Technical Appendix: Reproducibility Materials

1. Introduction

This document serves as a technical appendix to the extended research project report. This research aims to provide the precise details and materials a competent researcher would need to reproduce the research presented in the main report, from data collection to final analytical outputs, without needing to refer back to the report itself.

2. Data Sources

The primary data for this project was sourced from YouTube comment sections. The data were collected using a custom Python script and the YouTube Data API v3. The initial collection was based on policy-relevant search terms, including: "maths to 18," "Sunak education policy," "compulsory math 18," "Rishi Sunak maths," "UK education reform," "A-level reform," and "Core Maths."

The final dataset represents a refined corpus of 426 unique comments from UK-region, English-language videos discussing the "maths to 18" policy. The data was sampled to focus on top-level comments and was further filtered to ensure high relevance to the research questions.

3. Preprocessing and Data Cleaning

The sampled textual data was subjected to a rigorous preprocessing pipeline designed to improve signal clarity. This process involved the following sequential steps:

- **Lowercasing:** All text was converted to lowercase.
- **Tokenization:** Data was tokenized into unigrams and bigrams.
- **Noise Removal:** Hyperlinks, punctuation, emojis, and non-alphanumeric characters were removed.
- **Stopword Removal:** A dictionary of stopwords, extended with YouTube-specific terms (e.g., "video," "comment"), was used to filter common words.
- **Stemming:** Words were stemmed using the Snowball Stemmer to standardize vocabulary and mitigate data sparsity.
- **Relevance Filtering:** The final corpus was restricted to comments containing specific policy-related keywords, including: “math,” “maths,” “sunak,” “rishi,”

"tory," "A-level," "gcse," "compulsory," "educat," "school," "college," "student," and "learn."

4. Models and Analysis

All models and analyses were conducted using a hybrid computational social science approach. The methods and key parameters are described below.

Topic Modelling:

- **Methodology:** Non-negative Matrix Factorisation (NMF) was used for topic modelling to identify recurring themes. The optimal number of topics was determined to be seven through a coherence score sweep.
- **Equation:** NMF operates by decomposing the term-document matrix (V) into two lower-dimensional non-negative matrices (W and H) to minimize the Frobenius norm of the difference between V and WH.
- **Outputs:** The seven topics were interpretively consolidated into four major thematic domains: Practicality of Curriculum, Political Trust and Identity, Economic Utility and Global Relevance, and Student Autonomy and Educational Ethics.

Sentiment Analysis:

- **Methodology:** A custom-trained Logistic Regression classifier was used for sentiment classification (positive, neutral, negative).
- **Vectorization:** Comments were vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) technique.
- **Training:** The model was trained on a manually annotated sample of 100 comments with an 80/20 train-test split and five-fold cross-validation.
- **Performance:** The model achieved a robust accuracy of 85%.

5. Code Repository

All code and materials supporting the running of the code are hosted on a public GitHub repository. This repository is a central component for reproducibility.

- **GitHub Repository URL:**
- https://github.com/queenie-he/11531666_SAN_ERP.git

The repository is structured to support the analytical pipeline. It should include

Python scripts for data collection and preprocessing, Jupiter Notebooks for topic modelling and sentiment analysis, and the necessary requirements.txt file listing all dependencies.

6. Reproducibility Output

To facilitate verification, a copy of the key analytical outputs presented in the main report is included in this appendix. These outputs allow any researcher attempting to reproduce the work to verify whether their results match the original findings. This includes figures and tables:

