ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3); 247-259

Discovering Hidden Patterns: Applying Topic Modeling in Qualitative Research

Osman TAT* İzzettin AYDOĞAN**

Abstract

In qualitative studies, researchers must devote a significant amount of time and effort to extracting meaningful themes from large sets of texts and examining the links between themes, which are frequently done manually. The availability of natural language models has enabled the application of a wide range of techniques to automatically detecting hierarchy, linkages, and latent themes in texts. This paper aims to investigate the coherence of the topics acquired from the analysis with the predefined themes, as well as the hierarchy between topics, the similarity, and the proximity-distance between topics by means of the topic model based on BERTopic using unstructured qualitative data. This paper aims to investigate the coherence of the topics acquired from the analysis with the predefined themes, as well as the hierarchy between topics, the similarity, and the proximity-distance between topics by means of the topic model based on BERTopic using unstructured qualitative data. The qualitative data for this study was gathered from 106 students engaged in a university-run pedagogical formation certificate program. In BERTopic procedure, the paraphrase-multilingual-MiniLM-L12-v2 model was used as the sentence transformer model, UMAP was used as the dimension reduction method, and HDBSCAN algorithm as the clustering method. It was found that BERTopic successfully identified six topics corresponding to the six predicted themes in unstructured texts. Moreover, 74% of the texts containing some certain themes could be classified accurately. The algorithm effectively discerned which themes were analogous and which had significant distinctions from others. It was concluded that BERTopic is a procedure which is capable of identifying themes that researchers may not notice, depending on the data density in qualitative data analysis, and has the potential to enable qualitative research to reach more detailed findings.

Keywords: BERTopic, natural language processing, topic modeling

Introduction

Qualitative research, which has an important place in the field of social sciences, is based on a scientific methodology that enables in-depth examination and understanding of the phenomena by analyzing qualitative data obtained from words, pictures, or observations (Chwalisz et al., 1996; Rossman & Rallis, 2017). Researchers often use interviews, focus group discussions, observations, and documents as forms of data collection to draw comprehensive conclusions about the research topic. These techniques are very effective methods for understanding the underlying reasons behind the experiences, perspectives and actions of the participants and the way in which these actions occur (Dinçer & Yavuz, 2023; Wang & Heppner, 2011). In qualitative research, the process of analysis that enables meaningful inferences from the collected data is fundamental. Researchers follow a schematic approach that involves coding, categorizing, and interpreting data to identify patterns, themes, and relationships (Chang & Berk, 2009; Wildemann, 2023). During this iterative analytical process, researchers typically aim to gain a clearer understanding of the research findings by continuously analyzing and comparing their data with the developing conceptual framework (Levitt et al., 2018; Polkinghorne, 1994). The tasks of classifying, categorizing, and extracting relevant patterns and themes from data are complex and need significant attention. These operations are often done manually. However, with the extensive use of natural

* Asst. Prof.,Van Yüzüncü Yıl University, Faculty of Education, Van-Türkiye, osmantat@yyu.edu.tr, ORCID ID: 0000-0003-2950-9647
** Asst. Prof.,Van Yüzüncü Yıl University, Faculty of Education, Van-Türkiye, izzettinaydogan@yyu.edu.tr, ORCID ID: 0000-0002-5908-1285

To cite this article:
Tat, O., & Aydoğan, İ. (2024). Discovering hidden patterns: Applying topic modeling in qualitative research. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 247-259. https://doi.org/10.21031/epod.1539694
Received: 27.08.2024
Accepted: 17.10.2024

language processing, computers can now perform the difficult task of analyzing qualitative data with a proficiency comparable to that of humans.

The term natural language processing (NLP), which was introduced into our lives through the GPT language model and is rapidly gaining popularity, can be academically described as computers automatically processing both spoken and written human languages. This computational processing of human language enables computers to understand, interpret, and even generate documents or speeches in the target language (Aggarwal & Nair, 2012; Chowdhary, 2020; Pérez-Paredes et al., 2018). NLP includes tasks like part-of-speech tagging, chunking, named entity recognition, language modeling, and semantic role labeling (Tufféry, 2022). Topic modeling, one of the natural language processing methods, is a technique that allows for the detection of latent topics, trends, and themes in large textual datasets known as corpora. This approach is more flexible and effective than conventional techniques, including document clustering (Sudigyo, 2023; Yin & Yuan, 2022), as it allows one to find underlying themes in textual materials. Topic modeling helps researchers expose semantic patterns and structures within textual data, thereby facilitating a better knowledge of the underlying topics found in the data (Boussaadi et al., 2023; Özyurt, 2022; Shin et al., 2023).

Latent Dirichlet allocation (LDA) is the most frequently applied topic modeling technique in machine learning and natural language processing. LDA is a generative probabilistic model that uses unstructured documents as themes and represents each topic through a word distribution (Ekinci & Omurca, 2019; Foster, 2016). While LDA is an effective approach for extracting keywords connected to hidden themes in large collections of papers, the traditional LDA method lacks the capacity to use sentimental meanings during topic extraction (Im et al., 2019). Data sparseness and its incapacity to predict the order of words in documents (Ogunleye et al., 2023) are also known to cause inconsistent outcomes in this method (Weisser et al., 2022). There are also hybrid topic modeling techniques that can effectively overcome the aforementioned inadequacies of LDA. The most important of these is BERTopic, a modern topic modeling technique (Mendonça, 2024; Qiang et al., 2017). BERTopic (Grootendorst, 2022) is an advanced topic modeling technique that generates document embeddings using pre-trained language models (BERT), clusters these embeddings, and finally presents the topics through a class-based TF-IDF procedure. Unlike traditional methods, BERTopic supports multiple topic modeling, hierarchically reduces the number of topics to a fewer number of clusters and automatically determines the appropriate number of topics (Egger & Yu, 2022).

