ID: 24141153	Name: Abrar Sami Khan Alif	CSE424 Section: 01

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

Link: https://aclanthology.org/2022.nlp4pi-1.6/

Summary

Hypothesis

The study's goal is to investigate the application of natural language processing (NLP) tools, specifically S-V-O triple extraction and topic modeling, to evaluate narratives from students of low socioeconomic levels. The premise is that deriving meaningful connections from these tales will provide insights into the issues these students face and the solutions they implement.

Contribution

The paper presents four key contributions:

1. Introduces a novel method to extract S-V-O (Subject-Verb-Object) triples from unstructured narratives, focusing on the challenges and coping strategies of low socio-economic status (SES) students.
2. Provides a new approach to gaining structured insights from narrative data, making it easier to analyze unstructured experiences.
3. Demonstrates the effective use of NLP tools, such as Stanford CoreNLP and SpaCy, to analyze the experiences of marginalized student groups, offering a structured understanding of their struggles.
4. Publishes the code publicly, promoting further research and collaboration in the underexplored domain of NLP for low-SES student experiences.
5. Lays the groundwork for future research, applying NLP to study the life experiences of disadvantaged students—an unexplored topic in the educational and socio-economic research context.

Methodology

1. **S-V-O Triple Extraction:** To extract triples, Stanford CoreNLP's OpenIE tool was used, with redundant triples filtered out using SpaCy's lemmatization and cosine similarity comparison.
2. **String Matching and Topic Modeling:** To identify important contexts, extracted triples were matched with LDA model-generated topics. Two models were evaluated with coherence scores of 0.44 and 0.46, and the former outperformed the latter.
3. **Data Collection:** Manually obtaining narratives from Reddit. Despite a small dataset of 16 narratives, the authors discovered and characterized student challenges.

Conclusion

The study concluded that the model successfully identified relevant triples that reflect the experiences of low-income pupils. However, the authors noted limits in data quantity, topic modeling, and triple extraction. Despite these limitations, the technique shows promise for extracting insights from unstructured narrative data.

.Critiques or Limitations

1. **Small Dataset**: The study relied on only 16 narratives, limiting the generalizability of the findings. The manual search for relevant narratives was time-consuming and may have introduced bias.
2. **Limitations of Topic Modeling**: The LDA model does not account for word correlations, affecting coherence. A higher coherence score did not yield better results due to the small dataset.
3. **Triple Extraction Issues**: Stanford CoreNLP produced redundant and insignificant triples. Although filtering was applied, the model still missed some important relations.

Synthesis

1. Incorporating dynamic word embeddings like BERT or integrating WordNet-based semantic enrichment into the Open Information Extraction (OpenIE) pipeline to improve triple extraction accuracy and topic relevance, addressing the limitations of the LDA model.
2. Automate data collection using Pushshift Reddit API and apply semi-supervised learning techniques to identify relevant narratives from Reddit, enabling larger datasets, improved generalizability, and improved model performance.