DEI: Exploring academic reflections using natural language processing to create a roadmap of student success and foster inclusive engineering education

Nidhi Raviprasad, Rajvir Harshvardhan Vyas, Dr. Zoe Wood (Co-PI), Dr. Sumona Mukhopadhyay (PI)
Computer Science & Software Engineering Dept., College of Engineering, Cal Poly

CAL POLY College of Engineering

Abstract

Every year, the College of Engineering (CENG) students and faculty reach out to admitted students through "Text-a-Thon" programs to answer their questions about being a student at Cal Poly. In order to improve CENG outreach efforts, we analyzed these text conversations to predict the likelihood of an admitted student accepting an offer of admission from Cal Poly. Through our research, we discovered key factors that play a role in a student committing to Cal Poly through data-based insights. Additionally, we successfully used a human-on-the-loop system to help create Machine Learning (ML) models that predict satisfaction of response by way of sentiment analysis.

Background

The dataset contains text messages received from applicants and messages sent by faculty and students as part of different communication campaign efforts.

Our goal was to find a way to predict whether an engineering student admitted to Cal Poly would accept their offer as per these text message conversations. As a team, we were tasked to parse through the dataset and clean it up to present information clearly for use in data analysis. We decided to employ a human-on-the-loop system and an automated focus as well. This allowed us the best of both worlds as we utilized advanced data science techniques to streamline analysis as well as spent dedicated time for human review to verify results. All in all, we split the project up into three distinct sections- cleanup, natural language processing, and predictive modeling.

Data Cleanup and Preprocessing

The dataset was contained in an Excel spreadsheet, containing identifiers for the admitted student contacted (Contact ID)and unique message sent, as well as which outreach campaign the message was a part of. It was organized by date and time of the message being sent or received.

Here, we checked to see the ratio of incoming messages (replies from admitted students) to outgoing messages (initial outreach by students and faculty) by outreach campaign:

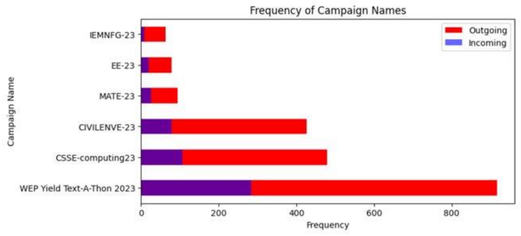

Figure 1: Frequency of Campaign Names

The ratio of incoming to outgoing messages is fairly proportional, with campaigns that reached out to more students experiencing more engagement. We also found the frequency of different majors mentioned in conversation:

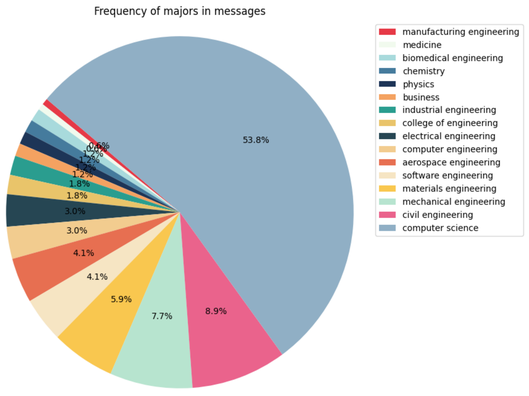

Figure 2: Frequency of Majors within Student Messages

We sorted these messages by Contact ID and then grouped conversations together, then separated conversations where the admitted student did not reply, as there was no conversation for us to analyze.

Natural Language Processing

Our first approach was to identify commonly occurring keywords to identify whether there was any association between the keywords mentioned in a conversation and the likelihood of a student accepting their offer.

We used Term Frequency and Inverse Document Frequency (TF-IDF), an informational retrieval algorithm that gives the relevance of terms in a corpus, from which we get the most common words and phrases across all conversations.

Figure 3: Most common keywords from the conversations

While this gave us an insight into the questions admitted students had, we could not find a direct correlation between specific keywords and the likelihood of an offer being accepted.

Predictive Modeling

We decided to try a different approach: predict the likelihood of accepting the offer directly from the conversation. We conducted a full manual review of the dataset, assigning satisfaction scores from 1-5 based on the following guidelines:

Counts	Satisfaction Score
92	5 (decided to commit/about to commit)
322	4 (a top choice/one or two doubts)
59	3 (still considering/waiting for other offers/unsure/uninished conversation)
9	2 (unlikely to attend)
60	1 (committed elsewhere)

Figure 4: Satisfaction Scores and Number of Conversations

Using our newly labeled data set, we split the data into two sections: whether the student was contacted by a faculty member or a current student, then found the average satisfaction.

	Count	Average Satisfaction (1-5)
Faculty	192	3.79811518
Student	352	3.601026
Total	544	3.695571956

Figure 5: Average satisfaction score for faculty versus students.

Interestingly, admitted students appeared to be more satisfied and likely to commit to Cal Poly when faculty, as opposed to current students, reached out. Figuring out why this is the case may prove to be helpful in the future.

Using our newly labeled dataset, we performed a training/test split for use in training sentiment analysis models. This way, in the future, we would be able to predict the satisfaction of an applicant in getting questions answered if Cal Poly continued operating their communications channels as is. We fed the data to supervised machine learning models such as Native Bayes Classifier, Random Forest and Support Vector Machine.

By generating classification reports that graded each model's accuracy, precision, recall, and f1 scores, we found the Support Vector Machine to be most relevant to our dataset.

```
Accuracy: 0.6697247706422018
              precision    recall  f1-score   support

           1       0.83      0.45      0.59        11
           2       0.00      0.00      0.00         1
           3       0.00      0.00      0.00        15
           4       0.63      0.98      0.77        60
           5       0.90      0.41      0.56        22

    accuracy                           0.67       109
   macro avg       0.47      0.37      0.38       109
weighted avg       0.61      0.67      0.60       109
```

Figure 6: Classification Report for Support Vector Machine Model

Outcomes

We were able to identify several factors that influenced an applicant's decision to attend Cal Poly such as types of questions they had, majors they were interested in, etc. We also observed from data-based insights effectiveness of campaigns and the effects of applicant's talking to faculty vs students.

There was also a solid foundation built for predictive analysis of applicant sentiments, which in conjunction to the insights would help inform an improved communication strategy with applicants.

Moving Forward

To improve the accuracy of our predictive model, we could utilize additional data sources such as Reddit posts or student demographics, so that we can identify other features that more accurately describe what applicants really care about.

This project can further be built upon to serve applicants through a dashboard or AI chatbot on the Cal Poly Admissions website that would help answer questions faster based on our observations.

ACKNOWLEDGEMENTS

This project was funded by CENG Dean's Club for Innovation DEI Funding.