# Mining College Subreddits for Unique Disclosure

**Dani Roytburg**
Department of
Quantitative Theory & Methods
Emory University
Atlanta, GA, USA
dani.roytburg@emory.edu

**Minh-Thy Tyler**
Department of
Quantitative Theory & Methods
Emory University
Atlanta, GA, USA
minh.tyler@emory.edu

## 1 Introduction

There are thousands of universities across the United States and each has a unique culture, system, and student body. One way that differences among colleges are expressed is through their corresponding subreddit. College subreddits contain submissions and comments that provide substantial language about the discourse and customs at the institution. As Emory students, we are curious about how our college experience and online discourse compare to other colleges geographically near us as well as to peers within the same tier. In particular, we seek to understand what concerns impact Emory students relative to other academic institutions. This research is important because it will provide insight into the similarities and differences among universities. In our project, we are comparing the subreddits of Emory, Georgia Tech, Georgia State University, and John Hopkins. We chose to include Georgia State and Georgia State University because they are big colleges near Emory. We chose John Hopkins because it is very academically similar to and in the same tier as Emory. Using PSAW: the PushShift API Wrapper for Reddit, we created our dataset of comments from submissions among the four colleges. We then use a topic modeling-guided analysis in order to determine how particular universities contribute to the language used in general social media discourse surrounding college life. In doing this, we look at the particularities of certain academic institutions to analyze the conversation surrounding particular topics. Further, we ran a sentiment analysis to quantify the emotional intensity of the comment in the college subreddits. We then average out and compare the polarity scores of the comments between each college subreddits. We expect to detect some evidence of polarity differences as well as unique topic specificity in the language used by each college. Most importantly, the results from this study will contribute to the understanding of the distinctive discourse among college subreddit communities.

## 2 Related Research

Melton et al. (2021) use public sentiment analysis and topic modeling regarding COVID-19 vaccines on Reddit. They collected data from 13 Reddit communities focusing on the COVID-19 vaccine from December 1, 2020, to May 15, 2021. The data was aggregated and analyzed by month to detect changes in any sentiment and latent topics. Although this article has a drastically different goal and topic from our project, we found it helpful as it uses a similar methodology (we are also analyzing different subreddits using topic modeling and sentiment analysis). This study investigates COVID-19 vaccine-related discussion and conducted a sentiment analysis and Latent Dirichlet Allocation topic modeling on textual data. In our project, we implemented an LDA topic model on the data we collected from the college subreddits. This paper's topic model revealed community members mainly focused on the side effects of the COVID-19 vaccine rather than conspiracy theories. Their polarity analysis allowed them to conclude that the 13 Reddit communities expressed more positive sentiments than negative regarding vaccine-related discussions. The interpretation of their results was helpful in regard to our interpretation of our results of the different topics and polarity scores we obtained from

running our methods on the college subreddits. Our research question contributes to studying general social media discourse, as it gives insight specifically into student and college experiences through the lens of Reddit.

We sought to examine topic modeling approaches to student writing, though unrelated to social media. In the 2013 paper submitted to the Conference on Empirical Methods in Natural Language Processing, an interdisciplinary team of psychologists and computational linguists from the University of Maryland utilized Latent Dirichlet Allocation (LDA) topic modeling to predict a student's propensity for depression given a stream of consciousness essay. This paper used Mallet, a long-standing Java library that implements vanilla LDA, in order to sort concerns and classify them in order of propensity for depression. This investigation also highlights rule-based methods that at the time represented the state of the art, such as the Linguistic Inquiry and Word Count (LIWC) categories. This uses a corpus of frequently used words (the size of this corpus changes and ranges in the thousands) and tags each word `with any number of 69 broadly constructed subcategories of speech, such as pronouns or swears words. The resulting many-to-many relationship between the 69 categories and elements of the glossary is a useful rule-based tool for classifying linguistic features. To be clear, many linguistic rules-based approaches have lost luster over time in the face of more flexible deep learning approaches. Nonetheless, the particular benefits of LIWC dovetail uniquely well with LDA modeling by providing cross-referential support for analyzing term usage in a body of text.

