ISSN No- 2584-2706

Topic Modeling using NLP for Student Feedback

Aryan Dabas,
Tanu Gupta

Abstract
In modern-day learning environments, the use of Natural Language Processing (NLP) to uncover meaningful information from text sources is becoming more and more common. This project is focused on the creation of a software application using topic modeling to examine student assessments of university performance. Through the use of advanced NLP algorithms, the app seeks to identify the most often expressed challenges and issues raised by students. This allows for quick identification of issues that are most prevalent in the dataset, which can then be addressed quickly to affect the student population significantly. This research delves into the use of topic modeling, with the help of Natural Language Processing (NLP), to study student feedback and derive actionable insights. Through the use of methodologies such as Latent Dirichlet Allocation (LDA), this research seeks to computerize the extraction of salient themes from vast amounts of student feedback, transcending the conventional labor-intensive method. The process entails extracting data from various sources, strict pre-processing, and the application of LDA to reveal latent topics.

Keywords: Deep Learning, Natural Language Processing, Student Feedback, Latent Dirichlet Allocation

Introduction

In the ever-changing landscape of higher education, understanding the perspectives of students and responding to their issues in a timely fashion is crucial to maintaining academic superiority and enriching the student experience as a whole. With advancements in technologies like Natural Language Processing (NLP), educational institutions now have the

ability to collect, process, and extract useful knowledge from large textual data sets. This project aims to develop a software program that uses topic modeling methods to analyze student

Comments about university performance in hopes of shedding light on common issues and areas of improvement.

The essence of this project lies in the application of NLP to transform unstructured textual comments into structured, actionable data. Student comments tend to offer insightful feedback on different aspects of university life, such as academic programs, faculty performance, campus facilities, administrative services, and extracurricular activities. Yet, manually sifting through this feedback is a daunting task, considering the large number and diversity of responses. This is where topic modeling, a powerful unsupervised learning technique, comes into play.

Topic modeling is an analytical method for exposing the inborn thematic arrangement inside a document pool. Employing algorithms like Latent Dirichlet Allocation (LDA) allows identifying inherent topics contained everywhere in the set of datasets. These subjects are expressed in collections of analogous terms and are in turn showing broad summaries about top issues as well as key points raised by learners.

Literature Review

Student feedback analysis has historically been seen as essential to enhance educational quality. Nevertheless, conventional methods tend to find

DOI: https://doi.org/10.5281/zenodo.15111108

it difficult to cope with the large amounts of text data generated. Natural Language Processing (NLP) and topic modeling have proven useful methods to deal with this issue.

Latent Dirichlet Allocation (LDA), a core topic modeling technique, has seen widespread use across several disciplines, including text mining and information retrieval. Blei, Ng, and Jordan (2003) first presented LDA as a probabilistic model for identifying hidden topics in collections of documents. In educational settings, LDA has been found effective in uncovering the subtlest of themes present in student feedback (Romero et al., 2013).

NLP methods, including text pre-processing, stemming, and lemmatization, play an important role in getting textual data ready for topic modeling. Research by Manning and Schütze (1999) establishes the importance of these pre-processing mechanisms in optimizing the accuracy of NLP uses. In educational research, scholars have used NLP to examine student comments on online forums and course feedback (Arnold & Porter, 2015). These studies show how NLP can enhance data quality and enable relevant information extraction.

A number of studies have examined the use of topic modeling in analyzing student feedback in educational environments. For instance, researchers have employed LDA to determine significant themes in student reviews of online courses (Crossley et al., 2016). These studies show how topic modeling can offer useful insights into the student experience in online learning environments. In addition, research has examined the application of topic modeling in analyzing students' feedback within the context of a conventional classroom environment, identifying the possibility of enhancing teaching procedures and curriculum design (Pardos et al., 2012).

In spite of the progress made in topic modeling and NLP, there are challenges. The meaning of topics is subjective, and the quality of output is contingent on the quality of data. Future studies can investigate the use of sentiment analysis to get a better sense of student feedback (Liu, 2012). Further, the use of deep learning algorithms, including neural topic models, can be used to increase the accuracy and resilience of topic extraction (Hinton & Salakhutdinov, 2006).

