Bridging the Socioeconomic Gap in Education: A Hybrid AI and Human Annotation Approach

Anonymous ACL submission

Abstract

Students' academic performance is influenced by various demographic factors, with socioeconomic class being a prominently researched and debated factor. Computer Science research traditionally prioritizes computationally definable problems, yet challenges such as the scarcity of high-quality labeled data and ethical concerns surrounding the mining of personal information can pose barriers to exploring topics like the impact of SES on students' education. Overcoming these barriers may involve automating the collection and annotation of high-quality language data from diverse social groups through human collaboration. Therefore, our focus is on gathering unstructured narratives from Internet forums written by students with low socioeconomic status (SES) using machine learning models and human insights. We developed a hybrid data collection model that semi-automatically retrieved narratives from the Reddit website and created a dataset five times larger than the seed dataset. Additionally, we compared the performance of traditional ML models with recent large language models (LLMs) in classifying narratives written by low-SES students, and analyzed the collected data to extract valuable insights into the socioeconomic challenges these students encounter and the solutions they pursue.

1 Introduction

Low socioeconomic status (SES) refers to a disadvantaged position in society determined by factors such as income, education, and occupation. Individuals with low SES typically have limited financial resources (Scott-Clayton, 2015), lower educational attainment (Titus, 2006), and reduced access to quality healthcare and academic opportunities (Adler and Newman, 2017). These disparities profoundly impact students' educational experiences, shaping their academic performance, career prospects, and long-term well-being. Students from low socioeconomic backgrounds often struggle with financial barriers, making it difficult to afford tuition and educational resources (Brown and Carr, 2013). They also have limited access to academic support and technology, which can impede their academic success. Additionally, these students may experience social isolation and psychological stress due to the pressure of competing with peers from more privileged backgrounds (Lee et al., 2008).

Research on students from low socioeconomic backgrounds is crucial for identifying and addressing the unique challenges they face in education. Understanding these struggles can inform policies and interventions that promote equity, ensuring that students receive the necessary support to succeed. Despite this importance, NLP research has largely overlooked socioeconomic status. A survey by (Curry et al., 2024) found only 20 papers in the ACL Anthology explicitly mentioning SES, highlighting a substantial gap in computational research. This lack of attention limits our understanding of how SES affects student life and contributes to the development of educational technologies that may not adequately address the needs of low-SES students, potentially widening the digital divide (Kelbessa et al., 2024).

A major challenge in computational research on low-SES students is the scarcity of high-quality labeled data. Most existing datasets rely on structured survey responses, which fail to fully capture the complexity of students' experiences. To address this, we analyze a dataset published by (Kelbessa et al., 2024), containing 74 narratives written by low-SES students on Reddit. These narratives offer valuable firsthand insights into the struggles and coping mechanisms of low-SES students, making them an important resource for NLP research. However, as the dataset was annotated by only two individuals and lacks gold-standard validation, it presents both an opportunity and a challenge for refinement and further analysis.

The remainder of this paper is structured as follows: In §2, we review previous research on SES and NLP. In §3, we describe our dataset, including its source and key attributes. In §4, we present our data evaluation and analysis methodology, which involves assessing data quality, filtering high-quality narratives, and applying linguistic metrics, sentiment analysis, and topic modeling. Finally, we summarize our findings (§5), discuss our key contributions (§6), explore the ethical and societal implications of our research (§7).

2 Related Work

Socioeconomic status (SES) influences various aspects of life, including education, health, and social mobility. Understanding SES disparities requires high-quality datasets that capture linguistic, behavioral, and demographic patterns. However, acquiring such datasets poses significant challenges, including ethical concerns, accessibility restrictions, and issues of representativeness.

Several datasets have been developed for SES research, particularly leveraging online sources and social media platforms. Twitter has been widely used due to its accessibility and large user base. For instance, (Lampos et al., 2014) used UK Twitter data to analyze how user-generated content predicts SES. Similarly, (Mentink, 2016) collected a dataset of 2.5 million Dutch Twitter users to infer their socioeconomic backgrounds. While Twitter data provides valuable large-scale insights, its brevity and informal nature limit its usefulness for in-depth socioeconomic analysis.

Beyond social media, narrative-based data provides rich contextual insights into SES struggles. (Kelbessa et al., 2024) compiled a dataset of 74 SES-related narratives as a foundation for further research. Unlike social media posts, these narratives offer longer, structured reflections on lived experiences, making them more informative for understanding the personal and systemic challenges faced by individuals from low-SES backgrounds.

Despite advancements in SES-related data collection, several challenges persist. Ethical concerns arise when gathering data from social media, as it raises privacy and consent issues (Stieglitz et al., 2018). Data accessibility is another major barrier, as many relevant datasets are either proprietary or require restrictive permissions. Additionally, existing SES datasets often overrepresent specific demographics, reducing their generalizability and limiting their applicability across diverse populations.

In this work, we aim to address these challenges by expanding SES-related narrative datasets with a focus on depth and representativeness. Our dataset mitigates the limitations of short-form social media text by collecting and analyzing longer, structured narratives. This work contributes to the growing body of research on SES in computational linguistics and provides a valuable resource for future studies examining socioeconomic barriers in education.

3 Data

Unlabeled Reddit Posts: This data is publicly available and was collected on 2019 for the ThinkPlayHack event hosted in July 2019 in Taos for Dr. Jo Guldi (Southern Methodist University (SMU), 2019). It contains over 1 terabyte of Reddit posts published from 2005 to 2013. To manage the extensive data, measured in terabytes, we initially selected posts exclusively from subreddits associated with low-SES, such as 'college', 'ApplyingToCollege', 'depression', 'askReddit', 'broke', 'financialaid', and 'fafsa'. After filtering for relevant subreddits and eliminating duplicates, the resulting dataset comprised 799,032 Reddit posts (total 106859972 words) with 7 average sentences and 134 average words per post.

Labeled Reddit Posts: (Kelbessa et al., 2024) gathered 74 low-SES narratives from Reddit. To ensure the validity of these narratives as data points, the following criteria were applied: 1) The narratives needed to shed light on the experience of being a low-SES student and attending higher education, focusing on financial, psychological, physical, or social struggles. 2) The narratives should describe the challenges faced by individuals with low-SES backgrounds, their efforts to improve their situation, and the outcomes of those efforts. 3) Narratives that primarily offered general commentary, described a condition, or provided advice were excluded, as they did not qualify as valid data points. Each narrative had to meet at least one of the first two criteria and also satisfy the third qualification to be included in the dataset. The narratives were doubly annotated by two annotators to ensure the consistency and quality of the data. To ensure the quality of this publicly available data (Kelbessa et al., 2024), we thoroughly applied the above criteria on the 74 narratives and identified 64 of them

met the criteria (background of low-SES).

To process 64 non-low-SES data points, we applied semantic textual similarity to determine which posts in the unlabeled Reddit Posts (mentioned above) had the weakest correlation with the ground truth 64 low-SES narratives. We explored several similarity measures, such as the Negative Euclidean Distance, Negative Manhattan Distance, and Cosine Similarity, using the top-ranked Sentence Transformer model from the Massive Text Embedding Benchmark (Muennighoff et al., 2022) leaderboard on Hugging Face. We identified the 64 posts and manually evaluated them that had the lowest similarity to the ground truth data points, treating these as the non-low-SES ground truth. An example for illustration is provided in the appendix A.1. The final dataset comprised 64 non-low-SES Reddit posts with 15 average sentences and 299 average words per post.

