Exploring the Interplay Between Student Responses and Academic Performance

An NLP-Driven Analysis of Student Data

Final Project Report for Natural Language Processing Elective

MSDS 2024 - LT2

Laylo, Rex Gregor

Vasquez, Ana Dominique

Zabala, James Daniel

Table of Contents

List of Figures

Introduction

In today's rapidly evolving educational landscape, the integration of high-tech solutions has transformed traditional learning environments. From digital platforms to innovative teaching tools, the use of technology in education provides an abundance of data that can be analyzed to gain deeper insights into student experiences and performance. This project focuses on data derived from an online learning platform, aiming to explore the relationship between student feedback and academic outcomes.

The dataset consists of three key components: student entries, student grades, and student responses. These dataframes contain critical details, such as the topic, scores, and textual responses, providing a comprehensive view of student interaction and performance. The goal of this project is to analyze this dataset using Natural Language Processing (NLP) techniques to uncover patterns and correlations between student responses and their grades.

Through this project, we will address the following key research questions: Is there a relationship between the sentiment of student feedback and academic performance? Does student engagement in online learning forums correlate with grades? What are the most common topics in student responses, and do these themes correlate with grades? By exploring these questions, we aim to gain a deeper understanding of how student perceptions and feedback might reflect or influence their learning success.

The project will involve data preparation, including anonymization and merging of datasets, followed by a series of NLP techniques such as sentiment analysis, topic modeling, and text classification. The findings will be presented

through visualizations and a detailed report outlining the methodology, analysis, and conclusions.

Objectives

The primary objective of this project is to analyze data from a high-tech educational environment using Natural Language Processing (NLP) techniques to explore relationships between student feedback and their academic performance. Specifically, the research aims to address the following objectives:

1. Examine the Relationship Between Response Length and Grades

   Analyze the relationship between the length of student responses and their grades. This objective seeks to determine whether more elaborate and detailed inputs from students are correlated with higher academic achievement, providing insights into how student engagement with reflective or evaluative feedback relates to their success in coursework.

2. *Classify Student Feedback Sentiment and Assess Its Impact on Grades*

   Use NLP techniques to classify student feedback into different categories (e.g., positive, negative, or neutral) and examine how these classifications relate to academic performance. The goal is to explore whether the sentiment or type of feedback students provide has any meaningful correlation with their grades, and to identify any patterns in how feedback type may be connected to learning outcomes.

These objectives will be achieved through data preparation, merging, and the application of NLP methodologies such as text classification and correlation analysis.

By fulfilling these objectives, the project aims to contribute to a deeper understanding of student behavior in an online platform educational setting and offer insights into the ways student engagement and feedback can impact academic success.

Data Description

For this project, two out of the three available datasets were utilized: *Student Grades Data* and *Response Data*. These datasets were merged based on the user ID and context ID, resulting in the *Merged Dataset,* to explore relationships between student feedback and academic performance. Below is a description of each dataset and the key features used in the analysis.

Response Data

Table 1. *Feature description for response dataframe*

Feature Name	type	Description
metadata_context_id	string	represents the subject or course context
body_user_id	string	identifies the user (student) who responded.
body_text	string	contains the actual response in the forum

Student Grades

Table 2. *Feature description for student grades dataframe*

Feature Name	type	Description
metadata_context_id	string	represents the subject or course context
body_user_id	string	identifies the user (student) who responded.
body_score	float	the score the student received for the task.
body_points_possible	float	the total points possible for the task.

Merged Dataset

The *Merged Dataset* is the result of merging both the *Response Data and Student Grades data*. This dataset comprises of the merging keys from the two

original datasets plus the addition of features extracted or engineered from the original set of features. The resulting dataframe consists of the following engineered features:

Table 3. *Feature description of the merged dataset*

Feature Name	**type**	**Description**
lang	string	the detected language in the feature *body_text*
avg_word_count	float	average word count of the responses of a user per course context
body_score_ratio	float	ratio of *body_score* and *body_points_possible*
entry_count	float	number of entries a user made for a course context

Exploratory Data Analysis

To have an initial overview of what the data is exploratory data analysis is performed on the dataset.

Since one of the objective is to check the effect of response length on grades, checking the distribution of response length would be beneficial. Figure 1 shows a right skewed distribution of word count for responses. This means that majority of the responses are less than 100 words. The distribution is skewed by a few responses reaching up to 1,200 words in length.

Regarding scores, examining the typical points possible for a response can indicate whether normalization is necessary, as shown in Figure 2. Since the maximum possible score varies from 10 to 100 points, normalizing the scores could be advantageous for comparative analysis, allowing responses with different scoring scales to be evaluated consistently.

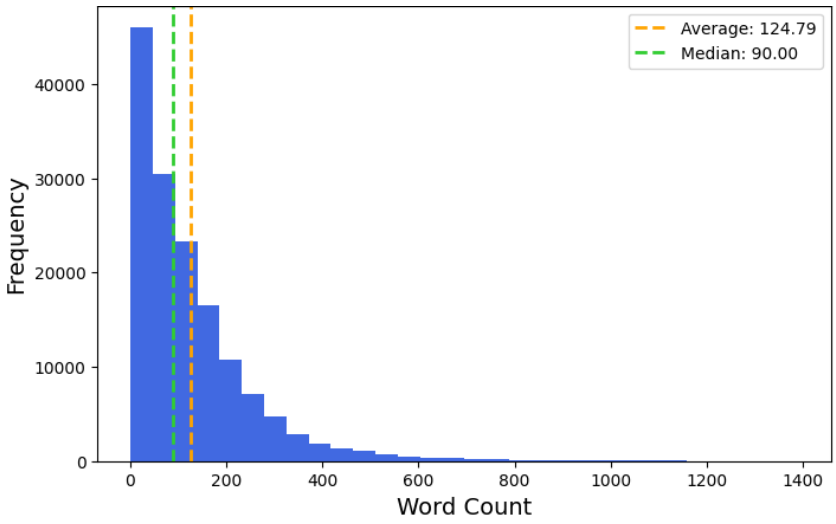


Figure 1. *Distribution of Word Count per response*

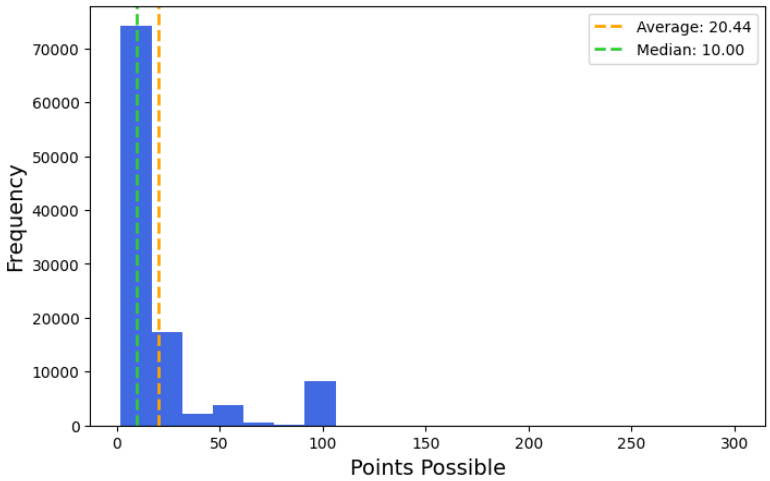


Figure 2. *Distribution of Points Possible*

Exploring the raw scores is also beneficial, as the goal is to examine whether the response features explain variations in the scores. The histogram in Figure 3 reveals a right-skewed distribution, mirroring the pattern observed in the distribution of possible points. This suggests that most responses cluster around higher scores, with fewer responses receiving lower evaluations.

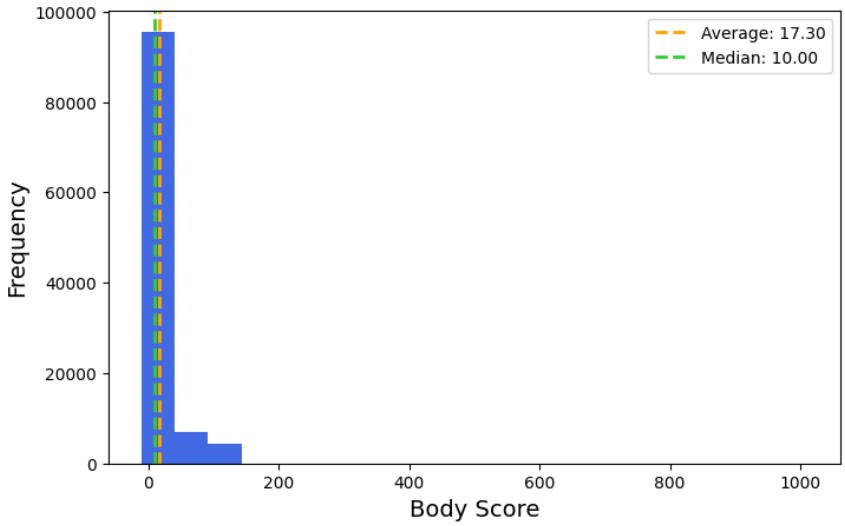


Figure 3. *Distribution of response scores*

The histogram in Figure 3 indicates a potential anomaly, displaying a right-skewed distribution with values reaching close to 1,000, even though the possible points cap at around 300. A quick glance at the boxplot in Figure 4 further highlights these anomalies in the response scores. Removing these outliers could improve the accuracy and reliability of the subsequent correlation analysis.

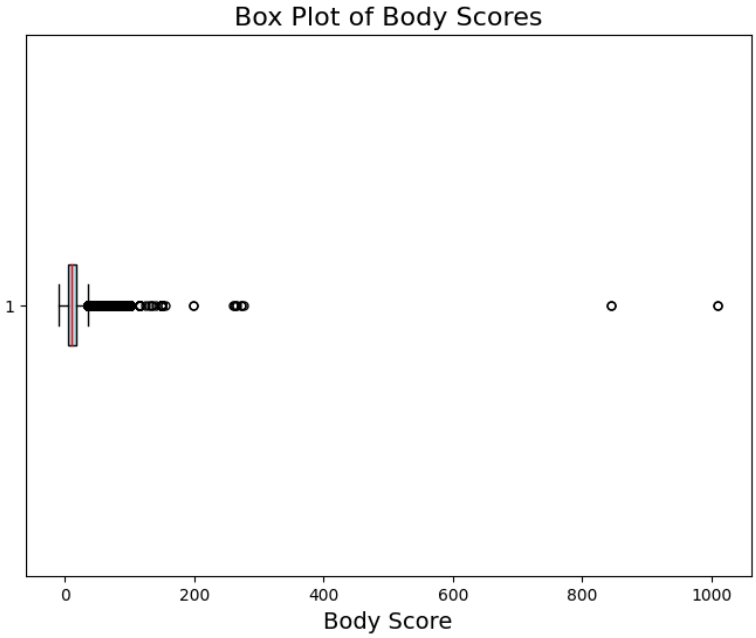


Figure 4. *Boxplot of response scores*

Methodology

The methodology for this project involves several key steps to ensure proper analysis of the student data. First, feature engineering is performed to create relevant features for analysis. Next, data merging integrates the textual data from multiple sources, ensuring a cohesive dataset for evaluation. Data anonymization is conducted to protect privacy, followed by comprehensive preprocessing, which includes removing non-English rows, converting text to lowercase, eliminating stopwords, and lemmatizing words. Subsequently, topic modeling is applied to uncover common themes in student feedback, while sentiment analysis determines the sentiment of responses. Finally, correlation analysis explores the relationship between response features and academic performance, providing insights into the impact of feedback characteristics on student grades.

