Impacts of Low Socio-economic Status on Educational Outcomes: A Narrative Based Analysis

Anonymous EMNLP submission

Abstract

Socioeconomic status (SES) is a metric used to compare a person's social standing comparing to others based on their income, level of education, and occupation. Students from low SES backgrounds are those whose parents have low income and have limited access to the resources and opportunities they needed to aid their success. Researchers have studied many different issues and solutions for students with low SES, and there is a lot of research going on in many different fields, especially in the social sciences. Computer science, however, has not yet as a field turned its considerable potential to addressing these inequalities. Utilizing Natural Language Processing (NLP) methods and technology, our work aims to address these disparities and ways to bridge the gap. We built a simple string matching algorithm including Latent Dirichlet Allocation (LDA) model and Open Information Extractor (open IE) to generate relational triples that are connected to the context of the students' challenges, factors contributed to those challenges, and the strategies they follow to overcome them. We manually collected 16 narratives about the experiences of low SES students in higher education from a publicly accessible internet forum (Reddit) and tested our model on them. We demonstrate that our strategy is effective in gathering contextual data about low SES students, in particular, about their difficulties while in an educational institution and how they improve their situation. A detailed error analysis suggests that improvement of the LDA model and quality of triples can help to get better results from our model. For the advantage of other researchers, we make our code available[1].

1 Introduction

An individual's or group's socioeconomic status is defined as their social rank or class based on metrics such as educational attainment, economic status and employment (Saegert et al., 2006). The definition however is not limited to the aforementioned; socioeconomic status can also be linked to factors such as a person's quality of life and privileges that are available to some people in society as opposed to others. There is an obvious inequality when discussing such a topic. Such inequality could manifest itself in the form of disparity in equal distribution of health services (Dickman et al., 2017), unequal educational outcomes (Morgan et al., 2009) find more here, and resource allocation (Aikens and Barbarin, 2008).

Prior work in the social sciences (Terenzini et al., 2001) (Rheinschmidt and Mendoza-Denton, 2014) has repeatedly demonstrated that students of low socioeconomic status, unlike their middle or high SES peers, attain lower levels of education and lack access to opportunities and resources that help them succeed in post-secondary institutions. However, this same abundance of research is not present in Computer Science and related fields such as NLP. There is of course some work that has been done, but most if not all of them incorporate the use of social science based structured data such as surveys, questionnaires, and focus groups to make predictions. For instance, a path based analysis of the educational attainment of low-SES students (Lee et al., 2008) and an analysis of STEM attitudes in low-SES students using descriptive statistics, confirmatory factor analysis, Ordinary Least Squares (OLS) regression, and path analysis (Ball et al., 2019). Their approach was almost purely computational, but the data points they based their work on were surveys.

Although it might seem that way, we are not trying to denigrate work made using structured data points in any way. In fact, structured data, such as questionnaires and surveys, make the act of data analysis straight forward because less time and resources are allocated to extract insights and bring about meaningful results. What we are proposing

[1] Code and data may be downloaded from anonymous link.

here and the motivation for our work is twofold: (1) Address the lack of research in Computer science, specifically NLP, pertaining to educational outcomes as a consequence of an individual's socioeconomic class, and (2) use unstructured narratives from internet forums (in our case Reddit) as a basis for our analysis.

To be more specific, we are identifying common patterns of struggles faced by low SES students in higher education and how those same students attempted to resolve their shortcomings. We ignore identifying the factors of having low SES because our data contain less information about it. As opposed to a close reading based approach which involves subjective analysis of certain each narrative, our whole approach is predicated on distant reading—-gathering generalizable insights and patterns within text in the most objective way possible. We use Genims' LDA model (Řehůřek and Sojka, 2010) to extract generalizable topics within our corpus that involve challenges faced by low-SES students and how these students attempt to remedy these challenges. We also use S-V-O triples extracted by CoreNLP's Open Information Extractor (Manning et al., 2014) to provide the necessary context behind the topic clusters identified by our LDA model.