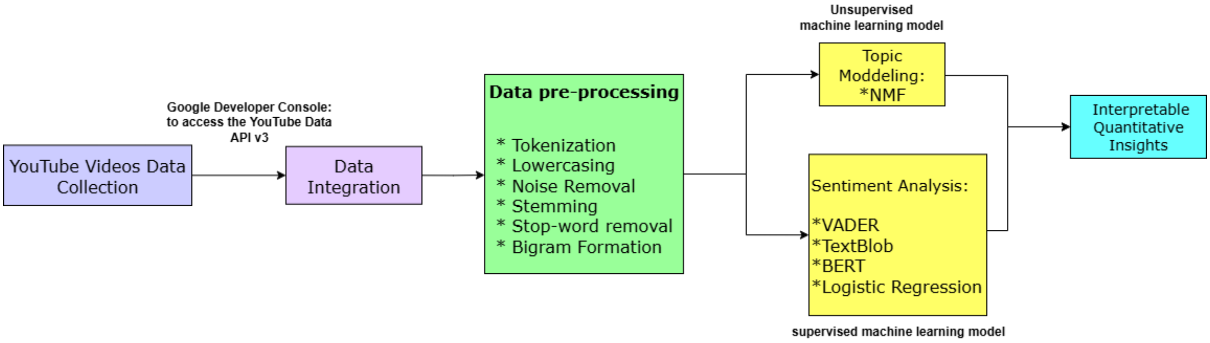

Figure 1: Methodological Pipeline Flowchart

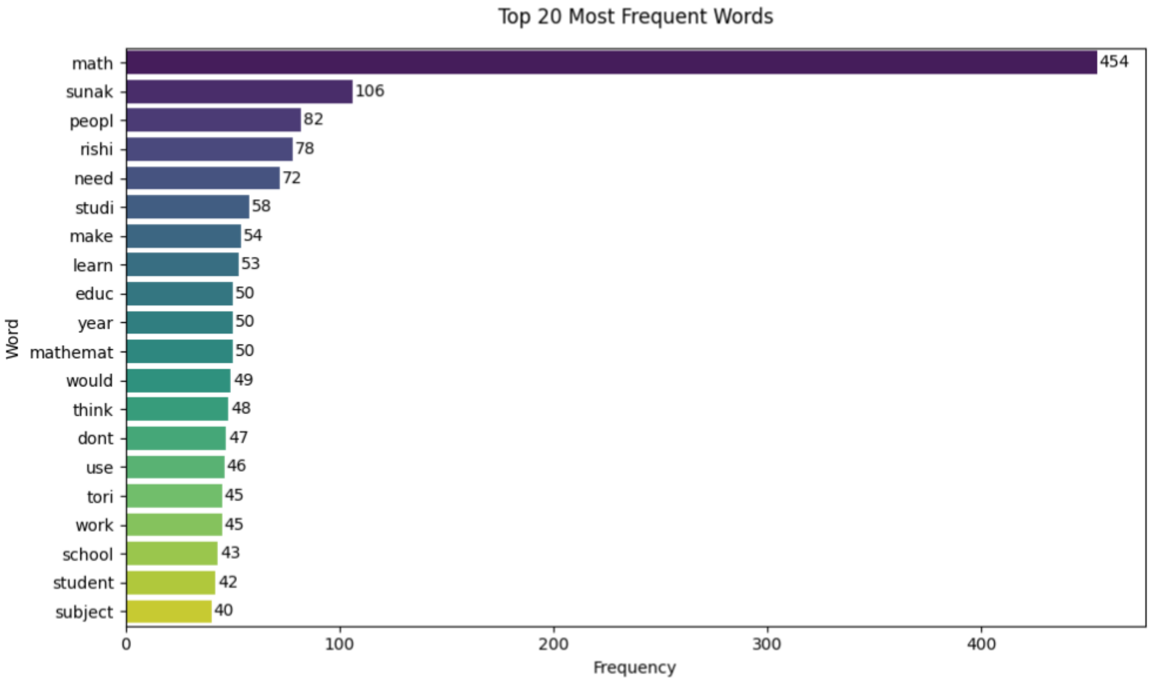

Figure 2: Top 20 Most Frequent Words.

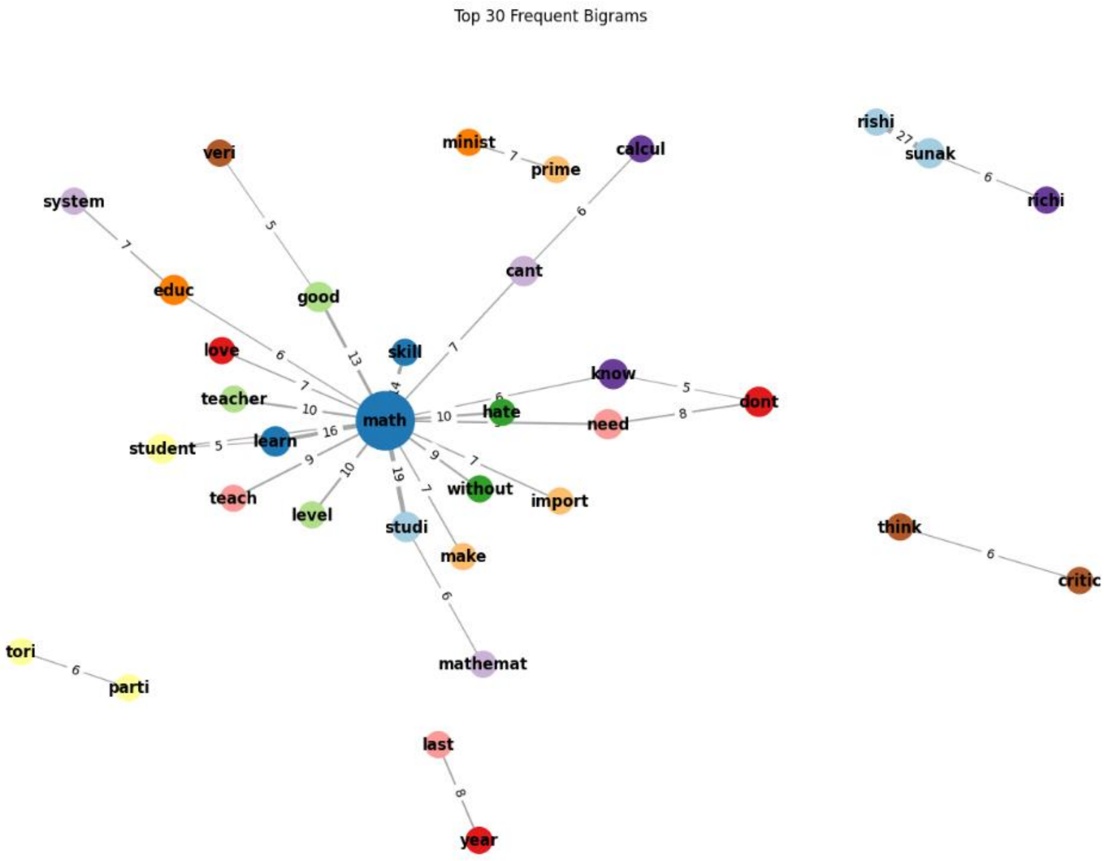

Figure 3: Top 30 Frequent Bigrams Network Graph.

