An Embedding Analysis of Diversity in College Opinion Writing

Anonymous Authors[1]

Abstract

Diversity is often celebrated as a sign of progress in society, and many colleges prioritize it in their admissions and on campus, but has this emphasis led to an increase in the diversity of ideas being discussed? We analyze student opinion writings in college newspapers from the past decade, using embedding vectors and topic modeling to measure the range of ideas. We then compare this data with the trend in the number of times diversity-related words are mentioned in the same newspapers. We find that explicit mentions of diversity have increased across colleges in the past decade, but the diversity of ideas (as measured by the size of the covered embedding space or the topics discussed) has decreased in most colleges during this period. This seems to indicate that while colleges are encouraging more discourse about diversity, the actual space of ideas being discussed in school newspaper opinion sections has been shrinking.

1. Introduction

Many believe that a diverse population will lead to a diverse range of perspectives and ideas. However, there have been concerns that some colleges are becoming increasingly homogenous in their thinking. Park et al. (2023) go as far as to argue that publications and patents are increasingly less likely to break with past ideas to push to new directions. Motivation for this current study came from Rasmussen's 2021 analysis of National Science Foundation (NSF) grants and his findings that ideas being funded have been becoming less diverse over the last few decades.

In colleges around the world, especially in the US, we are seeing a rise in awareness and representations of diversity. We want to know:

Has the recent emphasis on diversity in colleges led to a corresponding increase in the diversity of ideas and perspectives in these colleges?

To do this, we analyze student opinion writings in college newspapers from the past decade, using embedding vectors to proxy for the ideas discussed (Mikolov et al., 2013). An invariant across colleges is their student newspaper, and an "opinion(s)" section with contributions from students and staff. We selected a representative set of colleges across the US, with different institutional types, sizes, geography, such as liberal arts, single-gender, and religious colleges. In addition to comparing diversity measures across colleges, we also perform embedding analysis across time.

2. Methodology

This study aims to investigate the relationship between the mentions of diversity-related words in college newspapers and the diversity of ideas on college campuses. To do this, the following steps were taken:

1. Identify a "diverse" set of colleges representing those that are: prestigious research institutions, single-gender, religious, coastal, Southern, and liberal arts colleges.
2. A corpus of articles was created from each college newspaper's online editions using a custom made web-crawler.
3. The total mentions of diversity-related words in these corpora were measured and normalized by the size of each corpus.
4. The diversity of ideas in these corpora were also measured and normalized by the size of each corpus. We used embedding document vectors to proxy for ideas in these corpora.
5. Topics discussed across these corpora were modeled and measured across time.
6. The year-over-year changes across these measures were collected and analyzed.

The data was stored as parquet files in Amazon Web Services' Simple Storage Services and analyzed using Jupyter

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

{ ***'anti-discrimination', 'diversity', 'equality', 'equal opportunity', 'equity', 'inclusion', 'inclusive', 'inclusivity', 'intercultural', 'intersectional', 'intersectionality', 'multi-cultural', 'multicultural', 'racial justice', 'social justice'*** }

Figure 1. These are the "diversity-related words" that we use to count total mentions of. We use the lemmatized version of these words against lemmatized versions of corpus text.

Notebook and various Python libraries. We used a collection of 15 diversity related words and phrases, cleansed and lemmatized with NLTK, to represent explicit diversity awareness in a piece of writing (see Figure 1). For document embedding, we used Honnibal & Montani (2017)'s spaCy and the pre-trained model *en_core_web_lg*.

We represented the space of ideas in college discourses as document vectors, where each document (such as an opinion article) is represented by the average of its word vectors.

To ensure comparability across different colleges, we normalized the number of mentions of diversity-related words and the cosine distances.