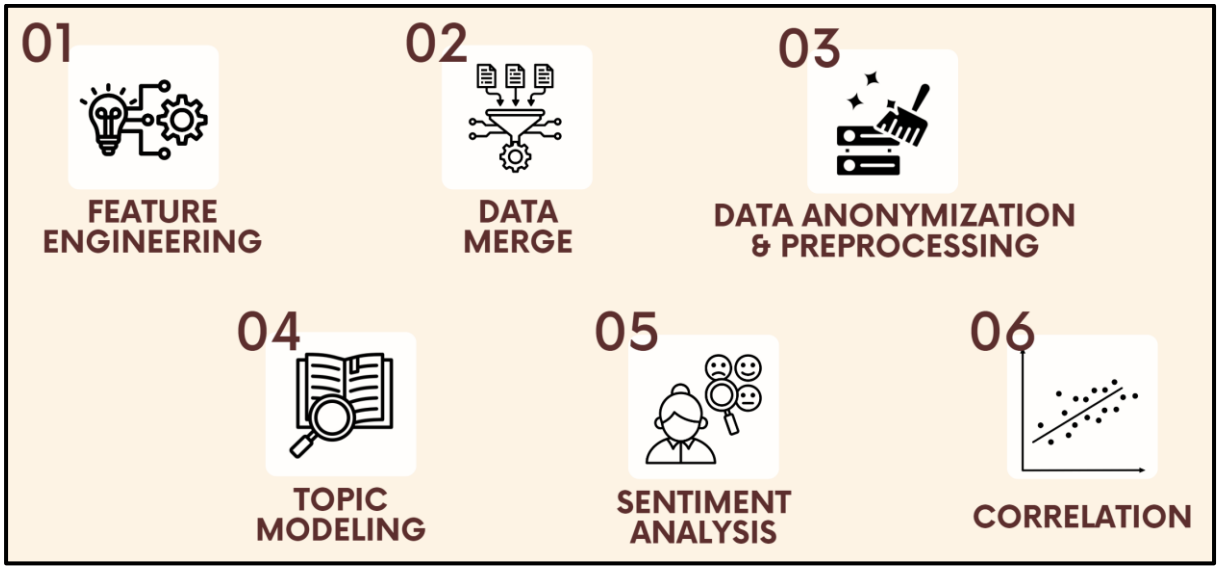


Figure 5. *Methodology Pipeline*

Feature Engineering

To enrich the dataset and provide more meaningful insights, several features were engineered from the existing data. These features were derived from the textual data in the responses dataframe and numerical data in the grades dataframe. These features are essential for further analysis, such as sentiment detection and correlation studies. The engineered features include the following:

- **lang**

  This feature represents the detected language of each student response. It was extracted using the *LangDetect* package, which helps identify and filter non-English entries during the preprocessing stage.

- **average_word_count**

  This feature captures the average number of words used in the student responses. It was calculated by first grouping the data by *body_user_id* and *metadata_context_id* and then aggregating the word count of each body_text within these groups. This provides insight into the verbosity and detail of the responses.

- **entry_count**

  This feature represents the total number of entries or responses made by each student. It was computed by performing a simple count of the body_text for each grouped body_user_id and context_id. This feature helps to quantify the frequency of each student's engagement across different contexts.

- **body_score_ratio**

This feature represents the ratio of a student's score (body_score) to the possible points for that response (body_score_possible_points). It provides a normalized measure of performance, allowing for comparison across different assignments with varying point values. Only valid score ratio between 0 and 1 is considered. This ensures that scores exceeding 1, possibly due to bonus points, and scores below 0, possibly due to deductions or errors, are excluded from the analysis to maintain data accuracy.

Data Merging

To analyze the relationship between student responses and academic performance, the response dataframe was merged with the student grades dataframe using *body_user_id* and *metadata_context_id* as the common key. An inner join was applied to ensure only entries with matches on both datasets are retained. This combined dataset provided a comprehensive foundation for exploring how features like response length and sentiment relate to academic performance.

Data Anonymization & PreProcessing

Data Anonymization

To safeguard the privacy of individuals in the dataset, Named Entity Recognition (NER) was applied to detect and anonymize personally identifiable information (PII). A transformer-based Named Entity Recognition (NER) model was used, with a confidence threshold of 0.85. This threshold was chosen to ensure a high level of confidence in identifying names by the NER model, although it may result in some names being missed, balancing precision and recall in NER.

Anonymization can also improve the results of topic modeling or sentiment analysis by providing more context on the statements. Without NER some important context-specific entities might be ignored or misinterpreted.

Data Preprocessing

1. *Language Filtering*

   Non-English responses were removed using the *LangDetect* package to ensure that only English responses were included in the analysis. This step helps enhance the accuracy and consistency of topic modeling by focusing on a single language, reducing noise and improving the quality of the results.

2. *Lowercase*

   To maintain consistency and improve the performance of text analysis, all responses were converted to lowercase. This step ensures that words are treated uniformly, preventing distinctions between uppercase and lowercase forms of the same word (e.g., "Student" and "student"). Lowercasing is essential for reducing noise in tasks like tokenization, sentiment analysis, and topic modeling.

3. *Removal of Stopwords*

   Stopwords, which are common words such as "the," "is," and "and," were removed from the text to focus the analysis on meaningful content. These words provide little value in natural language processing (NLP) tasks as they do not contribute to the overall meaning of the text. The *NLTK* library was used for this step, leveraging its comprehensive list of English stopwords. By removing these frequently occurring but uninformative words, the analysis can better capture the key themes and sentiments in the data.

4. *Word Lemmatization*

To normalize the text and reduce words to their base or root form, lemmatization was applied using *NLTK's WordNet* lemmatizer. This process helps group together different forms of the same word, improving the accuracy of tasks such as topic modeling and sentiment analysis by ensuring that variations of a word are treated as a single entity.

The final processed dataset comprises **18,115 rows** of cleaned and refined responses, ready for detailed analysis.

Topic Modelling

To identify and analyze the underlying themes in the textual data, the following steps were undertaken:

Document Vectorization with TF-IDF

The text data was converted into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. This method transforms the text into vectors based on the importance of each term in relation to the entire corpus, capturing the significance of words while reducing the impact of frequently occurring but less informative terms.

Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF) was chosen for topic modeling due to its interpretability. The word components in NMF are easier to understand because the values are additive (e.g., frequency-based), making the topics more meaningful and coherent. Metrics such as the Reconstruction

Error, Silhouette Score and Coherence Score were used to determine the best number of topics.

Other topic modeling techniques, such as LDA and BERTopic, were also applied for a comparative view and are detailed in the Appendix.

Sentiment Analysis

Sentiment analysis was conducted to assess the overall sentiment expressed in the student responses. TextBlob was utilized for this analysis due to its simplicity and effectiveness in extracting sentiment scores for long formal responses found in student forums. TextBlob provides polarity and subjectivity scores for each response, where polarity indicates the sentiment orientation (positive, negative, or neutral). The polarity scores were extracted for each row.

Correlation Analysis

A correlation analysis was conducted to explore the relationships between various features in the dataset, focusing on how sentiment scores and text-derived features relate to academic performance metrics, such as grades. Spearman's correlation was chosen for its robustness in handling non-normal distributions, making it well-suited for skewed data like student grades. This method was used to assess the strength and direction of the relationships, offering insights into whether positive or negative sentiments, along with other text features, significantly impact student performance. The findings were then used to generate insights on the topic.

Analysis and Findings

The topic modeling analysis identified nine topics as optimal, based on localized peaks in both silhouette and coherence scores as shown in Figure 6. These metrics demonstrated well-defined clusters and meaningful, coherent topics.

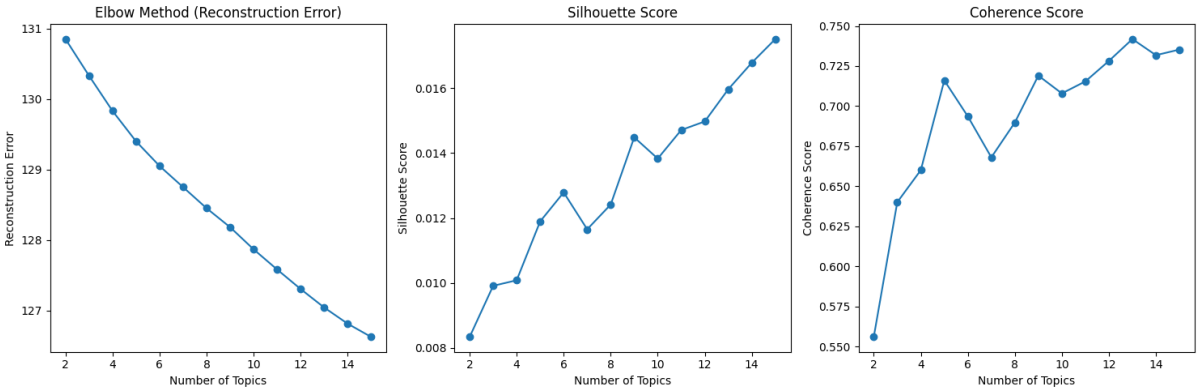


Figure 6. *Clustering metrics for Topic Modeling*

The topics would then be labeled based on the content of the text in each topic. To get a better visualization of the content of each topic, word clouds were generated for each topic as shown in Figure 7. The topics generated were inferred on the content of each topic, with the context of educational online platforms in mind. The topics derived are:

- Reflections
- Fitness
- Student Life
- Business
- Dance/ Music
- Legal Research Assessment
- Philippine Development
- Psychometric Testing
- Financial Scandals

Figure 7. *Word Cloud generated for each topic*

After identifying and labeling the topics, the responses are classified by assigning each of them to the topic with the highest probability score. This ensures that each response is grouped with the most relevant topic based on its content. This would enable us to count how many responses there are for each topic as shown in Figure 8.

Additionally, the histogram in Figure 9 displays the maximum probability scores for documents assigned to a particular topic, reflecting the confidence levels in topic assignment. This histogram reveals the variability in the strength of topic associations, offering insight into the consistency and reliability of the topic modeling outcomes.