We also investigated more sophisticated methods than the vanilla distribution algorithms outlined in these two papers. One paper, published this past year in Future Internet outlines a sophisticated two-phase pipeline for analyzing student content analysis, presented in this (simply exquisite) diagram:

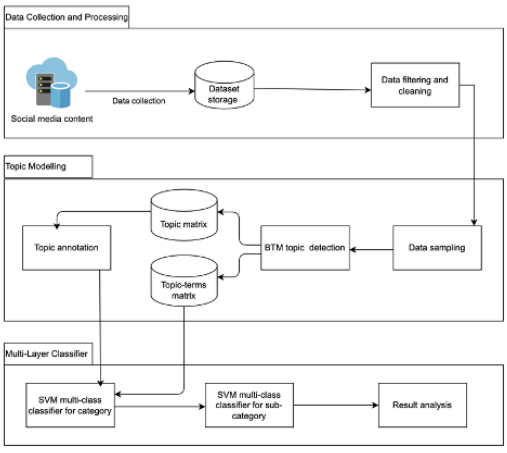



**Figure 2.** The structure of the method.

We might want to select particular subcomponents in this process for incorporation into a topic model. Using a biterm topic model, implemented in Python here, we might be able to provide better relational categorization of terms by focusing on two-term relations. This, along with one other survey of methodological approaches to topic modeling, should be consulted in addition to the initial LDA implementation. This much methodological attention is necessary; the

results from initial samples of Reddit modeling (both by our own work and in the literature review) seemed somewhat uncompelling, and as such we wish to find an approach that can draw clearer partitions between words.

**3 Dataset**

Our dataset composes of the subreddits r/emory, r/gatech, r/GaState, and r/jhu. We used the Pushift API Wrapper to collect the top 20,000 comments from each of the college subreddits which will be the primary component of our analysis. We put the comments into a pandas DataFrame that included columns corresponding with the comment's subreddit, author, id, score, and time posted (in UTC). From this, we created and downloaded our csv file. We are missing 287 entries, but the dataset constitutes 99.64% of the original query. PSAW keeps data on removed submissions (there are 1,164 in our case). The data fed into our model does not include those. During our data cleaning process, we decided not to remove spam or typos because topic modeling relies on term usage frequencies. Similar terms will pick up on differentiable topics and the broader topic approach's reliance on often-used vocabulary allows us to safely ignore that statistical nose. To define and process the data, we wrote two functions that we referenced from the class12-topic-modeling notebook. The tokenizing function iterate gensim's simple_preprocessing function which breaks the document into a list of lowercase tokens so that both uppercase and lowercase versions of the same word can be counted together. The iter_docs function iterates through our corpus and yields post['id'] and token.

**4 Process and Methods**

Throughout the process of creating our dataset and implementing our planned topic model method, we ran into many difficulties that we had to address. Originally we wanted to create our dataset by extracting the submissions of each of the college subreddits using PRAW: The Python Reddit API Wrapper. We did so by utilizing PRAW's "top" listings generator to procure the top 1,000 submissions from each subreddit. However, PRAW does not allow people to fetch more than 1000 items. Because of this, we would not have enough Reddit data to run our planned topic model. We backtracked and decided to use the PushShift API to obtain a larger dataset of subreddit comments specifically.

Topic modeling is a method that extracts themes or "topics" from substantial documents. Specifically for our project, topic modeling allows us to detect word patterns within comments and cluster the words into similar word groups that would best characterize them. To implement our topic model, we originally were going to use Mallet, but had difficulties connecting it to python. We decided to use the non-Mallet implementation of LDA topic modeling instead. Originally, we were going to structure our topics by university but concluded on working with a topic quantity that would return co-occurring terms. We decided on this in the case if each university held disjoint unrelated conversations, the topic model may have clustered five topics

per school. Having a large corpi would resolve that problem as it would create more interactions outside of a university domain. We decided to set our number of topics to twenty. LDA topic modeling is an unsupervised algorithm and is unaware of how the words it groups into topics connect to each other. So, after obtaining the twenty topics, we observe each of their meanings as later discussed in the results section.