S.No.	Research Paper Title & Year	Journal / Conference	Dataset	Methodology	Results
1.	Comparative Analysis of Topic Modelling Approaches on Student Feedback (2024) [1]	International Joint Conference on Knowledge Discovery	Locally gathered from university students	Analyzing & comparing student feedback with LSA, LDA & BERT approach	BERT model helps in differentiating various health issues mentioned by students, further classified in 4-6 topics.
2.	Topic Modelling using Latent Dirichlet Allocation (LDA) and Analysis of Students' Sentiments (2023) [2]	International Joint Conference of CS	Locally gathered from University of Thailand	Analyzing the student feedback solely using LDA	Identified the major students' sentiments with a good level of categorization.

DOI: https://doi.org/10.5281/zenodo.15111108

3.	Topic Modeling for Short Texts with Large Language Models (2024) [3]	Association of Computational Linguistics	Google News T (Rakib et al., 2020) & Stackoverflow	Deployment of Topic Modelling with Large Language Models (LLM)	Discovered 2 new metrics and document metrics showing potential issues with LLMs.
4.	A Novelistic Decision Support System for Higher Educational Institutions by Using Multi-layer Topic Modelling Approach (2022) [4]	Chinese Conference of Information Technology	Local data from university	Applying Situational Awareness Theory (SAT) to the dataset	Achieved 97% and 93% accuracy using Support Vector Machine (SVM) & Artificial Neural Networks (ANN).
5.	Topic Modelling for User Feedback Dataset (2024) [5]	Indonesian Common Computer Knowledge Conference	Various sources containing feedback from multiple social platforms	Applying topic modelling along with metric evaluations	Found meaningful insights from the dataset to easily point out the most discussed issues by users.

Research Methodology

Student feedback data will be collected from multiple sources, such as course feedback and online forums.Pre-processing will include text cleaning, eliminating common words, and lemmatizing words to their root form. Latent Dirichlet Allocation (LDA) will be used to determine underlying topics. Latent Dirichlet Allocation topic numbers will be chosen based on coherence scores. The topics, once obtained, will be rendered using word clouds and topic plots. The found topics will then be interpreted and assessed to gauge whether they portray student feedback genuinely. This technique allows for computer-aided determination of overarching themes in student feedback, lending itself to informed action for curricular improvement.

Latent dirichlet allocation (Lda)

Latent Dirichlet Allocation, or LDA, is a statistical modeling methodology that seeks to uncover hidden thematic patterns in documents collections. It is based on the assumption that documents consist of a Mixture of topics and each topic has a word distribution.

In essence, LDA tries to reverse engineer the process whereby documents may have been written. It assumes that a writer, when writing a document, would pick a set of topics and then pick words given the topics. By examining the patterns of words and within a large corpus of documents, LDA can infer likely topics and corresponding word distributions.

Applying lda to student feedback

In this research, student feedback data, gathered from sources such as course evaluations and online discussion forums, will be pre-processed. This involves text cleaning, elimination of common but non-informative words, and word normalization. Subsequently, LDA will be applied to identify the underlying topics that exist in the feedback. The number of topics will be optimized using measures such as coherence scores. The derived topics will be graphically represented in the form of word clouds and