The paper is organized as follows. We start by describing prior research (§2) on socioeconomic status in relation to educational outcomes in order to describe the motivation for our work. We then describe our corpus (§3) and our methodology (§4) for choosing specific data points. This is then followed by our approach in topic modelling using LDA and Subject-Verb-Object relation extraction. We present the results (§5) and make the limitations (§6) of our work clear, which leads us to discussions of future research. We conclude with our contributions (§7).

2 Related Work

In terms of educational outcomes in the realm of post secondary education, the socioeconomic strata into which an individual grew up in is directly correlated with their final educational and career outcomes (Jackson, 2018). Starting off, research has revealed that prospective college students from low-income families have restricted access to information about college (Brown et al., 2016). This could be information about financial aid, educational resources, and vocational development. On top of that, these same students are more likely to take on higher student loan debts that surpass the of national average (Houle, 2014). The aforementioned inequalities don't even consider the negative impacts that lack of resources and support have on the early literacy (Buckingham et al., 2013), academic achievement (Doerschuk et al., 2016), psychological outcomes (McLaughlin and Sheridan, 2016), and career aspirations (Diemer and Ali, 2009) of low-SES students before they enroll in any higher ed. institutions. When they do enter these institutions, low-SES students report a different sense of belonging (Ahn and Davis, 2020), experience financial stress that impedes their ability to succeed both academically and in social settings (Moore et al., 2021), and attain dissimilar levels of education as compared to their middle or high SES counterparts (Estep, 2016)(link)(link).

Previously mentioned research is also supplemented with multiple reports that address educational outcomes of low-SES students in post-secondary education as a function of their social class. One, for example, is College Board report based on prospective student profiles and survey data by Terenzini et al. (2001). It reports that low-SES students are less likely to complete a four-year degree once on an academic track, and are less likely to pursue further education after a bachelors. They attribute this reason to a list of disadvantages that low-SES students must confront when enrolling in higher education.

Other work has tackled educational outcomes and how they relate with class conditioned beliefs and social-class stereotypes. Rheinschmidt and Mendoza-Denton (2014) conduct 4 studies on students of diverse socio-economic status, and they found evidence that suggests that experimentally primed student beliefs about personal characteristics such as intelligence, effort, and sense of accomplishment predicted academic achievement in a college setting as a function of RS-Class (Rheinschmidt and Mendoza-Denton, 2014). Croizet and Claire 1998 extend the concept of Steele's stereotype threat (Steele and Aronson, 1995) to socioeconomic backgrounds as opposed to just racial and gender groups by the manipulating the instructions of tests administered to students of diverse SES during their study. Their results depict that, when tests instructions described the questions to measure intellectual ability, the performance of low-SES students was lower than their high-SES counter-

parts. The opposite results are observed when the test instructions mentioned nothing of measuring the student's intellectual capability. Thereby confirming that apprehension about confirming negative stereotypes about one's social-class affects academic performance.

Some cross field research that combines the social science and Computer Science also address the challenges and struggles that low-SES students face in higher educational institutions such as universities and 4-year colleges. One body of work, for example, addresses the challenges that underprivileged students, such as those from low-SES, face in integrating into post-secondary institutions even with the higher levels of reported cultural and socio-economic diversity in these institutions (Álvarez-Rivadulla et al., 2022). It uses a mixed method approach which involves an assortativity coefficient and a mean degree constrained model to test for preferential ties associated with attributes within student groups and test if those ties were related to the social class of students. Interviews were conducted to form ethnographic observations about campus culture and diversity. In more simple terms, they used both quantitative and qualitative methods of analysis to understand friendship networks at an elite university in order to determine what factors facilitate and what factors act as barriers to relations within the student body.

There is a limited amount of prior work done on low-SES students in a purely computational manner. Those we manged to find relied on structured data, such as surveys and questionnaires, for their analysis. Lee et al. (2008) , for instance, utilized a path based analysis model in order to investigate the long-term academic progress of students of low-SES. In this study, the ordinal variables acquired from the National Educational Longitudinal Study database were rescaled and linearized using an optimal scaling procedure to then implement a path analysis model.

