ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #44639

A Comparative Study of Topic Models for Student Evaluations

Joseph Carpenter Sheils, Marshall University

Joseph C. Sheils is an undergraduate researcher at Marshall University. With a background in statistics, he has conducted research on machine learning, probability theory, and natural language processing.

Dr. David A Dampier, Marshall University

Dr. Dave Dampier is Dean of the College of Engineering and Computer Sciences and Professor in the Department of Computer Sciences and Electrical Engineering at Marshall University. In that position, he serves as the university lead for engineering and computer sciences. He also serves as Director of the Institute for Cyber Security.

Dr. Haroon Malik, Marshall University

Dr. Malik is an Associate Professor at the Department of Computer Sciences and Electrical Engineering, Marshall University, WV, USA.

A Comparative Study of Topic Models for Student Evaluations

Joseph C. Sheils, Haroon Malik, and David A. Dampier
Department of Computer Sciences and Electrical Engineering
Marshall University
Huntington, WV 25703
Email: sheils9@marshall.edu, malikh@marshall.edu, dampierd@marshall.edu

Abstract

Deciphering valuable insights from unstructured written comments in student evaluations within higher education poses a significant challenge, calling for advanced analytical tools. Previous research has underscored the importance of employing topic modeling as a vital tool for unraveling complexities in large volumes of unstructured textual data and extracting dominant topics. While Latent Dirichlet Allocation (LDA) has been a longstanding choice in past studies, the emergence of new and modern topic modeling techniques demands a comparative analysis of their performance against LDA when used with student evaluations. This paper evaluates four topic modeling methods — Latent Dirichlet Allocation (LDA), Nonnegative Matrix Factorization (NMF), BERTopic, and Top2Vec — based on two key performance metrics: topic coherence and topic diversity. Using data from Rate My Professors, our assessment overwhelmingly endorses BERTopic, showcasing its superiority in producing the most informative topics, unmatched ease-of-use, and overall functional excellence for extracting valuable insights from written comments in student evaluations.

1. Introduction

Essential to the mission of higher education institutions is the ongoing evaluation of perceived educational quality and campus experiences. This evaluation is commonly conducted through the Student Evaluations of Teaching (SET) process at the conclusion of each academic semester. The SET process encompasses two distinct forms of student feedback: (a) quantitative Likert-scale ratings obtained through a structured questionnaire and (b) open-ended textual responses.

The resulting SET reports carry considerable weight, often serving as mandatory components in faculty applications for tenure and promotion. Despite their significance, these reports' data are typically kept private, leading students to turn to anonymous online platforms for disseminating evaluations of teachers and schools. Among these platforms, the Rate My Professors (RMP) website has evolved into a vast repository of student feedback data.[1] As of September 2023, the website boasted 23,841,831 teacher evaluations and 372,973 school evaluations, encompassing feedback for 1,997,515 teachers and 9,155 schools across the United States, Canada, and the United Kingdom.

To harness the wealth of information embedded in such text corpora, Natural Language Processing (NLP) techniques emerge as powerful tools. Researchers have extensively applied NLP techniques, particularly topic modeling, to conduct comprehensive studies on comments from various sources, including private SET collections, MOOC discussion forums, and the RMP platform.[2-4] Additionally, sentiment analysis, another prominent NLP technique, has been

employed to discern the emotional polarity of students towards their instructors.[5] Frequently, these two approaches are employed together to simultaneously extract insights from student feedback.[6]

Despite the established efficacy of topic modeling, especially in mining central topics and tracking discussions within student evaluations, prevailing studies predominantly lean on one of the oldest techniques—Latent Dirichlet Allocation (LDA). While LDA remains widely adopted, realizing optimal results with this unsupervised machine-learning technique necessitates meticulous preprocessing and hyperparameter tuning. Consequently, there exists a compelling imperative to explore and evaluate the performance of recently developed topic models that harness advancements in the field.

Towards this end, the paper makes the following contributions:

(a) **Comprehensive Data Sets** —— The paper stands as a pioneering effort by systematically crawling and mining the entire corpus of data from the crowd-sourced student evaluation site Rate My Professors (RMP) over a decade, spanning from 2010 to 2023. This extensive dataset comprises 23,841,831 evaluations of 1,997,515 teachers and 372,973 evaluations of 9,155 schools, encompassing both teacher and school-related information.