In qualitative studies, researchers must devote significant time and effort to extracting meaningful themes from large sets of texts and examining the links between themes, which are frequently done manually. The availability of natural language models has enabled the application of a wide range of techniques for automatically detecting hierarchies, linkages, and latent themes in texts. It might be argued that although topic modeling with the BERTopic approach has the highest potential among these methods, especially in scientific fields like educational sciences or psychology, where qualitative data is frequently used, its contributions have not yet been fully acknowledged. Therefore, conducting topic modeling based on BERTopic during the analysis of qualitative data is important for revealing possible new approaches that are appropriate to the nature of qualitative research. This paper aims to investigate the coherence of the topics acquired from the analysis with predefined themes, as well as the hierarchy, similarity, and proximity-distance between the topics by means of the topic model based on BERTopic using structured qualitative data. Seeking solutions to four research questions, the study focuses on the advantages of topic modeling in qualitative research. These are:

RQ1: Does the number of extracted topics match the predicted number of topics?
RQ2: To what extent are the main topics extracted consistent with the presumed themes?
RQ3: What is the hierarchy of extracted topics?
RQ4: How is the proximity-distance of the topics?

Literature Review

To the best of our knowledge, few studies examine how topic modeling can be used in qualitative research processes and the various possibilities it presents in terms of its contributions to educational sciences or psychology. Although few studies on topic modeling in the literature on educational sciences exist, it is observed that all of them were carried out using the LDA approach or other probabilistic methods. Çavuşoğlu et al. (2023) carefully examined the English teachers' techno-pedagogical subject knowledge in the Web of Science and Springer databases using the LDA methodology. Foster and Inglis

(2018) also examined the subjects and trends in the complete archive of two academic journals on mathematics education in the UK using the LDA approach.
Bent et al. (2021) examined the feedback written by peers on video recordings of pre-service teachers' teaching experiences with the help of LDA. The outputs of the method enabled them to track the development of pre-service teachers in teaching activities. Similarly, Hujala et al. (2020), using the LDA method, examined students' responses to open-ended feedback questions with the help of topic modeling. Wilson et al. (2024) examined secondary school students' perceptions of automated writing evaluation (AWE) with the help of a topic model based on their responses to open-ended interview questions. In addition to these studies, it is possible to come across many studies conducted with the topic model in the field of educational sciences. LDA algorithm has been used to examine how students access instructional materials without purchasing them (Mosia, 2024), to explore discursive contexts related to social institutions (Soysal & Baltaru, 2021), to analyze the views of faculty staff on the transition from conventional education to online (Casillano, 2022), and to study curriculum texts that define the skills to be taught (Kiener et al., 2023). As can be derived from the literature review, no study discusses how to analyze qualitative data in the field of educational sciences or psychology using BERTopic, an innovative approach that offers significant potential.

Methods

Population and Sample

106 students enrolled in the pedagogical formation certificate program run at Van Yüzüncü Yıl University, Faculty of Education, are the source of qualitative data used in this study. A simple random sampling strategy was used in the data collection process.

Data Collection Tools

The data utilized in this study were collected through online means. In addition to demographic variables, the online form included six open-ended questions that participants answered in writing. These questions were specifically related to six different aspects of the certificate program. The research collected students' views on various aspects of the certificate program, including class size, program planning, course effectiveness, competence of instructors, student-instructor communication, and the quality of measurement and assessment activities.

Data Analysis and Procedure

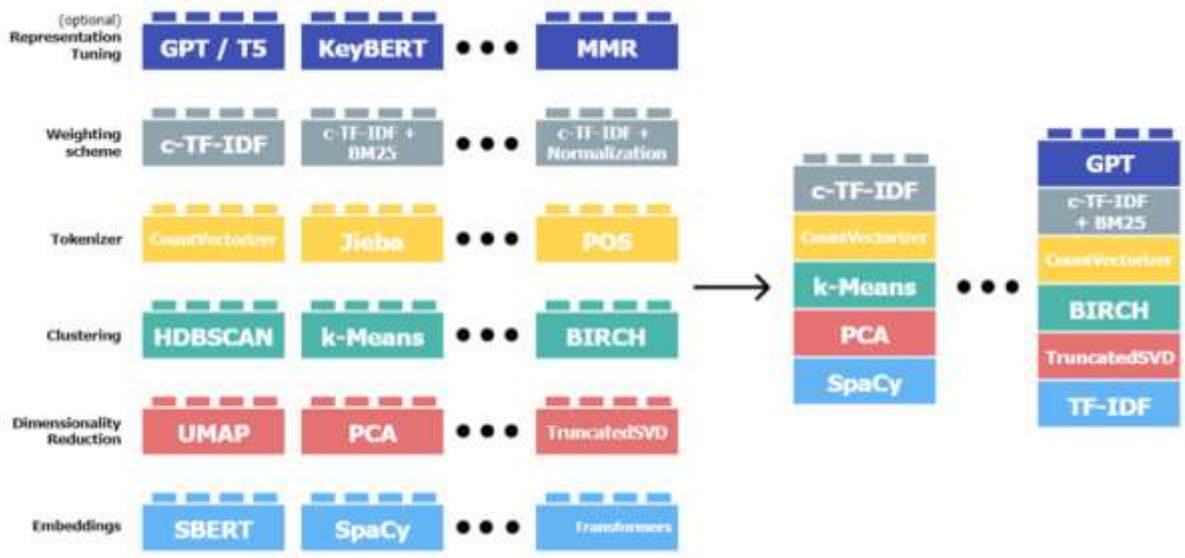
Before the data were analyzed with the topic model, punctuation marks, emojis, and any possible non-textual symbols were removed. Next, stop words that do not contribute to the semantic meaning of the text (e.g., I, me, down, also, etc.) were removed from the text. Afterward, the entire text was converted into lowercase to prevent the same word in uppercase and lowercase from being perceived as different tokens. The texts, consisting of answers to six open-ended questions, were combined into a single column, and the information regarding which text corresponded to which question was hidden from the BERTopic algorithm. When each of the 636 texts collected in a single column is considered as a document, the first step of the BERTopic algorithm is to calculate the embeddings of these documents. In this process, the sentence transformer model ‘paraphrase-multilingual-MiniLM-L12-v2’ (Reimers & Gurevych, 2019) was used. The model, which supports many languages including Turkish, transforms sentences or texts into 384-dimensional dense vector for use in semantic analyses. This methodology was chosen because it was found to produce more consistent outcomes in Turkish texts. UMAP (McInnes et al., 2018) was used to reduce the obtained sentence/text embeddings to a manageable number of dimensions. In the UMAP algorithm, the number of neighbors was set to 10, and the number of components was set to 7. The HDBSCAN (McInnes et al., 2017) algorithm was used to cluster the reduced dimensions by setting the smallest cluster size to 15. As a vectorization method, the count vectorizer was used in a word-based approach and the n-gram parameter was set to (1, 3). The Python libraries sentence transformers, umap, hdbscan, sklearn and BERTopic were used to analyze the data.
Prior to conducting topic modeling on the texts, qualitative analysis was not used to discover the themes. The texts were categorized based on the question to which they were written as a response, and this question served as the theme of the texts. For instance, if a response was provided to the inquiry regarding class size, it was presumed that the content pertained to the theme of class size. The writings

