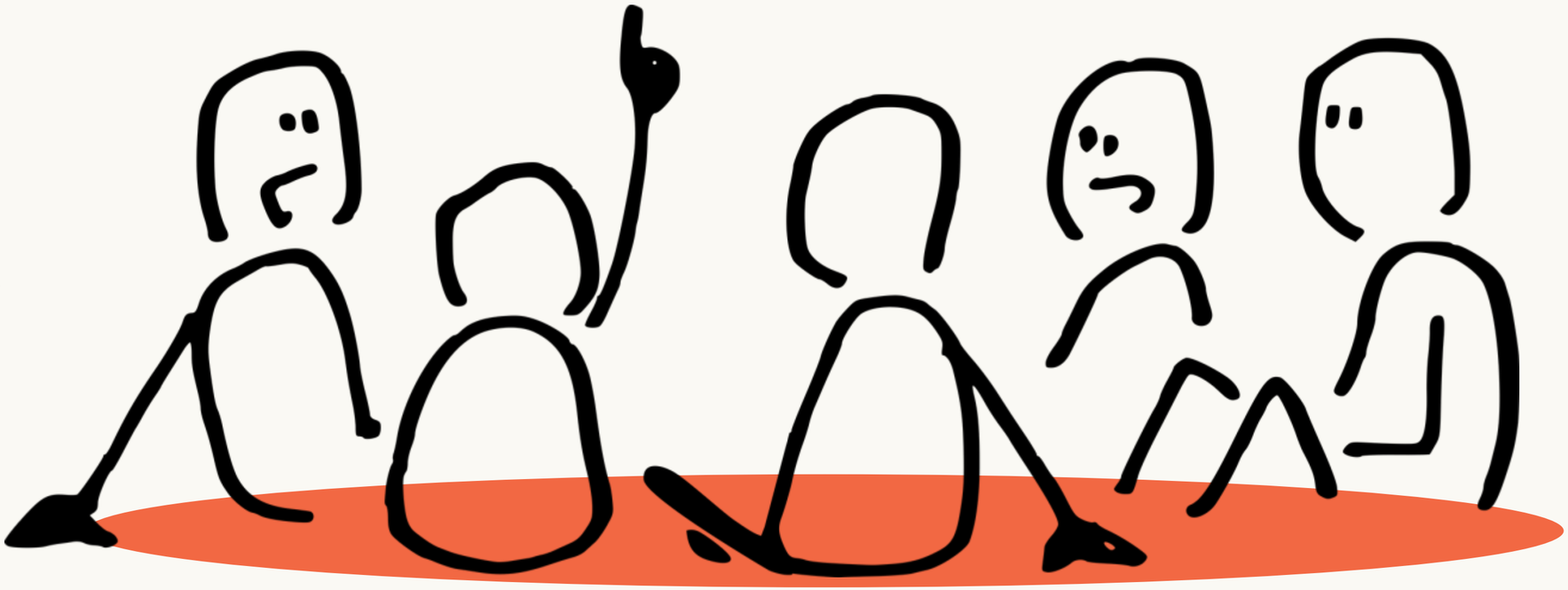
WHO'S THE REAL "MVP"?

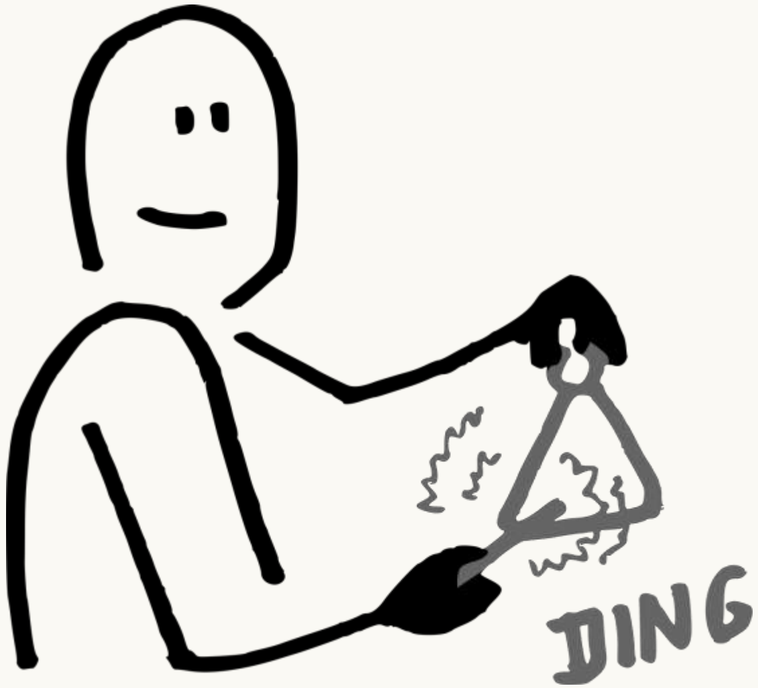
Presented by
Group 5



Module Lead by
Dr. Soham Ghosh

Leveraging Reddit Contents for University Strategies

Social Media Analysis



Agenda

- INTRODUCTION
- DATA WALKTROUGH
- KEY INSIGHTS
- STRATEGIC RECOMMENDATIONS

What is Reddit?