(b) **Comparison of Topic Modeling Techniques** —— The paper advances the field by constructing topic models on the RMP data using not only LDA, but also three recent and highly performing techniques: Nonnegative Matrix Factorization (NMF), BERTopic, and Top2Vec. This comparative analysis sheds light on their respective performances in extracting meaningful insights from the student evaluation data.

(c) **Open and Accessible Experimentation** —— The paper enhances research transparency and collaboration by making the experimental setup, data, and results publicly available. This commitment to openness aims to facilitate result reproducibility, enable the replication of studies by other researchers for comparison, and foster further research initiatives.

2. Methodology

Figure 1 presents a high-level overview of our methodology for creating diverse topic models, employing the techniques outlined in Section 1, and subsequently comparing them. The following section explains the steps comprising our methodology and provides a concise overview of the topic models selected for the comparative analysis.

2.1. Data Collection

In this study, we use a large corpus of data scraped from the Rate My Professors website using the 'Requests' package in JSON-format objects. For student evaluations of schools, 372,973 records were collected for 9,155 schools, posted from 2010 to 2023. The ability to rate teachers on RMP has existed since 1999, resulting in a much larger collection of 23,841,831 records. To avoid substantial computational expense during topic model training, a random sample of 1,000,000 records is selected for topic model comparison.

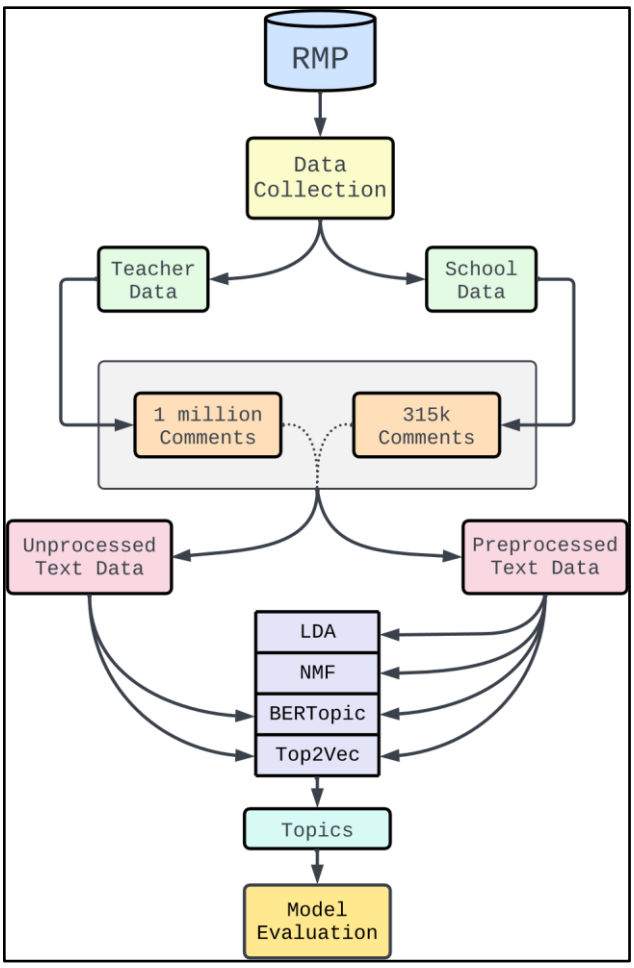

Figure 1: A high-level overview of the methodology for topic model comparison.

2.2. Pre-processing

Before topic modeling, the data goes through a comprehensive pre-processing phase, which involves the removal of non-descriptive comments (such as “No comment.” or “None.”), removal of punctuation and stop-words, lemmatization, and removal comments with less than 5 words. Excluding comments with fewer than 5 words in the preprocessing phase for topic modeling enhances the model's quality by prioritizing content richness, noise reduction, and improving interpretability. This strategy aims to ensure that the model focuses on more substantive text, contributing to robust and meaningful topic identification. As a result of preprocessing, 44% of the student comments on schools were removed. This is because most of the comments were short phrases such as “Great school”, “Love it here”, and “Amazing campus”.

When using topic models that leverage document embeddings, the holistic nature of a document's content is necessary for accurately learning its topic. As a result, model creators do not recommend stop-word removal, lemmatization, or punctuation removal, as these elements contribute to the overall linguistic context. Therefore, with the BERTopic and Top2Vec models, we opt to utilize both preprocessed and unprocessed written comments for comparison. This approach ensures that the models can effectively capture nuances of the language and context within comments, allowing for a more comprehensive and accurate representation of topics.