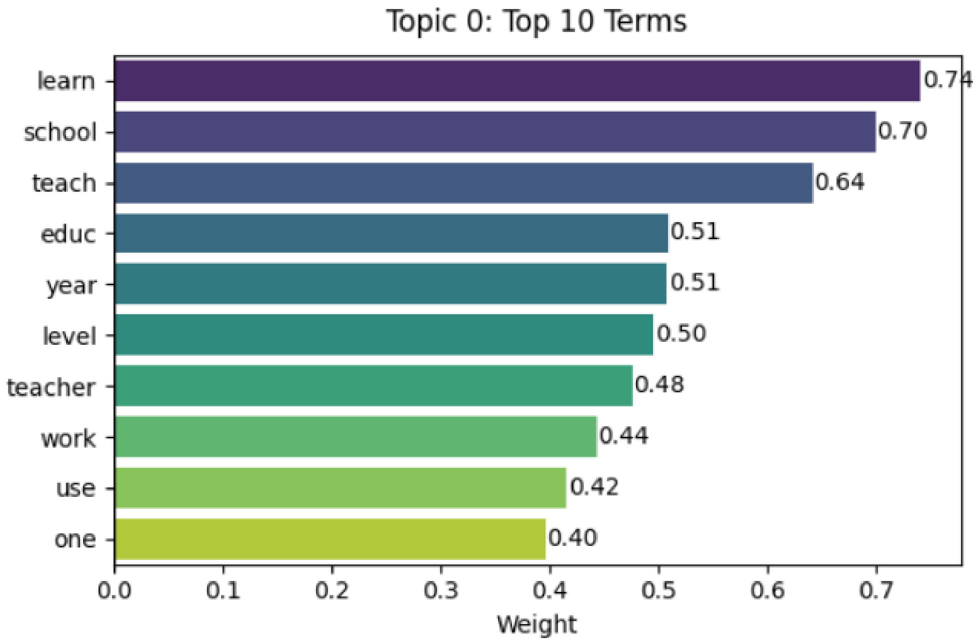

Figure 1: Top 10 Terms on Topic 0

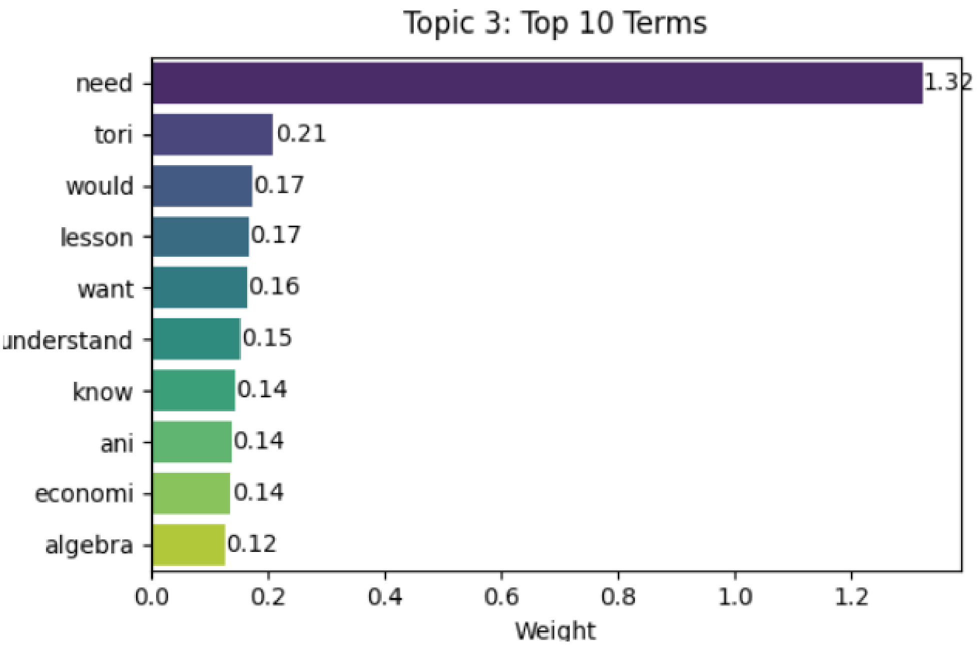

Figure 2: Top 10 Terms on Topic 3

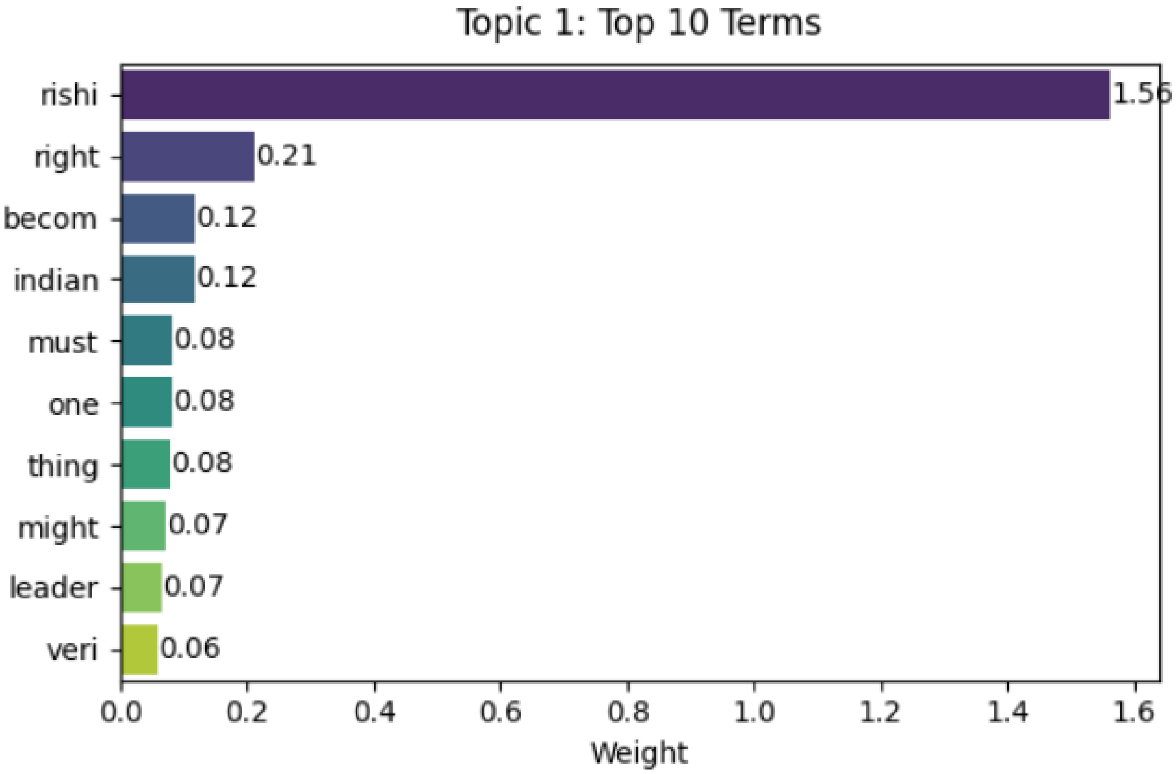