included in the study were not translated from the source language, Turkish, into any other language. Since the analyses were performed on Turkish texts, the visuals and table contents were also presented in Turkish.

BERTopic

BERTopic (Grootendorst, 2022) is a topic modeling method that uses transformer-based language models, namely Bidirectional Encoder Representations from Transformers (BERT), to extract coherent and meaningful topics from textual input. BERTopic differs from standard methods like Latent Dirichlet Allocation (Blei et al., 2003) by using BERT's context-sensitive embeddings to better capture the semantic meaning of words. This leads to more easily understandable subjects (Ding et al., 2023; Hamelberg, 2024). One of the key advantages of BERTopic is its adaptability, which arises from its flexible modeling technique that does not necessitate a predetermined number of topics (Cowan et al., 2022; Yang et al., 2023). BERTopic has additional features such as hierarchical clustering and interactive intertopic distance maps, which facilitate the exploration and analysis of a wide range of topics present in the data (Ramamoorthy, 2024). BERTopic is regarded as an ideal choice for various applications, since research has demonstrated its advantages over other topic modeling methods like as LDA in terms of accuracy and clustering performance (Scarpino et al., 2022; Watanabe & Baturo, 2024). BERTopic can be seen as a series of processes to be taken in order to build topic representations. This approach requires five essential steps to be completed while modeling topics. These steps include embedding generation, dimension reduction, clustering, tokenization, and weighting schemes. The presentation tuning stage is optional. BERTopic is a modular model in which each phase operates independently, and numerous approaches can be applied at each stage. When conducting topic modeling, any of these strategies may be preferred depending on the research goals. For example, HDBSCAN, k-means, and many other clustering algorithms can be used in the clustering stage, whereas GPT, BERT, and many other natural language models can be used in the representation of tuning stage. Thus, each researcher can create a unique BERTopic model tailored to their research (Grootendorst, 2022).

Figure 1.
BERTopic Components



The initial step in BERTopic is to transform input documents into numerical representations in a vector space. At this point, it is assumed that documents with the same topic exhibit semantic similarity. BERTopic employs the Sentence-BERT (SBERT) structure developed by Reimers and Gurevych (2019). This framework efficiently transforms words and paragraphs into vectors using pre-trained language models (Bianchi et al., 2020). Another key component of BERTopic is the process of reducing

the dimensionality of input embeddings. This is because embeddings with high dimensionality create challenges for clustering. One possible option is to reduce the dimensionality of the embeddings to a manageable dimensional space that can be used by clustering methods. As a dimension reduction technique, UMAP is employed as the default method in BERTopic because it has the ability to represent both the local and global high-dimensional spaces in lower dimensions (McInnes et al., 2018). Once the dimensionality of the input embeddings has been reduced, it is necessary to cluster them into groups of comparable embeddings in order to extract topics. The process of clustering is essential, as the effectiveness of clustering technique directly impacts the accuracy of topic representations. HDBSCAN is the typical choice for clustering in BERTopic. Thanks to its ability to effectively capture structures with varying densities (Wang et al., 2021; Zhang et al., 2018), this technique ensures that unrelated documents are not assigned to any cluster, thus enhancing the quality of topic representations.

The interpretability of topic representations plays an important role in topic modeling. Topic representations are generated based on the distribution of texts in each cluster and assigned to a single topic. This process depends on the identification of distinguishing factors between topics based on the word distributions in the clusters. In BERTopic, this process is performed by cluster-based Term Frequency-Inverse Document Frequency (c-TF-IDF), an adjusted version of the TF-IDF (Term Frequency and Inverse Document Frequency) metric, which measures the importance of a word in a document (Egger & Yu, 2022). In this method, instead of being represented as a set of articles, each cluster is transformed into a single document. Then, the frequency of word x in cluster c is calculated, where c corresponds to the previously created cluster. This results in a class-based term frequency representation. Finally, this representation is L1-normalised to adjust for differences in topic sizes (Grootendorst, 2022). For a term x within class c, c-TF-IDF.

$$W_{x,c} = \| tf_{x,c} \| \times \log(1 + \frac{A}{f_x}) \quad (1)$$

$tf_{x,c}$: frequency of word $\times$ in class c

f_x: frequency of word $\times$ across all classes

A: average number of words per class

Results

As a result of the unsupervised topic modeling, it was concluded that the unstructured text contains eight latent topics. The BERTopic algorithm assigns a value of -1 to any topic considered an outlier, which is not analyzed or interpreted as part of the analysis. In this scenario, it is more appropriate to analyze and comment on the details of the seven topics derived from the model. The results are displayed in Table 1 and Figure 2.

Figure 2.