After running the topic model, we ran a sentiment analysis to compare each college's polarity of subreddit comments. Sentiment analysis is a method of quantifying the sentiment of texts. Before running the sentiment analysis, we first converted the comments in the column of our csv file into type string. Then we pre-processed the data by lowercasing, which helps the process of normalization. We also removed the stop-words, as they have no predictive power. After doing all of this, we used the Natural Language Toolkit (NLTK) tool—VADER— to generate positive, negative, and neutral sentiment scores and the compound score for each comment (we added each score to its own column in the data frame). From this, we could use Python's groupby function to group the averages of each score by college subreddit in an easy-to-read graph. We go further in exploring these results below. We use a simple mathematical function to compute the relative propensity of the 20 topics to be either positive or negative.

**5 Results and Discussion**

Our model yielded the following 20 topics to characterize our dataset:

```
T0: thank, thanks, yes, good, yeah, ll, ur, luck, okay, waitlist, definitely, register, interested, advice, know,
T1: major, cs, courses, classes, research, med, science, pre, math, want, engineering, majors, bme, premed, course,
T2: food, building, library, open, floor, dining, center, th, buy, minutes, card, use, plan, meal, hall,
T3: gt, gsu, program, apply, application, transfer, applied, student, admissions, gpa, process, school, accepted, year, college,
T4: jhu, students, like, school, think, people, schools, better, high, college, faculty, emory, probably, things, especially,
T5: removed, sorry, post, link, comment, contact, message, account, ii, questions, bot, automatically, subreddit, portal, subject,
T6: use, app, google, man, stop, phone, blue, line, dude, gotta, damn, laptop, number, posts, budget,
T7: neuro, covid, problem, sis, cell, ppl, watch, guess, virtual, notes, peta, follow, section, design, wondering,
T8: got, email, know, lol, pm, free, parking, im, dm, hey, ok, let, sure, send, wait,
T9: oh, ams, charles, language, yep, enjoyed, hoping, desk, hell, ga, force, meant, cafe, favorite, air,
T10: health, information, op, mental, police, care, game, literally, anymore, mean, following, safety, fair, appointment, agree,
T11: classes, class, semester, time, year, taking, took, lab, summer, credit, fall, credits, ll, think, freshman,
T12: days, weeks, lmao, week, apartment, street, day, bad, ago, true, early, like, word, words, water,
T13: https, edu, amp, com, deleted, www, reddit, info, comments, check, page, services, org, website, studentaffairs,
T14: people, life, tech, social, join, clubs, love, group, club, want, fun, like, campus, groups, team,
T15: campus, live, housing, baltimore, walk, area, living, safe, park, city, homewood, month, home, car, place,
T16: people, like, ve, good, time, know, lot, friends, think, things, going, feel, heard, person, room,
T17: class, professor, like, good, think, course, grade, work, exams, easy, hard, pretty, professors, questions, exam,
T18: hopkins, school, research, work, job, grad, experience, undergrad, good, working, years, getting, know, phd, time,
T19: pay, need, student, money, help, aid, office, advisor, financial, hope, able, students, paid, gsu, ll,
```

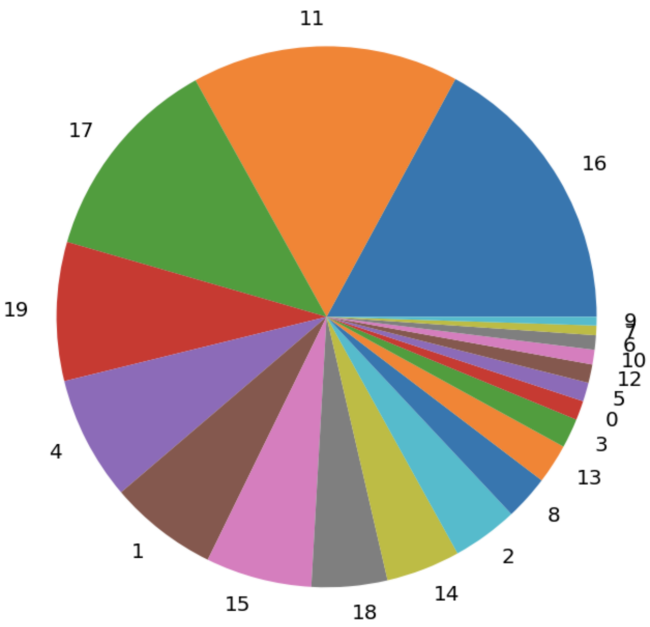
We analyzed the words in relation to each other and classified each topic as:

Topic 0: Class Registration
Topic 1: STEM Programs of Study
Topic 2: Campus Dining
Topic 3: College Admissions and Transfer Process

Topic 4: Comparison of Schools
Topic 5: Reddit Management (removing comments, moderator warnings, etc.)
Topic 6: Technology
Topic 7: Merges topics together (no clear topic)
Topic 8: Communication
Topic 9: Places
Topic 10: Health and Safety Services
Topic 11: Semester Registration
Topic 12: Apartment Life
Topic 13: Reddit Website Info (so not important to our analysis)
Topic 14: Campus Extracurricular Activities
Topic 15: Housing
Topic 16: Roommate and Social Perspectives
Topic 17: Exams and Testing
Topic 18: Graduate and Job Opportunities
Topic 19: Financial Aid

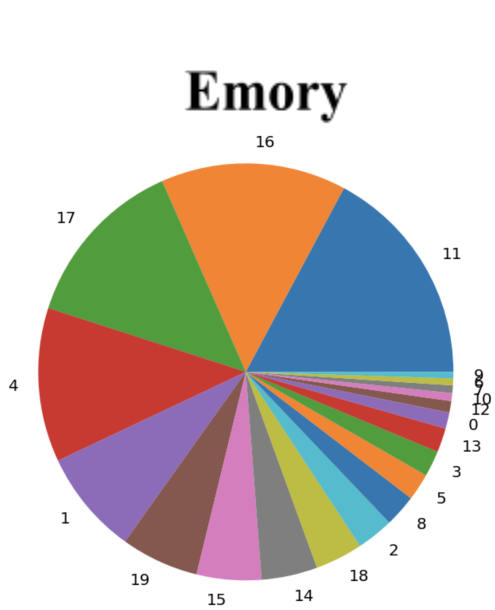
An interesting side note is that after we labeled each topic, we asked OpenAI's ChatGPT to label some of the topics as well. We wanted to see if the AI would corroborate our human interpretation, and it generated bizarrely similar labels for the majority of the topics.

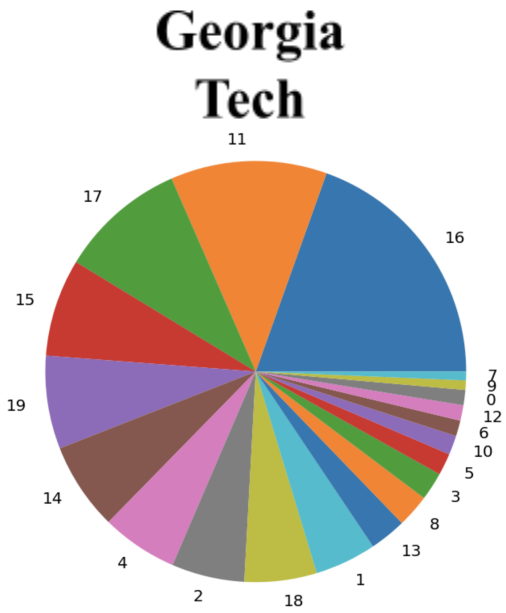
Below is a pie chart showing the total proportion of each topic:

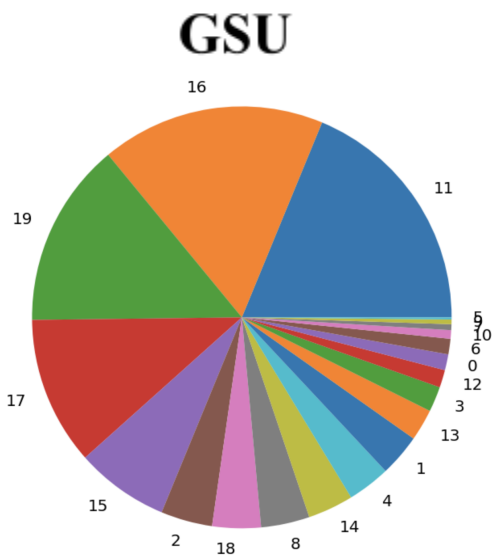


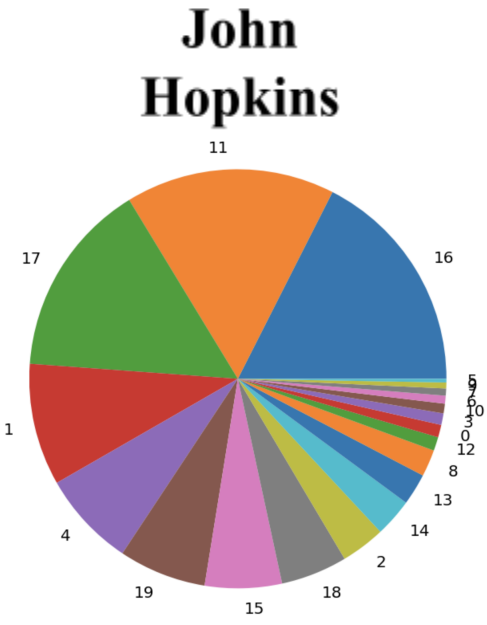
Topics 11, 16, and 17 are the biggest topics when we combine all of the college subreddits. When disaggregating by subreddit, we see that these topics are in fact the biggest topics across all colleges.

Below are pie charts of the proportion of each topic separated by colleges:




Georgia Tech and Georgia State University are classified within Topic 3. This topic represents college admissions and transfers as indicated by words like “application,” “transfer,” and

"accepted." The Georgia Tech and Georgia State University subreddits are similar in that they both have many submissions regarding university admission processes. John Hopkins and Emory are in Topic 4, but when looking at which subreddit individual posts in topic 4 came from, we saw that majority of them came from Emory. Reiterating that the Emory subreddit has a much higher amount of posts pertaining to the topic of comparing colleges than other college subreddits studied. Topic 1 for Emory and John Hopkins is significantly larger than for GSU and Tech. Topic 1 contains words like "pre-med," "research," "med," and "major." Emory and Hopkins are top pre-med schools, so it makes sense that this topic would be large for these two schools. Although topic 1 also contains "cs" and "engineering," it seems to contain more words relating to the pre-medical oriented sciences. Topic 18–which we labeled as Graduate and Job Opportunities–is sizably larger for Georgia Tech and John Hopkins than GSU and Emory, suggesting that a large proportion of discussion for Hopkins and GT relates to academic/career life after college. Compared to the other three schools, GSU has a much larger topic 19. In fact, topic 19–labeled Financial Aid–is one of the largest topics for GSU. This may be because GSU students have the lowest median family income out of the four schools, so there is more financial aid-related discussion on the college subreddit. We were surprised to see that sports were not a topic, particularly because Georgia State University has a large college sports culture.

The following table shows the average polarity scores of posts by college subreddit:

| subreddit | neg | neu | pos | compound |
|---|---|---|---|---|
| Emory | 0.056279 | 0.704573 | 0.234392 | 0.362206 |
| GaState | 0.066235 | 0.715203 | 0.211811 | 0.217214 |
| gatech | 0.075689 | 0.716319 | 0.204193 | 0.238707 |
| jhu | 0.057628 | 0.710993 | 0.225973 | 0.299742 |

In this table, we see that Emory's subreddit has the highest positive (.23) and lowest negative (.05) sentiment. Emory's polarity score is very similar to John Hopkins, which has a positive score of .22 and a negative score of .05. Unsatisfied with the results of a broader sentiment analysis across subreddits, we decided to turn towards the topics we had just generated to create an analysis along two dimensions. To do this, we iterated over the polarity scores for each post, which had already been classified to a most-proximate topic. Then, for each topic, we calculated a sentiment score based on the following formula: $S_n = 100 * \frac{\Sigma(s_i = 'positive') - \Sigma(s_i = negative)}{\Sigma(s_i = 'positive' \text{ or } 'negative')}$, where *i* is a post in topic *n*. This formula yielded the following scores:

| Topic | Score | Topic | Score |
|---|---|---|---|
| 0 | 21.06 | 10 | 4.35 |
| 1 | 1.91 | 11 | 2.05 |
| 2 | 1.19 | 12 | -0.68 |
| 3 | 1.22 | 13 | 3.20 |
| 4 | 2.32 | 14 | 4.22 |
| 5 | -0.22 | 15 | 1.71 |
| 6 | 0.00 | 16 | 4.84 |
| 7 | 0.80 | 17 | 3.30 |
| 8 | 3.20 | 18 | 3.10 |
| 9 | 2.22 | 19 | 2.11 |