4 Empirical Study

We started with the labeled dataset of 128 Reddit posts, consisting of 64 narratives from low-SES students and 64 from non-low-SES students. Over the course of three iterations, we curated and expanded this dataset, ultimately building a larger collection of narratives. Our primary goal is to collect enough data for future training and automation of the model. In each iteration, we added narratives in the training data that are newly labeled in the previous iteration as low-SES and then try to label the remaining unlabeled narratives. Every iteration followed a three-step process: first, we applied traditional binary ML classifiers and LLMs to categorize the unlabeled Reddit posts as either low-SES or non-low-SES; second, we used clustering techniques to identify and remove outliers from the dataset. Third, we manually annotated the narratives to perform the final evaluation. The numbers of the resulted labeled narratives at each step are shown at Table 2.

4.1 Step 1: Classification

We evaluated the performance of 22 traditional ML models and LLMs to distinguish between low-SES and non-low-SES texts. All the models we used are from scikit-learn (Pedregosa et al., 2011) and Hugging Face Transformers (Wolf et al., 2020). The evaluation followed a 5-fold cross-validation approach, with 70% of the dataset allocated for training, 15% for validation, and 15% for reporting the results. We conducted our experiments over three iterations, each with increasing complexity. In the first iteration, the dataset consisted of 64 low-SES narratives and 64 non-low-SES narratives, which were easily separable. This was due to our careful selection of the non-low-SES narratives, ensuring they were clearly distinguishable from the low-SES narratives. By the second iteration, the dataset had nearly doubled in size, and the classification task became more challenging. This time, the non-low-SES narratives were selected from the false positives of the first iteration, resulting in less clear separation between the classes. In the third iteration, the challenge intensified further as the dataset again doubled, with non-low-SES narratives chosen from the false positives of the second iteration. Consequently, the classes were significantly harder to separate, reflecting the increasing difficulty of the classification task. To address these increasingly complex classification tasks, we utilized a variety of fine-tuned pre-trained language models, models with few-shot capabilities, and traditional ML models. The overall results in three different iterations are shown at Table 1 and detailed parameters for all models are provided in Table 3 in Section A.2.

The Traditional models included Random Forest (RF), Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), Logistic Regression (LR), and Gradient Boosting (XGBoost). Overall, MNB, SVM, and LR demonstrated effective performance across the iterations, while RF and XGBoost struggled, particularly in the more challenging contexts. The fine-tuned pre-trained language models included Robustly Optimized BERT (RoBERTa), Robustly Optimized BERT-Large (RoBERTa-Large), Decoding-Enhanced BERT (DeBERTa), Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), A Lite BERT (ALBERT), eXtreme Language Model (XLNet), Text-to-Text Transfer Transformer (T5), Distilled BERT (DistilBERT), and Bidirectional Encoder Representations from Transformers (BERT). Overall, RoBERTa, RoBERTa-Large, and ELECTRA demonstrated strong and consistent performance across iterations, while ALBERT and XLNet had difficulty handling the increasing complexity of the task. The few-shot models included Open Pre-trained Transformer (OPT-13B), Claudia, LLM Meta AI (LLaMA-7B), and LLM Meta AI (LLaMA-1.3B). Few-shot models were not as ef-

Model	1st iteration			2nd iteration			3rd iteration			Model	1st iteration			2nd iteration			3rd iteration		
	P	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1		**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
RF	0.55	0.55	0.54	0.74	0.74	0.74	0.61	0.61	0.61	BERT	0.71	0.70	0.70	0.71	0.69	0.68	0.56	0.56	0.56
MNB	0.81	0.70	0.67	0.79	0.76	0.75	0.65	0.63	0.62	OPT-13B few	0.50	0.50	0.40	0.52	0.52	0.52	0.51	0.50	0.49
SVM	0.60	0.60	0.60	**0.87**	**0.87**	**0.87**	0.60	0.60	0.60	Claudia few	0.60	0.55	0.49	0.49	0.50	0.47	0.50	0.50	0.46
LR	0.60	0.60	0.60	**0.87**	**0.87**	**0.87**	0.64	0.64	0.64	LLamA-7B few	0.34	0.35	0.34	0.69	0.64	0.66	0.47	0.48	0.46
XGBoost	0.66	0.65	0.64	0.74	0.74	0.74	0.55	0.54	0.54	LLaMA-1.3B few	0.50	0.50	0.45	0.38	0.41	0.38	0.55	0.54	0.53
RoBERTa	0.75	0.75	0.75	0.76	0.76	0.76	0.73	0.67	0.65	LoRA DistillGPT	0.25	0.50	0.33	0.27	0.52	0.35	0.25	0.50	0.33
RoBERTa-large	0.75	0.75	0.75	0.74	0.74	0.74	**0.72**	**0.72**	**0.72**	LoRA LLaMA-1.3B	0.25	0.50	0.33	0.30	0.41	0.32	0.25	0.50	0.33
DeBERTa	0.77	0.75	0.74	0.74	0.74	0.74	0.56	0.55	0.55	LoRA GPT-2	0.25	0.50	0.33	0.27	0.52	0.35	0.25	0.50	0.33
ELECTRA	0.77	0.75	0.74	0.83	0.80	0.79	0.59	0.56	0.53	bart-large-mnli Zero	**0.80**	**0.80**	**0.80**	0.22	0.43	0.29	0.46	0.46	0.46
ALBERT	0.66	0.65	0.64	0.65	0.65	0.65	0.55	0.55	0.55	DistilBert	0.75	0.75	0.75	0.77	0.74	0.74	0.63	0.62	0.62
XLNet	0.60	0.55	0.49	0.63	0.61	0.60	0.66	0.65	0.65	T5	0.50	0.50	0.48	0.75	0.72	0.72	0.55	0.55	0.55

Table 1: Comparison of average Precision (P), Recall (R), and F_1 scores for both classes across three iterations for various classification models.

Iteration No.	Unlabeled Texts	Labeled (Classifier)	Labeled (Cluster)	Labeled (Human)
Iteration 1	799,032	13,635	289	110
Iteration 2	798,743	390	381	167
Iteration 3	798,362	5195	444	121

Table 2: Summary of the annotated narratives at three different steps- Classification, Clustering, and Human Annotation.

fective in distinguishing between increasingly similar narratives as the dataset complexity grew. The LoRA fine-tuning models included Distilled Generative Pre-trained Transformer (DistillGPT), LLM Meta AI (LLaMA-7B), and Generative Pre-trained Transformer (GPT-2). The results suggest that the LoRA fine-tuning models were not well-suited for the increasing complexity of the classification task, possibly due to their limited adaptation to more challenging data. The zero-shot model used was bart-large-mnli Zero (Yin et al., 2019). The sharp decline in performance in the second and third iterations suggests that the zero-shot model struggled to handle the increasing similarity between the low-SES and non-low-SES texts, as it lacked the fine-tuning capabilities of other models.

