AI Assistant for Socioeconomic Empowerment Using Federated Learning

Nahed Abdelgaber*[1], Labiba Jahan [1], Nino Castellano [1], Joshua R. Oltmanns [2], Mehak Gupta [1], Jia Zhang [1], Akshay Pednekar [1], Ashish Basavaraju [1], Ian Velazquez [1], Zerui Ma [1]
[1]Southern Methodist University, Dallas, TX 75205
[2]Washington University of St. Louis, MO 63130

Abstract

Socioeconomic status (SES) reflects an individual's standing in society, from a holistic set of factors including income, education level, and occupation. Identifying individuals in low-SES groups is crucial to ensuring they receive necessary support. However, many individuals may be hesitant to disclose their SES directly. This study introduces a federated learning-powered framework capable of verifying individuals' SES levels through the analysis of their communications described in natural language. We propose to study language usage patterns among individuals from different SES groups using clustering and topic modeling techniques. An empirical study leveraging life narrative interviews demonstrates the effectiveness of our proposed approach.

1 Introduction

Socioeconomic status (SES) is a key determinant of an individual's opportunities, well-being, and access to essential resources such as education, healthcare, and employment. Traditional SES assessments primarily rely on structured demographic data and self-reported surveys, which can be incomplete, biased, or intrusive. Many individuals may be reluctant to disclose their SES due to privacy concerns or social stigma, further limiting the effectiveness of such assessments.

Recent advancements in Natural Language Processing (NLP) provide an alternative approach by inferring SES from linguistic patterns in personal narratives. Research has shown that differences in word choice, discourse structure, and emotional expression correlate with socioeconomic background. However, most prior work has relied on structured social media text or survey responses rather than free-form narratives, restricting the depth of analysis.

This study introduces a federated learning-powered SES framework that preserves user privacy while analyzing life narratives. Federated learning (FL) enables decentralized model training without exposing raw personal data, addressing critical privacy concerns associated with SES inference. Our framework integrates NLP-based SES classification, topic modeling, clustering, and sentiment analysis to identify linguistic patterns linked to different SES levels. By leveraging a dataset of transcribed life narratives, we demonstrate that SES can be inferred effectively through language while maintaining privacy and scalability.

To evaluate our approach, we test multiple machine learning and transformer-based models, with RoBERTa after summarization achieving the highest performance. Additionally, we assess the model's generalizability by testing it on out-of-distribution (OOD) data. Our results highlight the potential of privacy-conscious SES classification for future applications in AI-driven social research and personalized support systems.

The remainder of this paper is structured as follows. Section 2 reviews prior research on identifying SES and section 3 presents our proposed federated learning framework, outlining its role in privacy-preserving SES identification. Then section 4 describes the dataset and preparation steps and section 5 details our methodology, including the machine learning models used for SES classification, topic modeling and sentiment analysis. Section 6 provides a broader discussion of the findings, their implications, and future research directions and Section 7 summarizes the study's key contributions. Finally, Section 8 examines ethical considerations and societal impact. The paper ends with a section on the limitations of the research in section 9.

*Major contributor. Contact: nabdelgaber@smu.edu

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 490–501
May 3-4, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Understanding socioeconomic status (SES) through language has been explored in computational social science, sociolinguistics, and NLP. Prior research has demonstrated that language use reflects socioeconomic differences, with variations in vocabulary, syntactic complexity, and discourse structures (Bernstein, 1971; Pennebaker, 2011). Lower-SES individuals tend to use more context-dependent language, while higher-SES individuals employ abstract and elaborative discourse (Bernstein, 1971). Additionally, studies in psycholinguistics have shown that SES influences cognitive framing and emotional expression (Snibbe and Markus, 2003; Kraus et al., 2017).

Traditional SES classification approaches rely on structured survey data or economic indicators such as income and education levels (Hennig and Liao, 2013; Balasankar et al., 2020). Computational methods have extended these approaches by analyzing text from social media to infer SES. For instance, (Lampos et al., 2016) classified Twitter users' SES using Gaussian Processes, achieving 82% accuracy. Similarly, (Levy Abitbol et al., 2019) trained Random Forest and XGBoost classifiers on Twitter data, reaching F1 scores of 0.70-0.73. Other studies have used Support Vector Machines (SVMs) and Naïve Bayes models to predict SES from online user profiles (Zhou, 2017). However, these studies primarily rely on structured social media data or metadata rather than free-form personal narratives, which provide deeper insights into lived experiences.

Beyond social media, NLP techniques have been applied to infer SES-related attributes from diverse sources. (Beckel et al., 2013) predicted household SES from electricity consumption data, while (Faroqi et al., 2018) used transit patterns to estimate SES indicators such as income. Despite these advances, few studies have explored SES classification from life narratives, which contain richer self-reflections and personal challenges.

Privacy concerns in SES classification have led to the exploration of Federated Learning (FL) as a decentralized and privacy-preserving approach (McMahan et al., 2017; Kairouz et al., 2021). FL has been widely applied in domains such as healthcare (Yang et al., 2019) and finance (Hardy et al., 2019), but its use in social science and SES inference remains limited. FL enables collaborative model training across decentralized devices without exposing user data, making it a promising solution for SES classification where individuals may be hesitant to share personal details. While prior studies have proposed FL for text classification tasks (Liu et al., 2021), this work is among the first to explore FL for SES inference from personal narratives.