A massive online forum where anonymous users share brutally honest opinions.

- 430M+ monthly active users.
- 2.5B+ monthly upvotes
- 100K+ subreddits

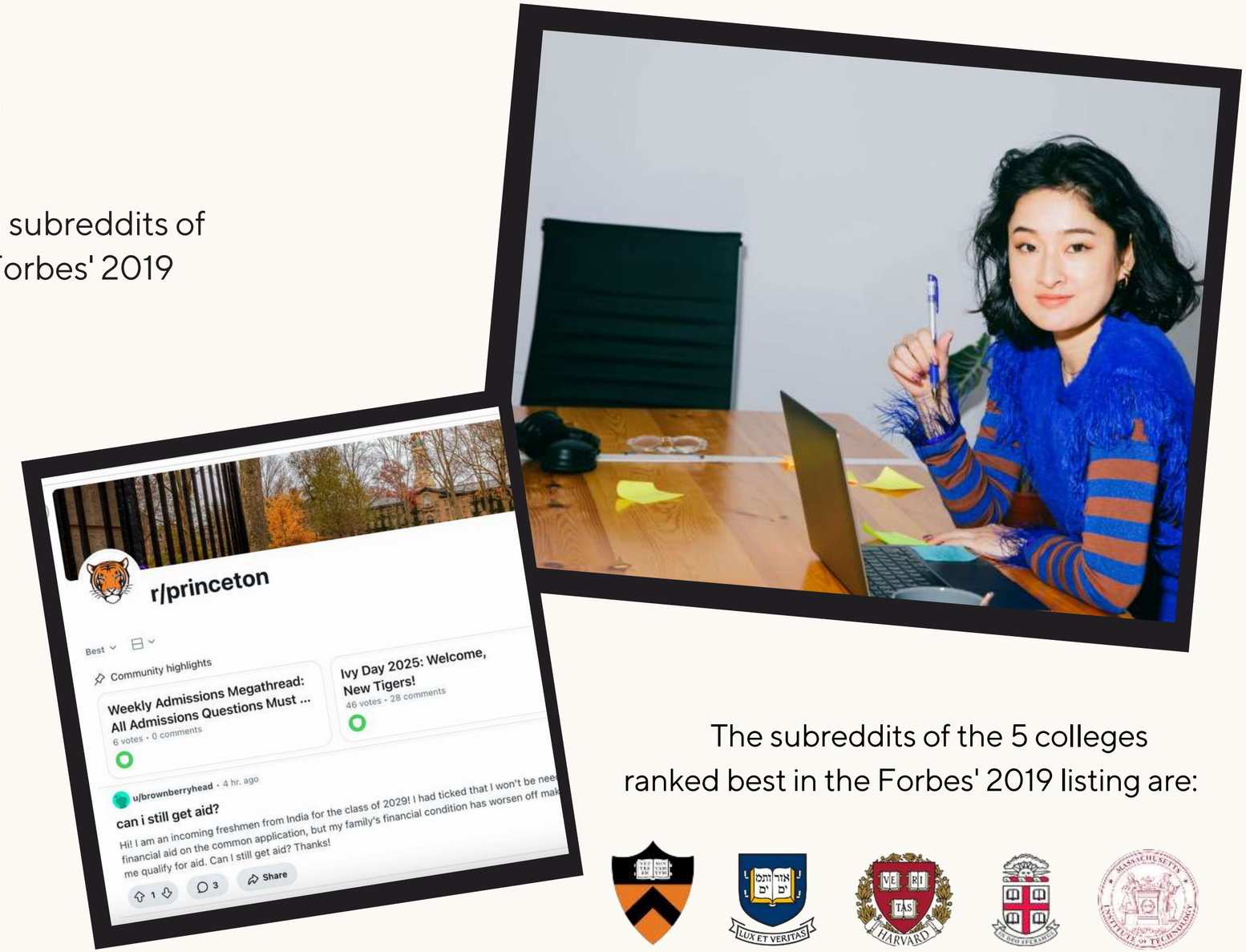
Why Reddit?

Over 70% of students use Reddit to fact-check marketing claims, seek ‘real’ student experiences and warn peers about red flags (Pew, 2022).