2.3. Topic Modeling

A topic model is an unsupervised machine learning technique to find clusters of similar words text corpus. The technique has been used to study the impact of the Covid-19 pandemic on students in Saudi Arabia and the effect of gender bias in student evaluations.[7,8] In other applications, topic modeling has been used to construct recommendation systems in Q&A sites, analyze developer discussions on Stack Overflow, explore posts on Twitter and Instagram, and understand the dining experience of tourists.[9-13]

Similar to numeric clustering methods, topic models discover patterns in unlabeled data. To find word clusters, various techniques can be used. In this paper, we study four topic models that use different implementation methods: (1) Latent Dirichlet Allocation, (2) Nonnegative Matrix Factorization, (3) BERTopic, and (4) Top2Vec.

2.3.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data such as text corpora.[14] It is defined as a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics, where each topic is modeled as an infinite mixture over an underlying set of topic probabilities. In our case, LDA models each student comment as a mixture of topics and assumes that each word in a comment is associated with a topic. As model output, each topic produced by LDA is a distribution of word probabilities. For input, LDA follows the bag-of-words approach, meaning that it disregards the order in which words appear in a text. Since the bag-of-words approach does not consider the contextual use of words within sentences, it may not provide a precise representation of documents.

Figure 2: Intertopic Distance Map for LDA with student evaluations of schools.

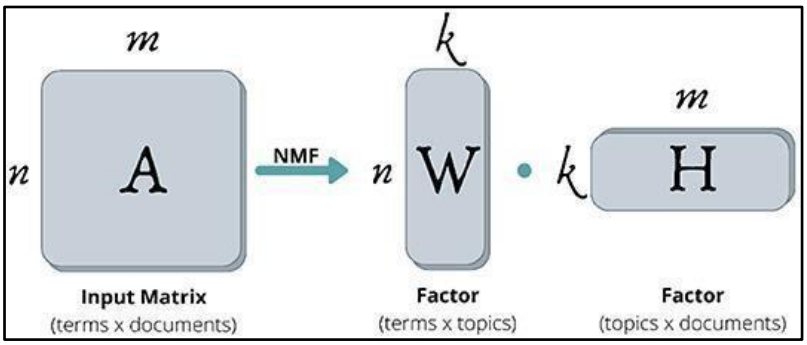

Figure 3: Nonnegative Matrix Factorization for topic modeling.[15]

Using the OCTIS package,[16] we run LDA with default hyperparameter values. Figure 2 demonstrates an Intertopic Distance Map created with LDAvis,[17] representing the 10 topics found by LDA in student evaluations of schools. Each topic is represented as a circle in two-dimensional space, allowing for a visual display of how similar a set of topics is.

2.3.2. Nonnegative Matrix Factorization

Nonnegative Matrix Factorization (NMF) represents an input text corpus as a term-document matrix *A*, following the bag-of-words approach as in LDA. In our case, each row represents a written comment, and each column is a unique word. The input matrix *A* is transformed using a TF−IDF (term frequency-inverse document frequency) weighting scheme to provide a relative measure of each word's importance to a corpus. The output matrices *W* and *H* are found by solving an optimization problem defined with the Frobenius Norm (a distance measure between two given matrices). Figure 3 demonstrates the decomposition of the input matrix *A* into two separate output matrices: (1) *W:* a term-topic matrix, and (2) *H:* a topic-document matrix. The term-topic matrix *W* shows which words comprise a topic, while the topic-document matrix *H* labels the documents associated with each topic.