Do note that VADER classified very few (<5%) of posts as not neutral, making the relative scale of data represented very small. This is likely because VADER does not use contextual embeddings to imbue words with contingent positivity and negativity. With more time, a BERT sentiment classifier might provide more posts that are either negative or positive, but the length of the submissions means that even this possibility could have limitations.

Nonetheless, these scores do provide more definition to our topics. For instance, Topic 0 has a positivity score that outpaces the rest by leaps and bounds. It is because of this that we can conclude that many of our comments consist *solely* of people thanking one another or demonstrating gratitude, as indicated by top words in the topic. Generally speaking, a bias will prevail towards positive posts because gratitude is often demonstrated and positively associated in Reddit comments. What this means is that negative scores might have a particularly salient meaning. Two topics meet this parameter. Topic 5 is from moderators policing their Reddits, so this makes sense given that a harsh, assertive tone must be used when engaging with potential offensive or unpermitted conduct. Topic 12 describes apartment and housing life and bears great similarity to Topic 15. However, the two have meaningfully different scores. Perhaps, then, Topic 12 refers to lamenting inconveniences of housing while Topic 15 provides more helpful advice. The data used does not represent the whole dataset to an extent where we would feel comfortable making this definitive claim, but it is useful exploratory analysis.

**6 Conclusion and Next Steps**

From our topic model, we see that Georgia Tech and Georgia State cluster together closer than John Hopkins and Emory. We also see that GT and GSU are more attentive to GPA and admission process than John Hopkins and Emory. It is not surprising that academic and course-related topics as well as rooming situations were among the biggest topic across all college subreddits. However, one of our most interesting findings was topic 19 (financial aid ) is the third biggest at GSU. Topic 19 was much smaller for the other schools (sixth biggest for Emory and Hopkins and fifth for Georgia Tech). It is heavily probable that because the average wealth of a student at GSU is lower than at the other three schools, they are more inclined to discuss finances on their subreddit. It would be interesting to conduct further analysis on how topic modeling can be used to help corroborate the economic disparities across different types of colleges (specifically private vs public).

In terms of the next steps, it would be interesting to specifically dive deeper into our topics and their relationship with time. Because our dataset only includes posts between 2020-2022, we felt as if we did not have enough data to properly look at any cyclical variations regarding the topics (especially because COVID-19 could have been a significant confounding variable when analyzing seasonal trends). Our project can be further refined if we collected data pre-2020 and analyzed how the pandemic could have affected the discourse within the college subreddit communities. For example, the topics regarding social culture (college clubs, going out, etc.) were minute across all four subreddits. This could be because the subreddits are inherently an academic community and people don't post about social culture as much, but it also could be because the pandemic limited much of the social, extracurricular, and in-person activities on the campuses. The post-pandemic subreddit data may have also been a reason why sports was not a topic. If we were to develop our project further, we would conduct a time series analysis and analyze the topics and sentiment scores seasonally as well as yearly. Finally, we might want to use more sophisticated, modern technology in making these groupings. Topic modeling can be enhanced with a classifier that uses contextual embeddings as opposed to word- based vectors. Relatedly, we might want to re-do our sentiment analysis with a transformer, such that we can obtain more relatively positive or relatively negative results.

Ultimately, however, our aim in this paper is not methodological perfection. In implementation, we see a clearly partitioned set of topics that might encapsulate the student experience and place words in relation towards one another. The partition along different university subreddits allowed us to prod at possible inherent differences in academic institutions, but clear differences did not emerge from this analysis. This suggests that, for instance, building a classifier across the four universities might not create meaningful results. Finally, our sentiment analysis suggests that such topics vary in positivity, suggesting further exploration into how internet topics in college

shape user responses. We have thus, to some degree of success, extracted novel insights from our dataset using quantitative methods.