Natural Language Processing for Education Research: Exploring Strategic Use of Traditional and Large Language Topic Models

Neha Kardam

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Chair Denise Wilson

Jennifer Hoffman

Sep Makhsous

Mahmood Hameed

Program Authorized to Offer Degree:
Electrical and Computer Engineering

University of Washington

Abstract

Natural Language Processing for Education Research: Exploring Strategic Use of Traditional and Large Language Topic Models

Neha Kardam

Chair of the Supervisory Committee:
Denise Wilson
Electrical and Computer Engineering

Engineering education research increasingly relies on qualitative analysis of short, open-ended survey responses to understand student experiences across courses and institutions, but extracting reliable themes from these texts at scale requires methods that balance computational efficiency with interpretive rigor. While Natural Language Processing (NLP) has been applied in education for automated grading and sentiment analysis, its systematic integration with qualitative thematic analysis for short, prompt-guided educational research texts has received limited attention. This dissertation addresses that gap by comparing five topic modeling methods on short student feedback on instructional support, and by developing the NLP-Assisted Thematic Analysis framework, a six-stage workflow that embeds domain expert judgment from data preparation through final validation.

Three survey datasets of undergraduate engineering student responses on faculty support, teaching assistant (TA) support, and peer support (1,667, 1,592, and 1,376 responses, respectively, for approximately 4,600 total) were processed through a standardized preprocessing pipeline and evaluated against expert-coded themes. Five methods were compared: k-means clustering, Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), BERTopic with MiniLM and MPNet sentence embeddings, and zero-shot classification (ZSC). Performance was evaluated using accuracy, macro and weighted F1, topic coherence, and inter-rater reliability (Cohen's κ). Ground truth was established by two

approaches: (a) a machine-led approach in which topic model keywords guided manual coding of the data; and (b) a human-led approach in which a domain expert coded the data independently.

Results varied by dataset, with no single method performing best across all three corpora. BERTopic MiniLM performed best on the concise, low-ambiguity peer support corpus (85% accuracy, 77% macro-F1), with LDA second at 78% accuracy and 67% macro-F1. BERTopic MPNet led on faculty support, where over one quarter of responses addressed overlapping themes (76.8% accuracy, 65.7% macro-F1), with NMF close behind on accuracy (76.76% accuracy, 62.82% macro-F1). TA support was the most challenging dataset due to higher thematic ambiguity and misalignment between model-generated topics and expert-identified themes. ZSC, applied to the peer support dataset, reached 85% accuracy and 60% weighted F1 when prompts used mainstream language, compared to 82% accuracy and 56% weighted F1 with domain-specific prompts.

The NLP-Assisted Thematic Analysis framework structures domain expert involvement across six stages of the analysis pipeline, from data preparation through final validation. Expert review consolidated nine algorithmic topics into five research themes, with inter-rater reliability (Cohen's κ) between 0.72 and 0.75 across all three datasets. Targeted interventions, including domain-specific stopword curation, hyperparameter selection, topic-to-theme bridging, and review of algorithmically uncertain responses, improved macro-F1 by up to 14 percentage points. The largest single gain arose from BERTopic outlier review on the TA support dataset, raising macro-F1 from 54.2% to 69.3%.

These results establish performance benchmarks for five NLP methods on short educational research text, identify where domain expert involvement has the greatest impact on accuracy and interpretive quality, and provide the NLP-Assisted Thematic Analysis framework as a reproducible, decision-guided protocol for researchers applying topic modeling to qualitative survey data in education and related fields.

Dedicated to my parents, Birendra Singh Kardam and Mithlesh Kardam; my husband, Mukul Jain; and my children, Kian and Suhaan.

ACKNOWLEDGMENTS

When I was in middle school back in India, I was fascinated by western culture and dreamed of coming to study in the United States. I did not know how I would ever study abroad, but I believed that one day I would. That dream stayed alive, and when I first visited the University of Washington, the beauty of the campus left me in awe. I quietly wished that one day I would study here.

That dream almost ended once. My first application to the Ph.D. program at the University of Washington was not successful, and for a while I convinced myself to let it go. I had just become a new mother; my son Kian was six months old, and life already felt full. But I refused to give up. An internal restlessness, a persistent determination, wouldn't let me quit. I remember returning from a trip with only a few days left before the application deadline. I sat down to write my statement of purpose and followed my heart, writing about my passion for understanding learning through education research. I reflected on my eagerness to transform the learning experiences I had been delivering during my time at Lake Washington Institute of Technology. When the offer came, I realized I had stepped into the life my younger self once imagined.