Another study, done by Ball et al. (2019), applied Expectancy-Value Theory (EVT) on survey data from a predominantly African American student district in southeastern USA in order to investigate the negative attitudes that students have toward STEM fields. Their analytical approach consisted of descriptive statistics (to gain better contextual understanding of data), confirmatory factor analysis (to confirm the independent variables' component structure within the data), Ordinary Least Squares (OLS) regression (to predict the potential of the EVT model and emotional cost variables), and path analysis (to understand the effects of the EVT constructs and emotional cost variables). Another study by Titus (2006) uses hierarchical generalized linear modelling (HGLM) to analyze variables in national survey data in order to understand the influence of institutional spending and revenue on college completion rates of low-SES students.

3 Data

Since it can be difficult to obtain sensitive and private information from students directly, surveys and interviews are time-consuming and challenging methods for gathering the data for our project. So, we decided to use an online forum to start our preliminary work. We chose the forum (cite reddit.com) since its users can express themselves freely and most posts are anonymous. We collected 30 narratives written by low SES students who discuss their monetary and familial challenges. For instance, some of the students discuss how they were raised without parental guidance, in abusive homes, with drug addictions, and without adequate financial support. They explain how these circumstances had a negative impact on their academic performance because they were forced to turn to working night shifts or two jobs to make ends meet, among other means of supporting their education. We then filtered out the less important narratives, where students do not discuss their challenges and strategies to overcome them as much; we only selected those in which the students primarily discussed their experiences as students from low SES backgrounds. The final number of the stories ended up at 16, and each one has an average of 15 sentences. We updated the narratives by removing symbols and personal identifying information (PII) before running our model on them. We decided not to disclose our data in order to maintain confidentiality of the narrators.

4 Approach

Our approach is based on this rationale: "If low SES students documented their post secondary education experience in these narratives, then it is safe to assume that they mentioned their struggles, what factors contributed to those struggles, and how those issues were resolved". Based on this rationale, we divided our approach into three parts, LDA Topic modelling, Subject-Verb-Obejct

(S-V-O) triple extraction, and String Matching between the topics and triples. With Topic modeling, we were able to identify common struggles within the low SES student community, factors such as poverty and lack of networking that contribute to such struggles, and solutions suggested within these stories that worked to alleviate these problems. S-V-O triples helped provide the necessary context behind the conclusions made by the LDA model. The relevance of data points between the S-V-O triples and topic clusters produced by the LDA model were addressed by string matching.

We first trained and optimized a Genism LDA Model on a pre-processed instance of the corpus to obtain relevant topics with improved coherence scores. Simultaneously, we used CoreNLP's Open Information Extractor to obtain S-V-O relation triples from the raw texts of our corpus. Then, we extracted the relevant S-V-O triples by string matching between the topics and triples.

4.1 Topic Modelling

We divided our LDA model implementation into three parts: (1) Pre-processing, (2) Topic Modelling, and (3) Model Optimization and Tuning.

Pre-processing: Besides training and tuning our model, we spent enough time on preparing the data and optimizing our pre-processing techniques. We emphasized on this step because our corpus was sampled from an internet forum, and therefore, it contained more colloquialisms and contractions than text sampled from a formal source. We implemented the data pre-processing as follows.

- **Tokenization and lemmatization:** To tokenize our initial corpus, we used *en_core_web_sm* from spaCy (make bib file for spaCy citation) to produce a doc object with filtered parts of speech, remove inflectional endings, and return the lemma of words; we kept the nouns, adjectives, verbs, and adverbs—the parser and name entity recognizer were not used. We considered Gensim's `simple_preprocess()`[2] to discard tokens that are either too long or too short (mention parameters), removed accent marks from all tokens, and once again removed stop words and short tokens after lemmatization was complete.