The Datasets

- **Description**: Comprehensive data on subreddits of the top 10 American colleges as per Forbes' 2019 list, up to Feb 21, 2022.

- **Diversity of content:**
  - 42,500 Posts
  - 178,000 Comments



The subreddits of the 5 colleges ranked best in the Forbes' 2019 listing are:

Source from kaggle

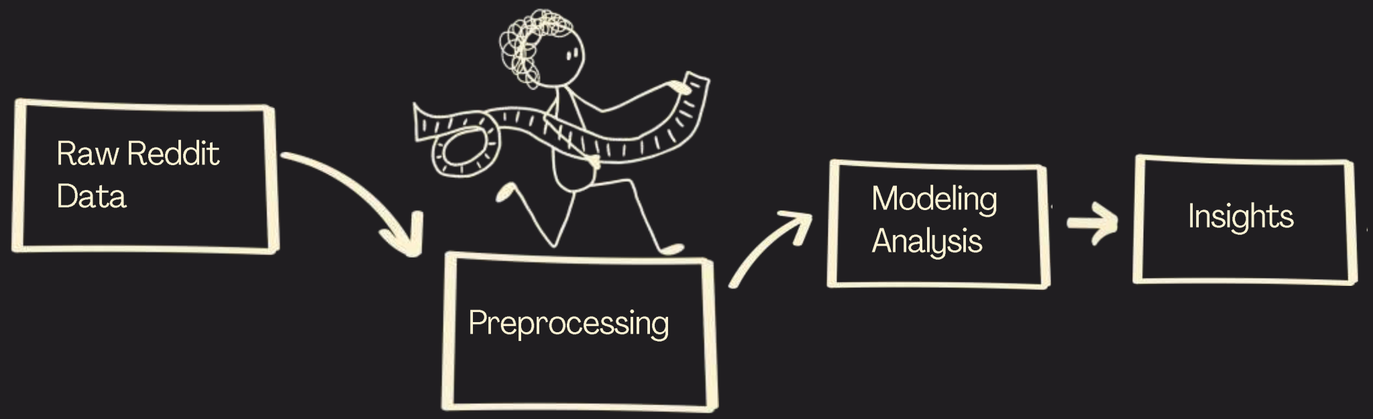
The Preprocessing

- Merge Dataset
- Data Cleaning
- Remove Stopwords & Tokenization (NLTK)
- Normalization

The Analysis

- Sentiment Analysis (VADER)
- Topic Modelling (BERT Based)
- Topic Classsification (KNN)
- Keyword Extraction (Cosine Similarity,TF-IDF)

Methodology Walk Through



Key Questions

Listen to the Silent Majority

Who's Talking the Most?