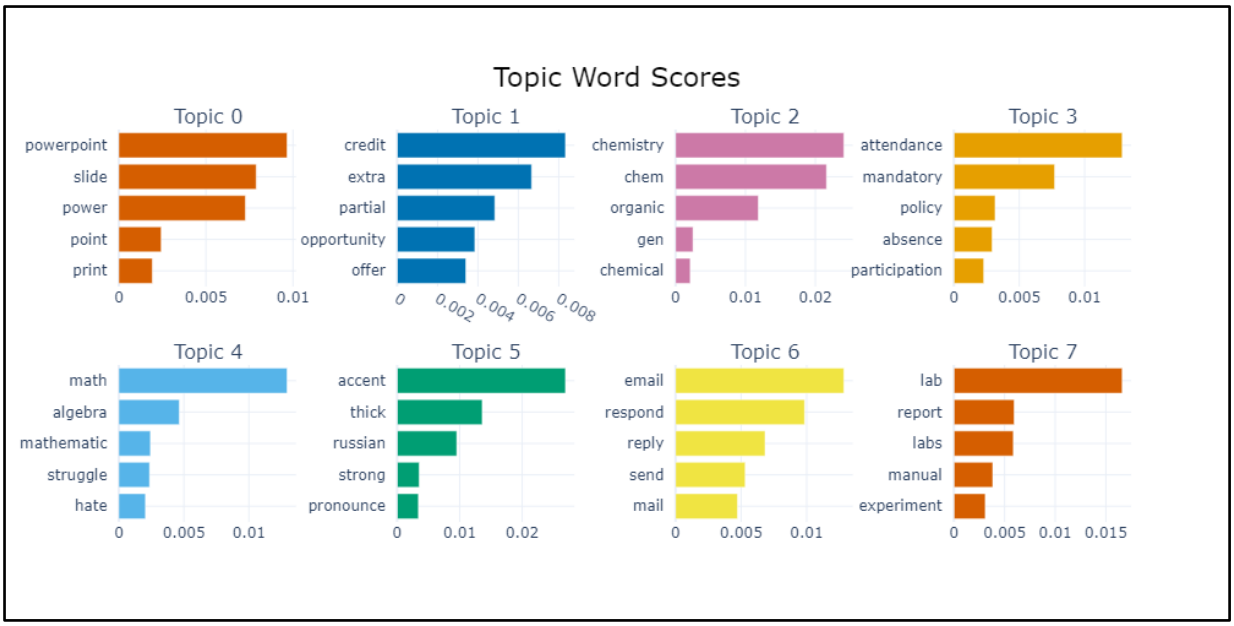

Figure 4: BERTopic word scores for topics in student evaluations of teachers

2.3.3. BERTopic

BERTopic uses pre-trained language models to convert sentences and paragraphs into numeric vectors in a process known as document embedding.[18] The representation of documents in a vector space allows for semantic comparisons that account for linguistic context. After clustering the embedded documents with HDBSCAN,[19] the dimensionality is reduced with UMAP.[20] To represent topics, a class-based variation of TF-IDF is used to model the relative importance of words within the clusters of embedded documents.

In this paper, we use the 'all-MiniLM-L6-v2' embedding model to create the embeddings that BERTopic learns topics from. BERTopic is compatible with any embedding technique pretrained on semantic similarity, as well as classical clustering and dimensionality reduction methods such as k-nearest neighbors (KNN) and principal component analysis (PCA).

2.3.4. Top2Vec

Top2Vec,[21] similarly to BERTopic, accounts for the semantic structure in a corpus by using an embedding approach. Building on the mechanisms of Doc2Vec and Word2Vec, Top2Vec creates word, document, and topic vectors together in high-dimensional space.[22,23] After finding clusters of documents, a topic is represented by the word vectors closest to the center of a document cluster, in contrast to the c-TF-IDF technique used in BERTopic.

Among topic modeling techniques, Top2Vec provides exceptional search functionality, such as the ability to (1) *query topics*: input any sequence of text and Top2Vec returns the topics closest to it, (2) *search documents by topic*: input a topic and Top2Vec returns the documents closest to it, (3) *search topics*: input a list of keywords and Top2Vec returns a choice of either the most similar or the most dissimilar topics. For a fair comparison with BERTopic, we also use the 'all-MiniLM-L6-v2' embedding model with Top2Vec.

2.4. Evaluation Metrics

To quantitatively compare the performance of the topic models, we use the NPMI topic coherence and topic diversity metrics.[24,25] For both metrics, computations are carried out with the OCTIS package. Topic coherence measures the average similarity of words in each topic as a score ranging from -1 to 1, where higher scores indicate greater semantic association within topics. The NPMI coherence score is chosen as it has been shown to provide the highest correlation with human evaluation.[26] However, previous work indicates that the NPMI topic coherence metric may not be suited to embedding-based topic models.[27]

The second evaluation metric we use is topic diversity, which measures the variety among topics as a score ranging from 0 to 1. The topic diversity is the percentage of unique words in all topics. High scores indicate that a topic model found a wide range of themes within a text corpus, while low scores indicate less variety among topics.