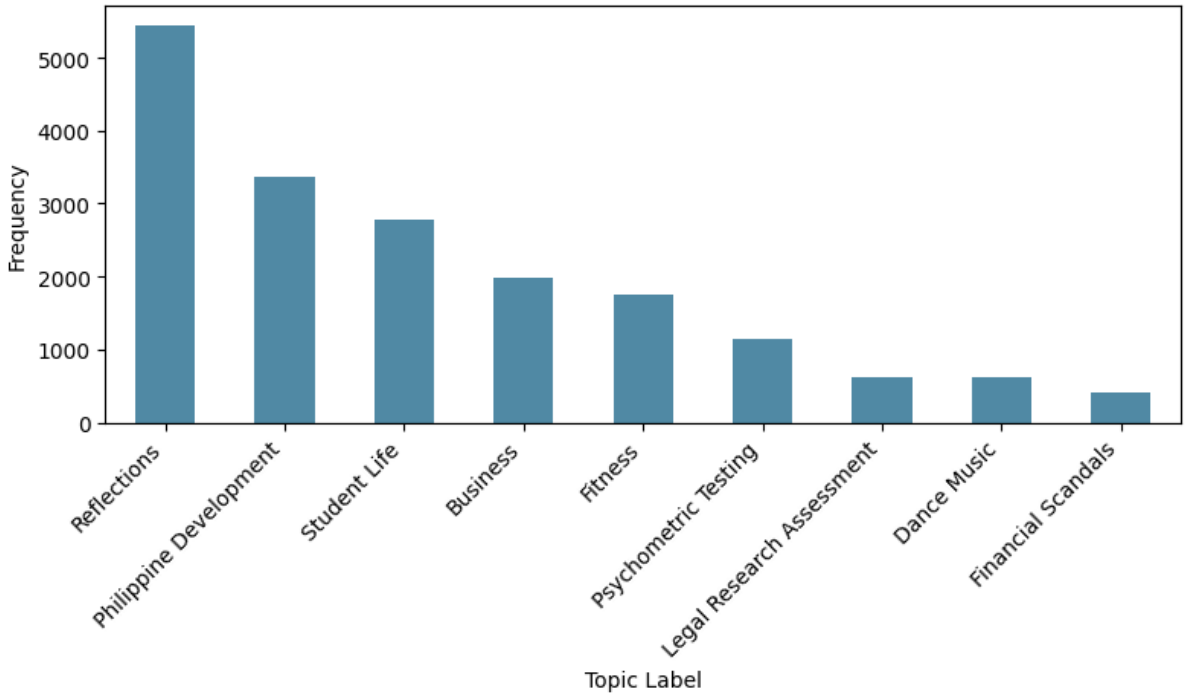


Figure 8. *Distribution of responses across topics*

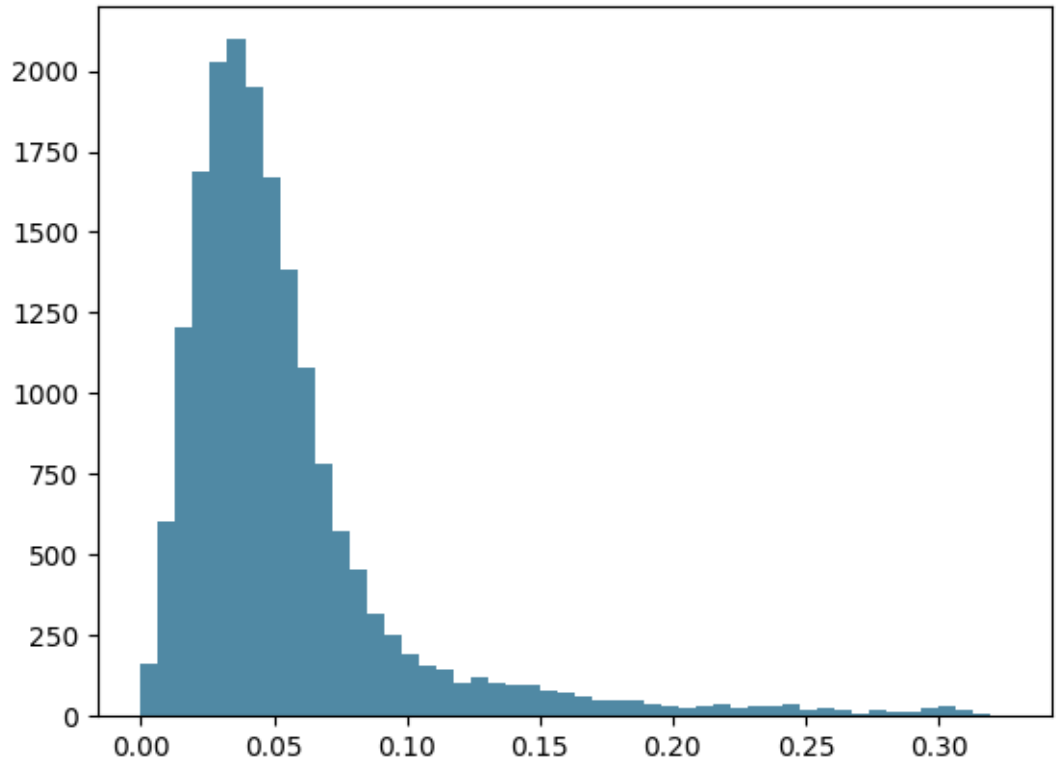


Figure 9. *Histogram of maximum probability scores for each response*

The results of our sentiment analysis are visualized in Figure 10, which shows the distribution of sentiment scores across the dataset. The figure clearly illustrates that the majority of responses exhibit positive sentiment, with scores clustering predominantly in the positive range. This indicates a generally favorable perception among the respondents. The distribution highlights that while there is a range of sentiment scores, the positive sentiment significantly outweighs the negative or neutral responses.

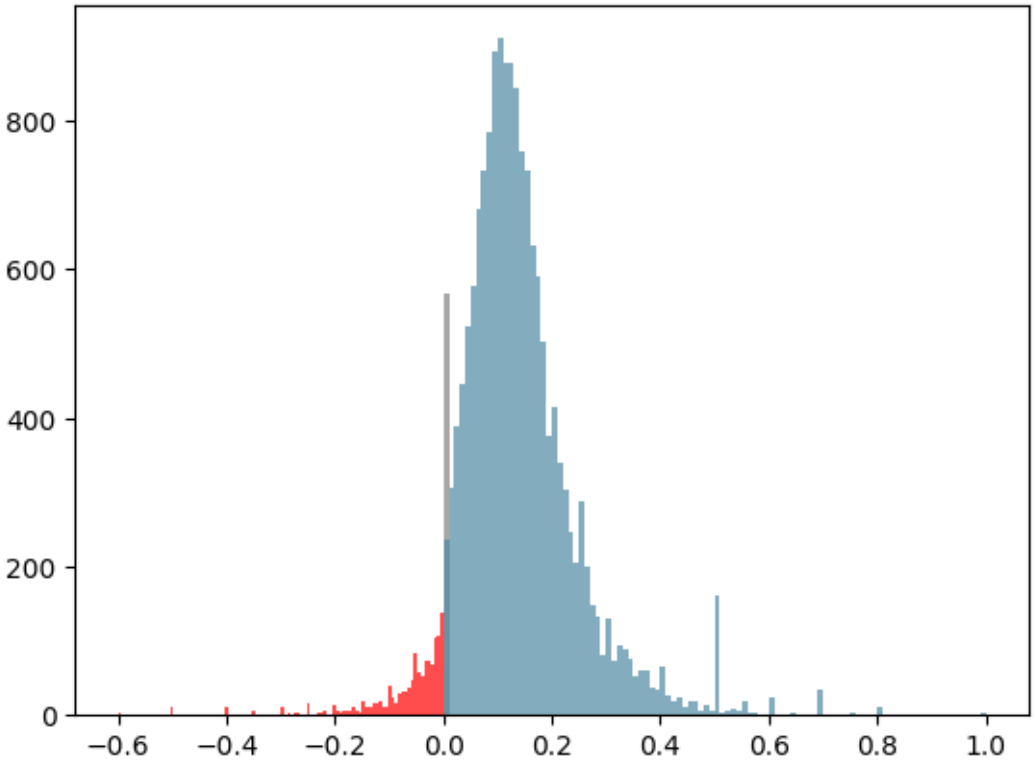


Figure 10. *Distribution of Sentiment Scores*

We can gain deeper insights by examining the sentiment score distribution for each topic, as illustrated in. This figure reveals that, overall, each topic's sentiment distribution reflects the general trend observed in the entire dataset, with positive sentiments prevailing over negative ones. However, notable variations emerge: the topic *"Fitness"* exhibits a higher proportion of negative sentiments, *"Dance Music"*

shows the highest proportion of positive sentiments, and *"Student Life"* contains the largest share of neutral statements. These distinctions provide a more nuanced understanding of sentiment across different topics.

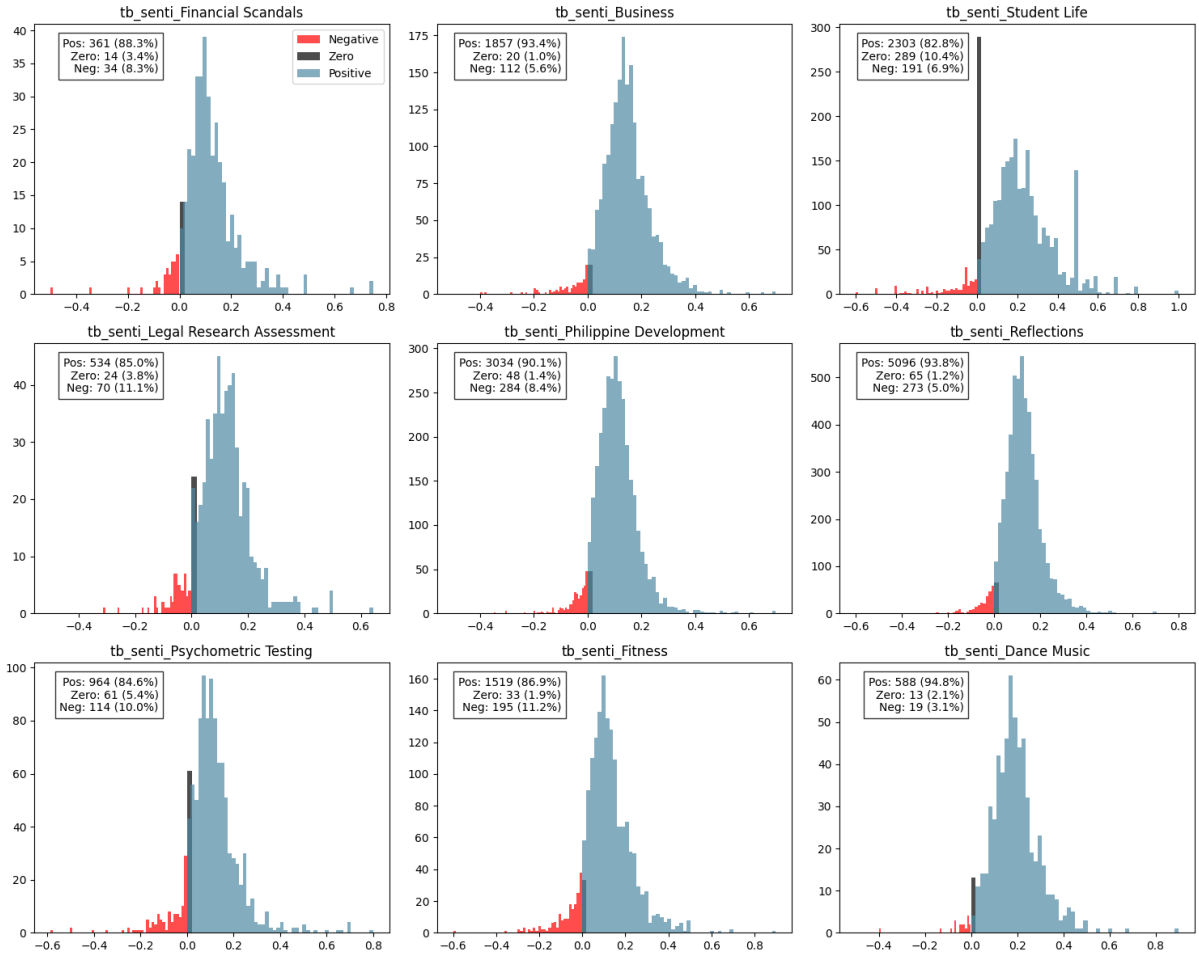


Figure 11. *Sentiment Score Distribution per Topic*

Some notable insights can be derived from the correlation analysis between scores and number of entries per topic. For example, in *Dance Music* **the number of responses of a student often corresponds to higher scores**. This could mean that for this topic, students who are more actively engaged in the discussion board tend to receive higher scores, suggesting that frequent participation might contribute to better performance or greater visibility in evaluations. If the scoring is based on the quality or impact of posts, this correlation might imply that students

who post more often are also contributing more valuable or relevant content, which is rewarded with higher scores.

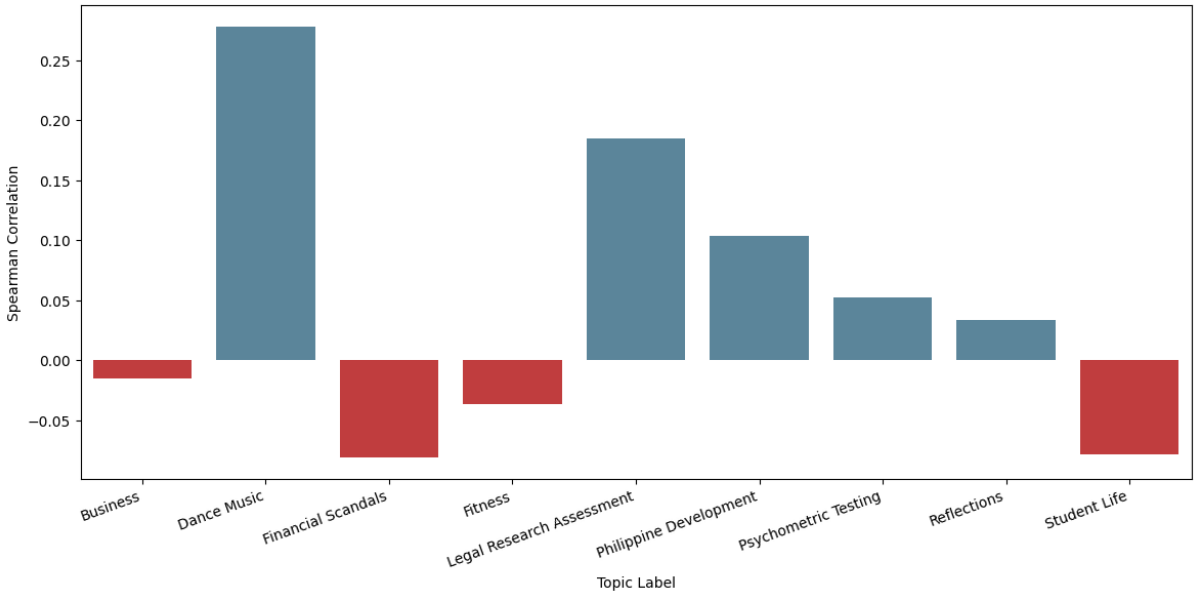


Figure 12. *Correlation between Scores and Number of Entries per Topic*

For certain topics, the correlation is negative, suggesting that increased responses may actually detract from the scores received. However, due to the low correlation in these cases, it's also possible that there is no significant relationship between the number of posts and the scores. This implies that, while more posts may sometimes be detrimental, the overall impact on scoring might be minimal or inconsistent.