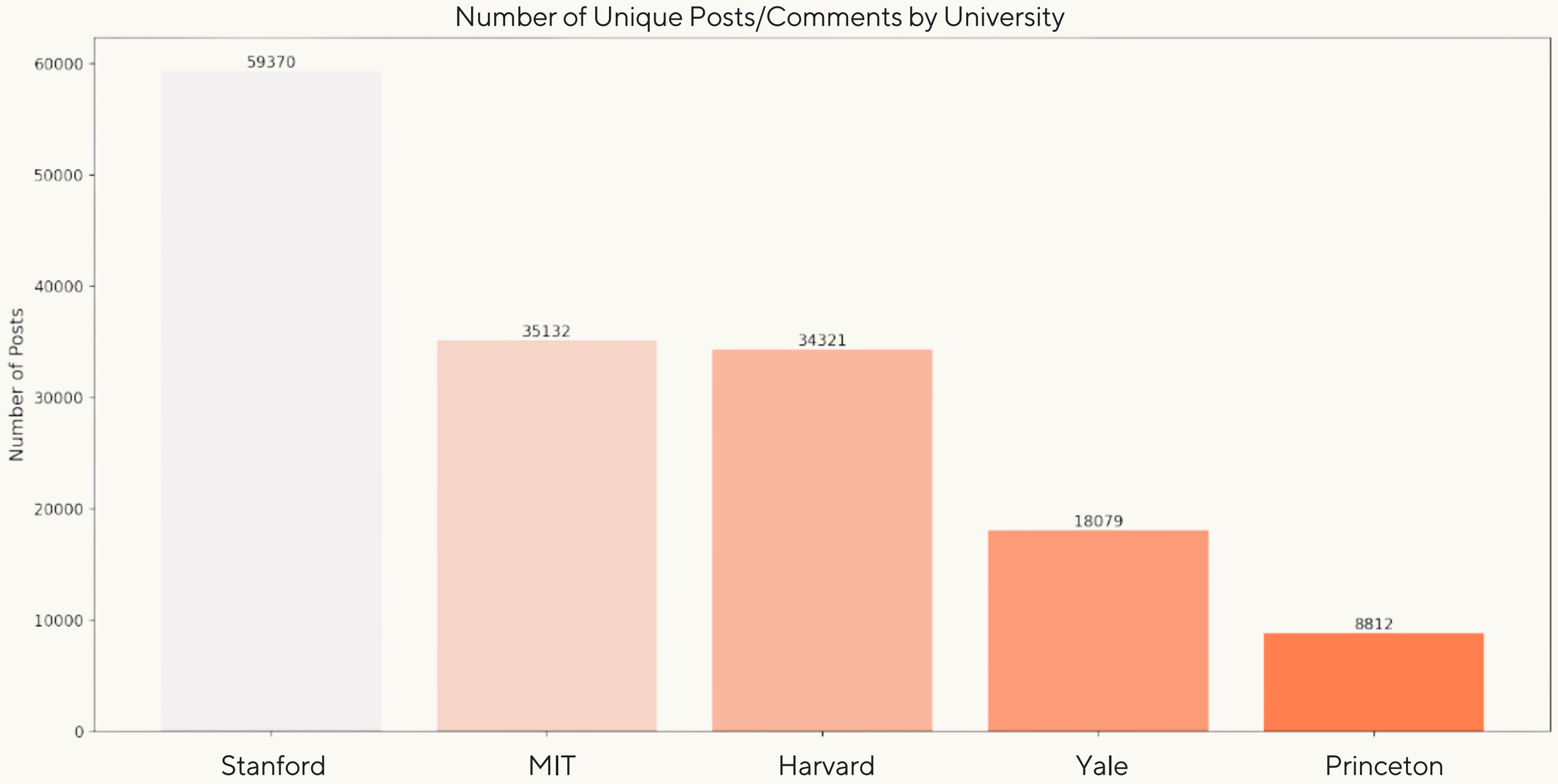
Decoding University Engagement

What Students Really Talk About?

The unspoken good and bad behind the buzzwords

Who Wins What?

Topic Deep Dive



Who’s Talking the Most?

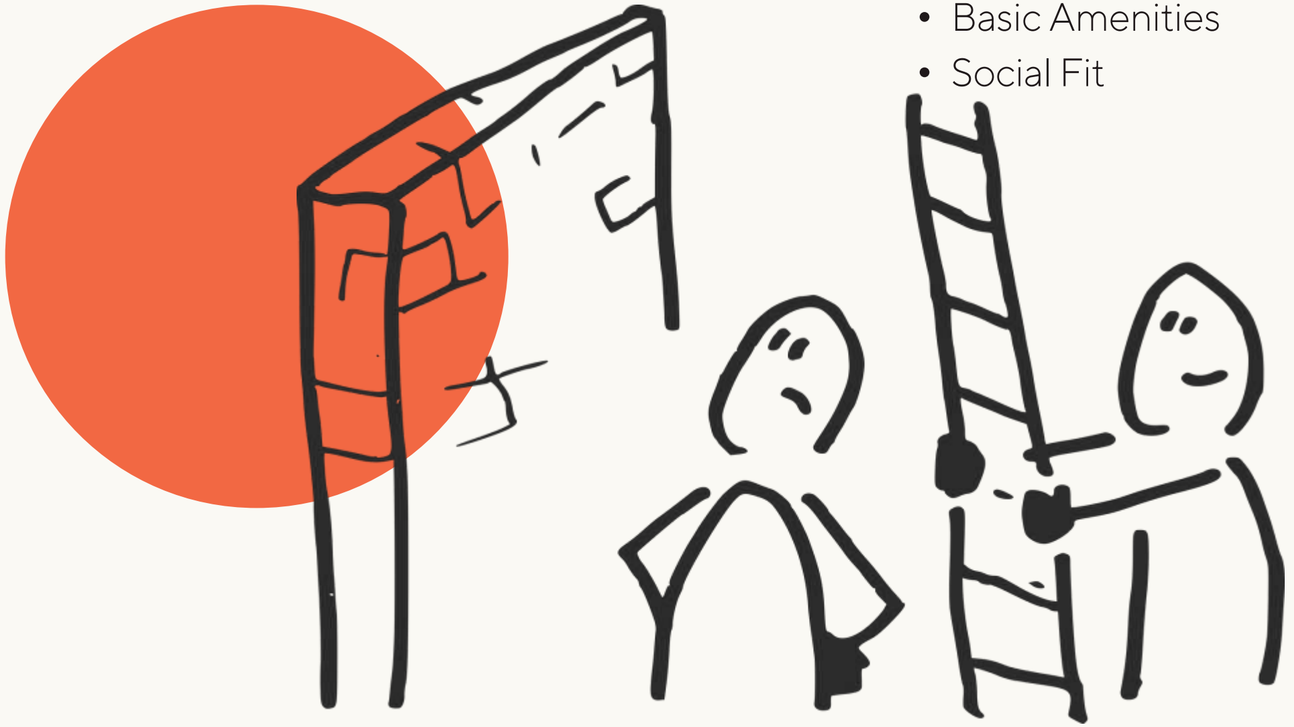
Engagement Metrics

Forbes Ranking v.s. Reddit Reality

The Forbes(2019)