3. Results

As the number of topics must be predefined when using LDA and NMF, we examine the coherence and diversity of topics produced when the number of topics varies. We run LDA and NMF with 10 to 50 topics, in increments of 10. Repeating this process over three runs for each increment, the models are each run a total of 15 separate times. The results in Table 1 show LDA and NMF topic coherence and diversity scores as an average of the 15 runs.

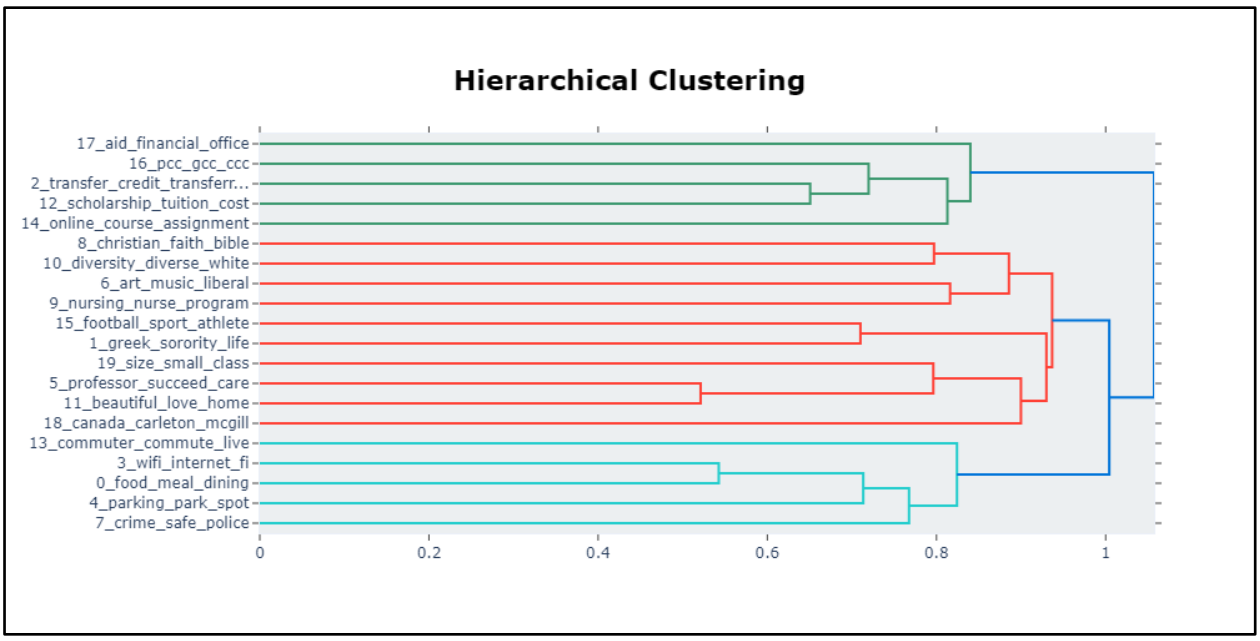

Figure 5: BERTopic hierarchical clustering of the top 20 topics in student evaluations of schools.

With BERTopic and Top2Vec, predefining the number of topics is not required to use the models. In both model implementations, the number of topics equals the number of dense clusters of embedded documents, which increases when large datasets are used. To rapidly narrow down the large set of topics, a hierarchical clustering method can be used. Figure 5 demonstrates the hierarchical topic reduction process for a BERTopic model trained on student evaluations of schools. The top 20 most frequent topics are clustered into 3 main themes: (1) *School Administration* in "aid, financial, office, transfer, credit, scholarship", (2) *Student Experiences* in "faith, diversity, art, sport, sorority, music, and class", and (3) *Campus Environment* in "wifi, internet, dining, parking, crime, safe". By clustering topics, we can quickly see the central features in a large collection of student evaluations.

For the embedding models discussed in this paper, we assess the NPMI coherence and topic diversity as an average across the top N most frequent topics of a corpus, with N ranging in increments of 10 from 10 to 50. We demonstrate all topic model performance scores for the preprocessed written comments from student evaluations of teachers and schools in Table 1, and BERTopic and Top2Vec performance for unprocessed comments in Table 2.