Figure 13 shows the relationship of the number of entries a student posts vs. the percentage score they get for their responses. From the visualization it seems that the percentage scores stabilize and increase towards the perfect score as the number of entries increase. This suggests that increased participation is associated with improved scores, and consistent engagement in posting may enhance overall performance in the discussion board.

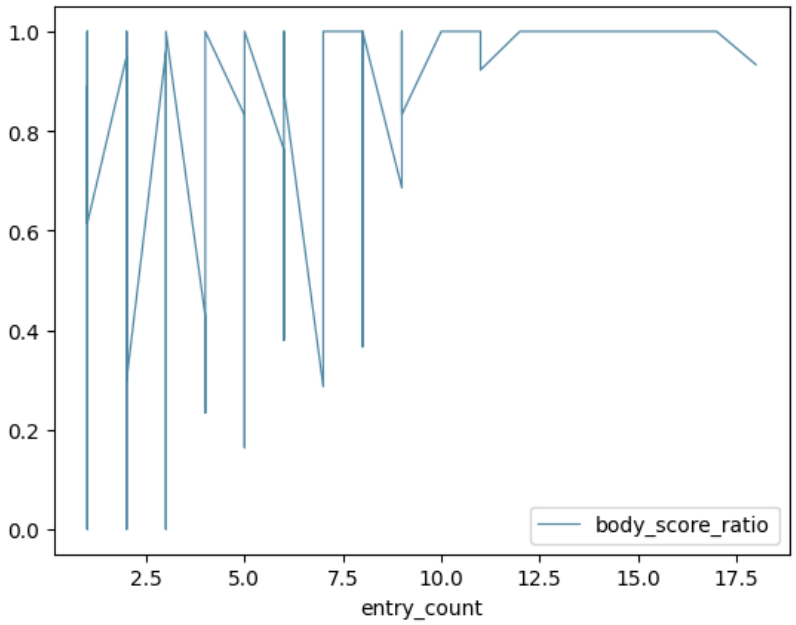


***Figure 13**. Plot of entry_count vs body_score_ratio*

The correlation analysis of word count shown in Figure 14 reveals that, for certain topics, **longer responses are associated with higher scores**. This trend is particularly evident in topics such as Financial Scandals and Legal Research Assessment. This suggests that, for these subjects, more detailed and extensive responses are likely to be valued and rewarded with higher scores. Consequently, a higher word count may contribute to more thorough and comprehensive answers, which are recognized and positively evaluated in these areas.

As shown in Figure 15, the correlation between average word count and percentage scores mirrors the relationship observed with the number of entries. The data suggests that percentage scores tend to stabilize and increase as the average word count rises. This indicates that, in most cases, longer responses are associated with higher scores, suggesting that more in-depth responses are generally rewarded with better evaluations.

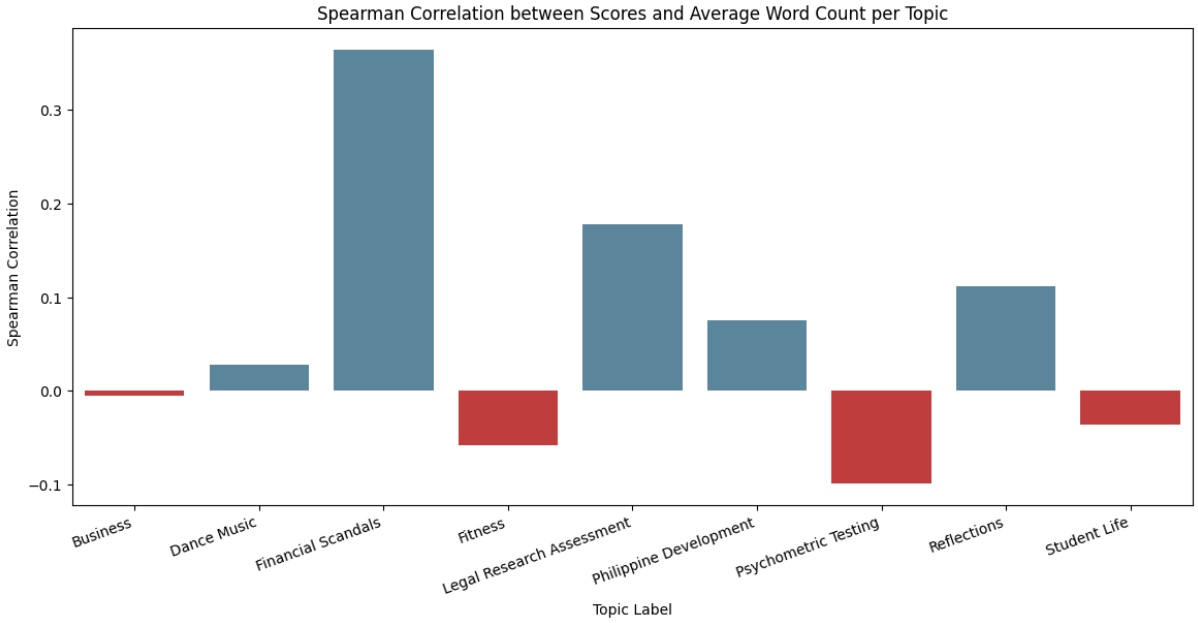


Figure 14. *Spearman Correlation of scores and avg word count per topic*

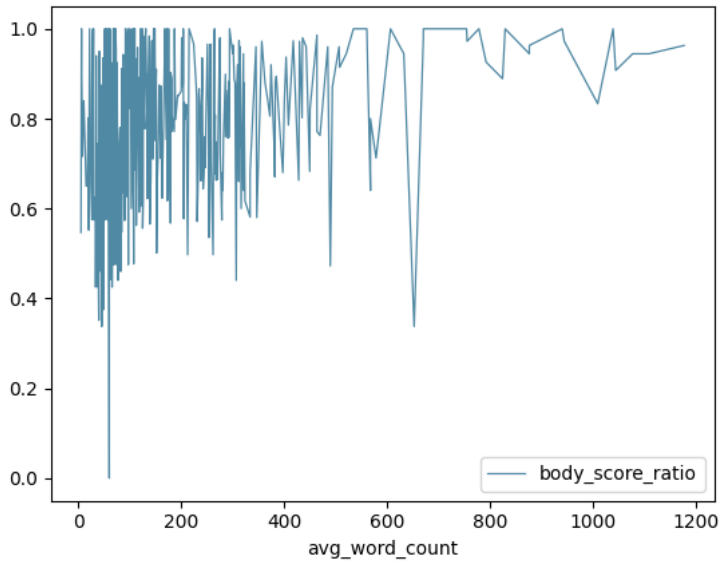


Figure 15. *Plot of avg_word_count vs body_score_ratio*

The correlation between response sentiment and student scores appears to be minimal as shown in Figure 16, indicating that sentiment has little to no impact on the scores students receive. This is an expected outcome, as the sentiment of a

response should not influence the objective evaluation of its content. This finding reinforces the notion that scoring should be based on the quality and relevance of the response, rather than its emotional tone.

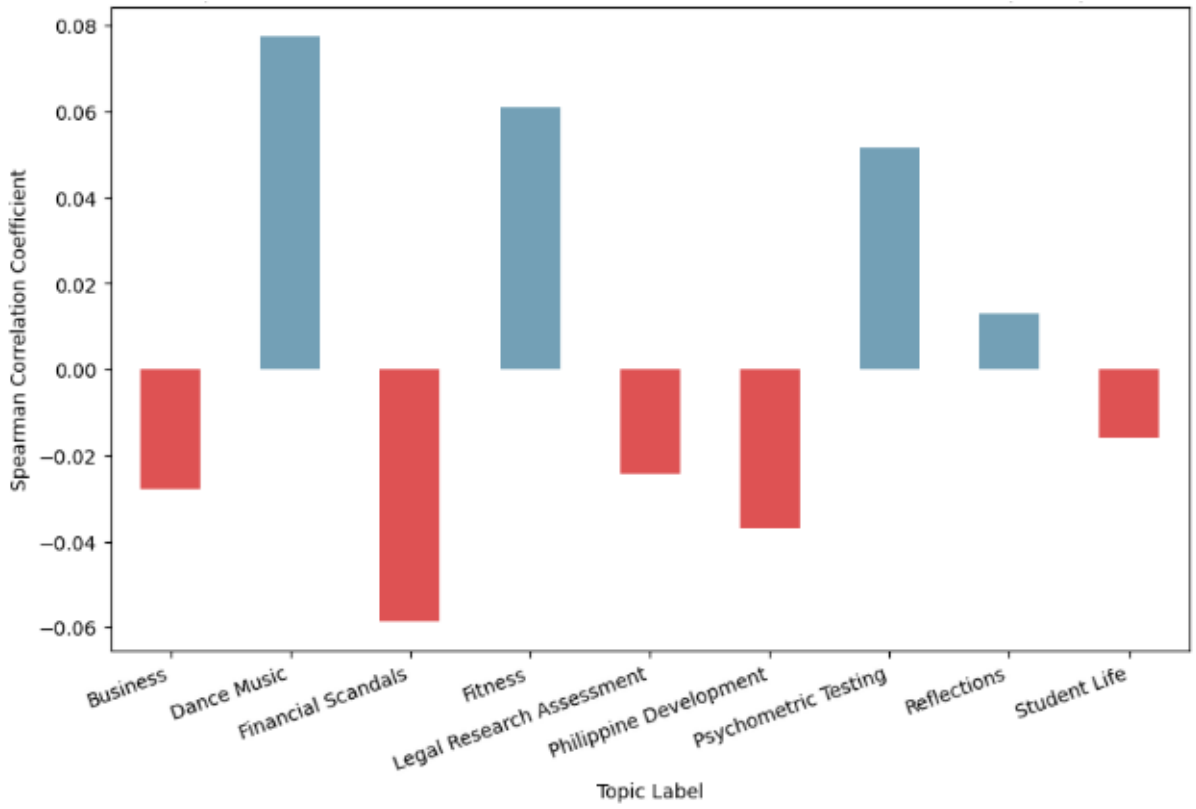


Figure 16. *Spearman Correlation between sentiment scores and response scores*

Figure 17 shows the relationship between student scores and sentiment scores. The correlation stabilizes as sentiment scores approach higher absolute values, indicating that the impact of sentiment on student scores becomes more consistent at extreme positive or negative sentiment levels.

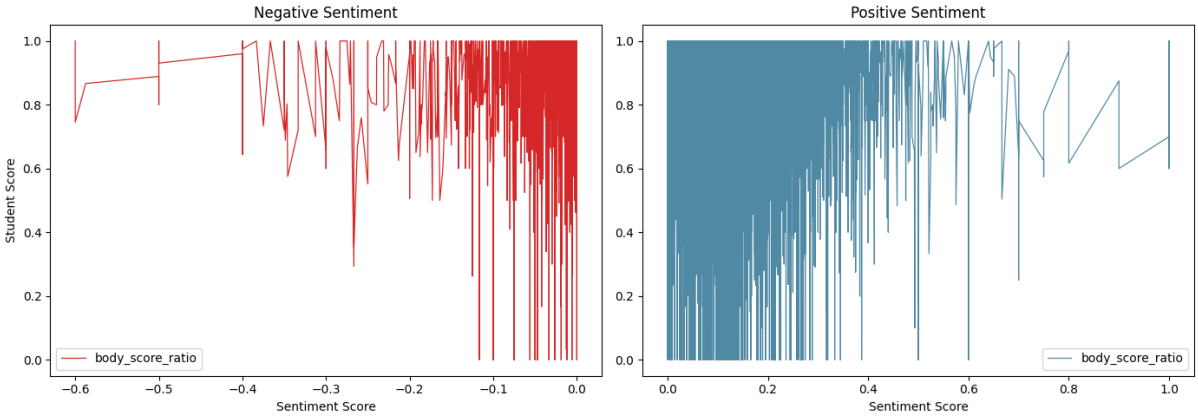


Figure 17. *Plot of sentiment scores vs student grade by Sentiment*

Conclusion

The correlation analysis conducted in this study yielded valuable insights into the relationship between student responses and their evaluations.