Figure 3: Top 10 Terms on Topic 1

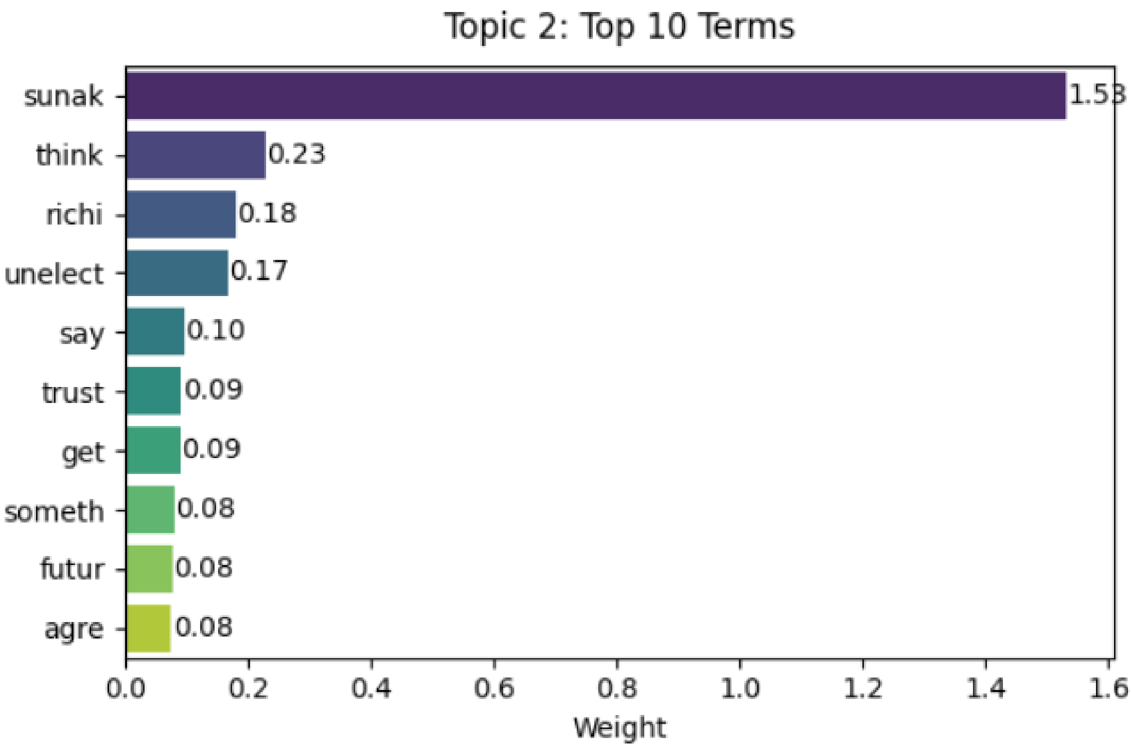

Figure 4: Top 10 words on Topic 2

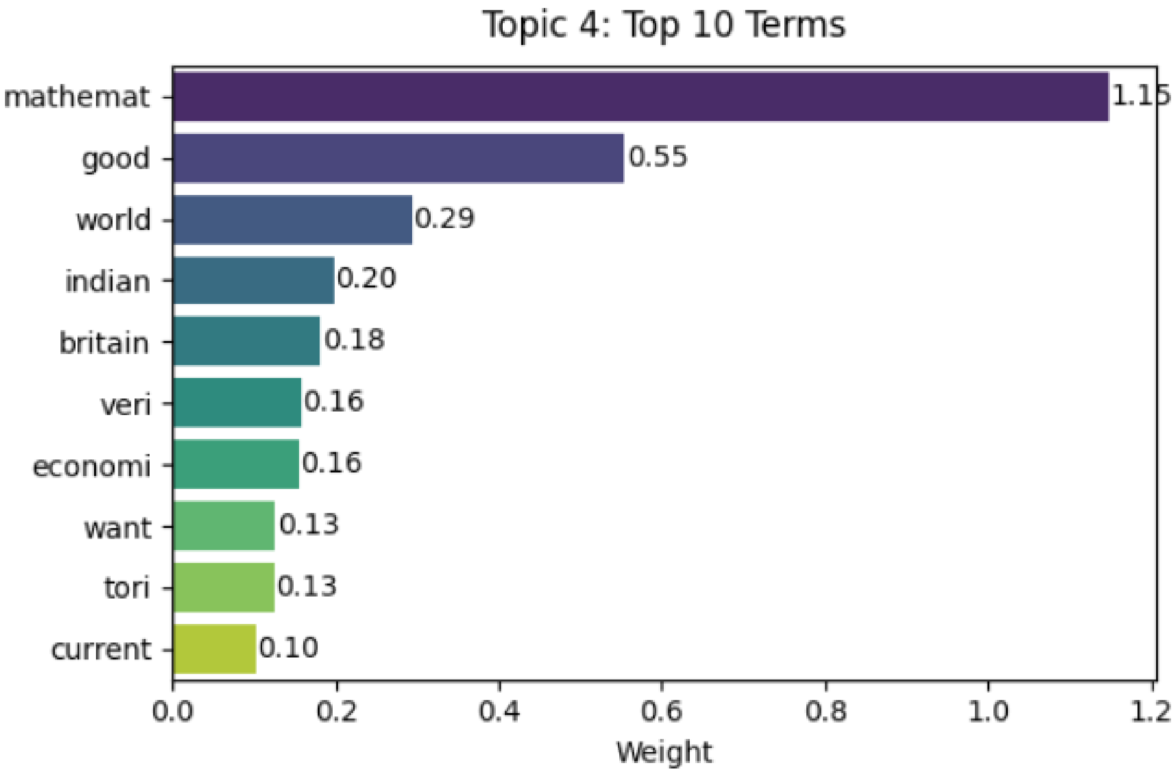