- **N-gram implementation:** For our implementation of N-grams we decided that Bigrams and Trigrams would be best based on previous trails. The two aforementioned N-grams were implemented using Genism's *model.phrases.Phrases* which we found to work best on our data as opposed to manually creating an N-gram function or using NLTK's *ngrams*.[3] We decided to set the parameters to low values because larger values failed to extract important N-grams from our limited data points. The N-gram implementation did not work very well on our data. The corpus used to train this model is a list of numerical bags of words containing 869 items (words) with their respective frequencies. Due to the highly informal and verbose nature of the language in our corpus, our demo algorithm prioritized words that occurred quite frequently yet contributed quite little to desired topics. Therefore, we decided to use *tf-idf* as a weighting factor in order to filter words in our corpus based on their relevance.

- ***Tf-idf:*** Our *tf-idf* model is implemented using the Gensim *tf-idf* module. We modified the input parameters for our data and experimented with different "low values" to determine the best fit—other parameters were left at default. We used the same bag-of-words we considered for our demo model as a corpus for our *tf-idf* model. Our *tf-idf* model checks for words that occur with an 'X' threshold (our low value); if a certain word within our corpus occurs with a certain frequency that lands it a *tf-idf* score below our low value X, then the algorithm will assume that it is so ubiquitous that it doesn't provide much value to our LDA model. The output from *tf-idf* model is then a numerical list of bag of words, which does not include words with scores below our threshold and words with zero scores. This output is then used to train the LDA model. However, we are aware of certain limitations of *tf-idf* in term weighing: lack of built-in lemmatization and semantic analysis, and inconsistent results when classifying non-uniform text.(Ramos et al., 2003; Fan and Qin, 2018/05) This will be further dis-

[2] *simple_preprocess* parameters were set to *deacc = True* and *min_len = 3*

[3] *model.phrases.Phrases'* parameters were set to *min_count = 2* (only for bigrams), *threshold = 10* (for bigrams) and *2* (for trigrams). The rest were left at default.

Table 1: A few selected topics generated by LDA Model

Topic 1	Topic 2	Topic 5	Topic 7
lot	feel	work	school
grow	well	job	friend
also	year	school	feel
poor	school	year	make
well	know	graduate	other
company	most	first	connect
good	push	well	never
career	mom	family	work
industry	only	hard	change
do	student	get	tool

cussed in our Limitations and Future Works section.

LDA Modelling: We decided to choose Gensim's LDA model for topic modelling because it did not require data labeling, which we did not have the resources for, and it fits within our time constraints. The model was trained with parameters set to *num_topic = 10*, *chunksize = 2000*, *passes = 20*, *iterations = 400*, and *eval_every = 0*. Besides the input parameters, the rest were either set to *'auto'* or left at default.

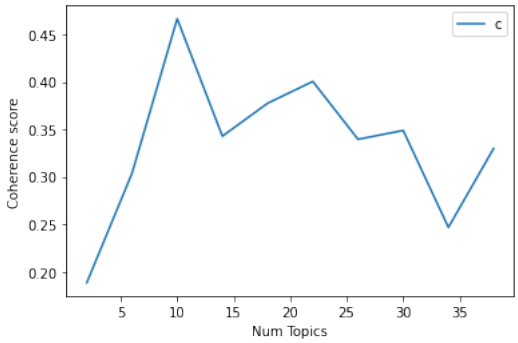
Table 1 presents the top ten terms for four selected topics after the model has been trained. Formally, the terms listed under the same topic in LDA Modelling are quite similar, and we observe the same trend in our model. For instance, Topic 1 seems to be about growing up poor and yearning for a good career in some industry and Topic 7 is about making connections with others at work and school. When using topic coherence to evaluate the semantic similarity between the top 10 words in the topics, our model had a score of 0.44. We used this score as a baseline for optimizing our model in 'Topic Modelling'.

Model Optimization and Tuning : We have developed two different models. Our first model only used Gensim's inbuilt version of the LDA algorithm. There is an LDA model that is more precise and produces better coherence scores but we decided not to use it due to some technical issues.

Instead of our initial optimization approach, we decided to tune the parameters to get better coherence scores for our second model. The two parameters we optimized for were `eval_every` (for minimizing log perplexity), and `num_topics`.