Response Length

The investigation found a positive correlation between response length and evaluation scores for certain subjects, suggesting that more detailed answers tend to receive higher ratings. This trend implies that thoroughness and depth are important for achieving better evaluations. Key insights include:

1. **Knowledge Demonstration**: Longer responses may reflect a deeper understanding of the material.
2. **Critical Thinking**: Extended answers often show more analysis and critical thinking.
3. **Attention to Detail**: Detailed responses may include more supporting evidence, which evaluators value.
4. **Effort and Engagement**: Lengthier responses might indicate greater effort and engagement.
5. **Subject-Specific Variations**: The correlation's strength can vary by subject, with detailed answers being more valued in some fields than others.
6. **Potential Drawbacks**: Quantity does not always equal quality; overly lengthy responses can be detrimental if they lack focus or clarity.

Number of Entries

Analysis revealed a positive correlation between the number of student entries and their overall scores. This suggests that more frequent participation is associated with better performance. Possible explanations include consistent engagement,

regular practice, effective use of feedback, and better time management. However, while quantity matters, the quality of submissions remains crucial for academic success.

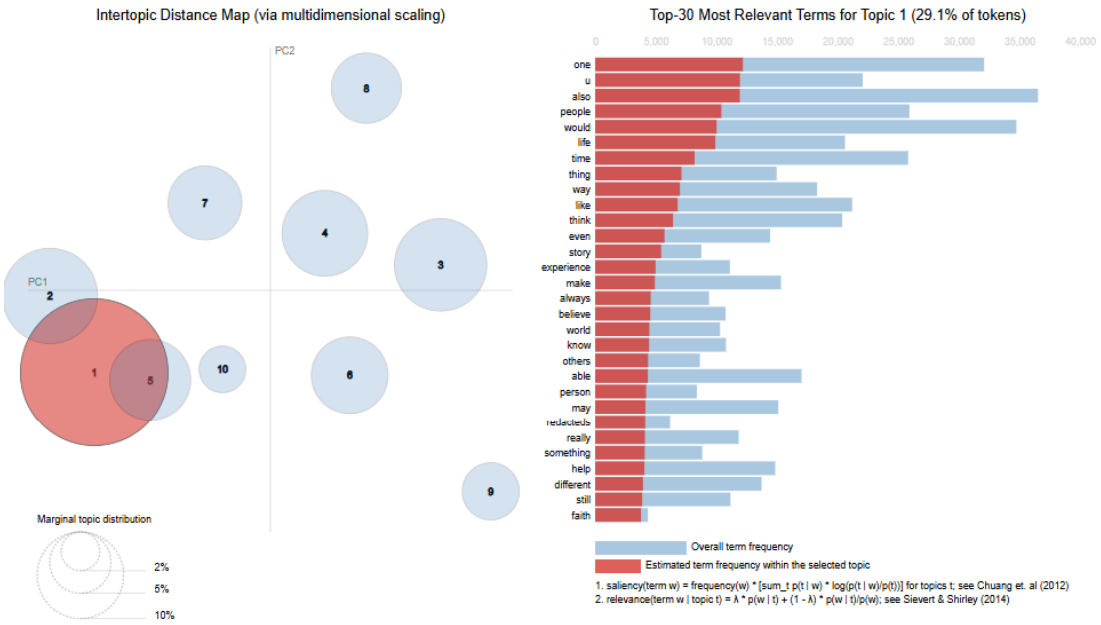
Response Sentiment

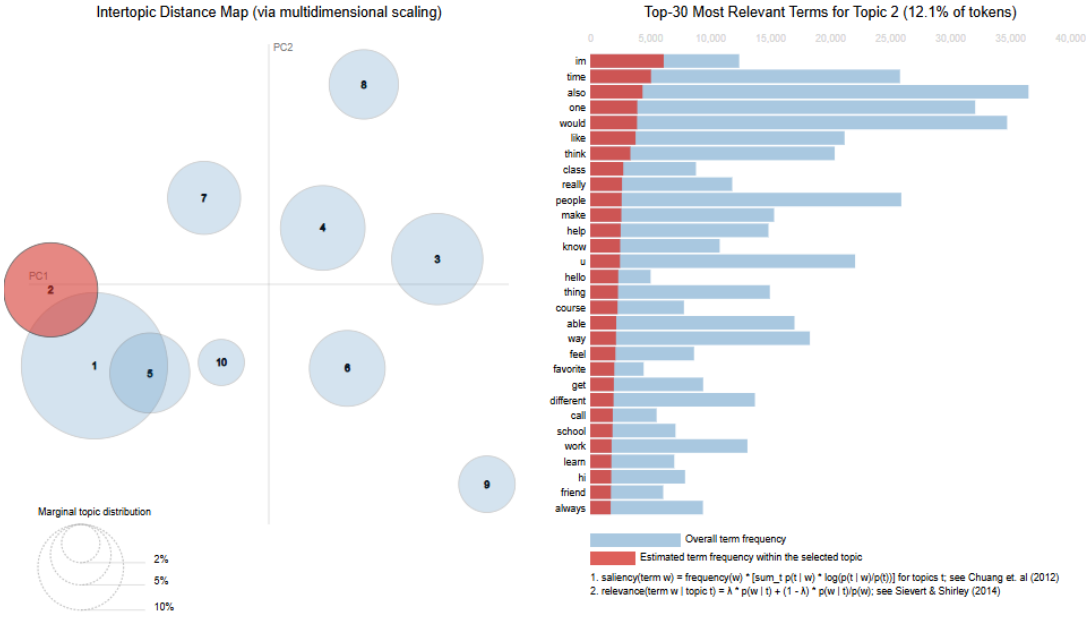
Analysis of sentiment scores versus evaluation outcomes revealed no significant correlation. This lack of relationship aligns with expectations, as the emotional tone of a response shouldn't ideally influence academic evaluations. The absence of correlation suggests that evaluators are appropriately focusing on content, accuracy, and relevance rather than the sentiment expressed in student responses. This finding reinforces the notion of objective assessment in academic settings.

Appendix

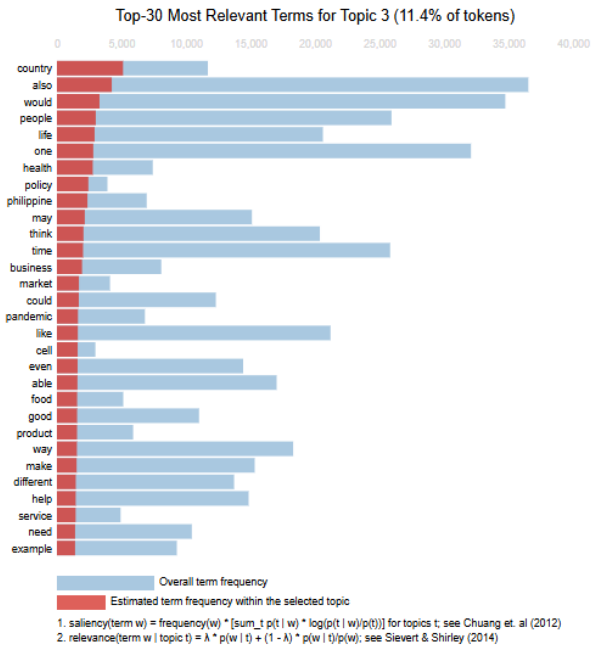
LDA: Topic Modeling

To compare NMF and LDA as topic modeling methods, LDA was used to generate 10 key topics from the student responses data.

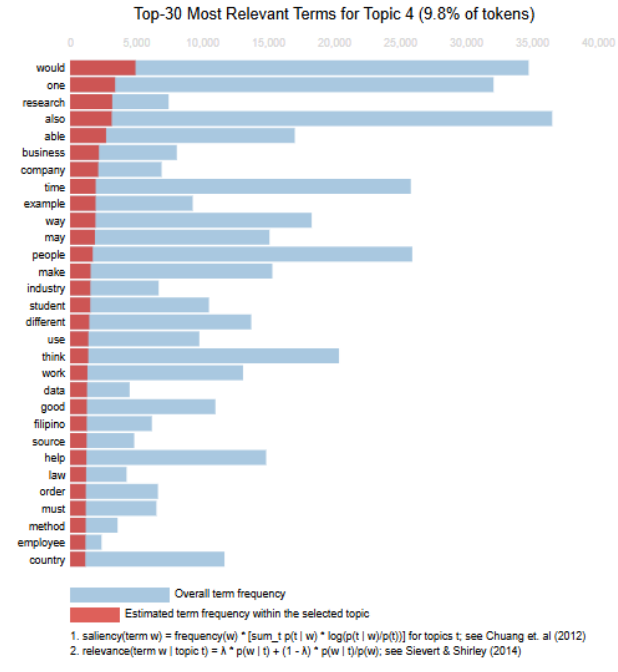


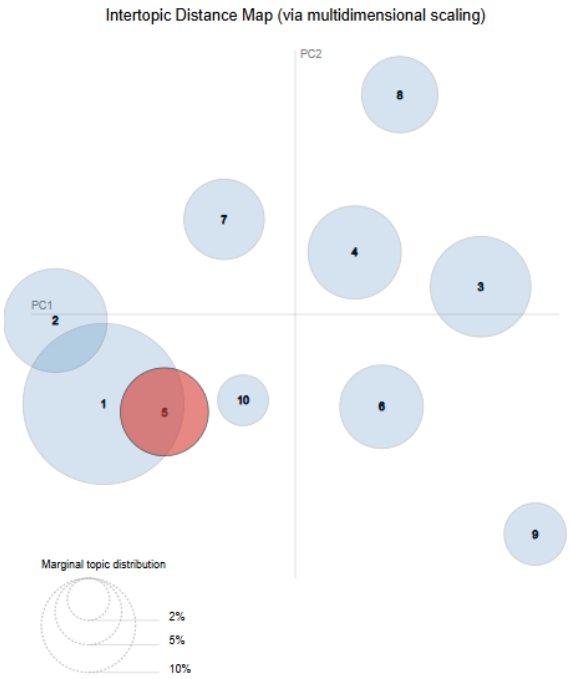


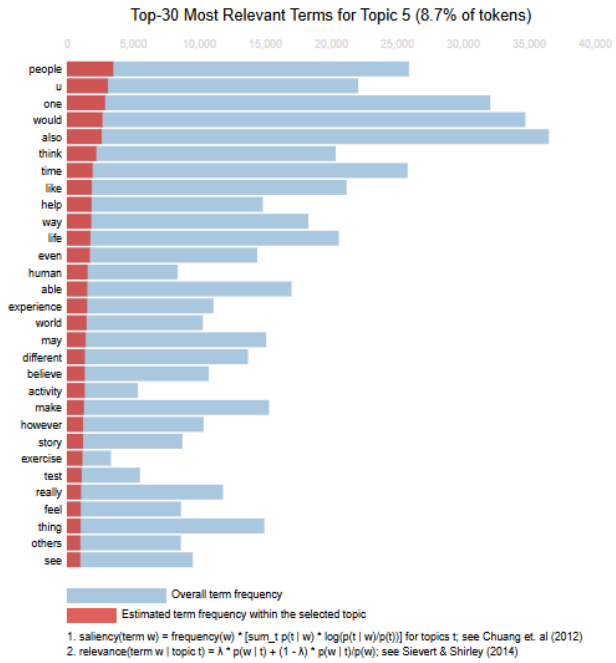

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
7
4
3
2
1
5
10
6
9
Marginal topic distribution
2%
5%
10%