	Teachers		**Schools**	
	TC	TD	TC	TD
LDA	.042	.806	.036	.807
NMF	.050	.538	.048	.560
BERTopic	.160	.942	.157	.911
Top2Vec	-.094	.497	-.098	.517

Table 1: Topic model performance on **preprocessed** comments from student evaluations. TC and TD measure Topic Coherence and Topic Diversity, which respectively score the average similarity of words *within* topics, and average variety of all topics.

	Teachers		Schools	
	TC	TD	TC	TD
BERTopic	.113	.769	.130	.911
Top2Vec	-.097	.304	-.183	.453

Table 2: BERTopic and Top2Vec model performance on **unprocessed** comments from student evaluations. TC and TD measure NPMI Topic Coherence and Topic Diversity, which respectively score the average similarity of words *within* topics, and average variety of all topics.

From Table 1, we observe that BERTopic achieves the highest topic coherence and diversity scores for both preprocessed datasets of comments in student evaluations. NMF demonstrates a slight advantage over LDA in topic coherence, but LDA demonstrates a stronger capability over NMF in creating diverse topics. Top2Vec scores the lowest for all evaluation metrics with both datasets, which could be due to possible incompatibility with the 'all-MiniLM-L6-v2' embedding model.

The results in Table 2 indicate that conducting basic preprocessing on corpora of student evaluations improves BERTopic and Top2Vec model performance. In comparison to results with preprocessed data, the models achieve lower scores across both sets of evaluation metrics and datasets. Although, with unprocessed comments, BERTopic achieves better model evaluation scores than LDA and NMF with preprocessed comments.

From the results, we suggest BERTopic when choosing a topic model to analyze written comments in student evaluations. Along with superior model scores, BERTopic provides exceptional ease-of-use, extensive documentation, built-in visualization features, and a flexible modular structure. As with Top2Vec, new, state-of-the-art embedding models can be used with BERTopic, allowing for progressive refinement as embedding models improve. Moreover, the ability to modify the underlying clustering technique contributes to an extensive range of possible customizations. Furthermore, integration with modern graphics processing units (GPUs) accelerates BERTopic, meaning that training a model and finding topics takes significantly less time than training with LDA.

4. Conclusion

In this paper, we evaluated the performance of four topic modeling techniques for students' written comments in evaluations of teachers and schools. Using Rate My Professors, we constructed comprehensive datasets of 23,841,831 evaluations collected for 1,997,515 teachers and 372,973 evaluations collected for 9,155 schools, spanning the entire RMP website up to September 2023.

We showed that for both datasets, BERTopic produces the best topic coherence and topic diversity when compared to Latent Dirichlet Allocation and Nonnegative Matrix Factorization with default hyperparameter values, as well as Top2Vec when used with the same embedding model.

We plan to experiment with BERTopic and Top2Vec with various embedding models, such as the Universal Sentence Encoder,[28] to discover possible improvements when conducting topic modeling for student evaluations. Additionally, we seek to calibrate the hyperparameters of LDA and NMF to compare the models in fine-tuned conditions.

5. Threats to the Validity

Generalizability: While our study leveraged textual data to construct topic models, specifically student comments extracted from a crowd-sourced site, it is essential to acknowledge the limitations in generalizing the performance of these topic model techniques to other crowd-sourced or social media platforms. The dynamics and characteristics of student evaluations on Rate My Professors might differ from those on other platforms, introducing variability in the performance of the employed topic models. Variations in language use, sentiment expressions, and thematic preferences among different online communities can influence the adaptability and generalizability of the chosen modeling approaches.

*Conclusion Validity***:** In our selection of three promising topic models for comparison with Latent Dirichlet Allocation (LDA), it is important to note that we did not extensively manipulate all their parameters to guarantee the complete optimization of the resultant models. Our rationale behind this approach was to compare the models in their foundational states, aiming to discern performance differences without additional tuning.

To address this potential threat to conclusion validity, future research endeavors will investigate the comprehensive exploration of each technique's parameters. This will not only shed light on the extent of effort required for optimization, but also quantify the potential performance deviations of these three models in comparison to LDA.

6. Future Work

Model Evaluation: We considered two topic model evaluation metrics. For further analysis of results, we plan to expand the evaluation criteria by considering more evaluation metrics, including but limited to topic diversity metrics such as Kullback-Leibler Divergence, topic similarity metrics such as word embedding based RBO matches and pairwise Jaccard similarity, and topic significance metrics such as uniform Kullback-Leiber divergence. The OCTIS package provides users with the ability to experiment with each of the evaluation metrics discussed.