Figure 5: Top 10 terms on Topic 4

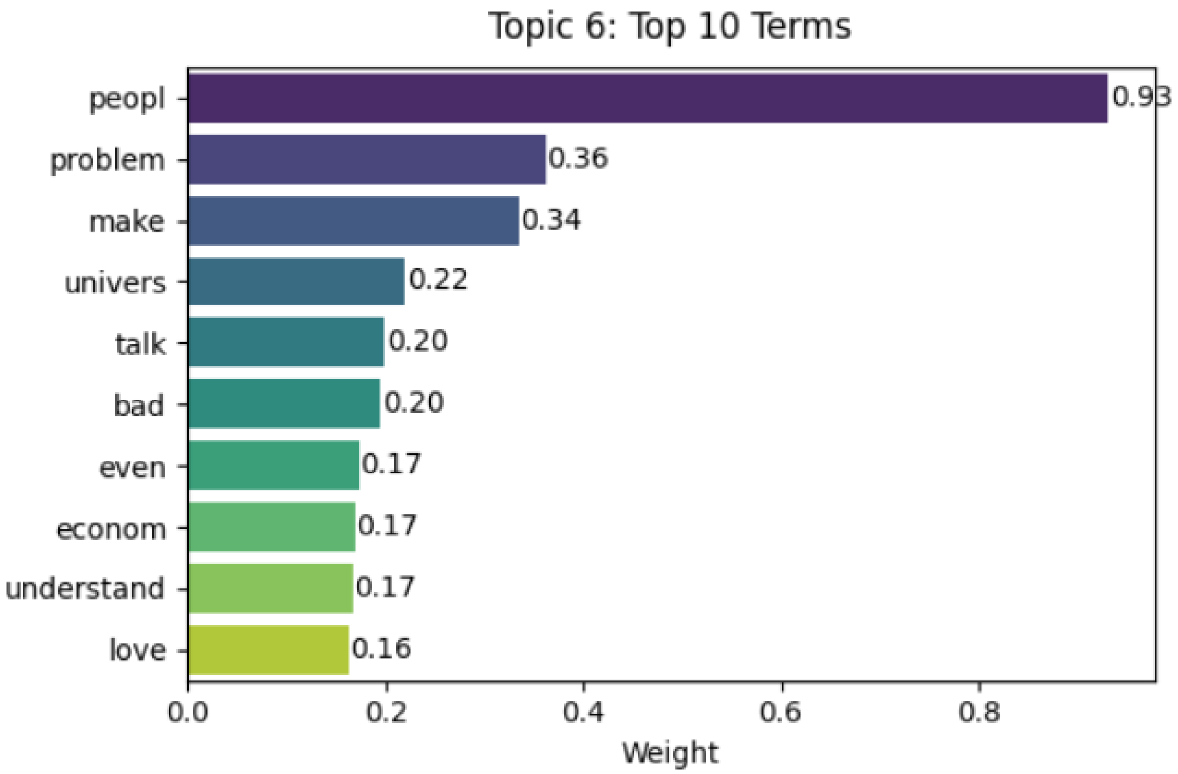

Figure 6: Top 10 Terms on figure 6

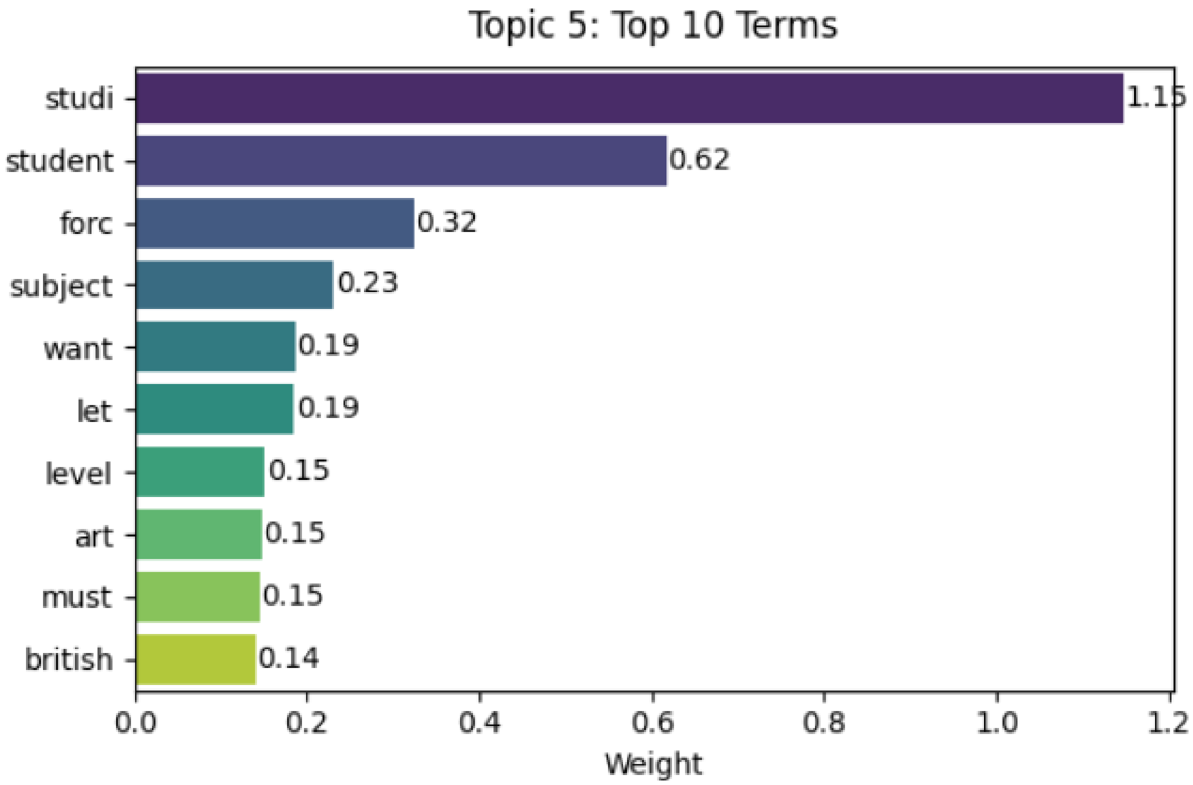

Figure 7: Top 10 Terms on Topic 5.