In the first iteration, the BART-large-mnli zero-shot model showed superior performance in terms of balanced metrics. We applied a confidence threshold of 0.7, meaning a text was classified as low-SES only if the model predicted it with confidence greater than 0.7. As a result, this model filtered the data, yielding 13,635 low-SES texts out of 799,032 unlabelled texts. In the second iteration, we employed an ensemble approach using SVM and LR, the two models that perform best in the first phase. Both models were trained separately and combined to improve classification accuracy. The SVM provided decision scores, and LR produced class probabilities, with a confidence threshold of 0.7 applied to both models. This approach ensured that a text was classified as low-SES only if both classifiers predicted low-SES with high confidence, resulting in 390 low-SES texts. In the third iteration, we applied RoBERTa-large, the best performer in this phase. A confidence threshold of 0.7 was used, yielding 5,195 low-SES texts out of 799,032 unlabelled texts.

Traditional ML Models Versus LLMs: In iteration 1, traditional models, such as SVM and LR, achieved balanced performance, with macro-averaged F_1 scores of 0.60. These models demonstrated robust precision and recall across both classes, although they did not outperform more advanced models. XGBoost performed slightly better, with an F_1 score of 0.66, particularly excelling in classifying non-low-SES texts.

Among LLMs, RoBERTa achieved an F_1 score of 0.75, showcasing strong performance with a balanced precision and recall across both SES classes. Similarly, DeBERTa and ELECTRA performed well, both achieving F_1 scores of 0.74. These models demonstrated better balance than some traditional models in classifying both low-SES and non-low-SES texts. However, other LLMs, such as ALBERT and XLNet, underperformed compared to their counterparts, with F_1 scores of 0.64 and 0.60, respectively. XLNet particularly struggled with the low-SES class, achieving an F_1 score of 0.31, highlighting its difficulty in accurately identifying low-SES narratives. Interestingly, zero-shot models like bart-large-mnli demonstrated strong performance, with an F_1 score of 0.80, matching the best-performing models in this iteration. In contrast, few-shot models, such as LoRA Fine-Tuned LLaMA and DistilGPT, performed poorly, indicating that few-shot learning in this context was less effective than fine-tuning. Detailed results are shown in section A.3 at Table 4.

In iteration 2, the comparison between traditional

ML and LLMs highlights the distinct strengths and weaknesses of each group. Traditional models, such as SVM and LR, outperformed most LLMs, achieving the highest F_1 scores of 0.87 across both low-SES and non-low-SES categories. These models demonstrated strong precision and recall, particularly in the non-low-SES class, underscoring their robustness in effectively balancing both classes. For example, SVM achieved a precision of 0.88 and recall of 0.85 for low-SES, while LR displayed a precision of 0.91 for low-SES and a recall of 0.93 for non-low-SES, making them reliable in these classification tasks.

In contrast, fine-tuned LLMs, such as RoBERTa and ELECTRA, also performed well but fell slightly behind the top traditional models. RoBERTa achieved a weighted F_1 score of 0.76, showing a strong balance between precision (0.74) and recall (0.82) for the non-low-SES class. However, its recall for low-SES (0.69) was lower compared to traditional models, meaning it missed more low-SES texts. ELECTRA achieved a higher F_1 score of 0.79 and performed exceptionally well in identifying non-low-SES texts, with a precision of 0.73 and recall of 0.96. Nevertheless, ELECTRA struggled with low-SES classification, showing a precision of 0.94 but a much lower recall of 0.62, indicating it missed more low-SES examples. DeBERTa displayed performance similar to traditional models like RF and XGBoost, with an F_1 score of 0.74. Its precision and recall were balanced across both classes but did not achieve the standout performance of models like RoBERTa or ELECTRA. Other LLMs, including ALBERT and XLNet, significantly underperformed compared to both traditional models and other LLMs, with F_1 scores of 0.65 and 0.60, respectively. XLNet particularly struggled with the non-low-SES class, achieving a recall of only 0.46, indicating significant difficulty in identifying non-low-SES texts accurately. ALBERT exhibited more balanced but lower performance across both classes.

These results for iteration 2 indicate that while LLMs have shown potential—particularly models like RoBERTa and ELECTRA—traditional models such as SVM and LR remain more reliable for tasks involving both low-SES and non-low-SES classification. Their superior balance between precision and recall across both categories demonstrates their robustness, whereas LLMs, though effective in certain areas such as precision for non-low-SES, may require further fine-tuning to achieve the same comprehensive balance seen in traditional models. Detailed results are shown in section A.3 at Table 5.

A further evaluation of traditional ML models and LLMs reveals interesting trends in model performance as the dataset complexity increases. While traditional models like SVM and LR continued to show stability, their dominance observed in earlier iterations has now been matched or exceeded by fine-tuned LLMs in certain aspects. Among the traditional models, LR demonstrated consistency, achieving an average 0.64 F_1 across both SES classes. While it excelled in the non-low-SES class, with a precision of 0.84 and recall of 0.93, it showed less robustness in the low-SES class, with precision and recall hovering around 0.64. Similarly, RF maintained stable performance, with an F_1 score of 0.61, although it underperformed compared to LR, particularly in the non-low-SES class (precision: 0.63, recall: 0.53).

Fine-tuned LLMs displayed notable improvements. RoBERTa-large emerged as one of the top performers, achieving the highest average F_1 score of 0.72 across both classes, surpassing traditional models like SVM and LR. This model exhibited a well-balanced performance with precision, recall, and F_1 scores closely aligned (precision: 0.70, recall: 0.76 for non-low-SES; precision: 0.74, recall: 0.67 for low-SES), indicating its capability to handle both classes. DeBERTa also showcased a solid performance, with an average F_1 of 0.55, though it struggled with the low-SES class (F_1: 0.49) compared to RoBERTa-large. Similarly, ELECTRA achieved an average F_1 of 0.53 but encountered challenges in classifying low-SES examples, where recall dropped to 0.31. These results suggest that while LLMs like RoBERTa-large are emerging as strong contenders, certain models such as ELECTRA and DeBERTa still require fine-tuning to handle the low-SES class. Few-shot models, such as Claudia and LLaMA-1.3B, presented mixed results. Claudia, in its few-shot configuration, achieved an average F_1 of 0.46, highlighting difficulties in identifying low-SES narratives (F_1: 0.32). On the other hand, LLaMA-1.3B fared slightly better, with an average F_1 score of 0.53, performing consistently across both SES classes. However, neither of these models surpassed fine-tuned LLMs or traditional models in overall performance. Fine-tuned models with LoRA (Low-Rank Adaptation), such as LoRA GPT-2 and LoRA LLaMA-1.3B, delivered uneven

outcomes. Both models demonstrated high precision for the non-low-SES class (0.50) but struggled significantly with the low-SES class, where they failed to capture any true positive examples (recall: 0.00, F_1: 0.00).