I owe that moment to Professor Denise Wilson. She saw purpose in my vision when I was unsure of it myself. Her empathy, patience, and guidance shaped how I grew as a researcher. I could not have asked for or imagined a better advisor. Her passion for her work and her way of leading by example inspired me every step of the way. Because of her, I believed I could grow beyond my engineering background, step into education research, and gain the endurance to tackle some of the most ambiguous research questions that reshaped my career.

I am grateful to my Ph.D. committee: Professor Jennifer Hoffman, whose qualitative research course, my first Ph.D. class, inspired me deeply and showed how engaging online

learning could be during the pandemic; Professors Sep Makhsous and Mahmood Hameed, whose thoughtful questions and discussions deepened my understanding of this work. I also thank Jennifer Huberman for patiently answering my questions and for her kind and consistent support.

I am thankful to UW for scholarship support, to Lake Washington Institute of Technology for funding my early Ph.D. coursework, and to Sonos for providing the time and flexibility to complete this degree and for valuing my research work.

There are people whose friendship carried me through the hardest days. Shruti, my first friend and peer at UW, guided me through the early months when everything felt uncertain. Sujata, my dear friend, offered advice and perspective that helped me think clearly about my career path and stay focused on my goals. To my friends beyond campus, thank you for your steady encouragement and for keeping my spirits up through the final stretch of this Ph.D.

To my parents, you planted this dream long before I understood it. This Ph.D. belongs to you as much as it does to me. Your love, blessings, and the strength I draw from your hard work and perseverance have guided me through every challenge. Your dedication over the years taught me to stay determined and to give my best, and that lesson has carried me to this point.

To my family, thank you for your steady support and for always checking in and sending positive vibes. It meant more than you know.

To my sons, Kian and Suhaan, thank you for the laughter that filled my study breaks and the hugs that reminded me what truly matters. Even on the hardest days, your joy kept me grounded. I promise to be more present in your moments ahead and to share with you what this journey has taught me about perseverance and following your dreams.

To my husband, this story would not exist without you. You carried the weight when I could not, watching the kids, making me tea every night so I could keep working, and holding faith when I began to lose mine. Those cups of tea became my fuel and your quiet way of saying 'keep going'. You made this possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Motivation and Problem Space

Natural Language Processing has transformed how researchers analyze text across many fields, but educational research presents unique challenges that existing methods struggle to address. One of these challenges is that modern learning environments and research endeavors generate large volumes of short, text-based responses such as reflections, discussion posts, and open-ended survey answers, but the deeper meaning within these texts often remains hidden and difficult to extract.

Educational applications of NLP can be broadly categorized into two distinct domains: *educational assessment* and *education research*. Educational assessment focuses on practical, course-specific improvements where educators analyze student feedback (such as course evaluations) to identify specific adjustments needed for their particular teaching context. In contrast, education research seeks to explore distinct patterns and to identify generalizable patterns across multiple courses, institutions, or contexts to advance theoretical understanding and inform broader educational practices.

This dissertation focuses specifically on *education research* directed toward the systematic analysis of short, prompt-guided survey responses used to understand student experiences and perspectives across contexts. Unlike educational assessment, which aims for course-specific actionable insights and summative evaluation measures, education research requires identifying themes that extend beyond individual contexts to support generalizable findings. These texts require qualitative interpretation rather than numerical scoring, and the goal is to discover patterns that can inform educational theory and practice broadly.

Topic modeling methods can organize and cluster words statistically, but their outputs must be aligned with human judgment to capture meaning accurately and ensure generalizability. This work explores how NLP can be integrated into education research to support

rigorous qualitative analysis of such data. It explores multiple types of topic models with the goal of developing decision and work flows in a human-in-the-loop topic modeling framework. This framework combines computational models with expert validation to transform short, semi-structured student responses into reliable and interpretable themes suitable for research contexts. Table 1.1 defines terms used throughout the dissertation.

Table 1.1: Key Terms and Definitions

Term	**Definition**
Topic	Word-weighted cluster from a topic modeling algorithm
Theme	Human-validated construct that explains meaning across texts
Semi-structured short text	Prompt-guided response of ≤ 200 words that uses domain-specific vocabulary
Human-in-the-loop	Domain expert input at specific stages of the analysis pipeline

1.2 Brief History of NLP

Natural Language Processing (NLP) is an interdisciplinary field combining linguistics, computer science, and artificial intelligence (AI). NLP enables computers to preprocess, analyze, and interpret large volumes of spoken or written language data. Its goal is to allow computers to process language similarly to how humans do to support various tasks across different fields [1]. Over time, NLP has evolved from early theoretical studies into a key part of modern AI systems. Figure 1.1 provides an overview of how NLP has grown from early rule-based systems to modern, sophisticated neural methods.