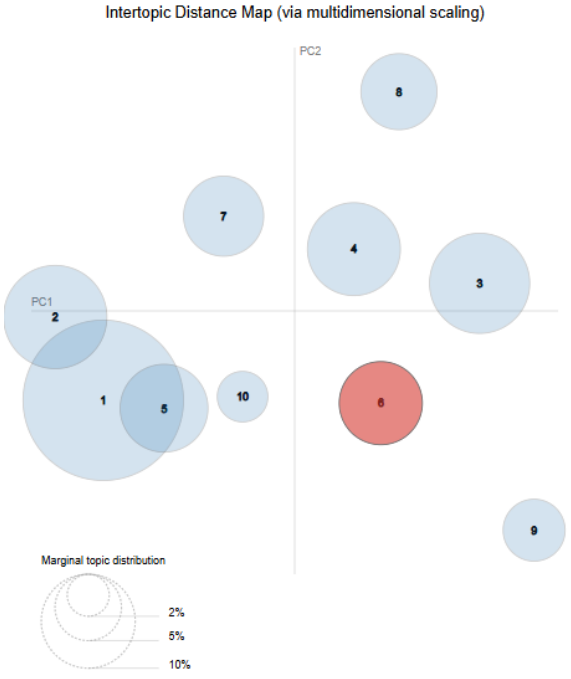

Top-30 Most Relevant Terms for Topic 3 (11.4% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
country
also
would
people
life
one
health
policy
philippine
may
think
time
business
market
could
pandemic
like
cell
even
able
food
good
product
way
make
different
help
service
need
example
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

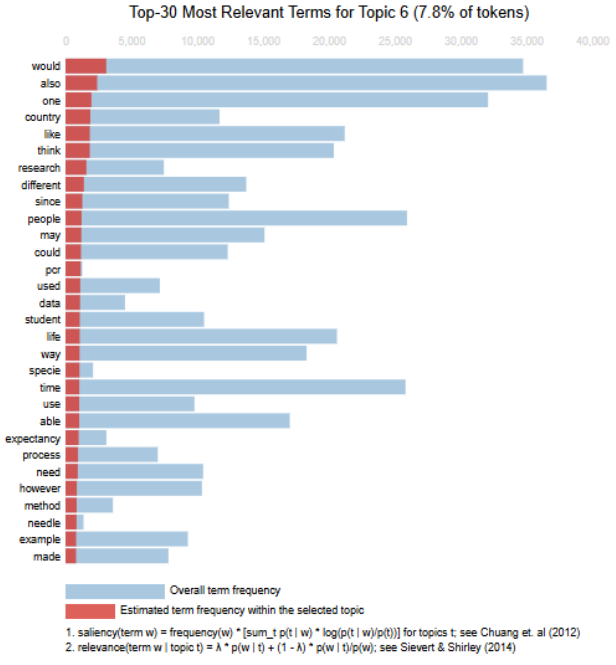

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
7
4
3
2
1
5
10
6
9
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 4 (9.8% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
would
one
research
also
able
business
company
time
example
way
may
people
make
industry
student
different
use
think
work
data
good
filipino
source
help
law
order
must
method
employee
country
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

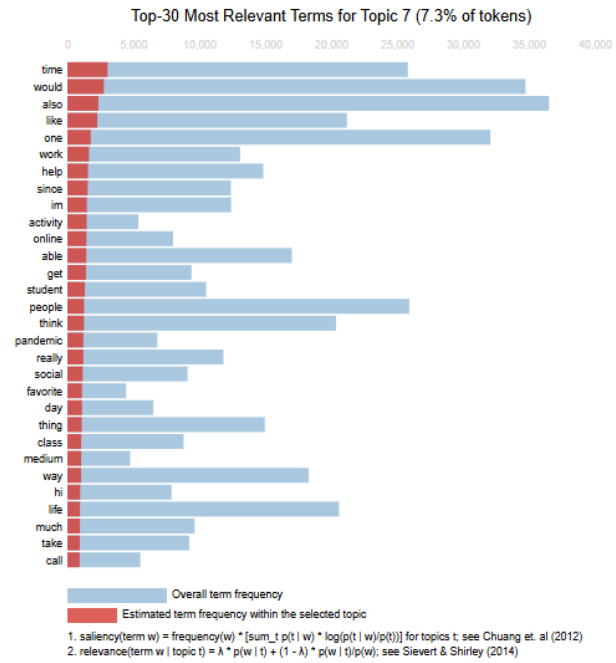

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
7
4
3
2
1
5
10
6
9
Marginal topic distribution
2%
5%
10%

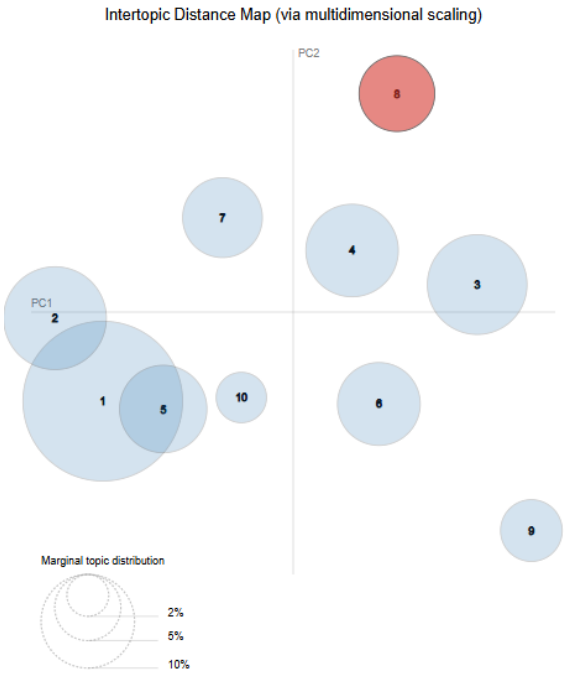

Top-30 Most Relevant Terms for Topic 5 (8.7% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
people
u
one
would
also
think
time
like
help
way
life
even
human
able
experience
world
may
different
believe
activity
make
however
story
exercise
test
really
feel
thing
others
see
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

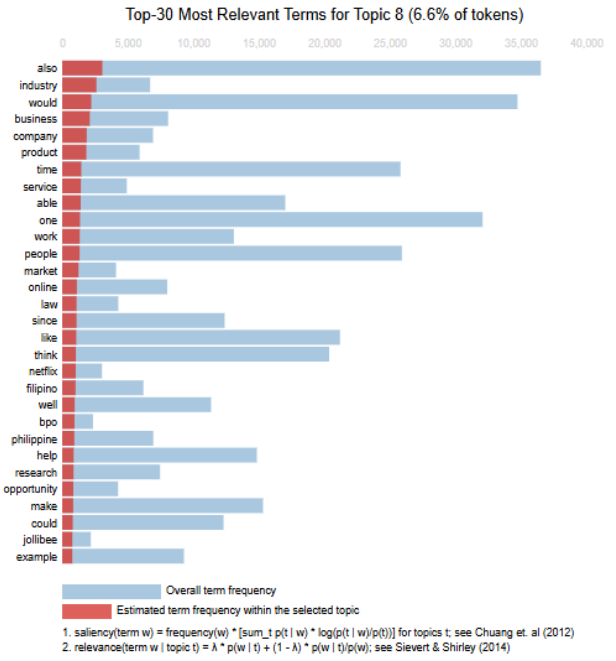

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
7
4
3
2
1
5
10
6
9
Marginal topic distribution
2%
5%
10%

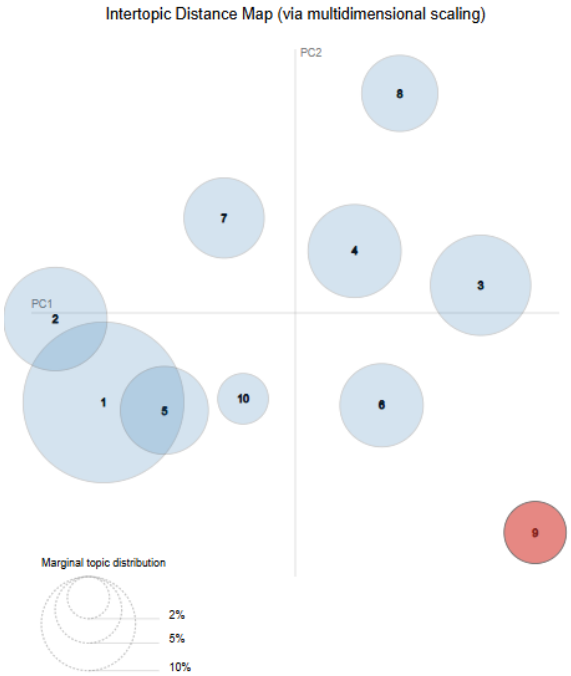

Top-30 Most Relevant Terms for Topic 6 (7.8% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
would
also
one
country
like
think
research
different
since
people
may
could
pcr
used
data
student
life
way
specie
time
use
able
expectancy
process
need
however
method
needle
example
made
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

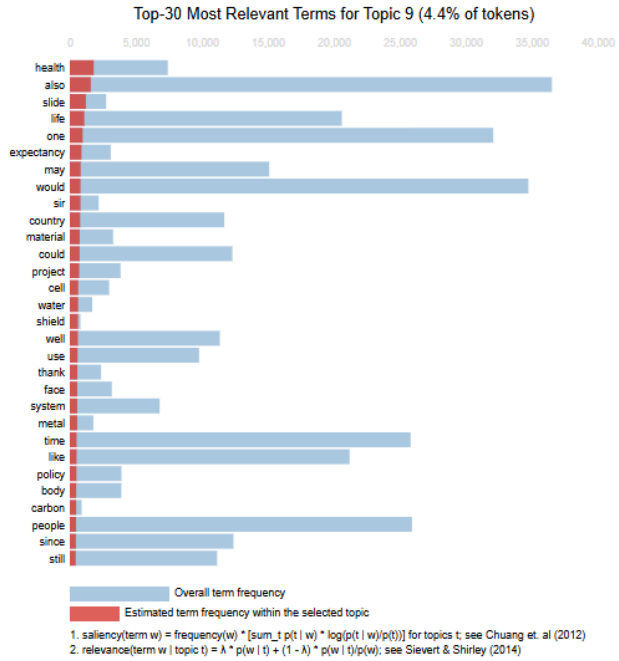

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
7
4
3
2
1
5
10
6
9
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 7 (7.3% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
time
would
also
like
one
work
help
since
im
activity
online
able
get
student
people
think
pandemic
really
social
favorite
day
thing
class
medium
way
hi
life
much
take
call
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

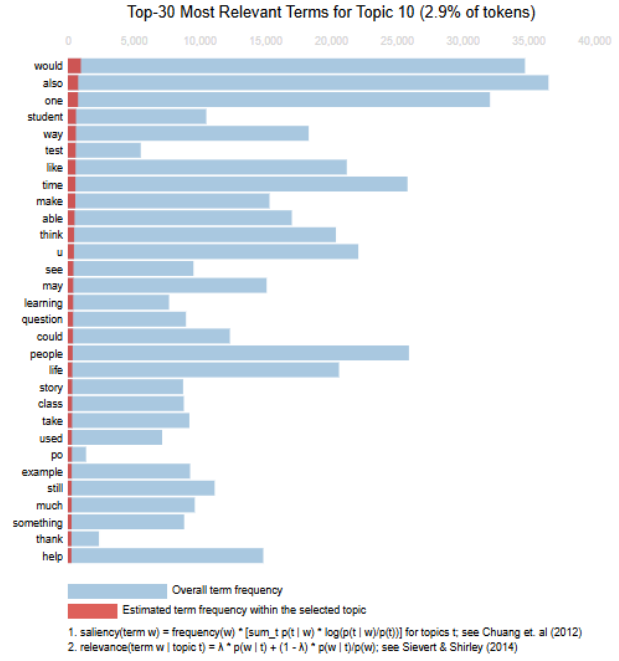

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
7
4
3
2
1
5
10
6
9
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 8 (6.6% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
also
industry
would
business
company
product
time
service
able
one
work
people
market
online
law
since
like
think
netflix
filipino
well
bpo
philippine
help
research
opportunity
make
could
jollibee
example
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