Iteration 3 reveals a growing strength of fine-tuned LLMs, particularly RoBERTa-large, which outperforms traditional models. This model's balanced precision and recall across SES classes emphasize its versatility. Models like DeBERTa and ELECTRA show that while LLMs are improving, they can still struggle with the low-SES class. Few-shot models and LoRA fine-tuned models exhibited less consistent results, often failing to achieve the comprehensive balance required for SES classification. This highlights the continued importance of fully fine-tuning LLMs for this task, as parameter-efficient models and few-shot learning may not yet match the robustness of more thoroughly fine-tuned counterparts like RoBERTa-large. Detailed results are shown in section A.3 at Table 6.

4.2 Step 2: Clustering

Once the classification step was completed, we used clustering to group similar texts and remove outliers. In the first iteration, we analyzed 13,635 posts identified as low-SES by the classifier. To compare the similarity between the ground truth samples and the collected posts, we applied PCA for dimensionality reduction using a tf-idf Vectorizer with a maximum of 1,000 features, visualizing the clusters in 2D space. Some outliers were detected in the ground truth data. These outliers were removed by applying the interquartile range (IQR) method, reducing the dataset to 55 points (Fig. 1a). Next, we computed cosine similarity between the normalized vectors of the collected data and the outlier-free ground truth data. A threshold of 0.9 was set to classify data points as similar or dissimilar. This process revealed that 289 from the collected data points met or exceeded the total similarity score(where the summation of the classified data is similar to ground truth data points) of 35, 940 data points had a similarity score of at least 34, and 10,821 data points had a score of 33 or less.

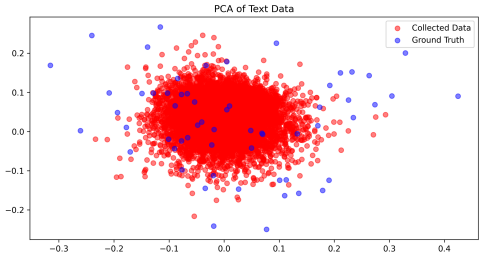
In the second iteration, after applying PCA to the collected data and the ground truth data, visualizing the results in a 2D space (see Fig. 1b). The collected data points and ground truth data were plotted to observe clustering patterns, allowing us to assess the similarity between the two datasets. Outliers were removed using the Interquartile Range (IQR) method, where values outside 1.5 times the IQR from the first (Q1) and third quartiles (Q3) were identified and excluded from both datasets. This process reduced the total number of collected data points from 390 to 381.

In the Third iteration, we analyzed 5195 posts identified as low-SES. To compare the similarity between the ground truth samples and the collected posts, we applied PCA for dimensionality reduction using a tf-idf Vectorizer with a maximum of 1,000 features, visualizing the clusters in 2D space. Some outliers were detected in the ground truth data. These outliers were removed by applying the interquartile range (IQR) method, reducing the dataset to 331 points (Figure 1c). Next, we computed cosine similarity between the normalized vectors of the collected data and the outlier-free ground truth data. We used the same threshold of 0.9 to classify data points as similar or dissimilar. This process revealed that 121 from the collected data points met or exceeded the total similarity score of 86, 260 data points with similarity score of at least 85, and 444 data points had a score of 84 or less.

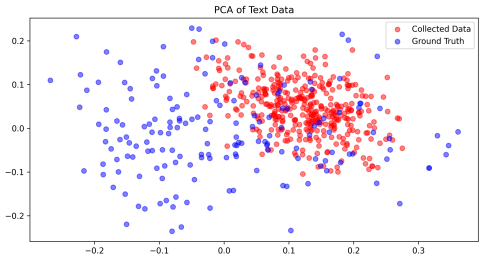
In iteration 1, the collected data predominantly covers the central portion of the ground truth spectrum, indicating that the initial classification managed to capture a concentrated part of the low-SES class but left much of the outer spectrum of the ground truth unexplored. Moving to iteration 2, the collected data begins to diverge, covering less of the ground truth spectrum compared to Iteration 1. This suggests that the classification in this iteration was more selective but also less comprehensive in capturing the full range of the low-SES data. Finally, in iteration 3, we see a significant improvement, with the collected data covering over 70% of the ground truth spectrum. This indicates a better alignment between the collected and ground truth data, suggesting that the classification in this iteration successfully captured a much broader range of the ground truth low-SES examples, resulting in a more balanced and comprehensive dataset.

5 Data Annotation

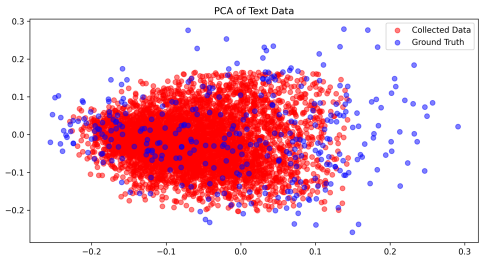
Human Annotation: The first three authors followed the criteria described in Section 3 to annotate the 289 texts from the clustering step at iteration 1 and getting the data points with a similarity score of at least 35. This process resulted in 110 texts



(a) Iteration 1 Clustering Without Outliers



(b) Iteration 2 Clustering Without Outliers



(c) Iteration 3 Clustering Without Outliers

Figure 1: Clustering of Ground Truth and Collected Data Using PCA Across Three Iterations Without Outliers

being annotated as low-SES out of 289. In the second iteration, we annotated 381 texts, of which 167 were classified as low-SES and in the third iteration, we annotated 121 texts that has a similarity score of at least 86, we got 46 low-SES Text.

In iteration 1, most of the texts annotated as not low-SES contained general advice and lacked the background indicative of being from a low-SES perspective. These texts often focused on providing broad recommendations rather than sharing personal experiences tied to financial difficulties. The absence of key indicators, such as struggles with income, debt, or reliance on social services, made it clear that these individuals did not face the same economic constraints typical of low-SES situations. As a result, their narratives were more aligned with middle or higher SES backgrounds, where financial stability was not a central concern. In iteration 2, as the task became more challenging, some of the texts annotated as not low-SES included background information and challenges but lacked personal experience that would validate them as low-SES. Upon further analysis, clear patterns emerged from the texts classified as low-SES. Many highlighted the need for financial aid, with individuals working multiple jobs or living in single-parent households with little to no income. Debt, particularly from educational loans or basic expenses, was a recurring theme, as was the lack of family support, indicating broader financial instability. These socio-economic markers—multiple jobs, debt, and minimal family support—are crucial for refining the model to better detect low-SES cases in future iterations. In iteration 3, after extracting 277 low-SES texts, some of the texts annotated as not low-SES described challenges and solutions but appeared to originate from individuals of medium SES rather than low-SES. We noticed that some texts annotated as not low-SES described challenges like juggling multiple jobs or balancing full-time work and studies. These individuals often discussed FAFSA loans or supporting a family due to a relative's disability. However, many of these cases appeared to stem from medium SES backgrounds, as they had access to basic loans or even an inheritance. This suggests that while these individuals faced financial difficulties, their situations were distinct from those typically associated with low-SES, highlighting the nuanced differences between SES classifications.