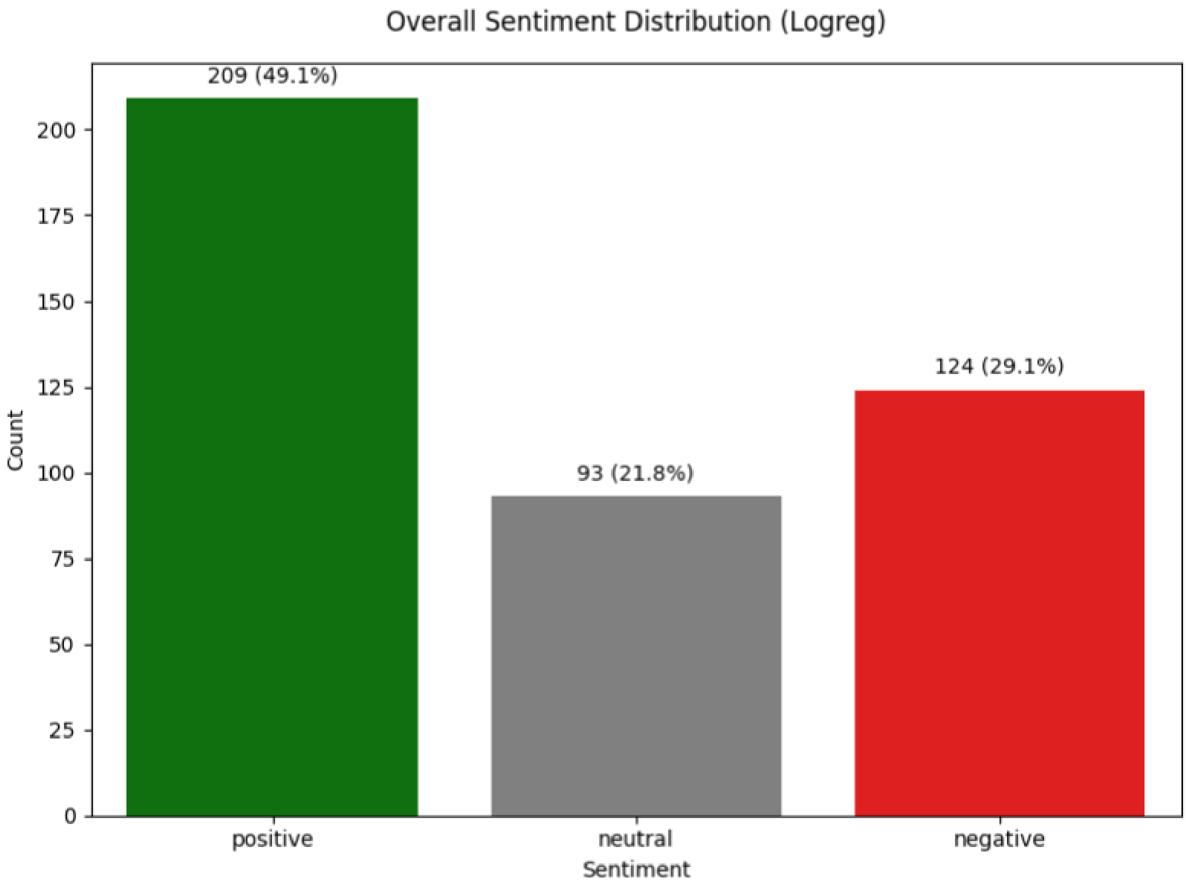

Figure 8: Overall Sentiment Distribution Bar Chart.