This study extends previous research by introducing a privacy-preserving SES framework that integrates FL with NLP-driven linguistic analysis. Unlike prior works that rely on structured SES indicators, our approach analyzes life narratives using transformer-based models while maintaining data privacy. Additionally, we evaluate our model on out-of-distribution (OOD) data, addressing a major gap in SES classification generalizability.

3 Proposed Framework

Traditional methods for socioeconomic status (SES) identification rely on centralized datasets and self-reported surveys, raising concerns about privacy, data availability, and scalability. This study introduces a federated learning-powered SES framework designed to preserve user privacy while allowing for decentralized model training. Unlike conventional approaches that require users to share raw personal data, federated learning (FL) enables collaborative model refinement by exchanging only model updates. This approach reduces privacy risks while preserving the overall effectiveness of the process.

This study implements and evaluates the SES classification and profiling stage within a simulated FL environment. The broader system envisions a privacy-preserving pipeline that integrates a knowledge graph (KG) to provide targeted recommendations based on SES profiling results. The focus of this study remains on demonstrating the feasibility of FL for SES classification and profiling, as well as assessing its generalizability.

The proposed framework consists of three primary components. The first is the federated SES profiling system, which applies machine learning techniques to infer SES-related patterns from life narratives. This component has been developed and evaluated in this study, demonstrating the viability of FL-based SES profiling. The second component, a knowledge graph, is intended to enhance the system by mapping SES-related factors to relevant support resources. While not implemented in this study, it represents a future direction for generating personalized recommendations based on an individual's linguistic markers and sentiment insights. The third component involves local model refinement on client devices, enabling continuous personalization and adaptation without exposing sensitive user data. This final component remains conceptual and will be explored in future work.

Figure 1 provides an overview of the proposed federated learning-powered framework. The process begins with the deployment of the SES classifier to client devices, where users classify their personal narratives without transmitting raw text. A proposed extension of this process would involve generating SES profiles and using the knowledge graph to provide context-aware recommendations. Users would then interact with the recommendations, offering feedback that refines the classifier, profiling system, and the recommendation system. The model updates generated from this interaction would be aggregated on a central server, improving classification and profiling system accuracy while preserving individual privacy. The simulated FL setup tested in this study captures only the model refinement process, demonstrating that SES classification and profiling can be performed in a decentralized environment without significant loss of accuracy.

Beyond classification, this study highlights the potential of integrating FL with SES profiling to support real-world AI-driven interventions. The next phase of this research will focus on refining model aggregation strategies, enhancing fairness in SES predictions, and developing an adaptive recommendation mechanism that aligns with users' socioeconomic contexts. As part of future work, we will explore real-world federated deployment and assess the effectiveness of AI-driven SES profiling in diverse settings.

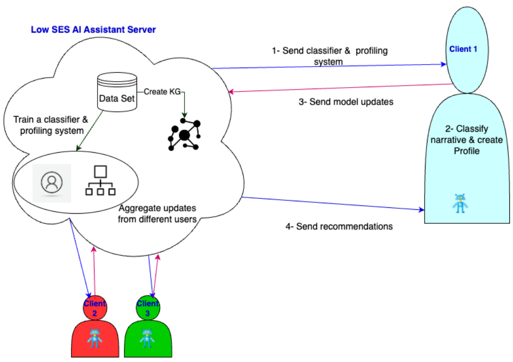


Figure 1: Proposed Federated Learning-powered SES Framework.

The federated learning approach offers several advantages for SES classification and profiling system. By keeping personal narratives on user devices, it eliminates ethical concerns related to direct SES data collection. The decentralized nature of the framework ensures that the system remains scalable and adaptable across different demographic groups. Additionally, the classifier and profiling system continuously improves as more users contribute model updates, enhancing its ability to detect linguistic markers of SES over time.

Despite these benefits, several challenges remain before real-world deployment is feasible. SES-related language varies significantly across individuals, introducing potential biases in model aggregation. The effectiveness of federated learning depends on user participation, as limited engagement in model fine-tuning could reduce the system's adaptability. Moreover, integrating a knowledge graph for SES-driven recommendations requires further research to establish meaningful connections between classified SES categories and actionable support interventions.

This study demonstrates the feasibility of federated SES classification and profiling system through simulation, highlighting its potential for privacy-preserving NLP applications. The broader framework, including real-world federated deployment and a knowledge graph-driven recommendation system, remains a direction for future research. Further exploration is needed to refine model aggregation strategies, improve fairness in predictions, and develop personalized recommendation mechanisms that align with users' socioeconomic contexts.

4 Data

4.1 Data Overview:

Data used for this study come from the St. Louis Personality and Aging Network (Oltmanns et al., 2014). Over 3.5 years, a representative community sample of 1,630 older adults were recruited from 100 square miles around the St. Louis area. Listed phone numbers and the Kish (Kish, 1949) method were used to identify a target for participation in a given household. Participants came to the laboratory and were interviewed for life history and other variables related to mental disorders and health status. Of the 1,630 participants, 1,408 participants had transcribed life narrative interviews for the present study.

STD	Low	Mid	High	Total
1.0	242	891	275	1408
0.5	474	541	393	1408
SES Class	**Total Texts**	**Avg. Sen.**	**Avg. Words**	**Total Words**
Low	474	113	1669	791145
Mid	541	105	1622	877750
High	393	101	1560	612951

Table 1: Summary of SES distribution and textual characteristics (STD = 0.5). Sen. = Sentences

We analyzed textual characteristics across SES classes, including the average number of sentences per text, words per text, and total words per class. Table 1 shows that while sentence and word structures remain relatively uniform across SES groups, the total content volume varies, potentially reflecting different levels of verbosity in narratives.

