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

Motti Kelbessa*
mottikelbessa21@augustana.edu

Ilyas Jamil*
ilyasjamil20@augustana.edu

Labiba Jahan
labibajahan@augustana.edu

Abstract

Socioeconomic status (SES) is a metric used to compare a person's social standing 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 need to aid their success. Researchers have studied many issues and solutions for students with low SES, and there is a lot of research going on in many 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) topic model and Open Information Extraction (open IE) to generate relational triples that are connected to the context of the students' 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 (from 37.50% to 80%) in gathering contextual data about low SES students, in particular, about their difficulties while in a higher educational institution and how they improve their situation. A detailed error analysis suggests that increase of data, improvement of the LDA model, and quality of triples can help 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 the privileges that are available to some people in society as opposed to others. When discussing such topics, there is an obvious inequality that has to be called out. 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), resource allocation (Aikens and Barbarin, 2008) and many more.

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 approaches were almost purely computational, but the data points they based their work on were surveys—structured data.

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 re-

*These authors contributed equally to this work.

[1]Code may be downloaded from https://github.com/MoRevolution/Low-SES_NLP

Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 59 - 69

sources are allocated to extract insights and bring about meaningful results. On the other hand, setting up surveys and interviews takes time, and the volume of data is always an issue. So, the motivation of 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 socio-economic 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. 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. We then use Subject-Verb-Obejct (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. For each narrative, our model produces a set of S-V-O triples that reflect the challenges of the student and solutions to them. These triples are helpful for summarizing the content of the corpus, for knowledge graph construction, in question answering systems, and many other functions in addition to providing us with insightful information.

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 S-V-O 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 has a direct correlation 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 of these students (Buckingham et al., 2013), their 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 educational 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).

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 statuses, 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 class-based reaction sensitivity (Rheinschmidt and Mendoza-Denton, 2014). Croizet and Claire 1998 extend the concept of Steele's stereotype threat, the risk of adhering to negative stereotypes about one's group (Steele and

Aronson, 1995), to socio-economic backgrounds as opposed to just racial and gender groups by the manipulating the instructions of tests administered to students of diverse SES in their study.

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.

There is 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. To the best of our knowledge, there is no prior work done on low SES students in the field of NLP.

3 Data

As mentioned prior, we demonstrate our approach on unstructured social media data from the internet forum page Reddit [2]. We were motivated to use social media data for our preliminary work because of two broad reasons: (1) the time and human resource constraints that we were working with, and (2) the scarcity of computational research that used unstructured data points. Since our topic entails the collection of sensitive and private information from students or alumni, either directly or indirectly, we anticipated that surveys and interviews would be time-consuming and challenging methods for gathering data for our research. With such constraints in mind, we decided to use Reddit as our preliminary source of unstructured data narratives because its users are able to express themselves in a relatively unimpeded manner, and it provided narratives that fit our qualifications best when compared to other online-forums and social media sites. In addition, the format of narratives we collected from Reddit were written in prose; this is of high importance to us since the approach we applied in our preliminary study could, with slight modification and improvement, be used for the next iteration of our work.

In the process of data gathering on social media sites and online forums, our qualification for a "good data point" were as follows:(1) the narratives should have the experience of being from a low-SES student and attending higher education as a focus; (2) the narratives should be about the struggles those students faced higher educational institutions and/or how they overcome those struggles, meaning no general commentary or advice; and (3) the narratives should at least be a paragraph long (150 words).

When looking for data, we found that Reddit provided the most data points that fit our criteria. Here are some Subreddits that we chose to gather our data points from: r/AskReddit, r/college, r/collegeadvice, r/science, r/psychology, r/socialwork, and r/personalfinance. At this stage of our research, we chose to manually search for posts and comments using a list of manually curated keywords that was inspired by terms from our related work section. Some keywords we used are:

[2] https://www.reddit.com/

"can't pay for school", "imposter syndrome", "college culture shock", "struggled growing up", "broken family", and "first-gen in college". We, however, came up with additional terms while searching.

We collected 30 narratives written by low SES students who discuss their monetary and familial challenges. For instance, some 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 less relevant narratives which didn't adequately discuss the challenges faced by these low-SES students. We believe the narratives we chose represent the experiences of low SES students because the students discuss how low their household income is and how they were attempting to improve their circumstances.

