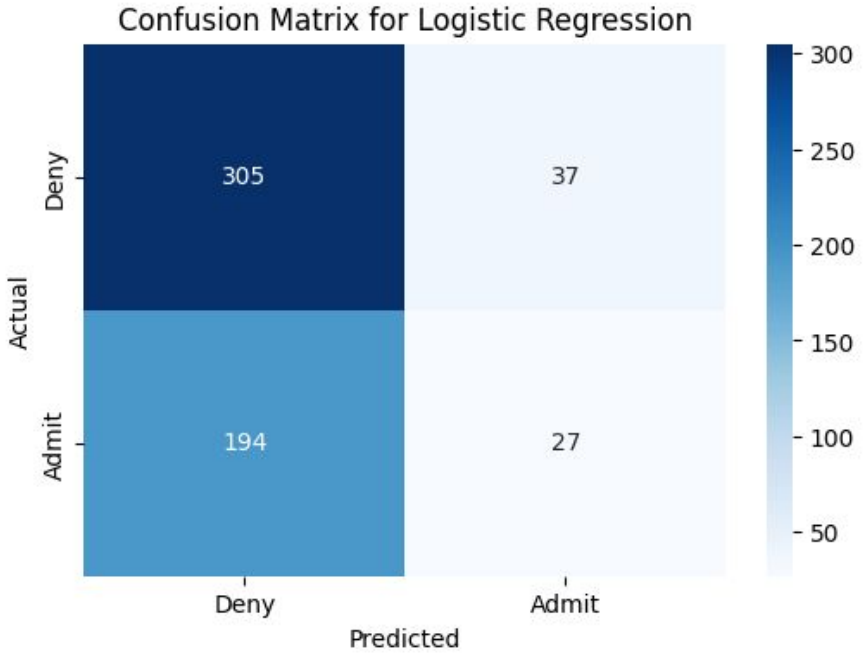
- We have run a variety of initial models for different types of predictions to better understand whether there is connection between the SOPs and the admission of students or the demographics (i.e. Citizenship)
  - First setup: prediction of admission with emotion, sentiment, grammar and spelling mistakes, and word count
  - Second setup: prediction of admission based solely on text data from SOPs
  - Third setup: prediction of citizenship based solely on text data from SOPs

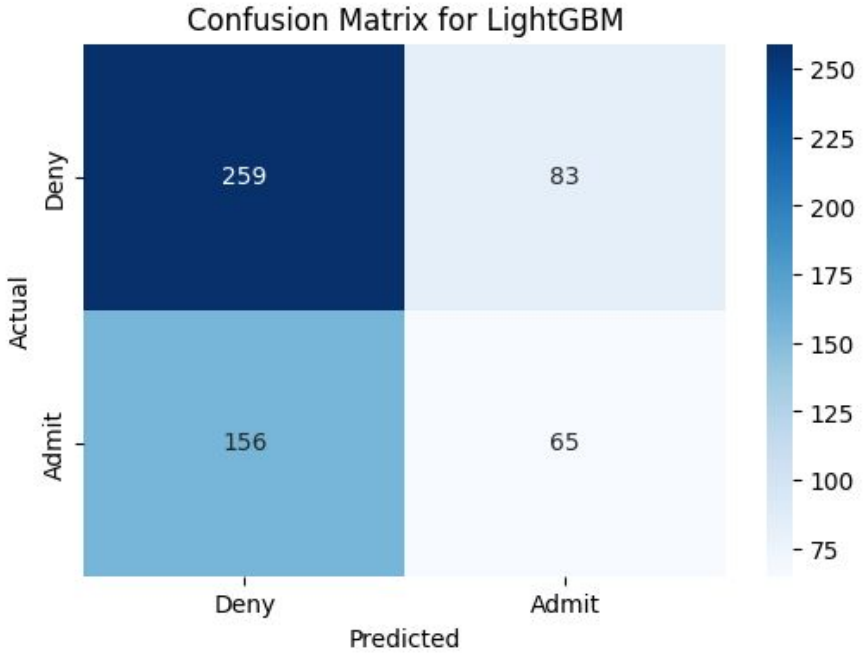
Setup 1 Results

- For Logistic Regression the accuracy on admittance is 0.12 and on denial 0.89. For Linear SVM it is 0.13 and 0.88 respectively
- The best performance on admittance though is from LightGBM with 0.29 and 0.76

Model	Accuracy
Random	0.5027
Logistic Regression	0.5897
Linear SVM	0.5879
Random Forest	0.5702
LightGBM	0.5755
XGBoost	0.5861