On our dispersion measures, let N represent the total number of documents in a given year for a particular college. With N, we can calculate the following for each college per year:

Total number of document pairs = $\binom{N}{2}$

One way to measure how different 2 documents are is to measure the "distance" separating their respective document vectors in the mapped 300-dimensional vector space. We use the cosine distance measure to calculate pairwise document-to-document distances:

$$1 - \frac{d_i \cdot d_j}{\|d_i\|_2 \, \|d_j\|_2}$$

For any given year in a college newspaper, one way to measure the span of all documents is to sum up that year's pairwise cosine distances, normalized by the total number of document pairs. We define a dispersion measure for the span of article embeddings in a given year this way:

Normalized Pairwise Dispersion =

$$\frac{2}{N \cdot (N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} 1 - \cos(d_i, d_j)$$

Rasmussen (2021) applied a similar document vector measure in studying NSF grants with insightful results.

For topic modeling, we used Grootendorst, 2022's BERTopic package and Reimers & Gurevych, 2019's SBERT pretrained *'all-MiniLM-L6-v2'* sentence embedding models. After fitting and transforming each college corpus to BERTopic, we parsed out topic embedding vectors for each year and analyzed how these topics evolve over time for each college. To measure the "spread" of topics for each year, we also applied the Normalized Pairwise Dispersion to the topic embedding vectors.

3. Experiments

Our experiments aim to answer not only how the space of ideas has evolved over time, they also aim to ask how this evolution differs across different types of colleges. For a fair comparison, we need to find an invariant data set across college campuses. Fortunately, US colleges all have campus newspapers, and an "opinion" section with contributions from students and staff.

We selected a representative set of colleges across the US, with varying institutional types, sizes, geography, liberal arts, single-gender, and religious colleges: Stanford, Harvard, U.C. Berkeley, Dartmouth, Wellesley, Liberty University, U.T. Austin, and Swarthmore College.

For each college, we created a customer web-crawler to crawl all articles in the campus newspaper opinion section. Each college corpus was analyzed for the mentions of diversity-related words as shown in Figure 1. We then performed embedding analysis on each corpus, and applied Normalized Pairwise Dispersion measure for articles published per year per college. Finally, we performed topic modeling for all articles for a given year using BERTopic. We also applied Normalized Pairwise Dispersion measures for the topic embeddings for each year per college. We then reported on the year-over-year change on these measures.

For each college, we calculated the total mentions of diversity-related words normalized by corpus size, as well as the diversity of ideas within each corpus, also normalized by corpus size.

Our hypotheses were: 1) more prestigious (usually research heavy) universities should exhibit a higher diversity of ideas; 2) liberal arts colleges, which prioritize expanding students' intellectual horizons, would also have a high diversity of ideas; 3) specialty colleges, such as religious colleges, should have a lower diversity of ideas because of the focus these specialty colleges have on their campuses; 4) coastal (and likely more liberal) institutions should have greater diversity of ideas compared to southern (and likely more conservative) colleges; and 5) larger colleges should have greater diversity of ideas compared to smaller colleges.

4. Results

In this study, we investigated the relationship between the frequency of diversity-related words mentioned in college newspapers and the actual diversity of ideas as indicated by word vectors.

Our dataset consists of text collected from online college newspaper opinion sections. Each college corpus has between a few hundred articles to a few thousand articles across the past decade. Some colleges only have online newspapers accessible from 2011 or even 2012 onward.

On average, the data showed an increase in the usage of diversity-related words over time, but a decrease in the actual diversity of ideas. Our analysis of Stanford, Harvard, UC Berkeley, UT Austin, Liberty University, and Swarthmore College revealed similar trends of increasing mentions of diversity but decreasing diversity as measured by word vectors. However, Wellesley College and Dartmouth College showed an increase in the diversity of ideas over time. Overall, our results indicated a disconnect between the mentions of diversity and its actual presence.