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. Besides, we are aware that making our data public will make it difficult to secure the narrators' ability to edit or remove their narratives.

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 modeling, S-V-O triple extraction, and String Matching between the topic clusters 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 Gensim 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 it therefore contained more colloquialisms and contractions than text sampled from a formal source. In addition, some of these preprocessing techniques help remedy the lack of built-in lemmatization and dimensionality problems in our *tf-idf* algorithm. 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()`[3] to discard tokens that are either too long or too short, 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

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

ngrams.[4] 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 discussed 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: Some topics generated by our first 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

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 the section below.

Model Optimization and Tuning: We have developed two different models. Our first model only used Gensim's inbuilt version of the LDA algorithm that uses Variational Bayes sampling method. Although fast, Variational Bayes Sampling method falls short in terms of precision, especially when compared to the LDA Mallet's Gibbs Sampling. Initially, our goal was to replace our first LDA model with the LDA Mallet model. However, we decided against replacing our model for two technical reasons: (1)Third party wrappers in Genism, which LDA Mallet was one of, were removed in the Gensim 4.0 release, and we fear that rolling back to older versions could introduce performance problems; and (2)The LDA Mallet model retains the mallet path and prefix path of the exact system it was trained on which makes it practically hard for us to test the model on different a system that the

[4] *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.

model wasn't initially trained on. [5][6]

Instead of our initial optimization approach of replacing our Gensim LDA model with LDA Mallet, we decided to tune the parameters to get better coherence scores. The two parameters we optimized for were `eval_every` (for minimizing log perplexity), and `num_topics` (to improve coherence scores while acquiring more subtopics).

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 generalization 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)* (Prabhakaran, 2018). 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. [7]

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 belive this is because of the small number of data points we used to train our model.

Comparing the topics generated by our topic-

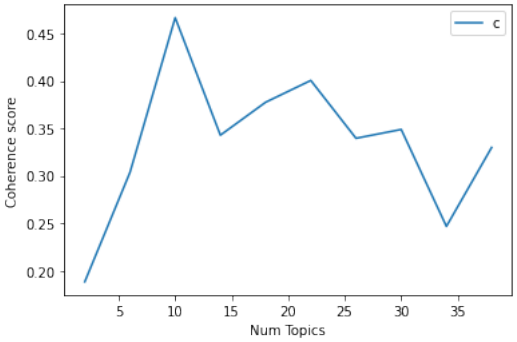

Figure 1: Coherence Score versus Number of Topics

number optimized model to our previous model, the coherence score improved by 6.38%. The difference in coherence scores might not be as substantial, but the terms produced by each model within a specific topic cluster are quite different: not only in terms of shared words within a topic cluster, but also in terms of how meaningful the terms in the topic cluster were to our corpus. This will be explored more in the results section.

4.2 S-V-O Triple Extraction

We used Stanford CoreNLP Open Information Extraction tool to get S-V-O relation triples from each narrative. Stanford CoreNLP has a tendency to produce repetitive triples, therefore, we filtered the triples using the SpaCy library (Honnibal et al., 2020).

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.

[5] https://groups.google.com/g/gensim/c/vVO0_t9jRUo/m/ZYFdq9_TBgAJ

[6] https://groups.google.com/g/gensim/c/_VO4otCV6cU?pli=1

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

4.3 String Matching

Finding the triples that best capture the context of the student's difficulties and their solutions is our ultimate objective, and finding the relevant topics is the first step in accomplishing it. However, when we examine the topics, we see that the majority of them represent the contexts we are interested in. We think this is as a result of the small size of our corpus and the little number of topics produced. Besides, the coherence score of the LDA model is low, and therefore, all of the terms in a specific topic are not much related to each other. So, if we do not consider a specific topic, we increase the chances of excluding related information. Thus, we decided to consider all of the topics and compare them with the triples extracted from the narratives. We repeatedly went through each triple, looking for any term that matched a topic on the list. If a match is found, the triple is taken into account for inclusion in the output list.

5 Results and Discussion

In order to evaluate our model, we first removed the S-V-O triples that did not include any elements related to low-SES, the issues that these students face, or solutions to those issues. Then, we made some inferences by comparing the triples we obtained from our model with these filtered triples. The detailed results are shown in Table 2 and a sample output is shown in Table 3 generated by our model from one of the narratives.