LLM Annotation: We used a pre-trained LLaMA model with 7 billion parameters (Touvron et al., 2023) to annotate the dataset of low-SES student narratives. The model was configured with a maximum of 200 tokens, a temperature of 0.5 for less randomness, and deterministic sampling. The goal was to extract information about students' background, struggles, and solutions related to financial, psychological, physical, or social challenges. Comprehensive descriptions of the prompts are provided in Section A.4. Although LLM was effective, the extraction of structured data was a challenge. The model occasionally produced extraneous text or improperly formatted outputs, requiring post-processing.

For visualization, we applied Principal Component Analysis (PCA) to reduce high-dimensional sentence embeddings on extracted background information from ground truth and collected datasets for visualization. K-means clustering was then used to group semantically similar sentences, identifying key thematic clusters. The resulting scatter plot (Figure 2) shows how the collected data expands the thematic coverage by displaying cluster distributions for both datasets. Collected data cluster 0

(Work Struggles) shows an expansion of themes around work experience and internships, academic support systems, and mental health or emotional struggles, which are less represented in the original data. Data cluster 1 (Family Dynamics) has more diverse and specific family backgrounds, struggles with independence and support, and impact of wider social and economic systems. Data cluster 2 (Mental Health) introduces more detailed reflections on emotional struggles and mental health challenges. Data cluster 3 (Societal Challenges) significantly enriches the thematic representation of challenges faced by low-SES students, particularly by introducing broader societal and personal insights that were underrepresented in the original dataset. Data cluster 4 (Systemic Critiques) reflects more detailed critiques of systemic issues affecting students, such as the cost of education, the student debt crisis, and the unrealistic promises of higher education as a golden ticket to success.

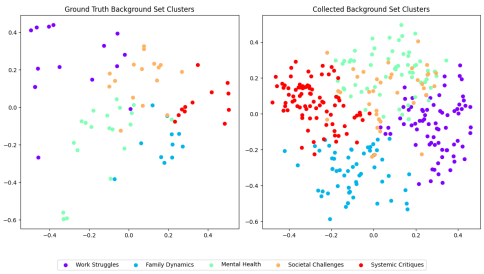


Figure 2: Ground Truth and Collected Data Clusters of Background Information

The final dataset began with a seed of 64 data points and, through the application of the proposed methodology, expanded to include 323 new data points. Both the ground truth and collected datasets were processed using prompt engineering with the LLaMA model to extract background information, struggles during higher education, and solutions students devised to improve their situations. While some data points have missing background, struggle, or solution information due to limitations in LLM extraction, we are actively working on refining the dataset through a rigorous gold-standard validation process. Additionally, we performed sentiment analysis on the entire text of both datasets using a sentiment fine-tuned model (Camacho-collados et al.). In the ground truth dataset, the sentiment distribution consisted of 10 positive, 29 neutral, and 25 negative texts. For the collected data, the distribution shifted to 37 positive, 138 neutral, and 148 negative texts. This significant increase in both neutral and negative sentiments in the collected dataset provides a broader scope for future exploration of the emotional landscape of low-SES students.

6 Contributions

First, we developed a dataset of unstructured narratives from low-SES students by semi-automatically collecting and expanding data from Reddit, producing a dataset five times larger than the initial seed. We will make our code and data public for the community. Second, we designed a hybrid model that combines ML and human insights to classify low-SES student narratives, comparing the performance of traditional ML models with recent LLMs. Third, given the challenges associated with data collection in this underexplored area, our work paves the way for fully automating this process, encouraging future research to focus on the educational barriers faced by low-SES students.

7 Ethical and Societal Impact

First, while we strive for fairness, it is challenging to ensure equal representation across geographic regions and genders in our dataset. This could lead to unintentional biases that affect the results and interpretations of our work. Second, although we will release the model under the appropriate license to ensure compliance with legal and ethical standards, there remains a risk of misuse. Specifically, the model could be used to classify low-SES individuals from publicly available narratives, potentially exposing them to harmful activities such as discrimination or exploitation. To mitigate this, we will enforce user agreements that explicitly prohibit harmful uses. Finally, although the data we collected is anonymous, it was sourced from public online forums, and we, as authors, cannot edit or delete this data once retrieved. This raises privacy concerns, as individuals may not have anticipated their posts being used for research purposes, even in an anonymized form. Additionally, although the narratives are public and anonymous, we still make sure we have IRB exempt status before publishing our collected narratives. Despite these concerns, we believe our work will have a positive societal impact. By providing a deeper understanding of the challenges faced by low-SES students, our findings could inform educational policies and initiatives aimed at addressing socioeconomic disparities. Ultimately, our research could contribute to greater equity and inclusion for marginalized communities.

References

Nancy E. Adler and Katherine Newman. 2017. The impact of socioeconomic status on access to healthcare: A review of the literature. *Social Science Medicine*, 181:25–33.

Jacob H. Brown and Michael L. Carr. 2013. The impact of financial aid on college enrollment and completion: Evidence from a randomized study.

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. TweetNLP: Cutting-edge natural language processing for social media.

Amanda Cercas Curry, Zeerak Talat, and Dirk Hovy. 2024. Impoverished language technology: The lack of (social) class in nlp. *arXiv preprint arXiv:2403.03874*.

Motti Kelbessa, Estephanos Jebessa, and Labiba Jahan. 2024. Addressing educational inequalities of low ses students: Leveraging natural language processing for impact. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 388–391.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 405–413.

Sang Min Lee, M Harry Daniels, Ana Puig, Rebecca A Newgent, and Suk Kyung Nam. 2008. A data-based model to predict postsecondary educational attainment of low-socioeconomic-status students. *Professional School Counseling*, 11(5):2156759X0801100504.

Fons Mentink. 2016. Machine driven predictions of the socio-economic status of twitter users. Master's thesis, University of Twente.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Judith Scott-Clayton. 2015. The role of financial aid in promoting college access and success: Research evidence and proposals for reform. *Journal of Student Financial Aid*, 45(3):3.

Southern Methodist University (SMU). 2019. Think-play-hack: World views.

Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. 2018. Social media analytics–challenges in topic discovery, data collection, and data preparation. *International journal of information management*, 39:156–168.

Marvin A Titus. 2006. Understanding college degree completion of students with low socioeconomic status: The influence of the institutional financial context. *Research in Higher Education*, 47:371–398.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

A Appendix

A.1 Data

You need to buy textbooks or find PDFs. Talk to your faculty adviser if you ever have questions, and make sure you're talking to the general university adviser too. Welcome, if you've ever been to the bank. Some teachers might inflate grades, but in the end, don't skip class or slack off on homework. Stop studying for exams the night before—it's a terrible habit. Get an internship as quickly as possible, and try to have a job lined up before graduation. Well, go to the career center and make friends with people who are getting jobs. Put together a serious resume and cover letter as soon as you can. No one is going to be impressed with fancy colors or formatting in the professional world. The career center can help you critique them for free. You're also going to get wrecked by student loans in a few years, so prepare wisely unless you're Richie Rich.