Model Calibration: The hyperparameters of topic models such as LDA and NMF can be optimized to achieve maximal values for given evaluation metrics. In the paper, we used the default hyperparameter values for both models, but those with experience in optimization may opt to find model hyperparameter values that yield maximal topic coherence or diversity measures.

Alternative Methods: Although we considered topic modeling techniques that encompass a wide range of implementation styles, as part of our future work, we will investigate and use more topic models such as Contextualized Topic Models (CTM),[29] Probabilistic Latent Semantic Analysis (PLSA),[30] ProdLDA,[31] Pachinko Allocation,[32] and CorEx[33]. Additionally, new and improved embedding models are frequently released on the Hugging Face platform, paving the way for further experimentation with embedding-based topic models.

References

[1]. [WEB] *Rate My Professors,* www.ratemyprofessors.com, Last accessed: 21st Jan 2024.

[2]. Jie Sun and Lu Yan. Using topic modeling to understand comments in student evaluations of teaching. Discover Education, 2(1), August 2023.

[3]. Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume, and Lise Getoor. Understanding mooc discussion forums using seeded lda. In Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, 2014.

[4]. Mahmoud Azab, Rada Mihalcea, and Jacob Abernethy. Analysing ratemyprofessors evaluations across institutions, disciplines, and cultures: The tell-tale signs of a good professor. In Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I 8, pages 438–453. Springer, 2016.

[5]. Heather Newman and David Joyner. Sentiment Analysis of Student Evaluations of Teaching, page 246–250. Springer International Publishing, 2018.

[6]. Ziqi Tang, Yutong Wang, and Jiebo Luo. Are top school students more critical of their professors? mining comments on ratemyprofessor.com, 2021.

[7]. Omer S. Alkhnbashi and Rasheed Mohammad Nassr. Topic modelling and sentimental analysis of students' reviews. *Computers, Materials amp; Continua*, 74(3):6835–6848, 2023.

[8]. Eric M. Dillon, Haroon Malik, David A. Dampier, and Fatma Outay. Fine-grained analysis of gender bias in student evaluations. In *2021 IEEE Integrated STEM Education Conference (ISEC)*, pages 306–309, 2021.

[9]. Wei Wang, Haroon Malik, and Michael W Godfrey. Recommending posts concerning api issues in developer q&a sites. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, pages 224–234. IEEE, 2015.

[10]. Anton Barua, Stephen W Thomas, and Ahmed E Hassan. What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical software engineering*, 19:619–654, 2014.

[11]. Roman Egger and Joanne Yu. A topic modelling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology,* 7, 886498, 2022.

[12]. Roman Egger and Joanne Yu. Identifying hidden semantic structures in instagram data: a topic modelling comparison. *Tourism Review*, October 2021.

[13]. Roman Egger, Angela Pagiri, Barbara Prodinger, Ruihong Liu, and Fabian Wettinger. *Topic Modelling of Tourist Dining Experiences Based on the GLOBE Model*, page 356–368. Springer International Publishing, 2022.

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

[15]. Roman Egger. Topic modelling: Modelling hidden semantic structures in textual data. In Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications, pages 375–403. Springer, 2022.

[16]. Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. Octis: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2021.

[17]. Carson Sievert and Kenneth Shirley. "LDAvis: A method for visualizing and interpreting topics." *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 2014.

[18]. Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. 2022.

[19]. Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates, page 160–172. Springer Berlin Heidelberg, 2013.

[20]. Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018).

[21]. Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[22]. Quoc V. Le and Tomas Mikolov. Distributed Representations of Sentences and Documents, 2014.

[23]. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013

[24]. Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.

[25]. Adji Bousso Dieng, Francisco J. R. Ruiz, and David M.Blei: "Topic modeling in embedding spaces". Trans. Assoc. Comput. Linguistics, 8:439–453, 2020.
[26]. Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 530–539.
[27]. Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is automated topic model evaluation broken? the incoherence of coherence. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc., 2021.
[28]. Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.
[29]. Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1676–1683, Online. Association for Computational Linguistics.
[30]. Thomas Hofmann. Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*, 2013.
[31]. Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models, 2017.
[32]. Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584, 2006.
[33]. Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542, December 2017.