Topic Word Score

Table 1

Topics and Representations

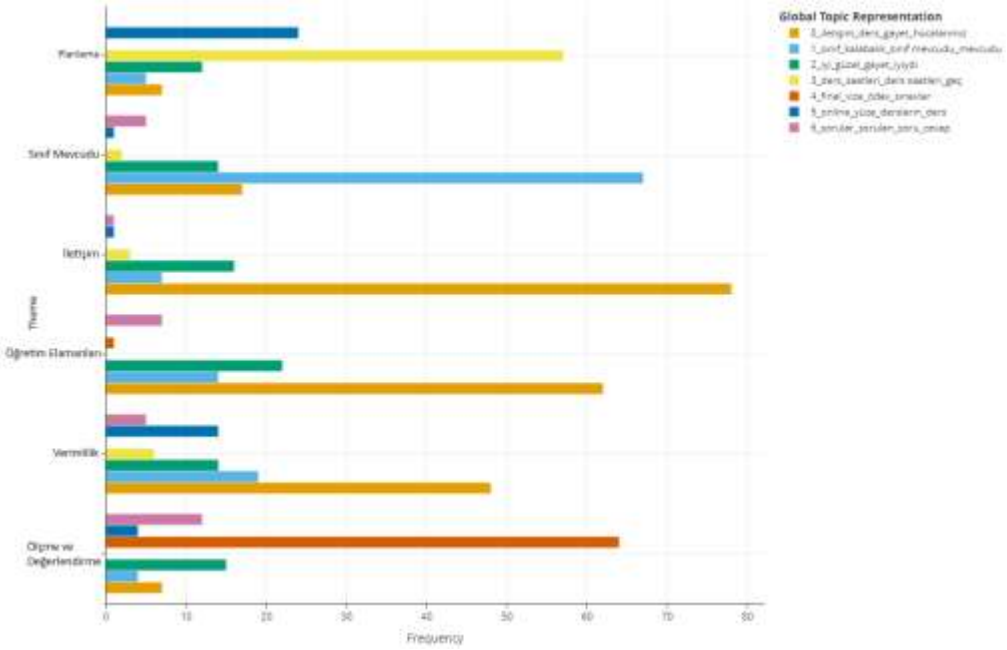
Topic	Count (n)	Name	Token/n-gram	Representative_Docs
-1	1	-1_değildir_yeterli_ doğru karar_karar	['değildir', 'yeterli', 'doğru karar', 'karar', 'dönemde', 'doğru', 'öğretmenlik', 'olağan', 'sorun yaşadım', 'varken formasyon']	"Çünkü 4 yıl eğitim görmek varken formasyon ile iki dönemde alıyoruz, bu yeterli derecede maalesef değildir. Fakat hocalarımızdan aldığımız bilgiler umarım bizler için yeterli olur çünkü buna öğretmenlik mesleğine başlamadan net bir şey söylemiyorum. Fakat yeterli değildir."
0	219	0_iletişim_ders_gayet_ hocalarımız	['iletişim', 'ders', 'gayet', 'hocalarımız', 'dersler', 'öğretim', 'iyi', 'verimli', 'akademik', 'hocalarımızın']	"Ders dışında ders ile alakalı olarak ihtiyaç duyulduğunda sınıf temsilcileri ile iletişim kurulması biraz garip olsa da gerekli açıklamalar kulaktan kulağa yapılmaktadır. Ders sırasında herhangi bir iletişim sorunu olmadan uygun açıklamalar yapılıyor."
1	116	1_sınıf_kalabalık_ sınıf mevcudu_	['sınıf', 'kalabalık', 'sınıf mevcudu', 'mevcudu', 'fazla', 'olumsuz', 'öğrenme', 'öğrenci', 'değildi', 'sınıf mevcudunun']	"Sınıf mevcudu ortalama 45 kişilik. Öğrenme sürecini olumsuz etkilediğini düşünmüyorum çünkü genel olarak herkesin yaşı itibari ile farkındalığından sınıf düzeni bozulmuyor çok fazla."
2	93	2_iyi_güzel_gayet_ iyiydi	['iyi', 'güzel', 'gayet', 'iyiydi', 'verimli', 'yok', 'yeterli', 'uygun', 'gerektiği', 'gayet iyi']	"İyi", "iyi", "iyi"
3	68	3_ders_saatleri_ders saatleri_geç	['ders', 'saatleri', 'ders saatleri', 'geç', 'akşam', 'gün', 'derslerin', 'saatler', 'sayısı', 'saatte']	"Örgün eğitim gören öğrenciler olduğundan dolayı formasyon öğrencilerinin akşam saatlerinde ders görüyor olmaları normal fakat ders saatleri olarak bakıldığında yeterli değildi. Öğretmenlik uygulaması ise benim açımdan verimli ve bilgi edindiğim bir program oldu."
4	65	4_final_vize_ödev_ sınavlar	['final', 'vize', 'ödev', 'sınavlar', 'sınav', 'ölçme', 'değerlendirme', 'vize final', 'ölçme değerlendirme', 'zor']	"Hocamız bu konu da vize ve finalde soru dağılımlarını çok iyi şekilde ayarlayarak şans başarısını düşürerek güvenilir ve tutarlı bir ders ve sınav yaşattı. Son olarak bu dersin bir dönem de aşılmasının doğru olmadığını da beyan ederim. Yani 4 yıl da alınmalıdır"
5	44	5_online_yüze_ derslerin_ders	['online', 'yüze', 'derslerin', 'ders', 'derslerin online', 'öğretmenlik', 'verimli', 'dersler', 'ders saatleri', 'saatleri']	"Ders saatleri uygun. Online dersleri pek etkili değil yüz yüze olması daha mantıklı. Online olan derslerin sınavları da online olması lazım.", "Bazı derslerin online olmasından çok memnunum. Hepsi online olsa çok daha iyi olur."
6	30	6_sorular_sorulan_ soru_cevap	['sorular', 'sorulan', 'soru', 'cevap', 'güzel', 'bence', 'olabilirdi', 'sorulan sorular', 'gayet', 'işlediğimiz']	"Hocalarımız sorulan sorulara eksiksiz cevap veriyor kafamızda soru işareti kalmıyor', "Hocalarımızın bilgi düzeyleri gayet makuldü konulara hakimiyet sorulan sorular karşısında verdikleri cevap tatmin edici."