4.2 Data Preparation:

We included the language of participants in the transcripted text and removed any other words spoken by the interviewer to reduce noise in the data. We converted all text to lowercase, tokenized the words, and removed stop-words using the Natural Language Toolkit (NLTK) (Bird and Loper, 2004). To create labels for our classifier, we defined the socioeconomic class as a composite of the means of parents' education, participant education, and annual household income (Iacovino et al., 2014). We classified the interviews into three socioeconomic classes—low, mid, and high—in two different ways: using 1 and 0.5 standard deviations from the composite mean that we calculated in table 1. This classification follows sociological research frameworks that stratify SES into three broad tiers rather than binary or more granular categories (Lampos et al., 2016). The data was standardized using the StandardScaler from scikit-learn (Buitinck et al., 2013) to normalize SES scores before classification.

5 Empirical Study

This section details the methodology and results of two key analyses conducted in this study: SES prediction using machine learning and topic modeling for thematic exploration. The SES prediction task evaluates multiple classifiers, including traditional machine learning models and transformer-based approaches, to determine the most effective method for inferring SES from textual narratives. Experimental results demonstrate that transformer-based models, particularly RoBERTa after summarization, achieve the highest classification performance.

In parallel, topic modeling is employed to uncover thematic patterns in the narratives across different SES groups. Using a combination of embedding-based clustering and sentiment analysis, we identify key topics related to socioeconomic experiences and examine their emotional tone. The results highlight both commonalities and distinctions in how different SES groups discuss various aspects of their lives.

5.1 SES Classification

The SES classification task involved training machine learning models on transcribed life narratives. A variety of classifiers were evaluated, including traditional machine learning models such as Random Forest, Naïve Bayes, XGBoost, Support Vector Machines (SVM), and Logistic Regression, alongside transformer-based models.

To represent textual data, we explored TF-IDF, Word2Vec, and Transformer-based embeddings, with RoBERTa-based models achieving the best performance. Three preprocessing strategies were tested to handle varying narrative lengths:

RoBERTa with Truncation: Input texts were tokenized with a 512-token limit, truncating longer texts. The model included a RoBERTa encoder, a dropout layer (rate 0.3), and a fully connected classification layer. This approach performed well across all SES categories, achieving macro and weighted average F_1 scores of 0.82.

RoBERTa with Chunking: Longer texts were split into 512-token chunks, processed separately, and classified by averaging predictions across chunks. However, this method yielded lower performance (F_1 = 0.66), suggesting that truncation and summarization were more effective.

RoBERTa after Summarization: To retain key information in long texts, we applied summarization using a fine-tuned LLaMA-2-7B model before classification. This approach achieved the best results (F_1 = 0.87), demonstrating that summarization preserved SES-related signals better than chunking and truncation.

Traditional models (Random Forest, XGBoost) produced competitive results but were outperformed by transformer-based approaches. Experiments with larger models (Longformer, LLaMA-2) resulted in overfitting due to the dataset's limited size.

All models were trained using cross-entropy loss with the AdamW optimizer and evaluated via precision, recall, and F_1 scores. Hyperparameter settings are detailed in Table 4.

5.1.1 Results and Evaluation

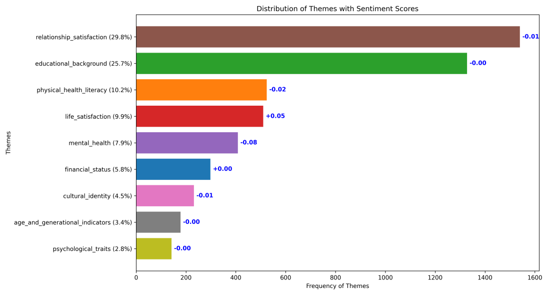
Table 2 presents the classification performance of different models. Among traditional classifiers, Random Forest and XGBoost achieved the highest weighted average F1 scores of 0.78 and 0.77, respectively. These models performed moderately well but struggled with capturing complex linguistic indicators of SES.

RoBERTa-based models demonstrated superior performance. The truncation-based RoBERTa classifier achieved an F1 score of 0.82, showing robustness across SES categories. The best results were obtained with the summarization-based RoBERTa model, which reached an F1 score of 0.87, highlighting the benefits of summarization in preserving key SES-related signals in lengthy narratives.

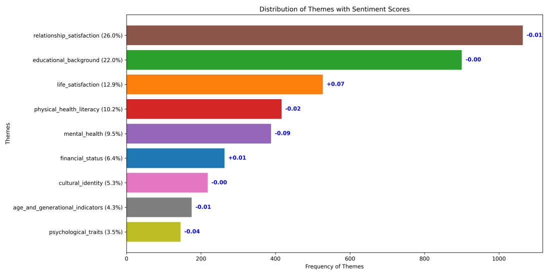
Evaluating Three-Classes Classifier: To assess the robustness of our models, we conducted an Out-of-distribution (OOD) evaluation using 74 low SES student narratives from (Kelbessa et al., 2024) and 74 manually selected non-low SES student narratives sourced from Reddit posts in *'college'* and *'ApplyingToCollege'*. The results are presented in Table 3.