The first college newspaper we analyzed was Stanford University's The Stanford Daily. We sampled 4914 opinion articles from 2010 to 2022 covering editorials, columns and letters. Our analysis, using a normalized measure for diversity mentions, revealed an increase of approximately 40% in the mentions of diversity from an average of 1.1 in 2010 to 1.9 in 2022. However, this trend was not reflected in the actual diversity of ideas as determined by our word vector analysis, which showed a decrease. Specifically, using our normalized cosine distance dispersion measure, we observed a drop of approximately 20% from 0.075 to 0.06 between 2010 and 2022. See Figure 2.

The next college newspaper we analyzed was Harvard University's The Harvard Crimson. We sampled 4893 opinion articles from 2010 to 2022 covering op-eds, editorials and columns. Our analysis, using a normalized measure for diversity mentions, revealed an increase of approximately 50% in the mentions of diversity from an average of 0.35 in 2010 to 0.7 in 2022. Similarly, this trend was not reflected in the actual diversity of ideas as determined by our word vector analysis, which showed a decrease. Specifically, using our normalized cosine distance dispersion measure, we observed a drop of approximately 7% from 0.075 to 0.07 between 2010 and 2022. See Figure 3.

The next college newspaper we analyzed was UC Berkeley's The Daily Californian. We sampled 6095 opinion articles from 2011 to 2022 covering editorials, columns and letters. Our analysis, using a normalized measure for diversity mentions, revealed an increase of approximately 60% in the mentions of diversity from an average of 0.2 in 2011 to 0.55 in 2022. Similarly, this trend was not reflected in the actual diversity of ideas as determined by our word vector analysis, which showed a decrease. Specifically, using our normalized cosine distance dispersion measure, we observed a drop of approximately 15% from 0.13 to 0.11 between 2011 and 2022. See Figure 4.

The next college newspaper we analyzed was UT Austin's The Daily Texan. We sampled 5713 opinion articles from 2010 to 2022 covering editorials, columns, forums and letters. Our analysis, using a normalized measure for diversity mentions, revealed an increase of approximately 60% in the mentions of diversity from an average of 0.25 in 2011 to 0.65 in 2022. Similarly, this trend was not reflected in the actual diversity of ideas as determined by our word vector analysis, which showed a decrease. Specifically, using our normalized cosine distance dispersion measure, we observed a drop of approximately 40% from 0.095 to 0.055 between 2010 and 2022. See Figure 5

The next college newspaper we analyzed was Liberty University's Champion. We sampled 1362 opinion articles from 2010 to 2022. Our analysis, using a normalized measure for diversity mentions, revealed an increase of approximately 60% in the mentions of diversity from an average of 0.075 in 2010 to 0.20 in 2022. Similarly, this trend was not reflected in the actual diversity of ideas as determined by our word vector analysis, which showed a decrease. Specifically, using our normalized cosine distance dispersion measure, we observed a drop of approximately 6% from 0.085 to 0.08 between 2010 and 2022. See Figure 6.

The next college we analyzed was Swarthmore College's The Phoenix. We sampled 1884 opinion articles from 2010 to 2022 covering columns, letters and editorials. Our analysis, using a normalized measure for diversity mentions, revealed an increase of approximately 80% in the mentions of diversity from an average of 0.15 in 2005 to 0.8 in 2022. This trend was not reflected in the actual diversity of ideas as determined by our word vector analysis, which showed a decrease. Specifically, using our normalized cosine distance dispersion measure, we observed a drop of approximately 10% from 0.09 to 0.08 between 2005 and 2022. See Figure 7.

The next college newspaper we analyzed was Wellesley College's The Wellesley News. We sampled 667 opinion articles from 2013 to 2022. Our analysis, using a normalized measure for diversity mentions, revealed an increase of approximately 10% in the mentions of diversity from an average of 0.8 in 2013 to 0.9 in 2022. Unlike the previous colleges analyzed, this trend was reflected in the actual diversity of ideas as determined by our word vector analysis, which showed an increase. Specifically, using our normalized cosine distance dispersion measure, we observed a slight increase of approximately 7% from 0.07 to 0.075 between 2013 and 2022. See Figure 8.