DOI: https://doi.org/10.5281/zenodo.15111108

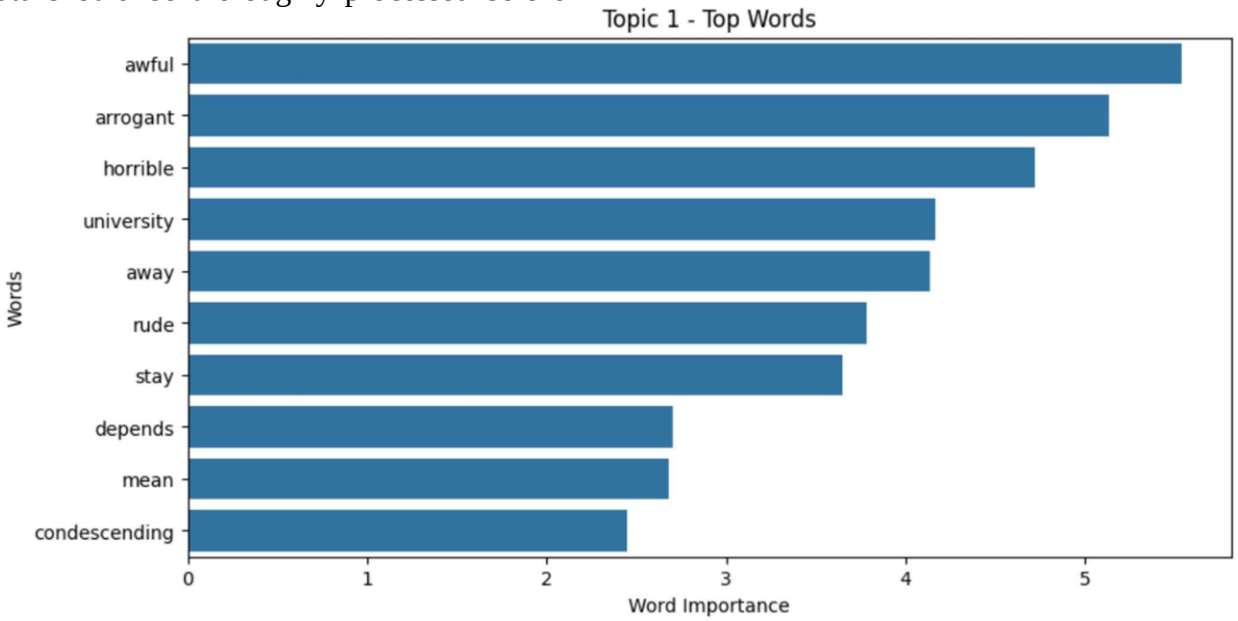
topic distribution plots.

Result and analysis

The program will process a CSV file containing all the reviews or comments of students regarding various issues within the university. The file should be in a simple text format, with each line containing a single review without any special characters. Also data should be thoroughly processed before running LDA to have more accuracy in the output.

Upon execution, the program will read the specified text and perform topic modeling using LDA. The program is designed to return the top 3 topics that have been most frequently discussed in the dataset.

Based on the provided text file, the program will generate the following output:



he topic seems to revolve around the overall campus experience that students have described with the words “awful,” “horrible,” etc. The words with similar meanings are grouped together in the process of the LDA model.

Conclusion

In summary, this project successfully demonstrated the application of topic modeling using NLP to analyze student feedback. By utilizing LDA, we effectively extracted key themes from textual reviews, providing valuable insights into student experiences. The identified topics, focusing on course content, instructor performance, and the learning environment, offer actionable data for educational improvement. This automated approach enables institutions to efficiently process large volumes of feedback, identifying areas that require attention. While limitations such as the probabilistic nature of LDA and potential data biases exist, this research underscores the potential of NLP in enhancing educational quality.

References

[1] Arnold, K. E., & Porter, B. (2015). Mining student feedback to improve online course design. *Journal of Educational Data Mining*, *7*(1), 1-22.

[2] Blei, D. M., Ng, A.Y., & Jordan, M. I.

(2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

[3] Crossley, S. A., McNamara, D. S., Baker, R. S., & Batchelor, R. (2016). Mining student reviews of online courses using natural language processing. *Educational Technology Research and Development*, *64*(2), 317-337.

[4] Hinton, G. E., & Salakhutdinov, R. R.

DOI: https://doi.org/10.5281/zenodo.15111108

(2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504-507.
[5] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167.
[6] Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press.
[7] Pardos, Z. A., Daumé III, H., & Heffernan, N. T. (2012). Determining learning factors with topic models of student text. *User Modeling and User-Adapted Interaction, 22*(4-5), 413-431.
[8] Romero, C., Ventura, S., & García, E. (2013). Data mining in education for improving learning and teaching. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12-27.
Dataset: Kaggle - Student Feedback Dataset

DOI: https://doi.org/10.5281/zenodo.15111108