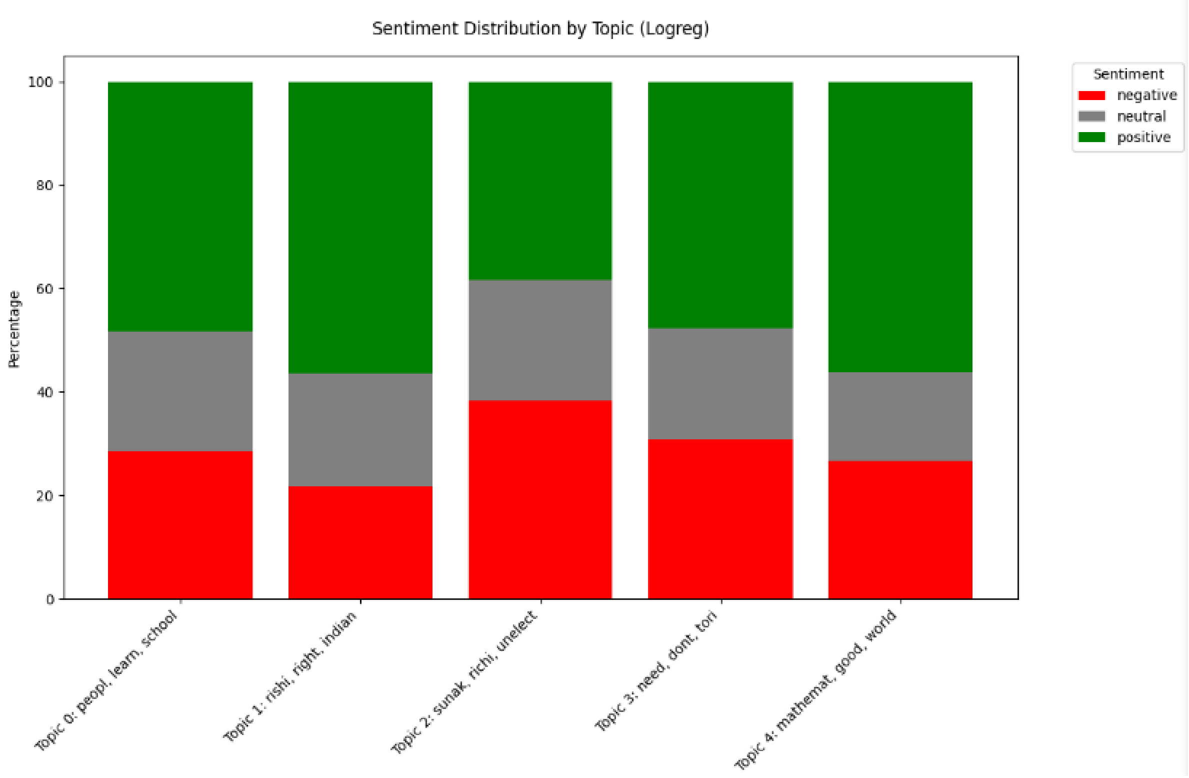

Figure 9: Sentiment Distribution by Consolidated Topic Stacked Bar Chart.