A.2 Classification

Model	Parameter values
RF	n_estimators = 100, random_state = 46. Random Forest model using 100 trees to ensure a balanced performance.
MNB	Default parameters, suitable for text data with TF-IDF representation. The Multinomial Naive Bayes assumes feature independence and is efficient for large-scale text data.
SVM	kernel = linear, probability = True, random_state = 46. A linear kernel is efficient for text classification, with probability estimates enabled for evaluation purposes.
LR	max_iter = 500, random_state = 46. LR with a limit on the number of iterations to ensure convergence.
BERT	Model = bert-base-uncased, MAX_LEN = 512, TRAIN_BATCH_SIZE = 32, VALID_BATCH_SIZE = 32, EPOCHS = 40, LEARNING_RATE = 1e-05, hidden size = 768, dropout = 0.3. Fine-tuned for binary classification with early stopping, patience = 8.
DistilBERT	Model = distilbert-base-uncased, MAX_LEN = 512, TRAIN_BATCH_SIZE = 32, VALID_BATCH_SIZE = 32, EPOCHS = 40, LEARNING_RATE = 1e-05, hidden size = 768, dropout = 0.3. Fine-tuned for binary classification with early stopping, patience = 8.
ALBERT	Model = albert-base-v2, MAX_LEN = 512, TRAIN_BATCH_SIZE = 32, VALID_BATCH_SIZE = 32, EPOCHS = 40, LEARNING_RATE = 1e-05, hidden size = 768, dropout = 0.3. Fine-tuned for binary classification with early stopping, patience = 8.
BART	Model = facebook/bart-large-mnli, MAX_LEN = 512, TRAIN_BATCH_SIZE = 16, VALID_BATCH_SIZE = 16, EPOCHS = 40, LEARNING_RATE = 1e-05, hidden size = 1024, dropout = 0.3. Fine-tuned BART with a binary classification head and early stopping, patience = 8.
DeBERTa	Model = microsoft/deberta-base, MAX_LEN = 512, TRAIN_BATCH_SIZE = 32, VALID_BATCH_SIZE = 32, EPOCHS = 40, LEARNING_RATE = 1e-05, hidden size = 768, dropout = 0.3. Fine-tuned for binary classification using CLS token with early stopping, patience = 8.
ELECTRA	Model = google/electra-base-discriminator, MAX_LEN = 512, TRAIN_BATCH_SIZE = 32, VALID_BATCH_SIZE = 32, EPOCHS = 40, LEARNING_RATE = 1e-05, hidden size = 768, dropout = 0.3. Fine-tuned for binary classification using CLS token with early stopping, patience = 8.
XLNet	Model = xlnet-base-cased, MAX_LEN = 512, TRAIN_BATCH_SIZE = 32, VALID_BATCH_SIZE = 32, EPOCHS = 40, LEARNING_RATE = 1e-05, hidden size = 768, dropout = 0.3. Fine-tuned XLNet model for binary classification with early stopping, patience = 8.
T5	Model = t5-base, MAX_LEN = 512, TRAIN_BATCH_SIZE = 32, VALID_BATCH_SIZE = 32, EPOCHS = 40, LEARNING_RATE = 1e-05, output = logits for binary classification with early stopping, patience = 8.
LLaMA	Model = princeton-nlp/Sheared-LLaMA-1.3B, MAX_LEN = 512, TRAIN_BATCH_SIZE = 8, VALID_BATCH_SIZE = 8, EPOCHS = 40, LEARNING_RATE = 1e-05, dropout = 0.3, with LoRA fine-tuning for binary classification with early stopping, patience = 8.
GPT-2	Model = gpt, MAX_LEN = 512, TRAIN_BATCH_SIZE = 2, VALID_BATCH_SIZE = 2, EPOCHS = 10, LEARNING_RATE = 1e-04, dropout = 0.3, early stopping, patience = 5.
DistilGPT-2	Model = distilgpt2, MAX_LEN = 512, TRAIN_BATCH_SIZE = 2, VALID_BATCH_SIZE = 2, EPOCHS = 10, LEARNING_RATE = 1e-04, dropout = 0.3, early stopping, patience = 5.
OPT-13B	Model = KoboldAI/OPT-13B-Erebus, MAX_LEN = 64, batch_size = 1, gradient checkpointing enabled, mixed precision used, early stopping, patience = 5.
LoRA GPT-2	Model = gpt2, MAX_LEN = 512, TRAIN_BATCH_SIZE = 2, VALID_BATCH_SIZE = 2, EPOCHS = 10, LEARNING_RATE = 1e-04, dropout = 0.3, with LoRA fine-tuning, early stopping, patience = 5.
LoRA LLaMA-1.3B	Model = princeton-nlp/Sheared-LLaMA-1.3B, MAX_LEN = 512, TRAIN_BATCH_SIZE = 8, VALID_BATCH_SIZE = 8, EPOCHS = 40, LEARNING_RATE = 1e-05, dropout = 0.3, with LoRA fine-tuning, early stopping, patience = 8.
LoRA Distill-GPT	Model = distilgpt2, MAX_LEN = 512, TRAIN_BATCH_SIZE = 2, VALID_BATCH_SIZE = 2, EPOCHS = 10, LEARNING_RATE = 1e-04, dropout = 0.3, with LoRA fine-tuning, early stopping, patience = 5.
Claudia few-shot	Model = Claudia few-shot, few-shot prompt-based learning, early stopping patience = 5, uses structured prompting with the dataset of examples.
LLaMA-7B few-shot	Model = LLaMA-7B, few-shot learning using a structured prompting with the dataset of examples.

Table 3: Summary of the architecture and parameters for each model used for classification.

Model	SES	Precision	Recall	F_1	Model	SES	Precision	Recall	F_1
Random Forest	not-Low	0.57	0.40	0.47	Multinomial Naive Bayes	not-Low	1.00	0.40	0.57
	Low	0.54	0.70	0.61		Low	0.62	1.00	0.77
	Avg.	0.55	0.55	0.54		Avg.	0.81	0.70	0.67
Support Vector Machine (SVM)	not-Low	0.58	0.70	0.64	Logistic Regression	not-Low	0.58	0.70	0.64
	Low	0.62	0.50	0.56		Low	0.62	0.50	0.56
	Avg.	0.60	0.60	0.60		Avg.	0.60	0.60	0.60
Gradient Boosting	not-Low	0.62	0.80	0.70	RoBERTa	not-Low	0.78	0.70	0.74
	Low	0.71	0.50	0.59		Low	0.73	0.80	0.76
	Avg.	0.66	0.65	0.64		Avg.	0.75	0.75	0.75
DeBERTa	not-Low	0.69	0.90	0.78	ELECTRA	not-Low	0.69	0.90	0.78
	Low	0.86	0.60	0.71		Low	0.86	0.60	0.71
	Avg.	0.77	0.75	0.74		Avg.	0.77	0.75	0.74
ALBERT	not-Low	0.71	0.50	0.59	XLNet	not-Low	0.53	0.90	0.67
	Low	0.62	0.80	0.70		Low	0.67	0.20	0.31
	Avg.	0.66	0.65	0.64		Avg.	0.60	0.55	0.49
T5	not-Low	0.50	0.30	0.38	OPT-13B fewshot	not-Low	0.50	0.90	0.64
	Low	0.50	0.70	0.58		Low	0.50	0.10	0.17
	Avg.	0.50	0.50	0.48		Avg.	0.50	0.50	0.40
Claudia fewshot	not-Low	0.53	0.90	0.67	LLamA-1.3B fewshot	not-Low	0.50	0.80	0.62
	Low	0.67	0.20	0.31		Low	0.50	0.20	0.29
	Avg.	0.60	0.55	0.49		Avg.	0.50	0.50	0.45
LoRA Fine-Tune GPT-2	not-Low	0.50	1.00	0.67	LLaMA-7B fewshot	not-Low	0.38	0.50	0.43
	Low	0.00	0.00	0.00		Low	0.29	0.20	0.24
	Avg.	0.25	0.50	0.33		Avg.	0.34	0.35	0.34
LoRA Fine-Tune DistilGPT	not-Low	0.50	1.00	0.67	LoRA Fine-Tune LLaMA-1.3B	not-Low	0.50	1.00	0.67
	Low	0.00	0.00	0.00		Low	0.00	0.00	0.00
	Avg.	0.25	0.50	0.33		Avg.	0.25	0.50	0.33
bart-large-mnli Zero Shot	not-Low	0.80	0.80	0.80	DistilBERT	not-Low	0.78	0.70	0.74
	Low	0.80	0.80	0.80		Low	0.73	0.80	0.76
	Avg.	**0.80**	**0.80**	**0.80**		Avg.	0.75	0.75	0.75
BERT	not-Low	0.75	0.60	0.67	RoBERTa-large	not-Low	0.78	0.70	0.74
	Low	0.67	0.80	0.73		Low	0.73	0.80	0.76
	Avg.	0.71	0.70	0.70		Avg.	0.75	0.75	0.75