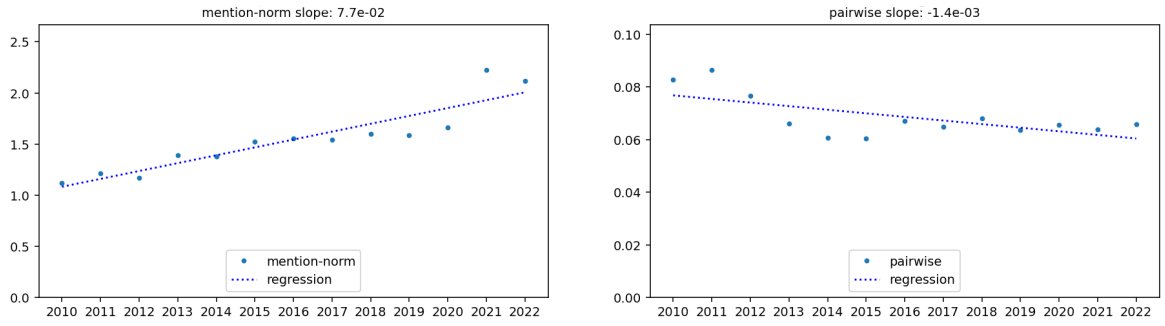

Figure 2. Trends of diversity measures for Stanford

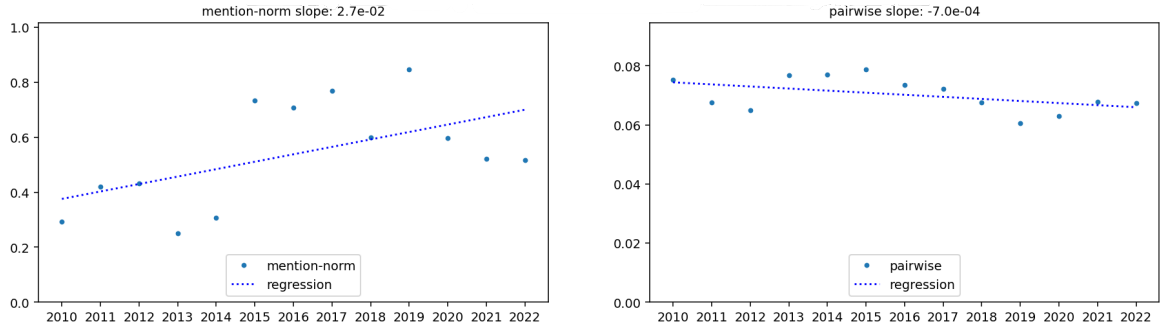

Figure 3. Trends of diversity measures for Harvard

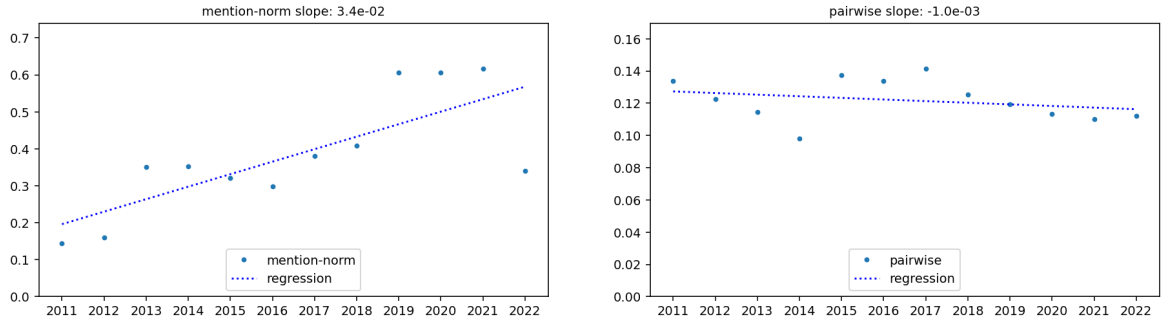

Figure 4. Trends of diversity measures for UC Berkeley

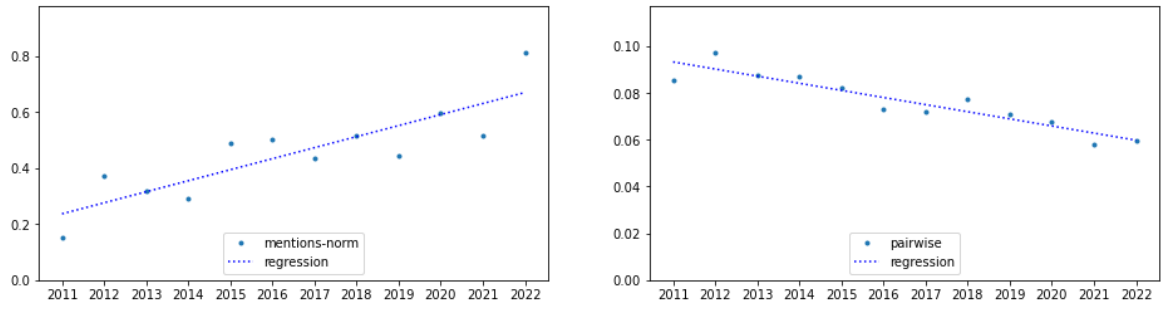

Figure 5. Trends of diversity measures for UT Austin

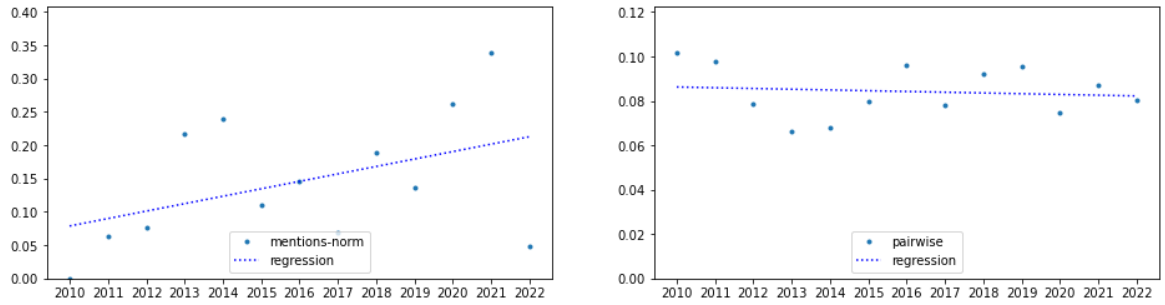

Figure 6. Trends of diversity measures for Liberty

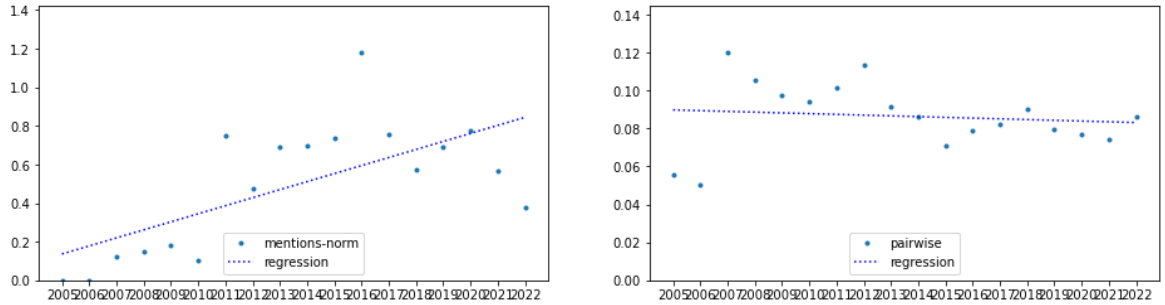

Figure 7. Trends of diversity measures for Swarthmore

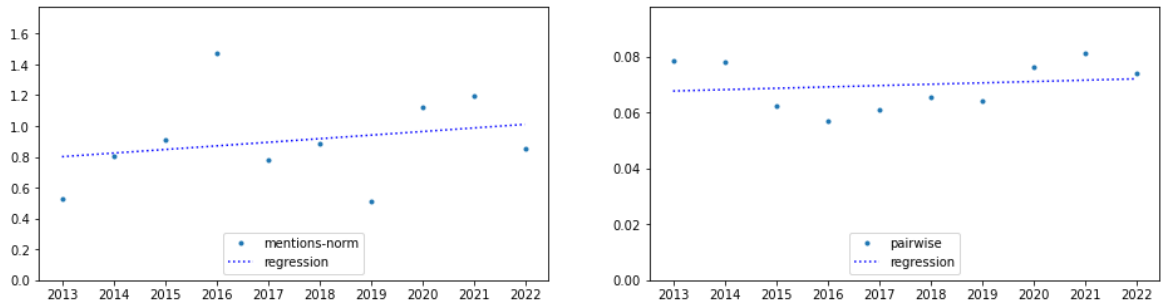

Figure 8. Trends of diversity measures for Wellesley

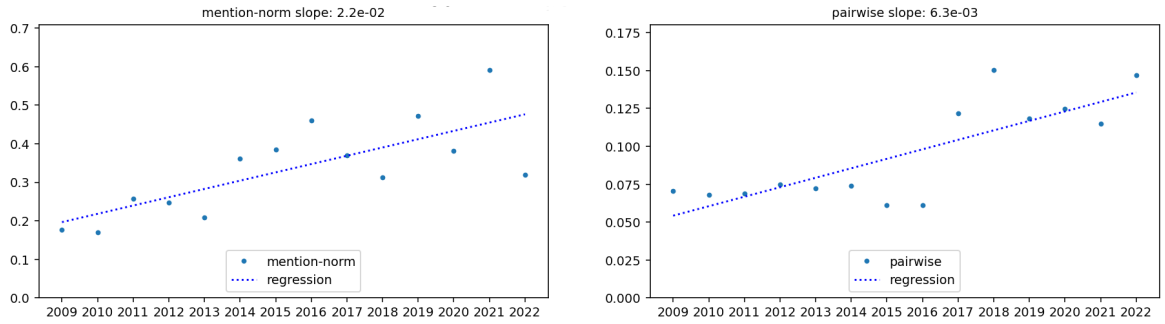

Figure 9. Trends of diversity measures for Dartmouth