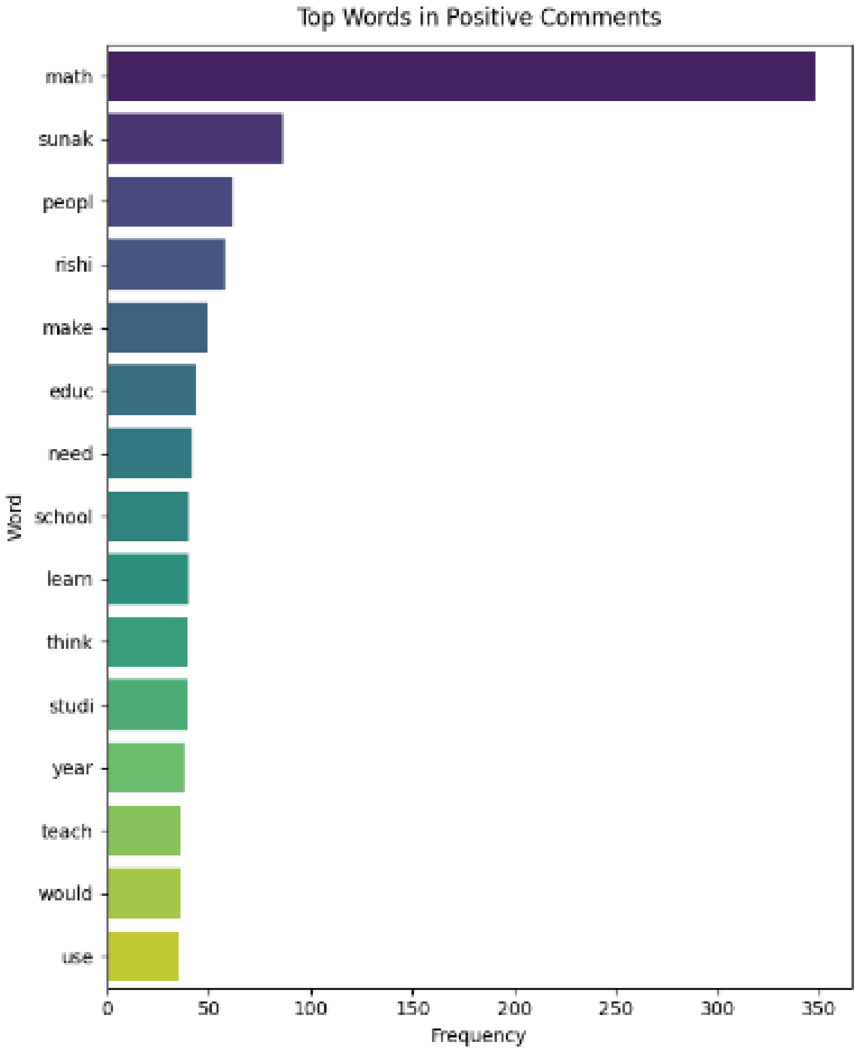

Figure 10: Top Words in Positive Comments by Sentiment Bar Charts

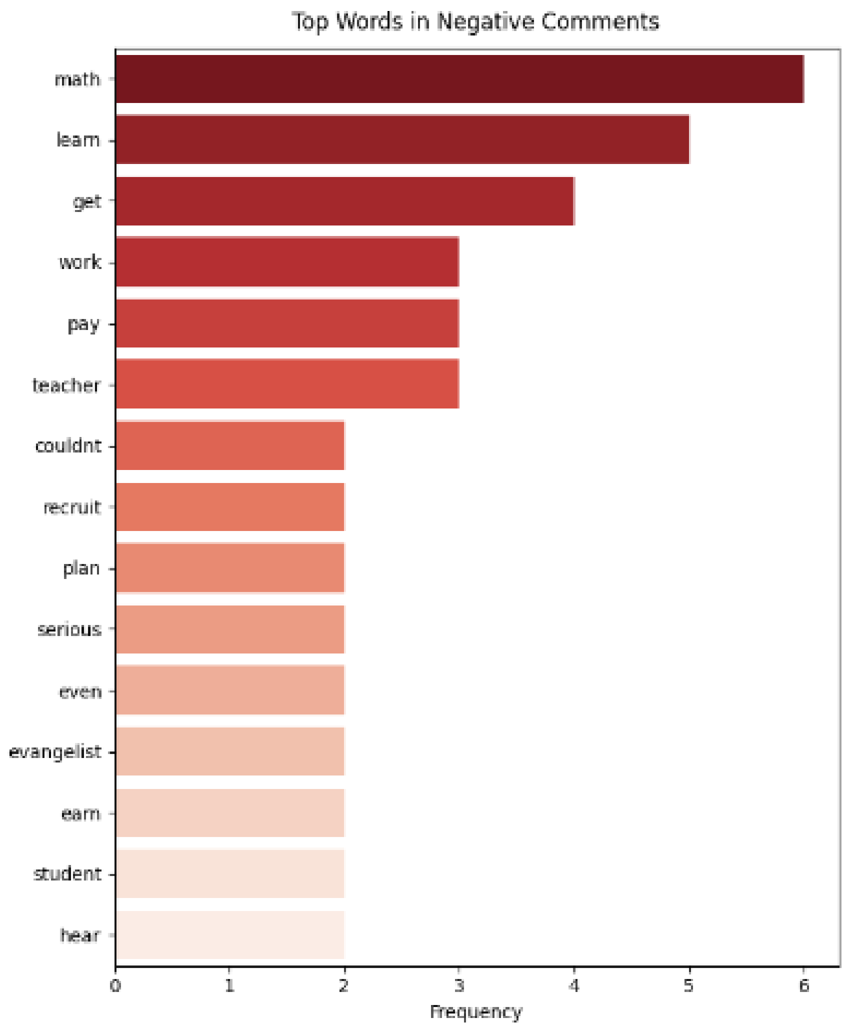

Figure 11: Top Words in Negative Comments by Sentiment Bar Charts.

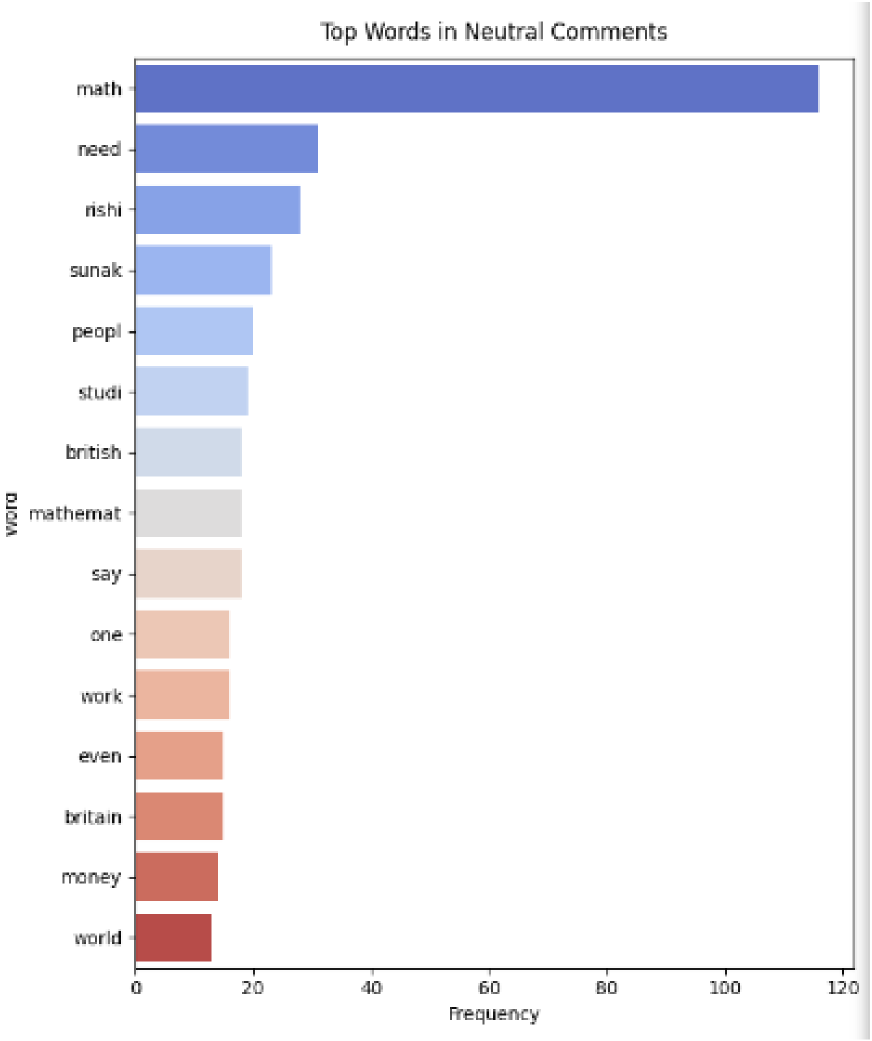

Figure 12: Top Words in Neutral Comments by Sentiment Bar Charts

Topic	**Top Terms**
Topic 0	learn, school, teach, educ, year, level, teacher, work, use, one
Topic 1	rishi, right, becom, indian, must, one, thing, might, leader, veri
Topic 2	sunak, think, richi, unelect, say, trust, get, someth, futur, agre
Topic 3	need, tori, would, lesson, want, understand, know, ani, economi, algebra

Topic 4	mathemat, good, world, indian, britain, veri, economi, want, tori, current
Topic 5	studi, student, forc, subject, want, let, level, art, must, british
Topic 6	peopl, problem, make, univers, talk, bad, even, econom, understand, love

Table 1: Top terms for each of the seven topics derived from the NMF model.

Theme	**Topic**	**Top Terms**
Practicality of Curriculum and Real-World Relevance	Topic 0	learn, school, teach, educ, year, level, teacher, work, use, one
	Topic 3	need, tori, would, lesson, want, understand, know, ani, economi, algebra
Political Trust and Identity	Topic 1	rishi, right, becom, indian, must, one, thing, might, leader, veri
	Topic 2	sunak, think, richi, unelect, say, trust, get, someth, futur, agre
Economic Utility and Global Relevance	Topic 4	mathemat, good, world, indian, britain, veri, economi, want, tori, current
	Topic 6	peopl, problem, make, univers, talk, bad, even, economi, understand, love
Student Autonomy and Educational Ethics	Topic 5	studi, student, forc, subject, want, let, level, art, must, british

Table 2: Thematic Consolidation of NMF Topics and Top Terms.