Setup 1 Results



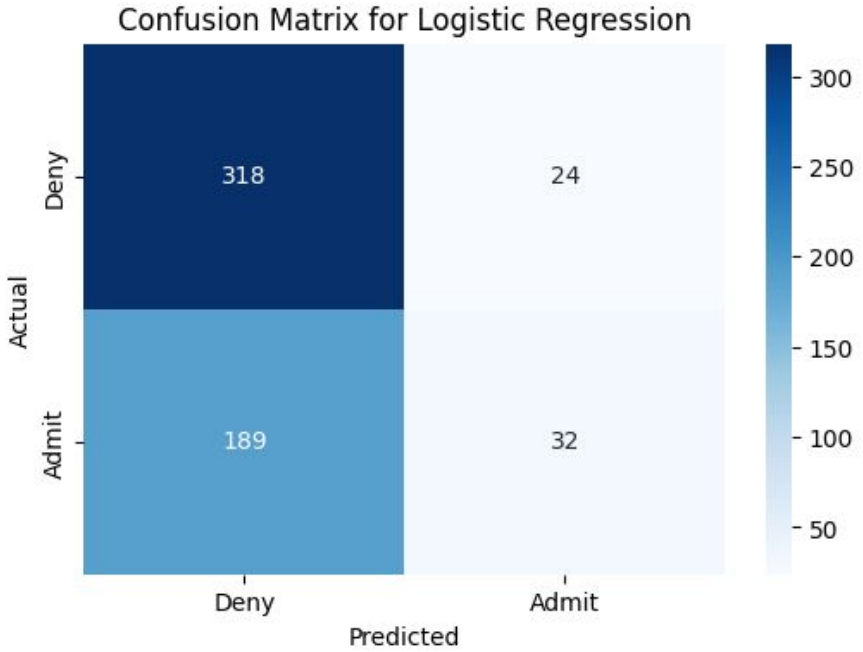


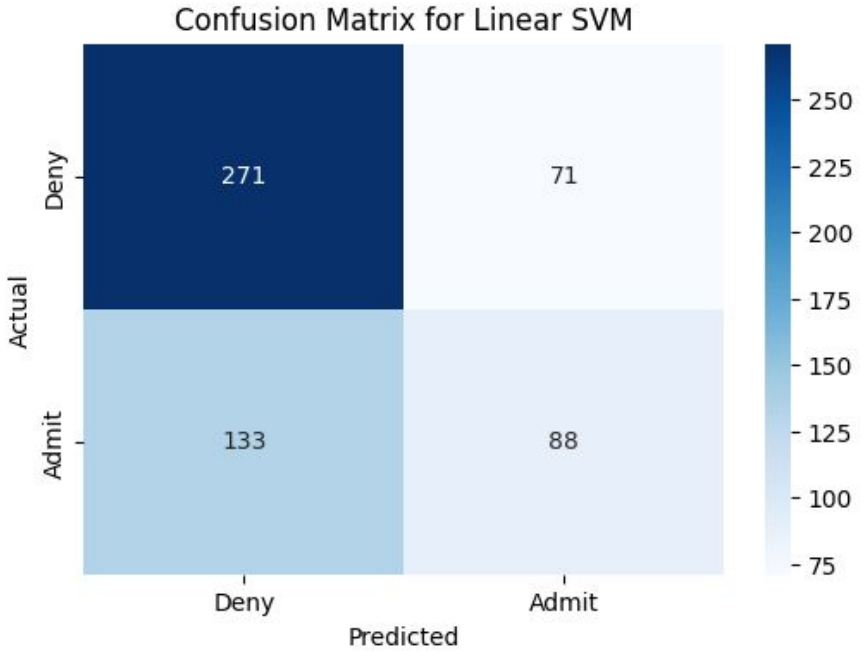
Setup 2 Results

- For Logistic Regression the accuracy on admittance is 0.14 and on denial 0.93. For Linear SVM it is 0.40 and 0.79 respectively which was the best performing model at predicting admittance

Model	Accuracy
Random	0.5027
Logistic Regression	0.6217
Linear SVM	0.6377
Multinomial Naive Bayes	0.6092
Random Forest	0.6128
LightGBM	0.6217
XGBoost	0.6128

Setup 2 Results





Setup 3 (Predicting Citizenship) Results

- The main issue here is still the class imbalance so results might not be very representative

Model	Accuracy
Logistic Regression	0.5648
Linear SVM	0.8721
Multinomial Naive Bayes	0.6767
Random Forest	0.6750
LightGBM	0.8273
XGBoost	0.8259

Results and Conclusions

- Overall from the emotion analysis portion of this project we did obtain results we were expecting as the majority of the emotions were positive, however the strength of the emotions detected was not as strong as we had hoped.
- The correlation analysis we ran showed us that there is not mot much that correlates the SOPs with the applicants demographics or admission.
- Finally, the predictive models did outperform the random prediction baseline, nonetheless the results were not strong enough to justify the prediction of admission of a student based solely on the SOP.

Challenges

- Our main challenge was connecting the SOPs to the admission status of an applicant
- When predicting the citizenship of an applicant one of the largest complications was caused by the class imbalance

Next Steps

- Validate our work
  - Do our results hold true with other departments? Other schools?
- Different approach: utilize tone or creativity analysis if possible for SOPs
- Tackle different projects
  - We are now the second group to have found little of substance from SOPs

Acknowledgements

We’d like to share our gratitude to Lisa Altman and the Goergen Institute of Data Science for the opportunity to work on this project.

We would also like to thank Prof. Cantay Caliskan for the support throughout the semester.

Thank you!