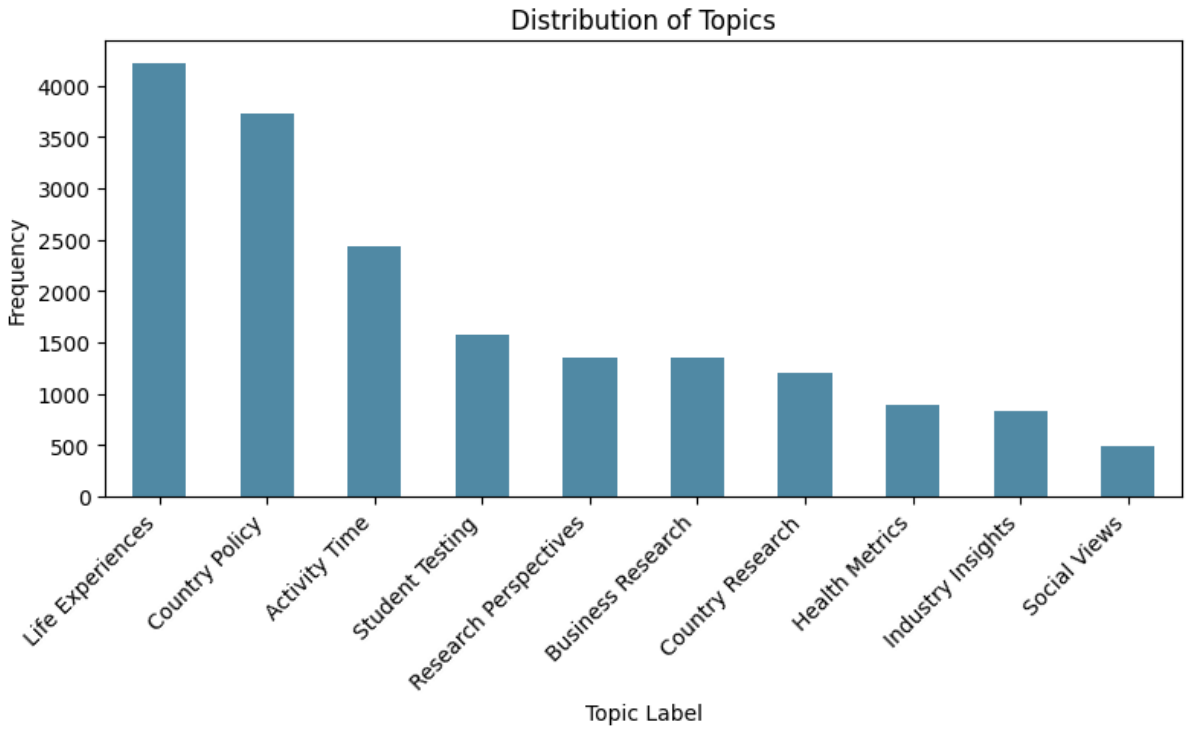

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
7
4
3
2
1
5
10
6
9
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 9 (4.4% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
health
also
slide
life
one
expectancy
may
would
sir
country
material
could
project
cell
water
shield
well
use
thank
face
system
metal
time
like
policy
body
carbon
people
since
still
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

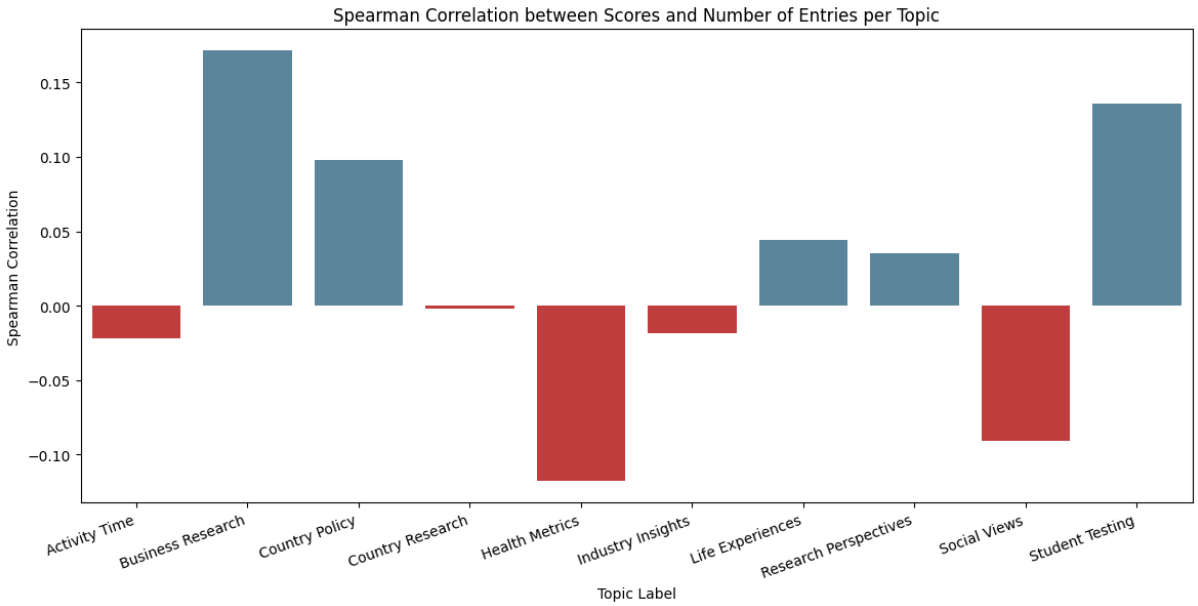

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
7
4
3
2
1
5
10
6
9
Marginal topic distribution
2%
5%
10%

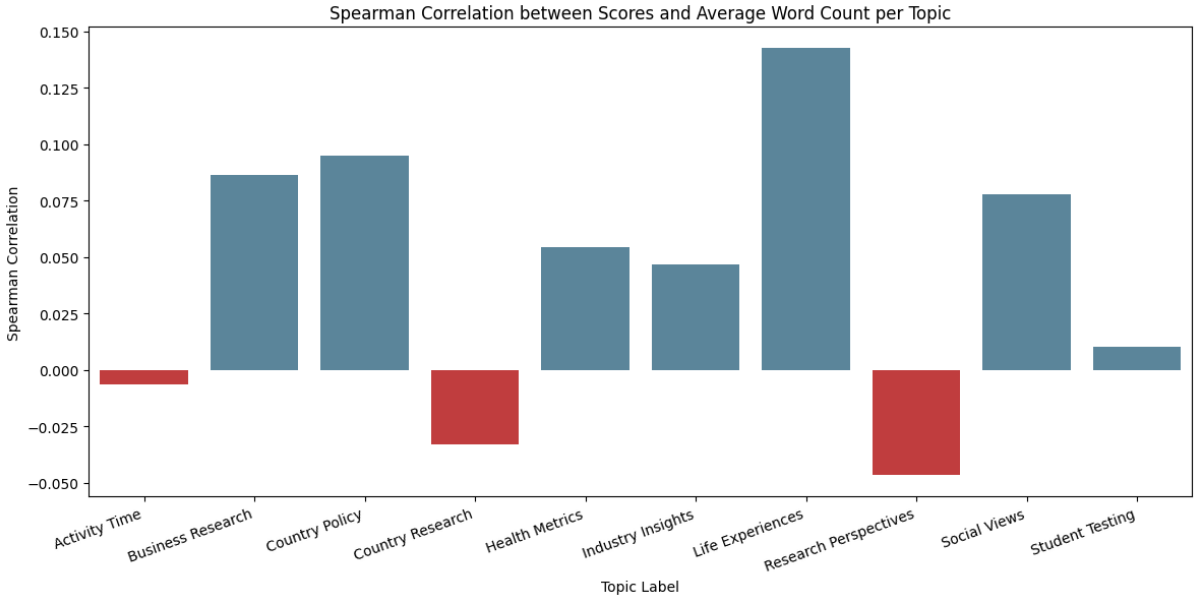

Top-30 Most Relevant Terms for Topic 10 (2.9% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
would
also
one
student
way
test
like
time
make
able
think
u
see
may
learning
question
could
people
life
story
class
take
used
po
example
still
much
something
thank
help
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