Harvard has been ranked No.1 out of 650 top universities evaluated

- Graduate Success
- Earnings Power
- Alumni accolades
- Graduation Efficiency



The Reddit

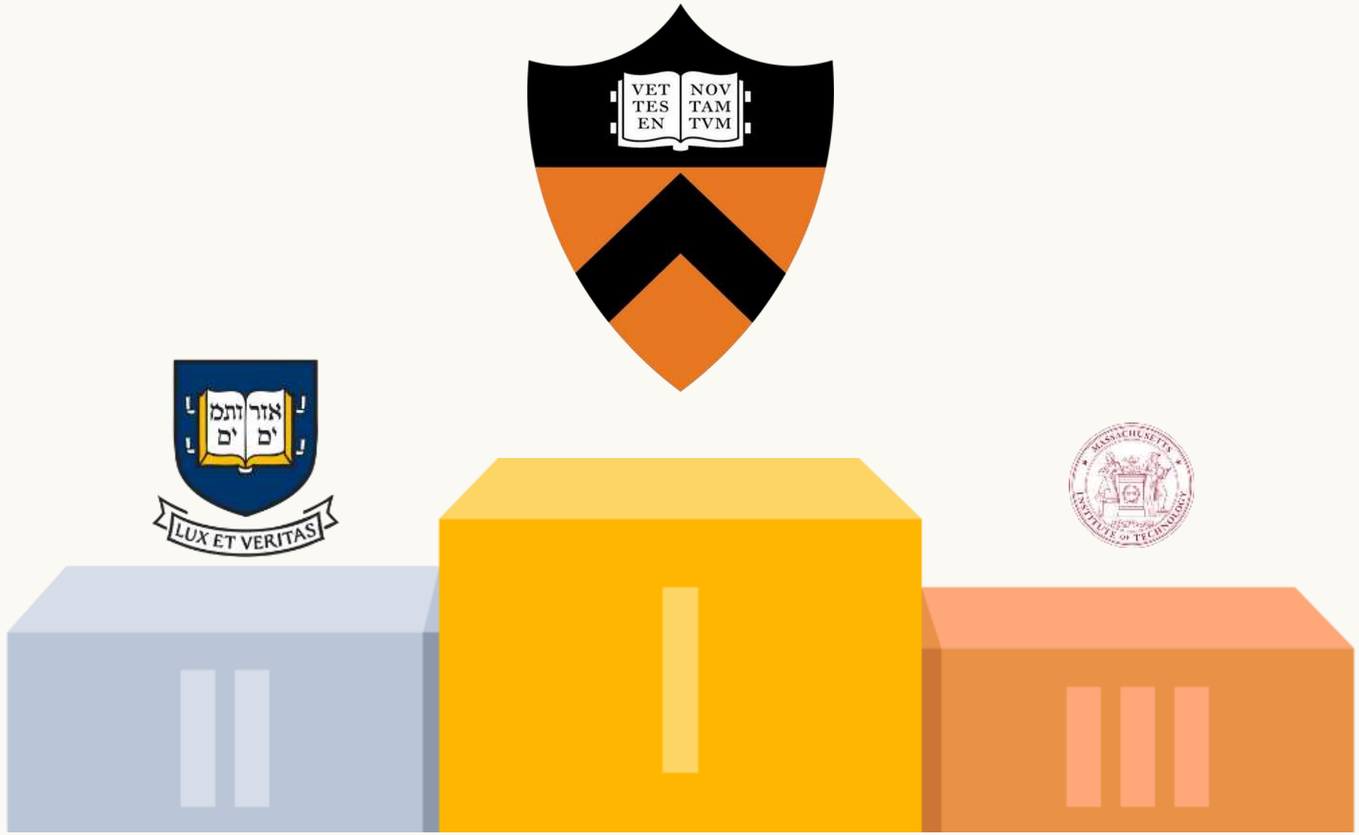
Forbes rewards prestige, not student experience. Broken dorms? Loneliness? Not on the scorecard.