Table 4: First Iteration Performance of different models for classifying socioeconomic classes. Avg. = Macro average.

A.3 Results

This Section has the results tables from the three iterations

Model	SES	Precision	Recall	F_1	Model	SES	Precision	Recall	F_1
Random Forest	not-Low	0.72	0.82	0.77	Multinomial Naive Bayes	not-Low	0.89	0.61	0.72
	Low	0.77	0.65	0.71		Low	0.69	0.92	0.79
	Avg.	0.74	0.74	0.74		Avg.	0.79	0.76	0.75
Support Vector Machine (SVM)	not-Low	0.86	0.89	0.88	Logistic Regression	not-Low	0.84	0.93	0.88
	Low	0.88	0.85	0.86		Low	0.91	0.81	0.86
	Avg.	**0.87**	**0.87**	**0.87**		Avg.	**0.87**	**0.87**	**0.87**
Gradient Boosting	not-Low	0.75	0.75	0.75	RoBERTa	not-Low	0.74	0.82	0.78
	Low	0.73	0.73	0.73		Low	0.78	0.69	0.73
	Avg.	0.74	0.74	0.74		Avg.	0.76	0.76	0.76
DeBERTa	not-Low	0.75	0.75	0.75	ELECTRA	not-Low	0.73	0.96	0.83
	Low	0.73	0.73	0.73		Low	0.94	0.62	0.74
	Avg.	0.74	0.74	0.74		Avg.	0.83	0.80	0.79
ALBERT	not-Low	0.66	0.68	0.67	XLNet	not-Low	0.68	0.46	0.55
	Low	0.64	0.62	0.63		Low	0.57	0.77	0.66
	Avg.	0.65	0.65	0.65		Avg.	0.63	0.61	0.60
T5	not-Low	0.84	0.57	0.68	OPT-13B fewshot	not-Low	0.53	0.57	0.55
	Low	0.66	0.88	0.75		Low	0.50	0.46	0.48
	Avg.	0.75	0.72	0.72		Avg.	0.52	0.52	0.52
Claudia fewshot	not-Low	0.51	0.71	0.60	LLamA-1.3B fewshot	not-Low	0.45	0.61	0.52
	Low	0.47	0.27	0.34		Low	0.31	0.19	0.24
	Avg.	0.49	0.50	0.47		Avg.	0.38	0.41	0.38
Lora finetune GPT-2	not-Low	0.52	1.00	0.68	LLamA-7B fewshot	not-Low	0.79	0.33	0.46
	Low	0.00	0.00	0.00		Low	0.74	0.78	0.76
	Avg.	0.27	0.52	0.35		Avg.	0.69	0.64	0.66
Lora finetune DistillGPT	not-Low	0.52	1.00	0.68	LoRA Fine-Tuning LLaMA-1.3B	not-Low	0.46	0.75	0.57
	Low	0.00	0.00	0.00		Low	0.12	0.04	0.06
	Avg.	0.27	0.52	0.35		Avg.	0.30	0.41	0.32
bart-large-mnli Zero Shot	not-Low	0.00	0.00	0.00	DistilBert	not-Low	0.85	0.61	0.71
	Low	0.45	0.88	0.60		Low	0.68	0.88	0.77
	Avg.	0.22	0.43	0.29		Avg.	0.77	0.74	0.74
BERT	not-Low	0.79	0.54	0.64	RoBERTa-large	not-Low	0 0.85	0.61	0.71
	Low	0.63	0.85	0.72		Low	0.68	0.88	0.77
	Avg.	0.71	0.69	0.68		Avg.	0.74	0.74	0.74

Table 5: Second Iteration Performance of different models for classifying socioeconomic classes on second iteration. Avg. = Weighted average by the number of narratives.

Model	SES	Precision	Recall	F_1	Model	SES	Precision	Recall	F_1
Random Forest	not-Low	0.63	0.53	0.57	Multinomial Naive Bayes	not-Low	0.70	0.45	0.55
	Low	0.60	0.69	0.64		Low	0.60	0.81	0.69
	Avg.	0.61	0.61	0.61		Avg.	0.65	0.63	0.62
Support Vector Machine (SVM)	not-Low	0.59	0.63	0.61	Logistic Regression	not-Low	0.84	0.93	0.88
	Low	0.61	0.58	0.59		Low	0.64	0.65	0.65
	Avg.	0.60	0.60	0.60		Avg.	0.64	0.64	0.64
Gradient Boosting	not-Low	0.53	0.65	0.58	RoBERTa	not-Low	0.61	0.92	0.73
	Low	0.56	0.44	0.49		Low	0.85	0.42	0.56
	Avg.	0.55	0.54	0.54		Avg.	0.73	0.67	0.65
DeBERTa	not-Low	0.54	0.69	0.60	ELECTRA	not-Low	0.54	0.82	0.65
	Low	0.58	0.42	0.49		Low	0.64	0.31	0.42
	Avg.	0.56	0.55	0.55		Avg.	0.59	0.56	0.53
ALBERT	not-Low	0.54	0.61	0.57	XLNet	not-Low	0.62	0.75	0.68
	Low	0.57	0.50	0.53		Low	0.69	0.56	0.62
	Avg.	0.55	0.55	0.55		Avg.	0.66	0.65	0.65
T5	not-Low	0.55	0.55	0.55	OPT-13B fewshot	not-Low	0.50	0.69	0.58
	Low	0.56	0.56	0.56		Low	0.52	0.33	0.40
	Avg.	0.55	0.55	0.55		Avg.	0.51	0.50	0.49
Claudia fewshot	not-Low	0.49	0.76	0.60	LLamA-1.3B fewshot	not-Low	0.53	0.71	0.61
	Low	0.50	0.23	0.32		Low	0.57	0.38	0.46
	Avg.	0.50	0.50	0.46		Avg.	0.55	0.54	0.53
Lora finetune GPT-2	not-Low	0.50	1.00	0.66	LLamA-7B fewshot	not-Low	0.48	0.65	0.55
	Low	0.00	0.00	0.00		Low	0.47	0.31	0.37
	Avg.	0.25	0.50	0.33		Avg.	0.47	0.48	0.46
Lora finetune DistillGPT	not-Low	0.50	1.00	0.66	LoRA Fine-Tuning LLaMA-1.3B	not-Low	0.50	1.00	0.66
	Low	0.00	0.00	0.00		Low	0.12	0.04	0.06
	Avg.	0.25	0.50	0.33		Avg.	0.25	0.50	0.33
bart-large-mnli Zero Shot	not-Low	0.45	0.45	0.45	DistilBert	not-Low	0.65	0.51	0.57
	Low	0.46	0.46	0.46		Low	0.60	0.73	0.66
	Avg.	0.46	0.46	0.46		Avg.	0.63	0.62	0.62
BERT	not-Low	0.56	0.55	0.55	RoBERTa-large	not-Low	0.70	0.76	0.73
	Low	0.57	0.58	0.57		Low	0.74	0.67	0.71
	Avg.	0.56	0.56	0.56		Avg.	**0.72**	**0.72**	**0.72**