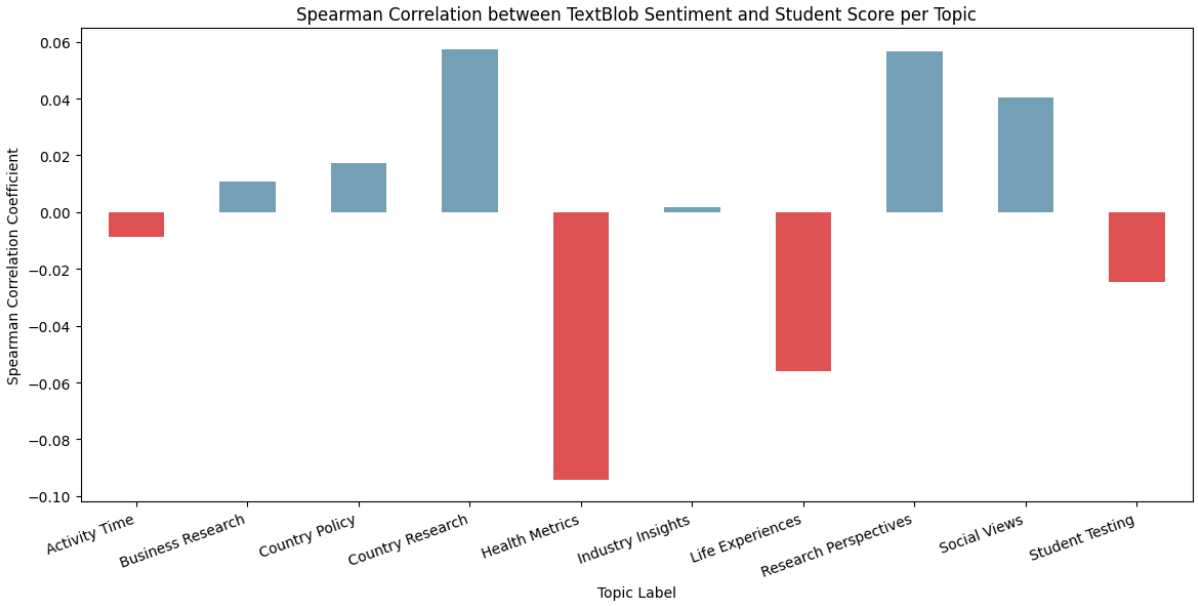
LDA: Count of Response per Topic



LDA: Correlation Analysis



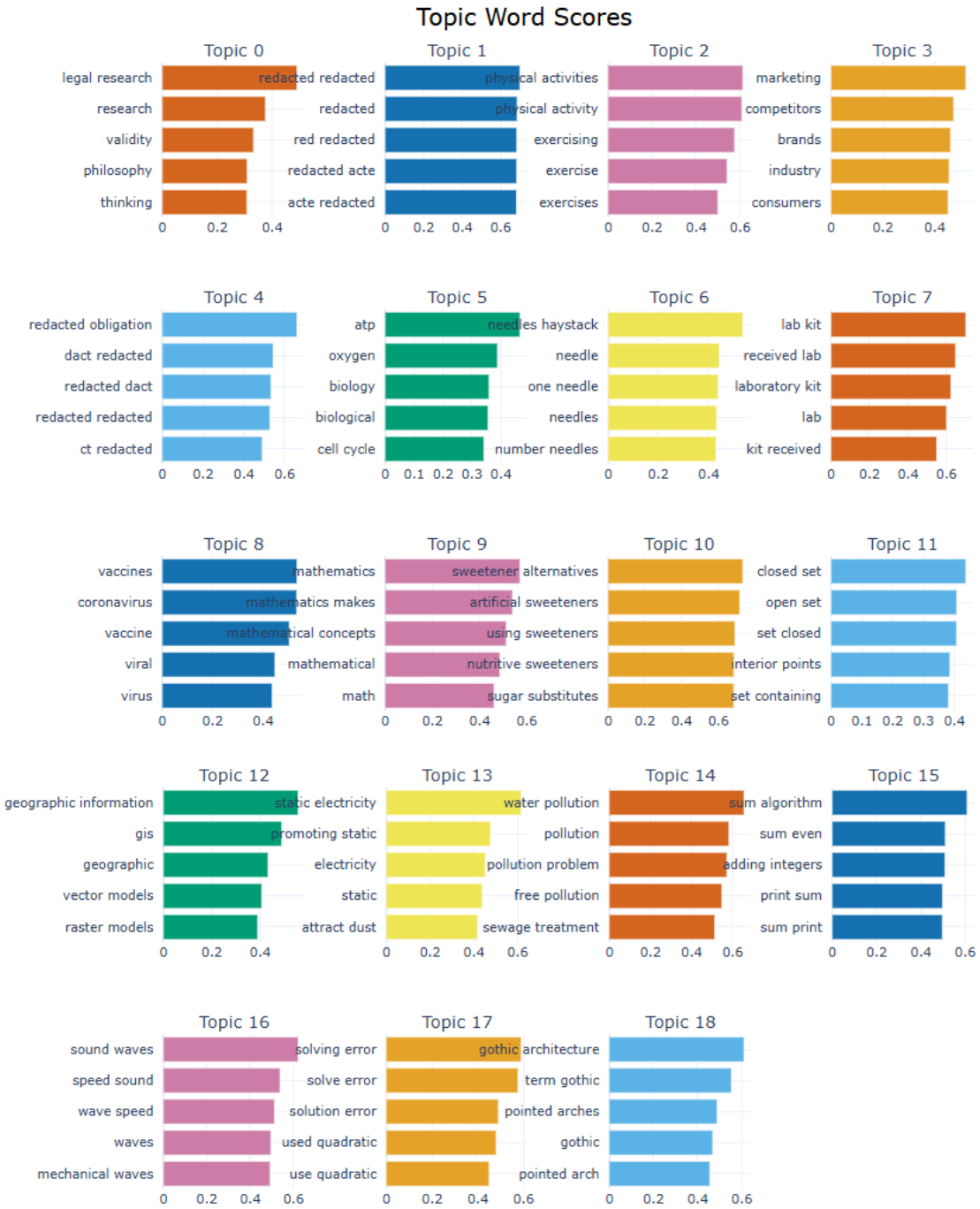


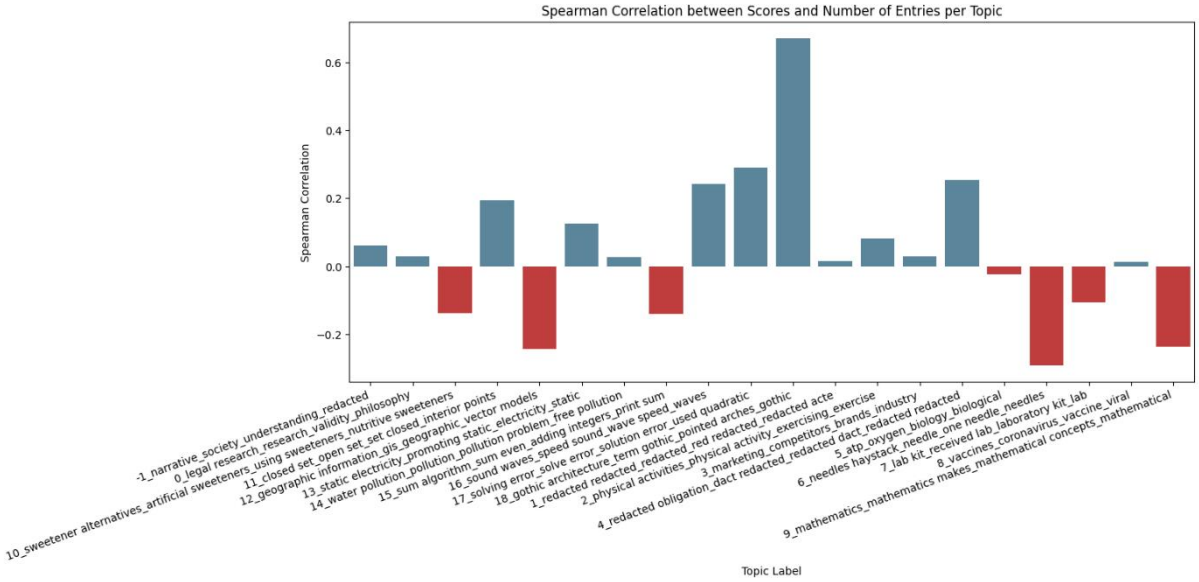


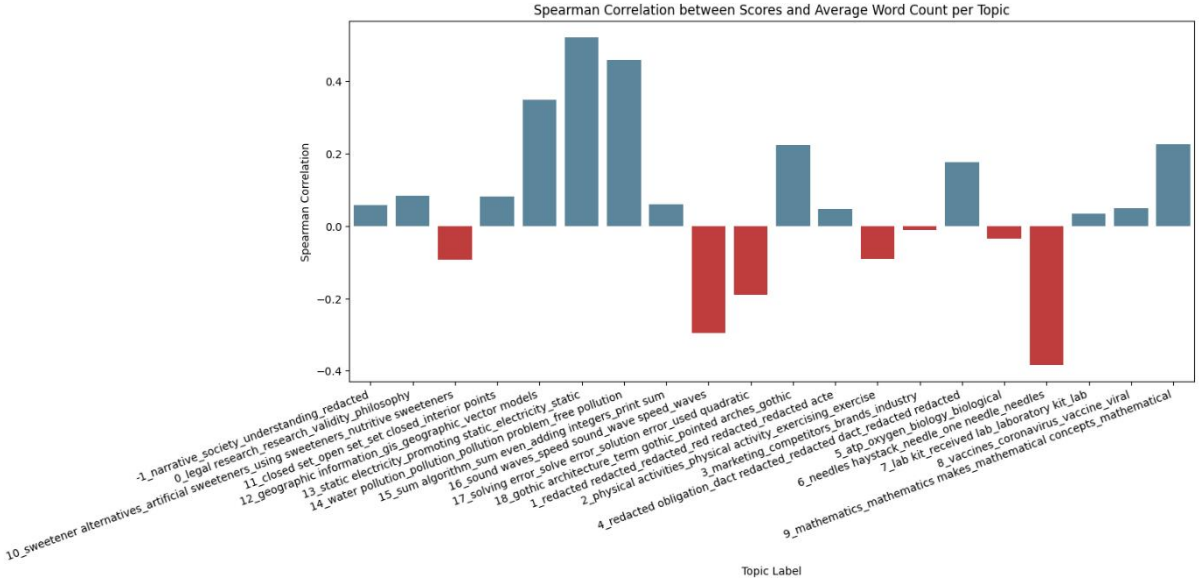
LDA shows similar or worse performance compared to NMF in terms of correlations. The analysis indicates that LDA does not offer any clear advantage over NMF in capturing relationships between topics and features. In some cases, LDA produces lower correlation scores, suggesting it may be less effective for this analysis. Overall, NMF provides more consistent results for the dataset.

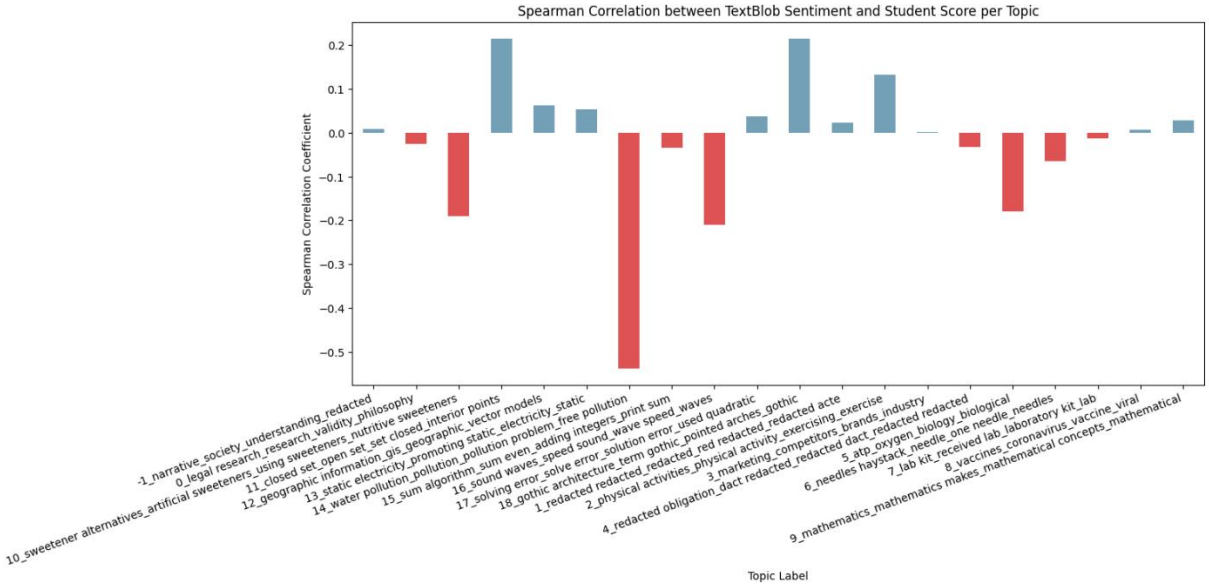
BERTopic: Topic Modeling

To compare classic methods such as NMF and LDA with SOTA methods, BERTopic was used to generate 20 topics to classify student responses.




Spearman Correlation between Scores and Number of Entries per Topic
Spearman Correlation
0.6
0.4
0.2
0.0
−0.2
-1_narrative_society_understanding_redacted
0_legal research_research_validity_philosophy
10_sweetener alternatives_artificial sweeteners_using sweeteners_nutritive sweeteners
11_closed set_open set_set closed_interior points
12_geographic information_gis_geographic_vector models
13_static electricity_promoting_static_electricity_static
14_water pollution_pollution_pollution problem_free pollution
15_sum algorithm_sum even_adding integers_print sum
16_sound waves_speed sound_wave speed_waves
17_solving error_solve error_solution error_used quadratic
18_gothic architecture_term gothic_pointed arches_gothic
1_redacted redacted_redacted_red redacted_redacted acte
2_physical activities_physical activity_exercising_exercise
3_marketing_competitors_brands_industry
4_redacted obligation_dact redacted_redacted dact_redacted redacted
5_atp_oxygen_biology_biological
6_needles haystack_needle_one needle_needles
7_lab kit_received lab_laboratory kit_lab
8_vaccines_coronavirus_vaccine_viral
9_mathematics_mathematics makes_mathematical concepts_mathematical
Topic Label


Spearman Correlation between Scores and Average Word Count per Topic
Spearman Correlation
0.4
0.2
0.0
−0.2
−0.4
-1_narrative_society_understanding_redacted
0_legal research_research_validity_philosophy
10_sweetener alternatives_artificial sweeteners_using sweeteners_nutritive sweeteners
11_closed set_open set_set closed_interior points
12_geographic information_gis_geographic_vector models
13_static electricity_promoting_static_electricity_static
14_water pollution_pollution_pollution problem_free pollution
15_sum algorithm_sum even_adding integers_print sum
16_sound waves_speed sound_wave speed_waves
17_solving error_solve error_solution error_used quadratic
18_gothic architecture_term gothic_pointed arches_gothic
1_redacted redacted_redacted_red redacted_redacted acte
2_physical activities_physical activity_exercising_exercise
3_marketing_competitors_brands_industry
4_redacted obligation_dact redacted_redacted dact_redacted redacted
5_atp_oxygen_biology_biological
6_needles haystack_needle_one needle_needles
7_lab kit_received lab_laboratory kit_lab
8_vaccines_coronavirus_vaccine_viral
9_mathematics_mathematics makes_mathematical concepts_mathematical
Topic Label



BERTopic generates more specific topics, which enhances correlation results for certain topics. Its detailed topic extraction improves alignment between topics and features, leading to stronger and more meaningful correlations. This specificity helps in uncovering precise patterns within the data.