- Mental Health
- Basic Amenities
- Social Fit

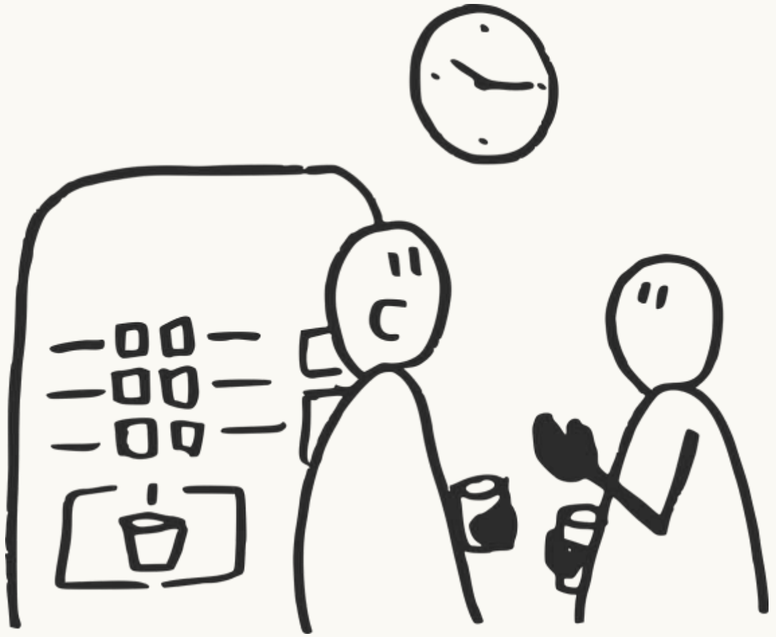
Princeton Wins Hearts !

University	Forbes Rank	Mean Sentiment Score
Princeton	#5	0.65
Yale	#2	0.57
MIT	#4	0.54
Harvard	#1	0.52
Stanford	#3	0.47

Social Media Awards Ceremony



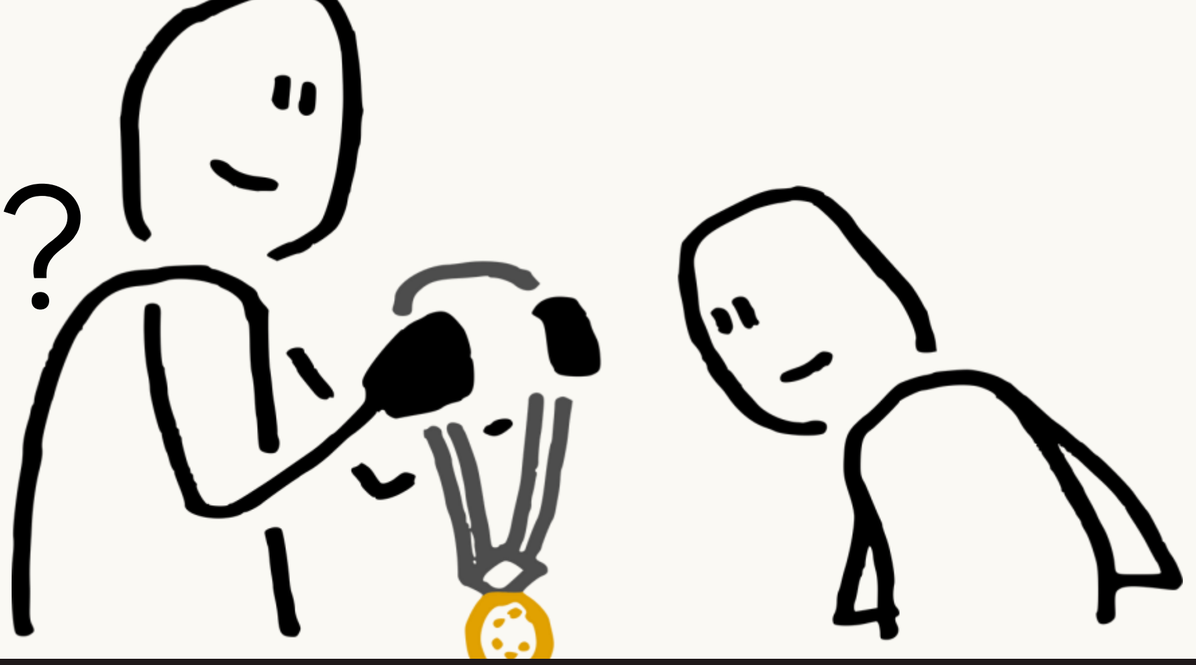
What Students Really Talk About?



- **Academics Dominate:** “courses”,” major”- but paired with “hard” and “ work”.
- **Social Paradox:** “People” and “social” trend, but context such as silent crices hidden
- **Future Anxiety:** “Graduate” and “want” hint at fears about jobs, debt, and uncertainty(Wang, 2018).

Who Wins What?