The first noteworthy finding is that Topic 2 was categorized as a separate topic because the algorithm could not associate the answers given to all questions with one or two words such as 'good', 'very good', or 'sufficient' with other topics. Since it is not possible to understand and interpret the context from this topic, it does not provide a significant finding for the research. When this topic is neglected, it can be asserted that the other topics precisely coincide with the pre-determined themes. Specifically, Topic 0 relates to the quality of communication between students and instructors, Topic 1 to the size of the class, Topics 3 and 5 to the planning and efficiency of courses, Topic 4 to assessment and evaluation activities,

and Topic 6 to the competence of instructors. Figure 3 presents the specific relationship between the texts related to the topics and the presumed themes.

Figure 3

Topic per Theme



The algorithm classified 74% (n=78) of the texts expected to belong to the communication theme as Topic 0. Similarly, 63% (n=67) of the texts belonging to the theme of class size were classified as Topic 1. 54% (n=57) of the texts known to belong to the theme of lesson planning were classified as Topic 3. 60% (n=64) of the texts known to belong to the theme of measurement and evaluation were classified as Topic 4. 13% (n=14) of the views on the efficiency of the courses were associated with Topic 5. Finally, 7% (n=7) of the texts belonging to the theme of the competence of the lecturers were classified as Topic 6. To examine the extent to which the texts are related to the topics, the BERTopic algorithm provides the classification probability of each text, which is a very useful feature. The average classification probability for each topic indicates how reliably the algorithm classifies each text. Accordingly, the average probability is 0.74 for Topic 0, 0.72 for Topic 1, 0.49 for Topic 2, 0.81 for Topic 3, 0.87 for Topic 4, 0.85 for Topic 5, and 0.64 for Topic 6. In this case, it is possible to say that the algorithm is not reliable enough when the texts are categorized into Topic 2, which consists of short answers such as ‘good’, ‘nice’, and so on.

Figure 4

Similarity Matrix

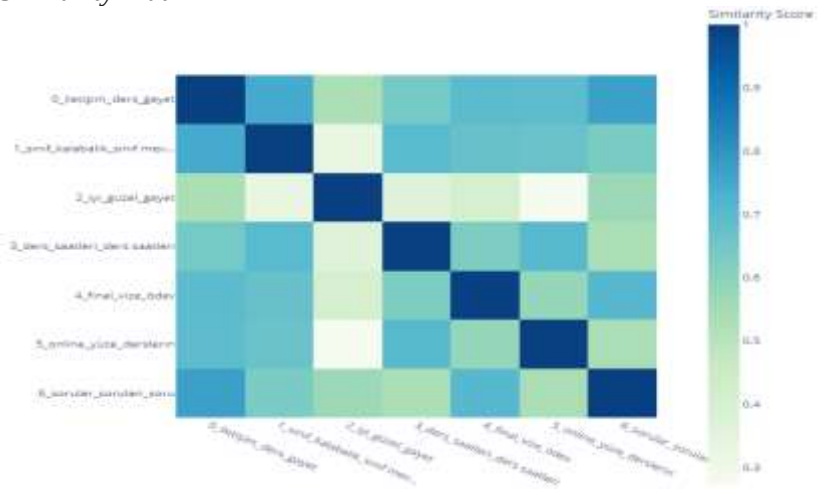
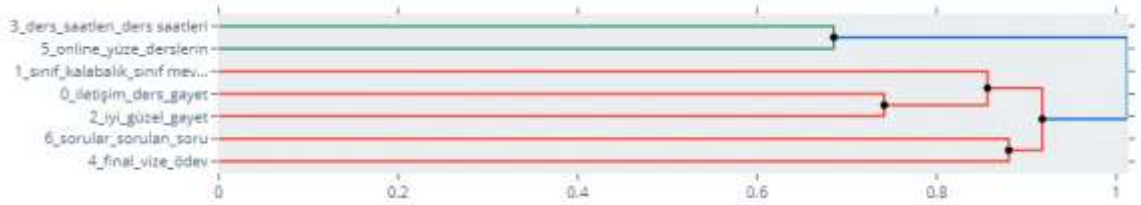