RoBERTa achieved an average accuracy of 76.35%, demonstrating strong performance on OOD data, particularly in distinguishing between SES categories. The model correctly classified 68.92% of low SES texts and 83.78% of non-low SES texts. In contrast, the Random Forest model exhibited high variance, performing exceptionally well on non-low SES texts (93.24% accuracy) but poorly on low SES texts (only 24.32% accuracy). This suggests that the Random Forest struggles to generalize to unseen low SES narratives, whereas RoBERTa maintains a balanced classification ability.

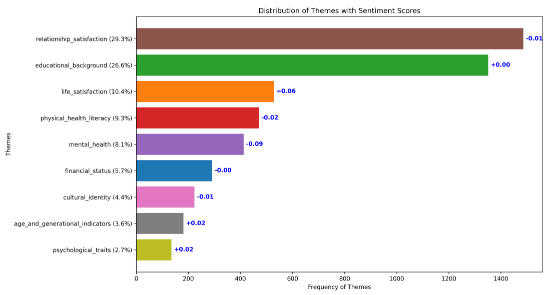
Evaluating Binary Classifier: A similar trend was observed in the binary classification task. RoBERTa outperformed Random Forest, achieving an overall accuracy of 80.00%, compared to 58.00% for Random Forest. RoBERTa classified 74.00% of low SES texts correctly, whereas Random Forest only managed 59.00%. Additionally, RoBERTa achieved an F1-score of 0.79 for Low SES, while Random Forest reached only 0.59, reinforcing that RoBERTa generalizes better across different SES distributions.



(a) Topic distribution in the low SES group. Relationship satisfaction and educational background dominate the discussions.



(b) Topic distribution in the medium SES group. The discussion remains balanced, with increased emphasis on mental health.



(c) Topic distribution in the high SES group. Financial status and cultural identity gain more prominence.

Figure 2: Comparison of topic distributions across SES groups based on the Biopsychosocial Model, incorporating sentiment analysis. The sentiment scores, shown in blue, reflect the emotional tone of each theme, providing further insights into SES-related discourse.

These findings highlight the superior generalization ability of RoBERTa, particularly in handling diverse and unseen text from different SES backgrounds. While Random Forest demonstrates high specificity in classifying non-low SES texts, its limited ability to classify low SES narratives reduces its effectiveness for this task.

5.2 Topic Modeling and Sentiment Analysis

To analyze themes from interviews, we implemented a topic modeling approach that integrated NLP techniques with clustering methods. Preprocessing steps included tokenization, stemming, and filtering out unnecessary terms to ensure that only meaningful content was retained. SentenceTrans-

Model	SES	Precision	Recall	F_1	Model	SES	Precision	Recall	F_1
Random Forest	High	0.89	0.68	0.77	Multinomial Naive Bayes	High	0.78	0.62	0.69
	Mid	0.79	0.79	0.79		Mid	0.76	0.78	0.77
	Low	0.71	0.83	0.77		Low	0.70	0.79	0.74
	Avg.	0.79	0.78	0.78		Avg.	0.74	0.74	0.74
Support Vector Machine (SVM)	High	0.72	0.63	0.68	Logistic Regression	High	0.80	0.59	0.68
	Mid	0.70	0.71	0.71		Mid	0.72	0.71	0.71
	Low	0.69	0.75	0.72		Low	0.67	0.80	0.73
	Avg.	0.70	0.70	0.70		Avg.	0.71	0.70	0.70
Extreme Gradient Boosting (XGB)	High	0.82	0.75	0.79	RoBERTa with Chunking	High	0.79	0.33	0.46
	Mid	0.74	0.79	0.77		Mid	0.55	0.80	0.66
	Low	0.75	0.79	0.78		Low	0.74	0.78	0.76
	Avg.	0.77	0.77	0.77		Avg.	0.69	0.64	0.66
RoBERTa with Truncation	High	0.75	0.74	0.74	RoBERTa with Summarization	High	0.89	0.85	0.87
	Mid	0.82	0.84	0.83		Mid	0.84	0.86	0.85
	Low	0.89	0.87	0.88		Low	0.87	0.88	0.88
	Avg.	0.82	0.82	0.82		Avg.	**0.87**	**0.86**	**0.87**

Table 2: Performance of different models for classifying socioeconomic classes. Avg. = Weighted average by the number of interview narratives.

3-Classes	Accuracy	Precision	Recall	F_1
RoBERTa (Avg)	76.35%	0.90	0.77	0.83
Low SES	68.92%	0.81	0.69	0.74
Not Low SES	83.78%	0.98	0.84	0.91
Random Forest (Avg)	77.66%	0.78	0.76	0.77
Low SES	24.32%	0.45	0.40	0.42
Not Low SES	93.24%	0.95	0.93	0.94
Binary	**Accuracy**	**Precision**	**Recall**	F_1
RoBERTa (Avg)	80.00%	0.80	0.80	0.80
Low SES	74.00%	0.83	0.74	0.79
Not Low SES	85.00%	0.77	0.85	0.81
Random Forest (Avg)	58.00%	0.58	0.58	0.58
Low SES	59.00%	0.60	0.60	0.60
Not Low SES	57.00%	0.56	0.57	0.56

Table 3: Performance of RoBERTa and Random Forest models on Out-of-Distribution (OOD) data.

former embeddings were used to create vector representations of the text, followed by dimensionality reduction using UMAP (McInnes et al., 2018). To identify distinct clusters of related text segments, we applied HDBSCAN (Rahman et al., 2016).