The final college newspaper we analyzed was Dartmouth College's The Dartmouth. We sampled 3635 opinion articles from 2010 to 2022. Our analysis, using a normalized measure for diversity mentions, revealed a significant increase of approximately 60% in the mentions of diversity from an average of 0.2 in 2009 to 0.45 in 2022. This trend was reflected in the actual diversity of ideas as determined by our word vector analysis, which showed an increase. Specifically, using our normalized cosine distance dispersion measure, we observed roughly doubling from 0.055 to 0.13 between 2009 and 2022. See Figure 9.

Topic Modeling Analysis

As most colleges in our corpora have seen a drop in pairwise cosine distance dispersion measures, we are interested to know how the topic space has evolved during this time period. We used BERTopic (Grootendorst, 2022) with pretrained 'all-MiniLM-L6-v2' SBERT model on the collected corpora. We will describe a few results for Stanford's The Stanford Daily here for qualitative analysis.

From 2011 to 2022, Figure 10 shows the evolution of online newspaper opinion topics identified by BERTopic from 2010 - 2022. A total of 85 topic clusters were found (the top 12 of which are shown on the right side of Figure 10). These topics are discussed in articles across these 12 years, with rise and fall shown in the figure below. Notable spikes of *"2_trump_clinton_election_republican"* topic in 2016, and *"16_covid_19_vaccine_pandemic"* topic in 2020.