Historically, NLP began in the 1950s with the goal of automating language translation. This early work laid the foundation for broader efforts in computational linguistics [1]. The first major milestones in the broader use of NLP came in the 1970s and 1980s, with the

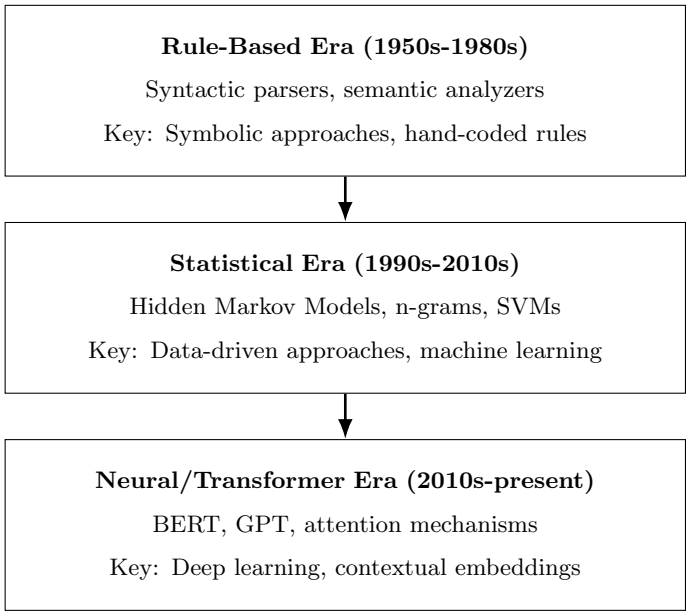


Figure 1.1: NLP evolution timeline: from rule-based systems to modern neural approaches.

creation of syntactic parsers and semantic analyzers [2]. In the 1990's, NLP experienced continued growth when statistical methods were introduced. These methods allowed more advanced language modeling for a wider range of practical applications.

NLP development has progressed through three main phases, as shown in Figure 1.1 [1, 3]. The first phase was rule-based, grounded in linguistic theory and symbolic AI, where systems relied on hand-written grammatical rules [2]. The second phase introduced statistical NLP, which shifted from rules to data-driven learning [3]. Researchers began using large collections of text to estimate how words and phrases occur together and improved the ability of computers to model real language use [4]. Later, neural language models extended these ideas by representing words as numerical vectors that capture their meanings and relationships and allow models to recognize similarities between words and handle longer text contexts [4]. The current and most advanced phase is based on neural networks and transformer models. Transformers use attention mechanisms to handle longer and more complex contexts in language data. This gave rise to large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers) [5] and GPT (Generative

Pre-trained Transformer) [6]. These models greatly improve the accuracy, versatility, and effectiveness of language understanding and generation, and this makes NLP increasingly valuable across many applications and domains.

Today, NLP continues to expand its reach across diverse domains (Figure 1.2), supporting tasks such as text summarization, sentiment analysis, machine translation, and conversational systems [1]. In healthcare, NLP assists with clinical documentation and patient data analysis. In business, it supports customer relationship management through sentiment analysis and automated customer support. These advances have made NLP increasingly relevant to fields that analyze large volumes of text, including education.

1.3 NLP in Education Research

The use of NLP in education has been extensive, particularly in assessing and categorizing student learning. Assessment determines the quality or level of learning, while classification identifies what students understand without judging its quality.

Automated assessment is an attractive solution for large student populations, and one of the most common applications of NLP in education is the assessment of student writing in the Test of English as a Foreign Language (TOEFL) [7]. NLP is used to evaluate grammar, mechanics, word usage, complexity, style, and organization of student essays. NLP-based assessments have demonstrated remarkable agreement with teacher grades, ranging from 70% to 90%, when combined with neural networks [8]. In terms of vocabulary assessment, NLP tools can explain 44% of the variance in students' vocabulary knowledge scores when compared to traditional vocabulary tests, and this result demonstrates that automated text analysis can capture a substantial portion of students' vocabulary proficiency [9]. These developments produced Automated Essay Grading (AEG) or Automated Essay Scoring (AES) systems, some of which have been adapted for engineering education and are particularly helpful for English as a Second Language (ESL) students [10, 11].

While NLP substantially reduces grading effort for instructors, it also supports identifying and categorizing what students learn. For example, NLP has been used to categorize student responses to a physics measurement problem by assigning student responses to one of three conceptual categories, with the same level of agreement as a human coder [12].