Minimizing Perplexity:When minimizing the perplexity score, we noticed that increasing the parameter by just one factor, increased the training time by 2X and made it impractical to pursue. However, we found that setting `eval_every = 1` substantially improved the generation performance of the model (Blei et al., 2003). Therefore, we decided that the value '1' for `eval_every` would be a good performance and output quality compromise.

Optimal number of topics: To find the optimal number of topics, We generated multiple LDA models with varied number of topics 'n' and chose the one with the highest coherence score to identify the ideal number of topics. This approach was adopted from Prabhakaran's article titled *Topic Modeling with Gensim (Python)* (pra). As in Prabhakaran's approach, we used the function `compute_coherence_values` that trains multiple models and returns the models with their respective coherence scores. Contrary to their approach, we decided against using LDA Mallet for the reasons mentioned above. We also modified the parameters to match our previous model with the modified `eval_every` value, and all other parameters were left at default. [4]



The number of topics 'n' marked at the peak offers the best results, in our case this was 10 topics with a coherence score of 0.47. Coincidentally, this is the same number of topics we picked for our unoptimized model by trial and error. As documented by Prabhakaran, picking a higher 'n' value could provide deeper insights with detailed subtopics, but that wasn't the case for us as the trend tends to drop off as shown in the line graph above. We think this is because of the small number of data points we used to train our model.

Comparing the topics generated by our topic-number optimized model to our previous model, the coherence score improved by 6.38%. The dif-

[4] *compute_coherence_values* parameters were set to *start=2, limit=40, step=4, chunksize = 2000, passes = 20, iterations = 400*, and $eval_every = 1$

ference in coherence scores might not be as unsubstantial, but the terms produced by each model within a specific topic are quite different.

4.2 S-V-O Triple Extraction

We used Stanford CoreNLP Open Information Extraction tool to get subject-verb-object (SVO) relation triples from each narrative. Stanford CoreNLP has a tendency to produce repetitive triples, therefore, we filtered the triples using the SpaCy library (cite it).

Triples extraction with CoreNLP: To get the S-V-O triples from our data, we annotated the content of the story line by line using the `client.annotate(line)` function of OpenIE. We then used the `line['Subject'] + line['Relation'] + line['Object']` feature to get the triples of each sentence as a string.

Triples filtering with SpaCy: To remove the repetitive triples that we received from our coreNLP model, we lemmatized the triples and removed the stop words, and then compared pairs of all the triples to check their similarity using the Cosine similarity feature of SpaCy. If the similarity score exceeds 0.8, the pair is added to a list of similar pairs. Then we addressed the index of the first triple in the pair and removed it. We repeated this process using recursion until there are no duplicate triples left.

4.3 String Matching

After we got the relevant topics out of the narratives using the LDA model, we decided to only keep the triples that are correlated to these topics. The topics were listed as ten different lists in a text file, and we also had a similarity text file that tells us which list a story is closely related to. We had two options: the first option was to compare each triple of a story to its related list of topics. The other option was to compare each triple with all the lists of topics. When we tried comparing the triple with the story's related list of topics, the results were not satisfactory opposed to comparing the triple with all of the topics and looking for a match. The next step was to iterate through all of the triples and check if any word in a triple matches with a topic in the list. If we find a match, we store the triple in a text file, with all of the other triples of that specific story that matched with a topic as well.

5 Results and Discussion

We filtered out the S-V-O triples from the original triples which do not contain any factors of having low SES, struggles of the students, or strategies to overcome the struggles. Then we compared the triples we get from our model with this filtered triples and drew some conclusions. The detailed results are shown in Table 2 and a sample output from one of the narratives is shown in Table 3.

We showed the results of two different models, one with the coherence score of 0.44 and the one is with 0.46. We expected to get better results from the second one but it turned out our first model outperformed the second. As our corpus contains only 16 narratives, the generated triples from the narratives are less in number. Therefore, with a high coherence score, our model extracted generalized topics which were not much helpful to filter contextual triples from the narratives compared to the first one. Also, due to the less number of data we couldn't be able to increase the number of the topics generated.