Trending of these topics discussed are plotted in the left hand graph below – fitted with a negatively sloped regression line. We applied the same technique of Normalized Pairwise Dispersion on the topic embedding vectors over time. The plot on the right of the Figure 11 shows trending of the dispersion of these topics in the embedding space. The regression line shows a negative slope – indicating that the size of the topics embedding space has been slightly decreasing over the last dozen years in Stanford's online newspaper opinion section.

Similar topic modeling was done to other colleges in our list with similar qualitative results. (We are working on how to quantitatively present these collective results succinctly.) By using topic modeling, we are able to reveal trending of specific discussion topics in campus newspaper to complement document embedding analyses above.

Overall, our experiments and analyses reveal that the mentions of diversity-related words has been on the rise over the last decade – indicating an increase in awareness of diversity in these school campuses. However, with the exceptions of Wellesley College and Dartmouth College, the diversity of ideas represented in our selection of college newspaper opinion sections has been decreasing over the same period.

5. Discussion

We aimed to assess the diversity of ideas among college students by analyzing the opinion sections of student newspapers. We hypothesized that more prestigious colleges would have a higher diversity of ideas, that liberal arts colleges would also have a high diversity of ideas, that specialty colleges would have a lower diversity of ideas, that coastal colleges would have greater diversity of ideas compared to southern colleges; and that larger colleges would have greater diversity of ideas compared to smaller colleges.

The opinion section of the school newspaper allows students to freely express their thoughts and ideas in an academic manner. Our main constraint was in getting access to openly available opinion articles. From our internet queries, we discovered that most colleges have their student newspapers online since around 2010. As a result, we were only able to collect text corpora across colleges from the current and previous decades.

As expected, "diversity" mentions have increased over time for all the colleges we sampled. However, our first hypothesis on more prestigious research colleges was proven incorrect. Stanford University, one the most prestigious research universities in the world, showed a decreasing trend in its diversity of ideas and had a pairwise dispersion in 2022 of 0.06 - the second smallest of the sampled colleges, with Harvard University at 0.07, having a decreasing diversity, UC Berkeley at 0.11, having a decreasing diversity, UT Austin at 0.055, having a decreasing diversity, Liberty University at 0.08, having a decreasing diversity, Wellesley College at 0.075, having an increasing diversity, Dartmouth College at 0.13, having an increasing diversity, and Swarthmore College at 0.08, having a decreasing diversity.

Our second hypothesis, that liberal arts colleges would have a greater diversity of ideas, held up well from our results on Wellesley College and Dartmouth College. However, analysis on the diversity of ideas from Swarthmore College (a prestigious liberal arts college) campus newspaper shows a general decrease over the last decade. That said, liberal arts colleges are the only ones in our analysis to have shown a positive trend on diversity of ideas compared to other types of colleges.

Our third hypothesis, although Liberty University had a decreasing trend in its diversity, its pairwise dispersion measure in 2022 was less than UC Berkeley's, but greater than Wellesley's, Stanford's, and UT Austin's. This result is intriguing because our initial guess was that religious specialty colleges should have a narrow scope of topics being discussed in their newspapers because of the guardrail on ideas imposed by the specialty nature of these institutions.

Our fourth hypothesis that coastal universities would have more diversity than southern colleges held up in this analysis.