Thematic Deep Dive



Campus life & Academic Environment

Keywords: dorms, freshmen, semesters, extracurriculars, grades, professor

Academic & Extracurricular Activities

keywords: extracurriculars, grades, internships, semesters, graduation

Curriculum & Career Development

keywords: majors, graduation, postgrad, curriculums, internships, gpa

"Not at all. It truly was a life changing opportunity for me and while there were some challenges, typically around academic workload stress, I overall loved it."

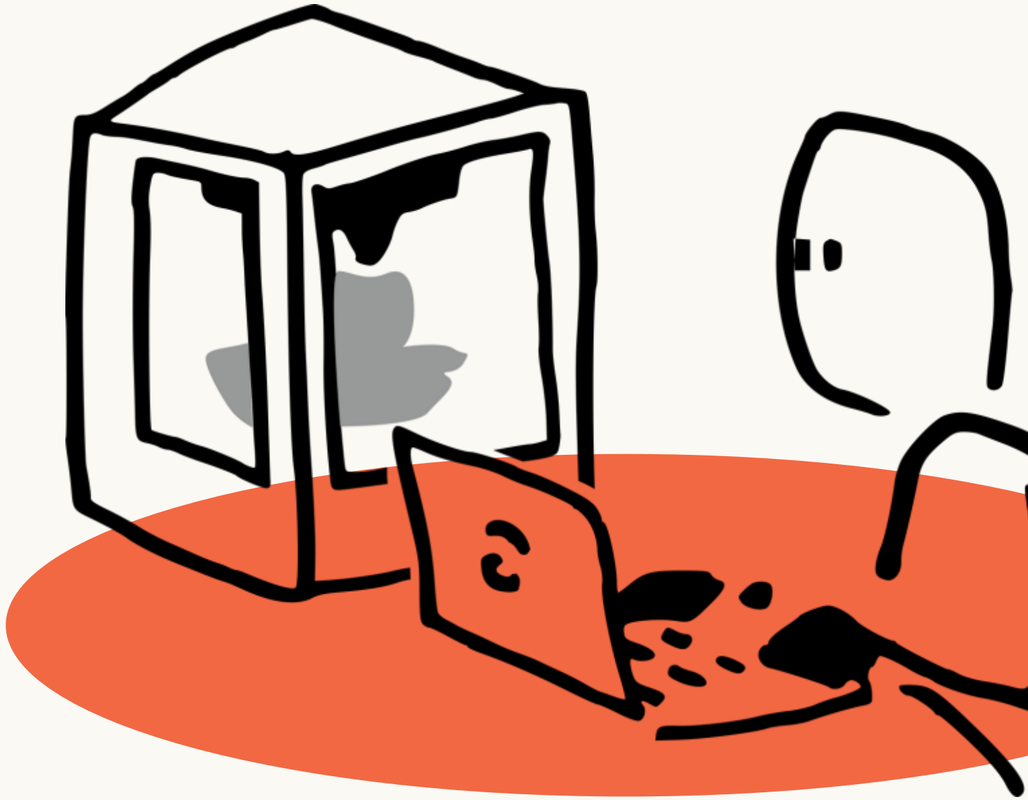
— Princeton Subreddit

"Take it easy and take it slow, one day at a time, relax - everything will be extremely overwhelming otherwise, especially academics. Don't give in to peer pressure"

— MIT Subreddit

"Got a B in Probability class. Am I doom for quant research jobs?"