If we look at Table 2, we see that for the most of the narratives, the matched triples are higher, more than 50%. The highest matched triples we found for narrative 6 which is 80% and the lowest is for narrative 3 which is 37.5%. On the other hand, the number of missed triples is also lower for this model, lowest is 20% for narrative 6 and the highest if 62.5% for the narrative 7. Although the number of missed triples is lower for the first model compared to the second one, the number of the additional triples is higher, 86 in total for the 16 narratives. We notice that this model extracts more triples than the second one, and that's why we get more informative triples as well as more additional triples than the other model.

On a contrary, we notice that there are more missed triples than matched triples. The lowest matched triples are for story 8, which had a percentage of 20%. And the story with the highest missed triples is story 8, with a percentage of 80% missed triples. This model produces less additional triples compared to the first model which 77 in total.

If we look at the sample output of our model in Table 3, we see that our model successfully generated the triples that contain common struggles of a student with low SES, for examples, having an alcoholic mother, coming from a low income family, and running out of money. Besides, some of the

triples provided information of how that student improved his socio-economic status, for examples, saving money, working full time, applying for jobs.

6 Error Analysis and Future Work

Error analysis of the results found some issues and limitations of within our methodology. These were based on limitations of the tools and the quantity of the data we utilized in our approach.

Preprocessing Limitations: To begin with, there are obvious limitations with our preprocessing techniques that ought to be addressed, particularly with the *tf-idf* algorithm. The most obvious constraint of *tf-idf* is the lack of built-in lemmatization features within it. This makes it hard for the algorithm to identify slight changes, tense or pluralization, within a word and makes the algorithm's output somewhat inconsistent (Ramos et al., 2003). These limitations, however, were easily overcame by lemmatizing and removing stop words from our corpus before running the over it. Additionally, this helps combat the dimensionality problem with a *Bag-of-Words* (BOW) approach.

Another limitation of our model is that, it doesn't capture semantic relationships between words and is also unable to check for co-occurence of words, given that it is based on a BOW model. To improve the performance of our *tf-idf* model in future iterations of our work, we plan to implement modified *tf-idf* weighing schemes used in text classification such as Decision Trees, Rule-based classifiers, Support Vector Machine (SVO) classifiers and Neural Network Classifiers (Kumar et al., 2015). Also, Dai's work reveals the limitation of a classic *tf-idf* approach when dealing with non-uniform text. We attempt to address this in our future work by using relative frequency algorithms (Dai, 2018/05) and incorporating Naïve Byes for improved class relationship classification (Fan and Qin, 2018/05)(Qaiser and Ali, 2018).

Finally, we are also considering using Dynamic Word Embeddings as a replacement for *tf-idf* as a weighting algorithm. This will be dependent on the results we get from modifying our current *tf-idf* model and comparing it to how a language model, such as Google's BERT (Bidirectional Encoder Representations from Transformers), will perform.

Data quality and quantity: -adding more data should give us more results, improve our topic optimization function and help us extract better subtopics. talk about maybe adding more data points–maybe even include stuff about being first-gen since it's a subtopic. (not sure about this part tho). Maybe even talk about our progress for getting data that relates to students getting pell grants.(raise this during the meeting)

(why did why we compared the triples to the whole corpus as opposed to comparing it with just one section)

Topic modelling in LDA: A key limitation of our LDA model is that it assumes that no correlation exists between the words and treats them as independent entities in a corpus. In addition to this, LDA modelling lacks built-in semantic analysis, which negatively affects the coherence score of our models. A good approach to solve this problem would be to use knowledge graphs such as Wikipedia [5] or ConceptNet [6] to link correlated topics with each other. Synonym relationships and name entity recognition could also be helpful to encourage that similar words be categorized in the same topic cluster.

An approach we are interested in implementing was suggested by Xie et al. in their study addressing the limitation of LDA models in detecting word similarities. They attempt to overcome this constraint by implementing a Markov Random Field (MRF) regularized Latent Dirichlet Allocation (LDA) model that incorporates word correlations knowledge within a topic while still providing flexibility for a word to be placed in different topic clusters. Their work addresses the topic relevance questions and importance questions raised in research that attempt to tackle the same word correlation problems of LDA.