To ensure a comprehensive understanding of SES-related experiences, we grounded our thematic analysis in the Biopsychosocial Model, which provides a holistic approach by integrating biological, psychological, and social dimensions of human well-being. This model, originally proposed by Engel (Engel, 1977), has significantly influenced medical and psychological research by emphasizing the interconnectedness of physical health, mental health, personality traits, social interactions, and cultural influences.

With the guidance of a psychology expert, we identified key markers aligned with this model and utilized them to define the themes extracted from the narratives. These markers encompass psychological and social indicators of well-being and life circumstances. Specifically, our predefined themes include the following two markers.

1. Psychological Markers: Indicators of physical health literacy, mental health, psychological traits, life satisfaction, and educational background that reflect an individual's health awareness, emotional regulation, personality dimensions, subjective well-being, and educational experiences.

2. Social Markers: Aspects of financial status, relationship satisfaction, cultural identity, and generational indicators, which capture financial stability, interpersonal relationships, societal belonging, and generational perspectives.

Using these markers—also referred to as themes or topics—we mapped narrative text clusters to predefined conceptual categories by calculating the cosine similarity between each cluster's centroid and a set of theme seed embeddings. This approach aligned the topic modeling results with well-established constructs from the Psychosocial Model, thereby enhancing interpretability compared to purely data-driven clustering.

In addition to topic extraction, we performed sentiment analysis using the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon (Roehrick, 2020) to assess the emotional tone associated with each theme. VADER is particularly effective for short text analysis and provides a compound sentiment score ranging from -1 (negative) to 1 (positive). By aggregating sentiment scores for each theme, we gained insights into the emotional tone expressed in narratives from different SES groups. This sentiment analysis enables a deeper contextual understanding of how individuals across SES levels discuss various aspects of their lives, from health concerns to financial stability and relationships.

These markers and sentiment insights will also serve as a foundation for the profiling system in future work. Beyond SES classification, the profiling system will utilize these dimensions and weight them to provide personalized insights and recommendations based on an individual's linguistic patterns. By leveraging markers from the Psychosocial Model alongside sentiment analysis, we aim to create an adaptive and interpretable system capable of contextualizing SES understanding within broader life experiences and psychological traits.

5.2.1 Results

Figure 2 presents the distribution of topics across low, medium, and high SES groups, highlighting key differences in their thematic focus. In addition to topic prevalence, we also analyzed sentiment scores for each theme using the VADER lexicon, which captures the emotional tone associated with each category. The sentiment values are indicated in blue to distinguish them from topic proportions.

The structured nature of the interviews provides an essential context for interpreting these distributions. All interviewees were asked the same set of questions designed to frame their life narratives into four distinct "chapters". These questions encouraged them to reflect on key aspects of their lives, including how they would divide their life story into major periods, the individuals who had the most positive and negative influence on their journey, and the activities, moments, or aspects of life that bring them the most happiness. Since these core topics were embedded in the interview design, it is expected that themes such as relationship satisfaction, educational background, and life satisfaction emerged consistently across SES groups. This structured approach naturally led to more evenly distributed thematic proportions, as all participants reflected on similar life-defining aspects.

For the low SES group (Figure 2a), the dominant themes are relationship satisfaction (29.8%) and educational background (25.7%), followed by physical health literacy (10.2%), life satisfaction (9.9%), and mental health (7.9%). The sentiment analysis reveals neutral to slightly negative sentiments: relationship satisfaction (-0.01), educational background (-0.00), physical health literacy (-0.02), and mental health (-0.08), while life satisfaction shows a mildly positive sentiment (+0.05). These patterns suggest that although themes like education and relationships are frequently discussed, they often carry a tone of concern, with life satisfaction offering the most optimistic outlook.

In the medium SES group (Figure 2b), the theme distribution appears balanced. Relationship satisfaction (26.0%) and educational background (22.0%) remain the most prominent themes, followed by life satisfaction (12.9%), physical health literacy (10.2%), and mental health (9.5%). Mental health again carries the most negative sentiment score (-0.09), indicating increased concern around stress, anxiety, or emotional well-being. Conversely, life satisfaction receives a notably positive sentiment (+0.07), suggesting more optimistic discussions within this theme.

For the high SES group (Figure 2c), the thematic range is relatively broad. Relationship satisfaction (29.3%) and educational background (26.6%) emerge as the dominant themes. These are followed by life satisfaction (10.4%), physical health literacy (9.3%), and mental health (8.1%), highlighting a strong focus on well-being and personal development. Financial status (5.7%), cultural identity (4.4%), age and generational indicators (3.6%), and psychological traits (2.7%) appear with lower frequencies. Sentiment analysis shows that life satisfaction (+0.06) and psychological traits (+0.02) are discussed positively, while mental health carries the lowest sentiment (-0.09), pointing to prevalent emotional challenges. Age and generational indicators also exhibit slightly positive sentiment (+0.02), reflecting thoughtful engagement with identity across generations.

Overall, all SES groups prioritize relationship satisfaction and educational background, though the similarity in topic proportions can largely be attributed to the structured interview format, which prompted responses along similar psychosocial dimensions. Sentiment analysis shows that while life satisfaction trends positively across groups, mental health consistently registers the most negative sentiment, particularly among medium and high SES groups (-0.09). A complete breakdown of theme-specific sentiment and word distributions is presented in Appendix A. To improve interpretability of relative theme differences across SES groups, numerical percentages and sentiment scores are used in the main text, as minor visual differences in bar lengths may not be perceptible in the figures alone.