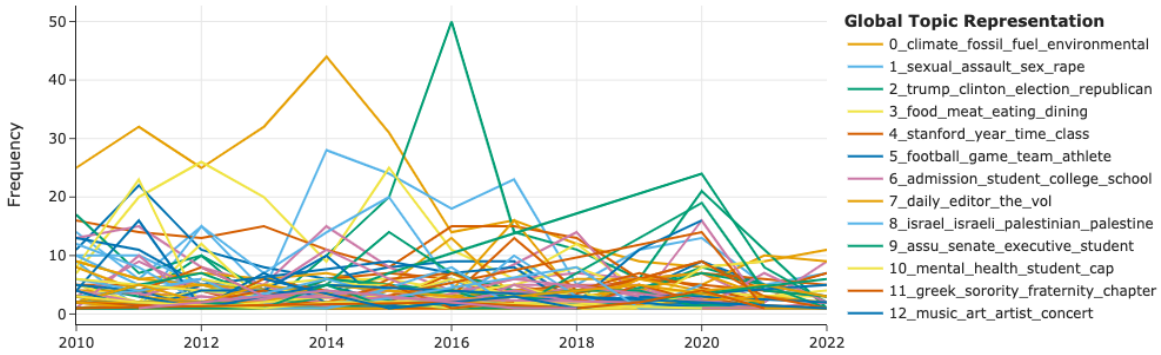

Figure 10. Trends of topics modeled for Stanford

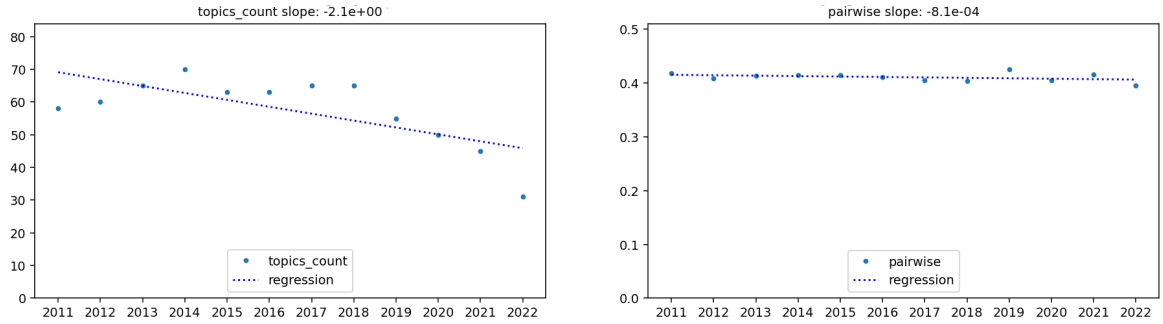

Figure 11. Trends of number of topics per year and topics pairwise dispersion for Stanford

UC Berkeley had a higher pairwise dispersion in 2022 than UT Austin, while Stanford had largely similar measures.

Our fifth hypothesis that larger colleges would have more diversity of ideas than smaller ones was proven incorrect. UT Austin, the largest of the sampled colleges with almost 52,000 students, had a pairwise dispersion of 0.055 which was the lowest measure of all of the sampled colleges including much smaller liberal arts colleges.

6. Conclusion

Explicit mentions of diversity have definitely increased across colleges in the past decade. Our initial hypotheses have been proven largely incorrect: that more prestigious colleges would have a higher diversity of ideas, that specialty colleges would have a lower diversity of ideas, and that larger colleges would have greater diversity ideas compared to smaller colleges. In fact, the diversity of ideas (as measured by size of embedding space covered or by number of topics discussed) has mostly been decreasing in most colleges in our study over the past decade. This seems to indicate that while these colleges are encouraging more discourse about diversity, the actual space of ideas being discussed in school newspaper opinion sections has unfortunately been shrinking.

In conclusion, while colleges proclaim their desire to promote diversity on their campuses, our research suggests that this desire might not always be reflected in the growth of diverse ideas on campus (as proxied by articles in college newspapers). To ensure that diverse voices and perspectives are not drowned out, it is important that we continue to shed light and address this issue.

References

Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Park, M., Leahey, E., and Funk, R. J. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942): 138–144, 2023. doi: 10.1038/s41586-022-05543-x.

Rasmussen, L. C. Increasing politicization and homogeneity

in scientific funding: An analysis of nsf grants, 1990-2020. 2021.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.