Figure 5

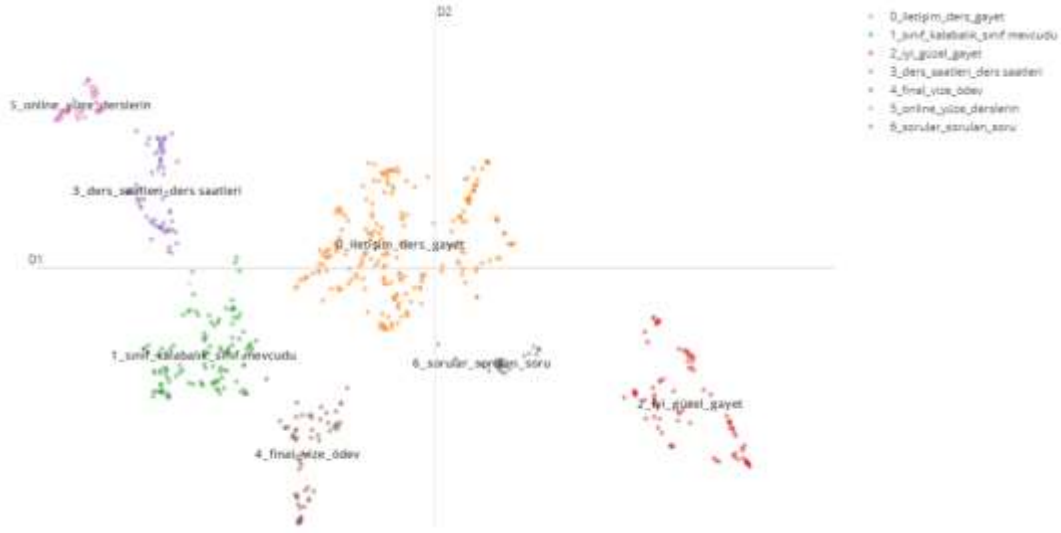
Hierarchical Clusters



One of the most important features of BERTopic for researchers is that the relationship (Figure 4) and hierarchy (Figure 5) between topics can be obtained after topic modeling. When the similarity matrix in Figure 3 and the hierarchy dendrogram in Figure 4 are analyzed, it can be said that there is a strong relationship between Topic 0 (communication theme) and Topic 1 (class size theme). The main reason for this is the large number of opinions in many texts stating that crowded classes make communication difficult, or that classes are not crowded and thus do not negatively affect communication. Similarly, it is possible to mention a strong relationship between Topic 0 and Topic 6. In most of the texts, students stated that they received satisfactory answers to the questions they asked to the lecturers both in and outside of class and that communication was at a sufficient level in this regard. Another notable relationship is observed between Topic 6 and Topic 4. The likely reason for this relationship is that some of the texts classified in Topic 6 are related to the questions asked to the instructors in the classroom, while some of them are related to the questions asked in the exams. Again, it is seen that Topic 3 and Topic 5 have a high level of similarity. This indicates that the themes of the time interval of the courses and the way the courses are organized (online or face-to-face) have many common aspects.

Figure 6

Document Distributions Based on Reduced Embeddings



Another important feature of topic modeling based on the BERTopic algorithm is that a scatter plot can be created from the reduced embeddings obtained from sentence transformers, which are reduced as a result of UMAP and clustering analysis. This graph visually illustrates how the topics are distributed on the two-dimensional surface and how the level of proximity-distance between the topics is. In Figure 6, it is firstly observed that the texts belonging to Topic 2 are quite distant from the other topics and are an isolated cluster with a weak relationship to the others. Again, it is seen that Topic 0, which contains the theme of the quality of communication, is in a more central position and is located especially close to Topic 1 and Topic 6. Similarly, it can be said that Topic 3 and Topic 5 are positioned very closely to each other.

Discussion

This work presents the advantages of utilizing the BERTopic (Grootendorst, 2022) method for extracting topics from unstructured texts. The algorithm integrates many components of natural language models, sophisticated dimensionality reduction, and clustering analysis to provide a more efficient approach to topic modeling (Abuzayed & Al-Khalifa, 2021; Kukushkin, 2022; Maryanto, 2024). As a result of the analysis, the algorithm detected that the texts contained seven hidden topics, except for the outlier topic (Topic -1). One of these seven topics (Topic 2) was found to contain very short answers such as 'good', 'good', 'good', 'enough', and its relationship with the other topics could not be fully determined due to the insufficient token diversity. Thus, it can be said that the topic model successfully identified six topics corresponding to the six predicted themes in unstructured texts. Topic 0 was found to represent the majority of the texts belonging to the theme of the quality of communication between teaching staff and students. Similar situations are also valid for other topics. For example, Topic 1 overlaps with the theme of class size, Topic 3 and Topic 5 with the themes of lesson planning and efficiency of lessons, Topic 4 with the theme of assessment and evaluation activities, and Topic 6 with the theme of competence of lecturers. The accuracy rate of the texts in the topics varies between 78% (Topic 0) and 7% (Topic 6). The lower-than-expected accuracy rates in some topics may be due to many texts containing opinions on more than one theme and a significant portion of the texts consisting of just one or two words. Texts expressing thoughts on multiple themes may have been assigned to the wrong topic, or texts with too few tokens may have been classified under another topic. Still, it can be argued that when the thoughts are detailed and the interview questions are effectively crafted, the BERTopic method can be highly successful.

Another advantage of topic modeling with BERTopic, as revealed in this study, is the ability to identify similarity levels between topics and the topic hierarchies obtained based on these similarity levels (Cheddak, 2024; Kousis, 2023). In this study, the level of similarity between topics was analyzed using a heat map based on the correlation between topics and a scatter plot of the texts generated from reduced embeddings. In light of these analyses, it was possible to determine which texts are more centrally located, i.e. which texts are likely to contain other topics, based on the relationship between topics. Additionally, it was also possible to determine which topics were clearly differentiated from other topics. To determine the hierarchy between texts, the results obtained from the cluster analysis can be analyzed with the help of dendrograms. It can be said that the intertextual hierarchy has many advantages for researchers. For example, identifying some topics as a whole of more than one sub-topic can enhance the researchers' understanding of the relationships between themes and thematization processes while conducting qualitative analyses. In conclusion, BERTopic is an advanced modern technique that combines many useful aspects of natural language processing, dimension reduction and cluster analysis techniques. It is thought that this algorithm will contribute to the automatic mining of qualitative data, which is frequently used in social sciences such as education and psychology, primarily by providing quantitative indicators based on qualitative data, and performing similarity and hierarchical structure analyses based on these quantitative values. In addition, BERTopic can identify themes that researchers may not have noticed, depending on the density of the data in qualitative data analysis, and enable qualitative research to reach more detailed findings.

Declarations

Conflict of Interest: The authors have no conflicts of interest to declare.

Ethical Approval: The study was ethically approved by the by Van Yüzüncü Yıl University Social and Human Sciences Ethics Committee dated 07/08/2024 and numbered 2024/16-17**.**

A part of this study was presented as an oral presentation at Vizyon Van 2024 Congress, Van, Turkey, 2024.

References

Abuzayed, A., & Al-Khalifa, H. S. (2021). Bert for Arabic topic modeling: An experimental study on BERTopic technique. *Procedia Computer Science*, 189, 191-194. https://doi.org/10.1016/j.procs.2021.05.096

Aggarwal, E., & Nair, S. (2012). NLP token matching on database using binary search. *International Journal of Computers & Technology*, 3(1), 140-143. https://doi.org/10.24297/ijct.v3i1c.2766

Bent, M., Velazquez-Godinez, E., & Jong, F. (2021). Becoming an expert teacher: Assessing expertise growth in peer feedback video recordings by lexical analysis. *Education Sciences*, 11(11), 665. https://doi.org/10.3390/educsci11110665

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual Contextualized Topic Models with Zero-shot Learning. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the European chapter of the association for computational linguistics*: *Main volume*,1676–1683. doi:10.18653/v1/2021.eacl-main.143

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(1), 993–1022.