6 Discussion and Future Work

This study establishes a foundational step in developing a privacy-preserving AI framework for SES understanding. By leveraging NLP techniques, topic modeling, sentiment analysis, and Federated Learning (FL), we demonstrate that SES-related themes can be identified from personal narratives while ensuring data privacy. The classifier developed in this work is a key component of the broader system, validating the feasibility of text-based SES inference and its generalizability across different datasets. Moving forward, we aim to implement the remaining components of the framework, extending beyond SES classification to a comprehensive profiling system. This system will integrate linguistic markers from the Psychosocial Model and sentiment analysis to create personalized insights and recommendations. By weighting these markers, the profiling system will adaptively assess an individual's SES-related discourse, providing a more nuanced understanding of their lived experiences. The next phase of development involves deploying the SES profiling system within a distributed FL environment, ensuring privacy while continuously improving model accuracy. Additionally, we will develop a dynamic SES knowledge graph to map socioeconomic challenges, available resources, and intervention strategies. This knowledge-driven system will support an AI-powered recommendation mechanism, offering tailored financial, educational, and mental health support based on individual needs. Beyond classification, this study highlights the potential for real-world applications of SES profiling in personalized AI-driven interventions. Future work will focus on refining model aggregation strategies, enhancing fairness in predictions, and developing adaptive recommendation mechanisms that align with users' socioeconomic contexts. Testing the complete system in diverse settings will be essential to assess its impact, ethical considerations, and effectiveness in supporting low-SES communities.

7 Contribution

We have three major contributions. First, we proposed a novel framework that integrates federated learning (FL) with NLP-driven SES classification, allowing SES inference from life narratives while preserving data privacy. Second, we conducted extensive experiments evaluating SES classification using both traditional machine learning and transformer-based models. Additionally, we assessed generalization through out-of-distribution (OOD) evaluations on unseen narratives. Finally, we introduced a topic modeling approach based on social and psychological markers.

8 Ethical and Societal Impact

First, while our data cannot be published or shared due to confidentiality agreements, we will release our trained model to enable others to classify SES levels from various types of life narratives. This approach ensures that the confidentiality of the dataset used in this study is maintained. The dataset itself was collected under an approved Institutional Review Board (IRB) protocol and has undergone thorough ethical review to ensure compliance with privacy and ethical standards. Second, to mitigate potential misuse, such as using the model to infer SES from publicly available narratives for targeting individuals or groups for economic harm, we will release the model under a proper license and user agreement. This agreement will explicitly enforce compliance with legal and ethical standards, limiting the model's application to research and socially beneficial purposes. Third, as part of our broader framework, we plan to integrate federated learning (FL), allowing decentralized model training while ensuring that personal data remains on user devices. Finally, beyond privacy, this research aims to positively impact society by advancing the understanding of SES-related challenges. The SES profiling system, combined with a knowledge graph, can support AI-driven interventions in education, financial assistance, and mental health. Future research will focus on transparency, ethical oversight, and collaboration with policymakers to ensure socially beneficial applications.

9 Limitations

First, while our data-driven approach has achieved promising results, our analysis revealed that some misclassified samples showed a low distinction between the narratives of low, medium, and high SES classes. This suggests that certain narratives contain overlapping linguistic features that blur the boundaries between SES classifications. To address this, future work will explore incorporating a weighting system based on social markers to better differentiate SES classes within text narratives. Second, although RoBERTa with summarization provided the best performance, our findings indicate that summarization can lead to a loss of nuanced information. Similarly, truncation and chunking approaches, while practical for handling lengthy narratives, lose different types of contextual data. In future studies, we plan to explore advanced context-preserving methods. Finally, the private and sensitive nature of the data means it cannot be published or shared. However, we will make the trained model publicly available under a proper license to ensure its ethical use.

Acknowledgements

This work was supported by Computer Science Dept. at Lyle school of Engineering, Southern Methodist University (SMU). We would also like to thank SMU Office of Information Technology team for their assistance in using SMU AI SuperPOD.

References

V Balasankar, SV Penumatsa, and PRV Terlapu. 2020. Intelligent socio-economic status prediction system using machine learning models on rajahmundry ap, ses dataset. *Indian Journal of Science and Technology*, 13(37):3820–3842.

Christian Beckel, Leyna Sadamori, and Silvia Santini. 2013. Automatic socio-economic classification of households using electricity consumption data. In *Proceedings of the fourth international conference on Future energy systems*, pages 75–86.

Basil Bernstein. 1971. *Class, Codes and Control: Theoretical Studies Towards a Sociology of Language*. Routledge & Kegan Paul, London.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

George L. Engel. 1977. The need for a new medical model: A challenge for biomedicine. *Science*, 196(4286):129–136.

Hamed Faroqi, Mahmoud Mesbah, and Jiwon Kim. 2018. Inferring socioeconomic attributes of public transit passengers using classifiers. In *Proceedings of the 40th Australian transport research forum (ATRF)*.

Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Ben Edwards, Wray Buntine, and Leslie Cann. 2019. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4816–4823.

Christian Hennig and Tim F Liao. 2013. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 62(3):309–369.

Juliette M Iacovino, Joshua J Jackson, and Thomas F Oltmanns. 2014. The relative impact of socioeconomic status and childhood trauma on black-white differences in paranoid personality disorder symptoms. *Journal of abnormal psychology*, 123(1):225.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210.