— Stanford Subreddit

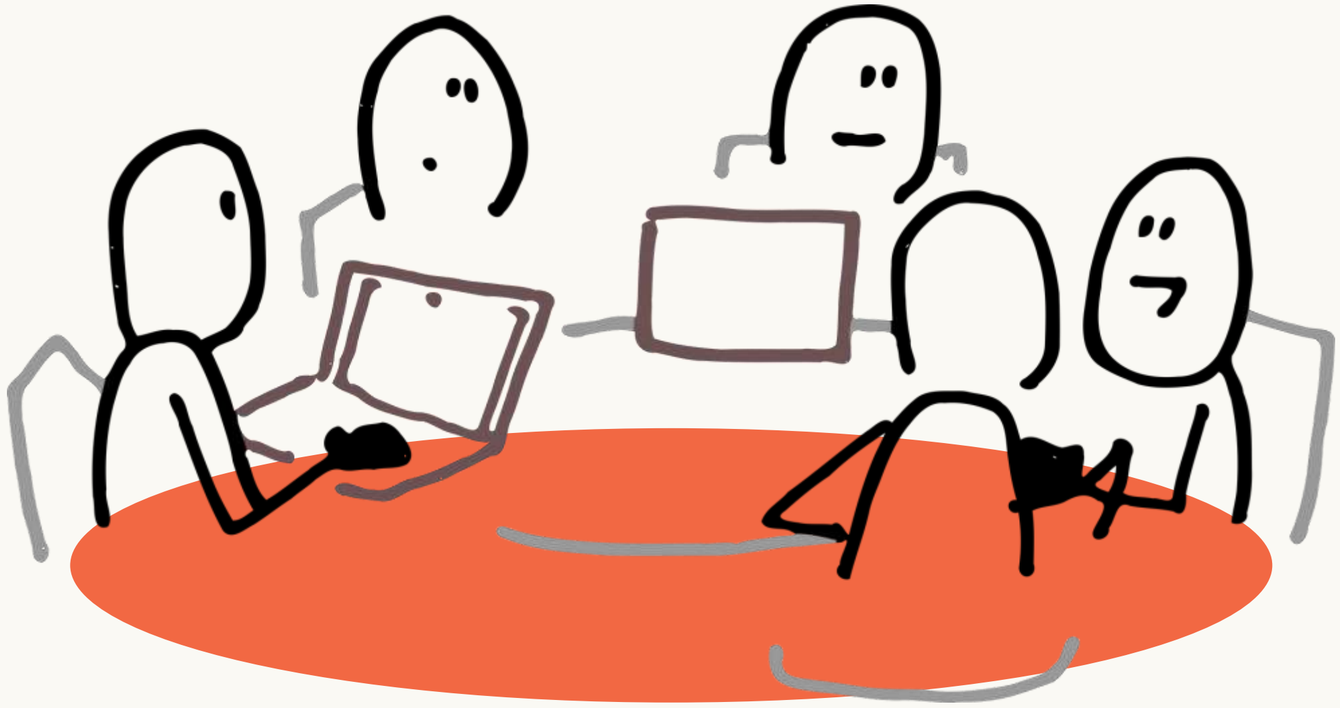


Student Voice Spotlights

Strategic Recommendations

Redefine “ Success ” Beyond Just Rankings

- Fix the “ Little Big Things”
- Turn Neutrality Into Advoacy
- Weaponized the Social Dynamics



Challenges & Limitations

Data Limitations

Stanford contributing disproportionately more content than others, which can skew the analysis and reduce generalizability (Baumgartner, 2020).

Sampling Bias

Reddit user data suffers from self-selection bias, where a small vocal minority produces the majority of content, creating skewed perceptions (Weninger, 2013).

Sentiment Model Limitations

Tools like VADER and BERT-based sentiment models struggle to accurately classify sentiment in the presence of sarcasm, slang, or domain-specific discourse common in student-generated content (Hutto and Gilbert, 2014).

Improvements & Solutions

Rebalance the Dataset

Techniques such as downsampling and weighting can be employed to mitigate the dominance of overrepresented classes (Buda, Maki and Mazurowski, 2018).

Cross-Platform Validation

Incorporate data from other platforms (e.g., Twitter comments) to validate whether findings are consistent across different online communities（Tufekci, 2014）.

Involve Students in Feedback Loop

Incorporate qualitative feedback from real students to validate topic modeling and sentiment accuracy, bridging the gap between algorithm and lived experience（Creswell and Miller, 2000）.

Listen to students' whispers before they become screams.

Thank you x Group 5

Appendix: Literature Review (Partial)

Pew Research Center. (2022)Teens, Social Media and Technology 2022. https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/

Wang, R. et al. (2018) 'Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing', Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 2(1), p. 43:1-43:26. Available at: https://doi.org/10.1145/3191775.

Baumgartner, J. et al. (2020) 'The Pushshift Reddit Dataset', Proceedings of the International AAAI Conference on Web and Social Media, 14, pp. 830–839. Available at: https://doi.org/10.1609/icwsm.v14i1.7347.

Chen, C. et al. (2013) 'Battling the internet water army: detection of hidden paid posters', in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York, NY, USA: Association for Computing Machinery (ASONAM '13), pp. 116–120. Available at: https://doi.org/10.1145/2492517.2492637.

Hutto, C. and Gilbert, E. (2014) 'VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text', Proceedings of the International AAAI Conference on Web and Social Media, 8(1), pp. 216–225. Available at: https://doi.org/10.1609/icwsm.v8i1.14550.

Buda, M., Maki, A. and Mazurowski, M.A. (2018) 'A systematic study of the class imbalance problem in convolutional neural networks', Neural Networks, 106, pp. 249–259. Available at: https://doi.org/10.1016/j.neunet.2018.07.011.

Tufekci, Z. (2014) 'Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls'. arXiv. Available at: https://doi.org/10.48550/arXiv.1403.7400.

Creswell, J.W. and and Miller, D.L. (2000) 'Determining Validity in Qualitative Inquiry', Theory Into Practice, 39(3), pp. 124–130. Available at: https://doi.org/10.1207/s15430421tip3903_2.