Boussaadi, S., Aliane, H., & Abdeldjalil, O. (2023). Using an explicit query and a topic model for scientific article recommendation. *Education and Information Technologies*, 28(12), 15657-15670. https://doi.org/10.1007/s10639-023-11817-2

Casillano, N. F. B. (2022). Discovering sentiments and latent themes in the views of faculty members towards the shift from conventional to online teaching using VADER and latent dirichlet allocation. *International Journal of Information and Education Technology*, 12(4), 290-298. https://doi.org/10.18178/ijiet.2022.12.4.1617

Çavuşoğlu, D., Kıncal, R. Y., & Kartal, O. Y. (2023). Systematic review of research conducted on the techno-pedagogical content knowledge of English teachers. *Journal of Family Counseling and Education*, 8(2), 170-192. https://doi.org/10.32568/jfce.1269034

Chang, D. F., & Berk, A. (2009). Making cross-racial therapy work: A phenomenological study of clients' experiences of cross-racial therapy. *Journal of Counseling Psychology*, 56(4), 521-536. https://doi.org/10.1037/a0016905

Cheddak, A. (2024). BERTopic for enhanced idea management and topic generation in brainstorming sessions. *Information*, 15(6), 365. https://doi.org/10.3390/info15060365

Chowdhary, K. R. (2020). Natural language processing. Fundamentals of Artificial Intelligence, 603-649. https://doi.org/10.1007/978-81-322-3972-7_19

Chwalisz, K., Wiersma, N., & Stark-Wroblewski, K. (1996). A quasi-qualitative investigation of strategies used in qualitative categorization. *Journal of Counseling Psychology*, 43(4), 502-509. https://doi.org/10.1037/0022-0167.43.4.502

Cowan, T., Rodriguez, Z., Granrud, O., Masucci, M., Docherty, N., & Cohen, A. (2022). Talking about health: A topic analysis of narratives from individuals with schizophrenia and other serious mental illnesses. *Behavioral Sciences*, 12(8), 286. https://doi.org/10.3390/bs12080286

Dinçer, P., & Yavuz, H. (2023). Behind the screen: a case study on the perspectives of freshman EFL students and their instructors. *Education and Information Technologies*, 28(9), 11881-11920. https://doi.org/10.1007/s10639-023-11661-4

Ding, Q., Ding, D., Wang, Y., Guan, C., & Ding, B. (2023). Unraveling the landscape of large language models: A systematic review and future perspectives. *Journal of Electronic Business & Digital Economics*, 3, 3-19. https://doi.org/10.1108/jebde-08-2023-0015

Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Frontiers in Sociology*, 7. https://doi.org/10.3389/fsoc.2022.886498

Ekinci, E., & Omurca, S. (2019). Concept-LDA: Incorporating Babelfy into LDA for aspect extraction. *Journal of Information Science*, 46(3), 406-418. https://doi.org/10.1177/0165551519845854

Foster, A. (2016). An extension of standard latent dirichlet allocation to multiple corpora. *SIAM Undergraduate Research Online*, 9. https://doi.org/10.1137/15s014599

Foster, C., & Inglis, M. (2018). Mathematics teacher professional journals: What topics appear and how has this changed over time?. International Journal of Science and Mathematics Education, 17(8), 1627-1648. https://doi.org/10.1007/s10763-018-9937-4

Grootendorst, M. (2022). BERTOPIC: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arxiv.2203.05794

Hamelberg, K., de Ruyter, K., van Dolen, W., & Konuş, U. (2024). Finding the right voice: How CEO communication on the Russia–Ukraine war drives public engagement and digital activism. *Journal of Public Policy & Marketing*. https://doi.org/10.1177/07439156241230910

Hujala, M., Knutas, A., Hynninen, T., & Arminen, H. (2020). Improving the quality of teaching by utilizing written student feedback: A streamlined process. *Computers & Education*, 157, 103965. https://doi.org/10.1016/j.compedu.2020.103965

Im, Y., Park, J., Kim, M., & Park, K. (2019). Comparative study on perceived trust of topic modeling based on affective level of educational text. *Applied Sciences*, 9(21), 4565. https://doi.org/10.3390/app9214565

Kiener, F., Gnehm, A., & Backes-Gellner, U. (2023). Noncognitive skills in training curricula and nonlinear wage returns. *International Journal of Manpower*, 44(4), 772-788. https://doi.org/10.1108/ijm-03-2022-0119

Kousis, A. (2023). Investigating the key aspects of a smart city through topic modeling and thematic analysis. *Future Internet*, 16(1), 3. https://doi.org/10.3390/fi16010003

Kukushkin K., Ryabov Y., & Borovkov A. (2022). Digital Twins: A Systematic Literature Review Based on Data Analysis and Topic Modeling. *Data*, 7(12):173. https://doi.org/10.3390/data7120173

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA publications and communications board task force report. *American Psychologist*, 73(1), 26-46. https://doi.org/10.1037/amp0000151

Maryanto, M. (2024). Hybrid model for extractive single document summarization: Utilizing bertopic and bert model. *IAES International Journal of Artificial Intelligence* (Ij-Ai), 13(2), 1723. https://doi.org/10.11591/ijai.v13.i2.pp1723-1731

McInnes, L., Healy, J. J., & Astels, S. (2017). HDBSCAN: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018). UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29), 861.