Finally, we would also like to address the debate between text classification and LDA topic modelling as a way to make to obtain insights from our corpus.

S-V-O Triples: Although we filtered the triples generated by Stanford CoreNLP to remove the repetitive ones, this toolkit often produces insignificant and less important triples. Using a better Open IE library can result in better triples and better performance for our model.

-And to expand the amount of meaningful triples we get from our model, a possible way would be to use a tool like Nltk Wordnet to get synonyms of the topics we generated from our LDA model.

[5] https://www.wikipedia.org/
[6] https://conceptnet.io/

Narrative	Model 1					Model 2				
	Matched Count	Matched %	Missed Count	Missed %	Addi-tional	Matched Count	Matched %	Missed Count	Missed %	Addi-tional
1	8	61.50	5	38.50	12	5	38.50	8	61.50	6
2	3	75.00	1	25.00	0	3	75.00	1	25.00	3
3	6	50.00	6	50.00	5	3	25.00	9	75.00	3
4	5	41.70	7	58.30	3	4	33.30	8	66.70	2
5	5	62.50	3	37.50	8	4	50.00	4	50.00	6
6	8	80.00	2	20.00	12	7	70.00	3	30.00	10
7	3	37.50	5	62.50	3	3	37.50	5	62.50	3
8	2	40.00	3	60.00	7	1	20.00	4	80.00	6
9	3	60.00	2	40.00	3	4	80.00	1	20.00	4
10	19	52.00	17	47.20	14	13	36.10	23	63.90	10
11	4	50.00	4	50.00	6	4	50.00	4	50.00	7
12	13	40.60	19	59.40	4	13	40.60	19	59.40	3
13	10	76.90	3	23.10	3	7	53.80	6	46.20	5
14	10	71.40	4	28.60	2	7	50.00	7	50.00	2
15	4	66.70	2	33.30	6	4	66.70	2	33.30	5
16	4	57.10	3	42.90	1	4	57.10	3	42.90	2

Table 2: Performance of **Model 1** and **Model 2**. 'Matched' denotes how many triples matched with the originally annotated triples, 'Missed' denotes how many triples did not match with the originals, and 'Additional' denotes how many triples are not present in the original annotated triples, but our model addressed them.

Sample output from Model 1
My mom struggling alcoholic
My mom was unable
My mom help out high school
residence halls was last minute option
I go to college
I come from low income family of substance abusers
it 's headed my freshman year of college
me feel like I did not belong in school
I was working full time trying
My GPA was at time less than 2.3
I work to save
I work for year
my bachelor ran out money
I applied at_time past year with pandemic
my sober mom is in audience
I walking at_time time
you push through anything life

Table 3: The triples obtained from the first version of our model

7 Contribution

This paper makes three contributions. First, we develop a model that can generate relational triples from narratives of the students with low SES; these which are important to get the insights of the life experiences of the students, specifically their struggles and strategies to overcome those struggles. Second, we make a conclusion that we can employ NLP tools and technologies to understand the unstructured narratives of the students from low SES background. Third, we make our code public to the community. Finally, to the best of our knowledge, there is no prior work done in NLP about low SES students, our work will pave the way for other possible NLP research in this area of study.

Acknowledgements

Anonymized for peer review.

References

Topic modeling in python with gensim.

Mi Young Ahn and Howard H. Davis. 2020. Students' sense of belonging and their socio-economic status in higher education: a quantitative approach. *Teaching in Higher Education*, 0(0):1–14.

Nikki L Aikens and Oscar Barbarin. 2008. Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. *Journal of Educational Psychology*, 100:235–251.

María José Álvarez-Rivadulla, Ana María Jaramillo, Felipe Fajardo, Laura Cely, Andrés Molano, and Felipe Montes. 2022. College integration and social class. *Higher Education*, pages 1–23.