We showed the results of two different models, one with a coherence score of 0.44 and the other 0.46. We expected to get better results from the second model, but it turns out that 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 very helpful to filter contextual triples from the narratives compared to the first one. Additionally, we weren't able to generate more useful topics without compromising the relevance of topic clusters because of the small number of data that our model was trained on.

If we look at Table 2, we see that in Model 1, the matched triples are higher, more than 50% for the most of the narratives. 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 when compared to the second, the number of additional triples here are higher, 86 in total for the 16 narratives. We notice that the first model extracts more triples than the second one; this is why we get more informative triples as well as more additional triples than the other model.

Additionally, we notice that there are more missed triples than matched triples in Model 2. 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 triples provided information of how that student improved his socio-economic status, for example, saving money, working full time, and 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.

6.1 Data quality and quantity

We believe that the biggest constraint within our present work is the quantity of narratives we used as data points for our model. As mentioned in the (§3) section, we found it difficult to manually search for narratives that qualify as valid data points in our research: we only had 16 data points to train our models on. Many narratives we initially found were either too short or strayed towards being informational posts instead of topically relevant narratives. We believe the small quantity of data points contributed negatively to the generalizability of our LDA model.

Admittedly, all of our data hunting methods were manual and were therefore subject to human biases, were inefficient, and time consuming. We chose to manually search Reddit instead of using a Web

Narrative	Model 1					Model 2				
	Matched Count	Matched %	Missed Count	Missed %	Additional	Matched Count	Matched %	Missed Count	Missed %	Additional
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
Average	6.7	57.7	5.4	42.3	5.6	5.4	49.0	6.7	51.0	4.8

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

Scraping tools, such as Selenium [8] or Scrapy [9], for two main reasons: (1) since narratives are unstructured in nature, we lacked data samples that we could use as references for our filtering parameters during web scraping; and (2) even with the use of general keywords as filtering parameters, we don't have enough people on our team to go through and check the qualifications and relevance of the narratives presented to us by the scraping tool.

We now believe, however, that the results of our primary work, after addressing some limitations in our current approach, could provide us with samples or keywords that we could use to automate our data collection methods. We also intend on using the Pushshift Reddit API[10] as a tool to search for Reddit posts and comments, because it offers more search and filter features as compared to Reddit's search bar. As mentioned before, a major reason we did not automate our data collection process was because of the problem of relevance, "How appropriate are the narratives for our kind of work?". Sure, a web scraping bot could find posts and comments with keywords that pertain to low-SES students, but the posts and comments it finds might not be as useful to us. To address this problem, we propose using an LDA modelling as an additional filtering layer that we could use for managing the relevance problem.

6.2 Topic Modelling

6.2.1 Pre-Processing 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 that it does not capture semantic relationships between words and is also unable to check for co-occurence of words,

[8] https://www.selenium.dev/

[9] https://scrapy.org/

[10] https://github.com/pushshift/api

given that it is based on a Bag of word 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 (SVM) 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).

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.

6.2.2 LDA Modelling

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 [11] or ConceptNet [12] 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 vs LDA topic modelling as a way to obtain insights from our corpus. In essence, this is almost an argument between supervised versus unsupervised learning as our approach. Without getting into the weeds of this debate, we chose an unsupervised approach for the following reason:

- Unsupervised learning is much less resource intensive as compared to a supervised approach. Due to the lack of personnel on our team to label each of the data points in the corpus, a less resource intensive approach in unsupervised learning seemed the most appropriate—especially once we obtain more data points to train our topic model.

6.3 S-V-O Triples

Although we filtered the repetitive triples generated by Stanford CoreNLP, Stanford CoreNLP often produces insignificant and less important triples. We believe that 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 WordNet (Fellbaum et al., 1998) to get synonyms of the topics we generated from our LDA model.

7 Contribution

This paper makes four contributions. First, we develop a model that can generate relational triples from narratives of the students with low SES; 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

This work was supported by two funding sources from Augustana college, 'New Faculty Research Award' and 'Larry P. Jones Faculty Fellowship Grant'.

[11] https://www.wikipedia.org/
[12] https://conceptnet.io/

References

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.

Christiane Fellbaum et al. 1998. Wordnet: An electronic lexical database mit press. *Cambridge, Massachusetts*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

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.

Selva Prabhakaran. 2018. Topic modeling in python with gensim.

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.