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*, PETRA '24, page 388–391, New York, NY, USA. Association for Computing Machinery.

Leslie Kish. 1949. A procedure for objective respondent selection within the household. *Journal of the American statistical Association*, 44(247):380–387.

Michael W. Kraus, Jacinth J. X. Tan, and Joseph P. Park. 2017. Signs of social class: The experience of economic inequality in everyday life. *Psychological Review*, 124(4):546–573.

Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. 2016. Inferring the socioeconomic status of social media users based on behaviour and language. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 689–695. Springer.

Jacob Levy Abitbol, Eric Fleury, and Márton Karsai. 2019. Optimal proxy selection for socioeconomic status inference on twitter. *Complexity*, 2019(1):6059673.

Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng. 2021. Federated semi-supervised medical image classification via inter-client relation matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 159–168. Springer.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282.

Thomas F Oltmanns, Merlyn M Rodrigues, Yana Weinstein, and Marci EJ Gleason. 2014. Prevalence of personality disorders at midlife in a community sample: Disorders and symptoms reflected in interview, self, and informant reports. *Journal of psychopathology and behavioral assessment*, 36:177–188.

James W. Pennebaker. 2011. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press, New York.

Md Farhadur Rahman, Weimo Liu, Saad Bin Suhaim, Saravanan Thirumuruganathan, Nan Zhang, and Gautam Das. 2016. Hdbscan: Density based clustering over location based services. *arXiv preprint arXiv:1602.03730*.

Katherine Roehrick. 2020. *vader: Valence Aware Dictionary and sEntiment Reasoner (VADER)*. R package version 0.2.1.

Alison C. Snibbe and Hazel Rose Markus. 2003. You can't always get what you want: Social class, agency, and choice. *Journal of Personality and Social Psychology*, 85(5):989–1007.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19.

Ying Zhou. 2017. *Statistical Analysis of Classification Algorithms for Predicting Socioeconomics Status of Twitter Users*. Ph.D. thesis, Carleton University.

A Appendix

Model	Parameter values
RF	Tuned with max depth = 30, n estimators = 100, max features = 'sqrt', min samples leaf = 5, and min samples split = 10. These settings balance model complexity, ensuring diverse feature selection, while preventing overfitting by limiting tree depth and requiring minimum samples for splits and leaves.
MNB	Tuned with $alpha$ = 0.0001 and fit prior = False, controlling the likelihood estimate's smoothness and reducing bias from the prior class distribution. This enhances the model's ability to detect subtle differences in word frequencies across classes.
XGB	Tuned with learning rate = 0.1, max depth = 6, n estimators = 100, reg $lambda$ = 1, and subsample = 0.8. This configuration balances complexity and regularization, enhancing generalization.
SVM	C = 100 with a linear kernel, offering simplicity in interpreting decision boundaries and computational efficiency, suitable for high-dimensional text data.
LR	Tuned with C = 100, penalty = 'l2', and class weight = 'balanced', ensuring appropriate regularization and class balance while reducing bias by focusing on closely fitting the data.
RoBERTa (T)	Includes a pre-trained RoBERTa encoder, a dropout layer (rate 0.3), and a fully connected layer mapping the 768-dimensional output to three classes. Trained with cross-entropy loss and AdamW optimizer (learning rate 1e-5) for 50 epochs, batch size = 32, with early stopping. Avg. training time: 343.89 sec (NVIDIA A100-SXM GPUs).
RoBERTa (C)	Uses RoBERTa (T) model to process all chunks and averages results across the chunks to capture the full interview context. Avg. training time: 597.42 sec (NVIDIA A100-SXM GPUs).
RoBERTa (S)	Includes a dropout layer and a fully connected layer mapping RoBERTa's output to three classes. Early stopping was applied to prevent overfitting. Avg. training time: 1193 sec (NVIDIA A100-SXM GPUs).

Table 4: Summary of the architecture and parameters for each model. RF = Random Forest, MNB = Multinomial Naive Bayes, XGB = Extreme Gradient Boosting, SVM = Support Vector Machine, LR = Logistic Regression, RoBERTa (T) = RoBERTa with Truncation, RoBERTa (C) = RoBERTa with Chunking, RoBERTa (S) = RoBERTa with Summarization.

Top Words per Theme Across SES Groups

This section presents the top 20 words for each theme extracted from the narratives of low, medium, and high SES groups. These words were identified using a similarity-based clustering approach.

Theme	Top 20 Words
Educational Background	education, student, school, academic, college, university, schooling, studying, study, teacher, classroom, learning, semester, colleges, teach, undergraduate, graduate, lecture, schoolwork, attending
Physical Health Literacy	medicine, health, illness, doctor, healthcare, illnesses, medication, med, physician, surgery, patient, disease, sickness, hospital, hospitalize, drug, surgeon, diseases, clinic, pain
Relationship Satisfaction	relationship, marriage, partner, spouse, relationships, married, marrying, lover, conflict, together, companionship, marriages, fight, affair, sex, marries, girlfriend, argue, marry, fiancee
Psychological Traits	personality, trait, confidence, ego, introvert, outspokenness, attitude, outspoken, attraction, ability, intelligent, courage, demeanor, characterize, temperament, **insecurity**, introspect, **insecure**, shy, **inferiority**
Age and Generational Indicators	youth, teenager, teen, teenage, teens, adolescent, older, age, juvenile, adulthood, younger, adolescence, adult, youngster, childhood, grandpa, maturity, grandson, grandchildren, **grandchildrens**
Life Satisfaction	success, fulfilling, satisfaction, accomplishment, fulfillment, achievement, outcome, progress, blessing, reward, hopeful, succeed, accomplish, fulfill, achieve, contentment, praise, accolades, joy, **victory**
Financial Status	spending, money, finance, budget, income, debt, cash, economy, monies, rich, afford, wealthy, funding, expense, revenue, spend, spends, prosperity, **poverty**, fund
Mental Health	happiness, emotion, stress, therapy, mental, psychology, mentality, misery, sadness, feeling, mood, therapist, discomfort, frustration, anger, thinking, desire, fear, dying, **disorder**
Cultural Identity	culture, heritage, slang, civilization, territory, race, style, fashion, **german**, immigration, diversity, prejudice, assimilation, land, **italian**, mafia, citizen, **jewish**, renaissance, **white**