Christopher Ball, Kuo-Ting Huang, R V Rikard, and Shelia R Cotten. 2019. The emotional costs of computers: an expectancy-value theory analysis of predominantly low-socioeconomic status minority students' stem attitudes. *Information, Communication Society*, 22:105–128.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Michael Brown, Donghee Wohn, and Nicole Ellison. 2016. Without a map: College access and the online practices of youth from low-income communities. *Computers Education*, 92-93:104–116.

Jennifer Buckingham, Kevin Wheldall, and Robyn Beaman-Wheldall. 2013. Why poor children are more likely to become poor readers: The school years. *Australian Journal of Education*, 57(3):190–213.

Jean-Claude Croizet and Theresa Claire. 1998. Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, 24(6):588–594.

Weisi Dai. 2018/05. Improvement and implementation of feature weighting algorithm tf-idf in text classification. In *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*, pages 583–587. Atlantis Press.

Samuel L Dickman, David U Himmelstein, and Steffie Woolhandler. 2017. Inequality and the health-care system in the usa. *The Lancet*, 389(10077):1431–1441.

Matthew A. Diemer and Saba Rasheed Ali. 2009. Integrating social class into vocational psychology: Theory and practice implications. *Journal of Career Assessment*, 17(3):247–265.

Peggy Doerschuk, Cristian Bahrim, Jennifer Daniel, Joseph Kruger, Judith Mann, and Cristopher Martin. 2016. Closing the gaps and filling the stem pipeline: A multidisciplinary approach. *Journal of Science Education and Technology*, 25(4):682–695.

Tiffany M. Estep. 2016. The graduation gap and socioeconomic status: Using stereotype threat to explain graduation rates.

Huilong Fan and Yongbin Qin. 2018/05. Research on text classification based on improved tf-idf algorithm. In *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*, pages 501–506. Atlantis Press.

Jason N. Houle. 2014. Disparities in debt: Parents' socioeconomic resources and young adult student loan debt. *Sociology of Education*, 87(1):53–69.

C. Kirabo Jackson. 2018. Does school spending matter? the new literature on an old question. Working Paper 25368, National Bureau of Economic Research.

Sandal Kumar, Christopher Columbus, and Research Scholar. 2015. Various improved tfidf schemes for term weighing in text categorization: A survey. *International Journal of Engineering Research*, 10:11905–11910.

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*, 11:2156759X0801100504.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Katie A. McLaughlin and Margaret A. Sheridan. 2016. Beyond cumulative risk: A dimensional approach to childhood adversity. *Current Directions in Psychological Science*, 25(4):239–245. PMID: 27773969.

Andrea Moore, Annie Nguyen, Sabrina Rivas, Ayah Bany-Mohammed, Jarod Majeika, and Lauren Martinez. 2021. A qualitative examination of the impacts of financial stress on college students' well-being: Insights from a large, private institution. *SAGE Open Medicine*, 9:20503121211018122. PMID: 34094560.

Paul L Morgan, George Farkas, Marianne M Hillemeier, and Steven Maczuga. 2009. Risk factors for learning-related behavior problems at 24 months of age: Population-based estimates. *Journal of abnormal child psychology*, 37(3):401–413.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Michelle L Rheinschmidt and Rodolfo Mendoza-Denton. 2014. Social class and academic achievement in college: The interplay of rejection sensitivity and entity beliefs. *Journal of Personality and Social Psychology*, 107(1):101.

Susan C Saegert, Nancy E Adler, Heather E Bullock, Ana Mari Cauce, William Ming Liu, and Karen F Wyche. 2006. Report of the apa task force on socioeconomic status. *Retrieved from the American Psychological Association website*.

Claude M. Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of african americans. *Journal of Personality and Social Psychology*, 69(5):797–811.

Patrick T. Terenzini, Alberto F. Cabrera, Patrick T. Terenzini, Alberto F. Cabrera, Elena M. Bernal, Patrick T. Terenzini Is Professor, and Senior Researcher. 2001. Swimming against the tide: The poor in american higher education.

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(4):371–398.

Pengtao Xie, Diyi Yang, and Eric Xing. 2015. Incorporating word correlation knowledge into topic modeling. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: human language technologies*, pages 725–734.