Table 6: Third Iteration Performance of different models for classifying socioeconomic classes on second iteration. Avg. = Weighted average by the number of narratives.

A.4 Prompt Engineering for Extracting Background, Solutions, and Struggles Information

This subsection provides a detailed explanation of the prompt engineering techniques used to extract background information, solutions, and struggles from the narratives of low-SES students. By constructing specific prompts and using the LLaMA model, we ensured the precise extraction of information in a structured format, relying only on direct quotes from the texts. These prompts aim to assist in understanding the challenges and efforts described by low-SES students, ensuring that no additional information is added or altered during extraction.

A.4.1 Model Pipeline Setup

We utilized the `transformers` library from Hugging Face to create a pipeline for text generation and extraction. The LLaMA model was fine-tuned for generating outputs that align with our prompt design. The following configuration was applied to the pipeline for all tasks:

- **max_new_tokens=300**: Sets the maximum number of tokens to generate during extraction. This ensures that the output is concise and focused.
- **do_sample=False**: Sampling is disabled to provide deterministic and consistent outputs from the model.
- **temperature=0.5**: A lower temperature value ensures less randomness in the output, resulting in more controlled and accurate text generation.
- **device**: The model was configured to run on either GPU (if available) or CPU, ensuring flexibility in processing.

The prompts were specifically designed to elicit structured information, such as family background, solutions, and struggles, from the students' narratives. Below, we describe each function used to extract these key elements.

A.4.2 Extracting Background Information

The first step was to extract background information, particularly focusing on the family situations described in the narratives. The goal was to identify direct quotes that describe the family context of the students, such as financial hardships or living conditions.

The following function was designed to handle this task:

```
def extract_background(text):
    prompt = f'''
    All the texts provided are written by low-SES (SES) students
        who are writing about their struggles.
    ...
    Important: Extract the following information exactly from the
        text without adding or changing any words:
    - background or any texts about family situations (directly
        quoted from the text)

    Text: {text}

    valid JSON Output (only with direct quotes from the text):
    '''

    output = llama_pipeline(prompt, max_new_tokens=200, do_sample=
        False, temperature=0.5)
    generated_text = output[0]['generated_text']
    # Process output for background quotes
    ...
```

This prompt ensures that only direct quotes describing the students' family background are extracted and returned in a valid JSON format.

A.4.3 Extracting Solutions Information

In addition to background information, we extracted the solutions that students employed to address their struggles. These solutions may involve actions taken to overcome financial or social barriers, as well as any efforts to improve their academic or personal circumstances.

The function below is responsible for extracting the solutions from each text:

```
def extract_solutions(text):
    prompt = f'''
    All the texts provided are written by low-SES (SES) students
        who are writing about their struggles.
    ...
    Important: Extract the following information exactly from the
        text without adding or changing any words:
    - Solutions or actions they took to address these struggles (
        directly quoted from the text)

    Text: {text}

    valid JSON Output (only with direct quotes from the text):
    '''

    output = llama_pipeline(prompt, max_new_tokens=200, do_sample=
        False, temperature=0.5)
    generated_text = output[0]['generated_text']
    # Process output for solutions quotes
    ...
```

This function captures the strategies or actions the students took to manage or overcome their struggles, returning the data in a structured JSON format for analysis.

A.4.4 Extracting Struggles Information

The third aspect of our extraction was to focus on the specific struggles described by the students. These struggles include financial, psychological, physical, or social hardships. The function uses a similar approach, instructing the model to identify and extract direct quotes related to the students' difficulties.

The function for extracting struggles is as follows:

```
def extract_struggles(text):
    prompt = f'''
    All the texts provided are written by low-SES (SES) students
        who are writing about their struggles.
    ...
    Important: Extract the following information exactly from the
        text without adding or changing any words:
    - Struggles they faced (directly quoted from the text)

    Text: {text}

    Output valid JSON with only direct quotes related to struggles:
    '''

    output = llama_pipeline(prompt, max_new_tokens=300, do_sample=
        False, temperature=0.5)
    generated_text = output[0]['generated_text']

    # Process and return the generated text as JSON
    ...
```

This function extracts the struggles faced by the students and returns them as direct quotes in a JSON structure.

A.4.5 Post-processing and Valid JSON Output

In all cases, after the output is generated by the LLaMA model, the generated text is processed to extract the relevant information in JSON format. The output is validated to ensure it contains the correct fields

(e.g., background, solutions, or struggles), and any parsing errors are handled gracefully by returning a fallback structure if needed.

The extracted data is then consolidated into a structured format for further analysis. This structured data helps in understanding the key themes and experiences described by the low-SES students.

A.4.6 Conclusion

By employing these prompt engineering techniques, we were able to extract detailed and structured information regarding the backgrounds, struggles, and solutions described by the students in their narratives. The use of precise prompts, alongside the LLaMA model, allowed for accurate extraction of direct quotes, preserving the authenticity of the students' experiences. This extracted data provides valuable insights into the challenges faced by low-SES students and their efforts to overcome them.

A.4.7 Limitations

We acknowledge several limitations in our current research that we plan to address in future work. First, although our dataset offers valuable insights into the experiences of low-SES students, it is limited to narratives from a specific time frame. Expanding the dataset to include narratives from a broader range of years will provide a more comprehensive view of the evolving challenges faced by low-SES students. Second, while our data were annotated semi-automatically, it has not yet undergone a rigorous double-annotation or gold-standard validation process, which we are currently working on to enhance the dataset's reliability. Implementing this more precise annotation method will improve the reliability of our results. Finally, although our semi-automatic data collection model showed promising results, future work will focus on refining this model to ensure it can operate independently and efficiently at scale and using active learning.