Mendonça, M. (2024). Topic extraction: BERTopic's insight into the 117th congress's twitterverse. *Informatics*, 11(1), 8. https://doi.org/10.3390/informatics11010008

Mosia, M. (2024). Data-driven insights into non-purchasing behaviours through latent dirichlet allocation: Analysing study material acquisition among university students. *Journal of Culture and Values in Education*, 7(1), 72-82. https://doi.org/10.46303/jcve.2024.5

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of topic modelling approaches in the banking context. *Applied Sciences*, 13(2), 797. https://doi.org/10.3390/app13020797

Özyurt, Ö. (2022). Empirical research of emerging trends and patterns across the flipped classroom studies using topic modeling. *Education and Information Technologies*, 28(4), 4335-4362. https://doi.org/10.1007/s10639-022-11396-8

Pérez-Paredes, P., Guillamón, C. O., & Jiménez, P. A. (2018). Language teachers' perceptions on the use of oer language processing technologies in mall. *Computer Assisted Language Learning*, 31(5-6), 522-545. https://doi.org/10.1080/09588221.2017.1418754

Polkinghorne, D. E. (1994). Reaction to special section on qualitative research in counseling process and outcome.. Journal of Counseling Psychology, 41(4), 510-512. https://doi.org/10.1037//0022-0167.41.4.510

Qiang, J., Chen, P., Wang, T., & Wu, X. (2017). Topic modeling over short texts by incorporating word embeddings. *Advances in Knowledge Discovery and Data Mining*, 363-374. https://doi.org/10.1007/978-3-319-57529-2_29

Ramamoorthy, T., Kulothungan, V., & Mappillairaju, B. (2024). Topic modeling and social network analysis approach to explore diabetes discourse on twitter in India. *Frontiers in Artificial Intelligence*, 7. https://doi.org/10.3389/frai.2024.1329185

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Retrieved from http://arxiv.org/abs/1908.10084

Reimers, N., & Gurevych, I. (2019). Sentencebert: Sentence embeddings using siamese BERTnetworks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Procesessing Association for Computational Linguistics*.

Rossman, G., & Rallis, S. F. (2017). *An introduction to qualitative research: Learning in the field*. SAGE Publications. https://doi.org/10.4135/9781071802694

Scarpino, I., Zucco, C., Vallelunga, R., Luzza, F., & Cannataro, M. (2022). Investigating topic modeling techniques to extract meaningful insights in italian long covid narration. Biotech, 11(3), 41. https://doi.org/10.3390/biotech11030041

Shin, M., Ok, M. W., Choo, S., Hossain, G., Bryant, D. P., & Kang, E. (2023). A content analysis of research on technology use for teaching mathematics to students with disabilities: Word networks and topic modeling. *International Journal of STEM Education*, 10(1). https://doi.org/10.1186/s40594-023-00414-x

Soysal, Y., & Baltaru, R. (2021). University as the producer of knowledge, and economic and societal value: The 20th and twenty-first century transformations of the UK higher education system. *European Journal of Higher Education*, 11(3), 312-328. https://doi.org/10.1080/21568235.2021.1944250

Sudigyo, D., Hidayat, A. A., Nirwantono, R., Rahutomo, R., Trinugroho, J. P., & Pardamean, B. (2023). Literature study of stunting supplementation in Indonesian utilizing text mining approach. *Procedia Computer Science*, 216, 722-729. https://doi.org/10.1016/j.procs.2022.12.189

Sutton, J., & Austin, Z. (2015). Qualitative research: Data collection, analysis, and management. *The Canadian Journal of Hospital Pharmacy*, 68(3). https://doi.org/10.4212/cjhp.v68i3.1456

Tufféry, S. (2022). *Deep learning: From big data to artificial intelligence with r*. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119845041.ch9

Wang, L., Chen, P., Chen, L., & Mou, J. (2021). Ship AIS trajectory clustering: An HDBSCAN-based approach. *Journal of Marine Science and Engineering*, 9(6), 566. https://doi.org/10.3390/jmse9060566

Wang, Y., & Heppner, P. P. (2011). A qualitative study of childhood sexual abuse survivors in Taiwan: Toward a transactional and ecological model of coping. *Journal of Counseling Psychology*, 58(3), 393-409. https://doi.org/10.1037/a0023522

Watanabe, G., Conching, A., Nishioka, S. T., Steed, T., Matsunaga, M., Lozanoff, S.,…& Noh, T. (2023). Themes in neuronavigation research: A machine learning topic analysis. *World Neurosurgery: X*, 18, 100182. https://doi.org/10.1016/j.wnsx.2023.100182

Watanabe, K., & Baturo, A. (2024). Seeded Sequential LDA: A Semi-Supervised Algorithm for Topic-Specific Analysis of Sentences. Social Science Computer Review, 42(1), 224-248. https://doi.org/10.1177/08944393231178605

Weisser, C., Gerloff, C., Thielmann, A., Python, A., Reuter, A., Kneib, T., … & Säfken, B. (2022). Pseudo-document simulation for comparing LDA, GSDMM and GPM topic models on short and sparse text using twitter data. *Computational Statistics*, 38(2), 647-674. https://doi.org/10.1007/s00180-022-01246-z

Wildemann, S. (2023). Bridging qualitative data silos: The potential of reusing codings through machine learning based cross-study code linking. S*ocial Science Computer Review*, 42(3), 760-776. https://doi.org/10.1177/08944393231215459

Wilson, J., Zhang, S., Palermo, C., Cordero, T. C., Zhang, F., Myers, M. C., … & Coles, J. (2024). A latent dirichlet allocation approach to understanding students' perceptions of automated writing evaluation. *Computers and Education Open*, 6, 100194. https://doi.org/10.1016/j.caeo.2024.100194

Yang, L., Shi, J., Zhao, C., & Zhang, C. (2023). Generalizing factors of covid-19 vaccine attitudes in different regions: A summary generation and topic modeling approach. *Digital Health*, 9. https://doi.org/10.1177/20552076231188852

Yin, B., & Yuan, C. (2022). Detecting latent topics and trends in blended learning using LDA topic modeling. *Education and Information Technologies*, 27(9), 12689-12712. https://doi.org/10.1007/s10639-022-11118-0

Zhang, D., Lee, K., & Lee, I. (2018). Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. *Expert Systems with Applications*, 92, 1-11. https://doi.org/10.1016/j.eswa.2017.09.040