Table 5: Top 20 words per theme in the Low SES group. Bold words indicate unique terms for this SES group.

Theme	Top 20 Words
Educational Background	education, student, school, academic, college, university, **teaching**, **academics**, study, teacher, classroom, **universities**, **educator**, learning, semester, teachers, teach, undergraduate, class, **schoolteacher**
Physical Health Literacy	medicine, health, illness, doctor, healthcare, illnesses, medication, med, physician, **treatment**, surgery, patient, disease, sickness, hospital, hospitalize, **cure**, drug, surgeon, **injury**
Relationship Satisfaction	relationship, marriage, partner, spouse, **love**, **wife**, **husband**, relationships, married, marrying, lover, conflict, together, companionship, marriages, fight, affair, sex, girlfriend, **loving**
Psychological Traits	personality, **personalities**, trait, confidence, traits, confident, ego, introvert, **extroverted**, attitude, **egos**, **qualities**, **attitudes**, attraction, appearance, ability, intelligent, **smartness**, **aggressiveness**, courage
Age and Generational Indicators	youth, teenager, youthful, teen, teenage, teens, adolescent, older, generation, age, juvenile, adulthood, youngness, younger, midlife, adolescence, adult, youngster, demographic, **retirement**
Life Satisfaction	**gratitude**, success, **optimism**, satisfaction, accomplishment, fulfillment, **admiration**, achievement, outcome, successes, progress, blessing, satisfying, reward, succeed, accomplish, **gratification**, **relive**, **happy**, **content**
Financial Status	spending, money, **wealth**, finance, budget, income, debt, cash, economy, **budgeting**, monies, afford, wealthy, funding, **expenditure**, expense, revenue, spend, spends, poverty
Mental Health	**depression**, happiness, emotion, **anxiety**, stress, therapy, mental, mentality, **melancholy**, misery, **stressful**, sadness, feeling, **empathy**, **psychiatrist**, **suffering**, mood, **distress**, **comforting**, **depress**
Cultural Identity	culture, **ethnicity**, **nationality**, country, **accent**, nation, immigrant, tradition, **european**, territory, race, **indian**, style, translation, **african**, racist, **originate**, fashion, american, translate

Table 6: Top 20 words per theme in the Medium SES group. Bold words indicate unique terms for this SES group.

Theme	Top 20 Words
Educational Background	education, student, school, academic, college, university, schooling, teaching, **academia**, teacher, classroom, educator, semester, teach, undergraduate, **schoolteacher**, **professor**, graduate, lecture, schoolwork
Relationship Satisfaction	relationship, marriage, partner, spouse, relationships, marrying, lover, conflict, **divorcee**, marriages, fight, affair, sex, girlfriend, argue, marry, fiancee, **intimate**, **divorce**, **romance**
Life Satisfaction	success, fulfilling, satisfaction, accomplishment, fulfillment, **appreciation**, **rewarding**, **blessings**, achievement, outcome, successes, progress, blessing, satisfying, reward, hopeful, succeed, accomplish, **relive**, achieve
Physical Health Literacy	medicine, health, illness, doctor, healthcare, illnesses, medication, med, physician, treatment, patient, disease, sickness, hospital, hospitalize, **cure**, drug, surgeon, **clinic**, pain
Mental Health	happiness, emotion, **emotional**, anxiety, stress, mental, psychology, mentality, misery, **psychologist**, **stressful**, feeling, **psychiatrist**, **psychiatry**, **wellbeing**, therapist, distress, **stressor**, depress, **miserable**
Cultural Identity	immigrant, **spanish**, civilization, **mexican**, territory, style, citizens, fashion, translate, **tribe**, **dutch**, **asian**, **hispanic**, **german**, immigration, **folk**, diversity, **belonging**, **antique**, **ruling**
Financial Status	spending, money, wealth, finance, budget, income, debt, cash, economy, monies, rich, afford, funding, expense, spend, spends, poverty, fund, **currency**, **economics**
Psychological Traits	personality, trait, confidence, confident, ego, intelligence, introvert, extroverted, attitude, characterizes, attraction, ability, courage, demeanor, characterize, temperament, **perfectionism**, **insecurity**, **insecure**, shy
Age and Generational Indicators	youth, teenager, teen, teenage, adolescent, older, generation, age, juvenile, adulthood, younger, adult, youngster, childhood, grandpa, maturity, grandson, grandchildren, **kiddos**, **grandfather**

Table 7: Top 20 words per theme in the High SES group. Bold words indicate unique terms for this SES group.