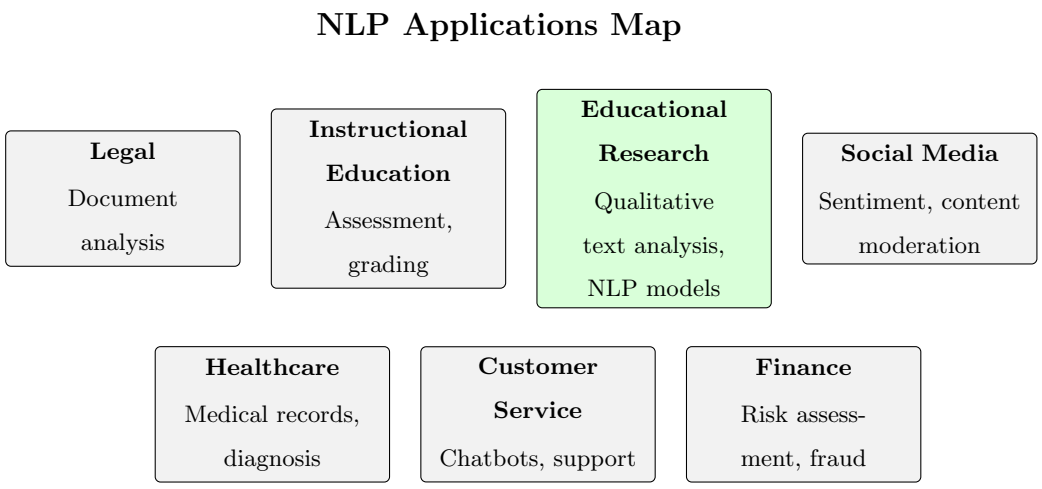


Figure 1.2: NLP Applications across Domains

However, NLP was unable to perform finer classifications using subcodes used by human graders [12]. In engineering, NLP has been applied to classify documents produced by student design teams, to predict the success of these teams, and to analyze student writing in various disciplines [13, 14, 15]. These efforts reinforce the notion that NLP is not only useful in producing a numerical or quantitative judgment of student work (i.e., a grade) but can also assist teachers in understanding differences among students and their learning.

These applications focus on artifacts created within classrooms for instructional or grading purposes. *Education research*, however, seeks insights that extend beyond a single course or cohort or institution. It aims to generate generalizable knowledge about learning processes. In this broader research context, NLP analyzes text that students produce in response to research instruments rather than coursework. Such instruments include structured or semi-structured surveys designed to study constructs such as motivation [16, 17], identity [18], goals [19], and persistence [20].

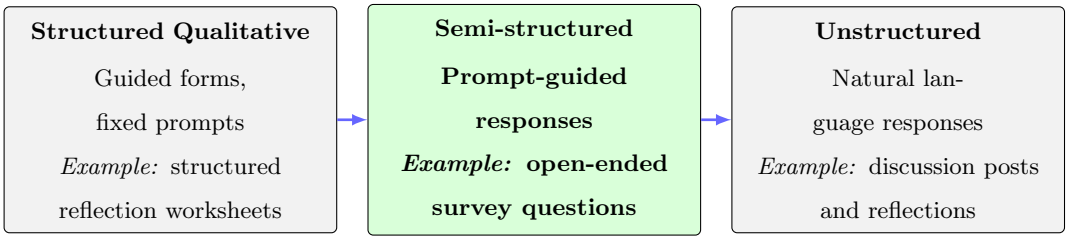
NLP has also advanced educational research through sentiment and affect analysis. Studies use sentiment classifiers to detect emotional tone in student reflections and feedback and have reported high accuracy levels, ranging from 75% to 99% when compared to traditional human approaches [21, 22]. More recently, sentiment analysis has been applied to

a finer grained analysis of sentiment detecting tones of joy, fear, sadness, anger, analytic, confident, and tentative in student generated stories of their lived experiences [23]. These findings help researchers understand students' affective states, yet emotion alone does not explain the reasoning behind their perspectives. Educational research therefore requires analysis that interprets the content and meaning of student text, not only its sentiment. NLP can assist by identifying linguistic patterns or conceptual groupings within qualitative text, which human researchers can then interpret through established frameworks [24, 25].

Most NLP work in education emphasizes grading or sentiment classification, while fewer studies link computationally derived text clusters to researcher-interpreted qualitative categories. This dissertation addresses that gap by examining how NLP models can interact with expert judgment (i.e., human-in-the-loop workflows) to analyze short, prompt-guided responses collected for educational research. Because the structure of these responses determines which computational and qualitative methods are appropriate, the next section describes the properties of these research texts.

1.4 Data in Educational Research

Data used in educational research differs significantly from data typically seen in other contexts, such as social media or short text communications [26]. Figure 1.3 illustrates the range of short text formats used in qualitative educational research and identifies the semi-structured short texts that are the focus of this study.



This figure addresses short text formats (≤ 200 words). Longer-format qualitative data (interviews, focus groups, and essays) are outside the scope of this study.

Figure 1.3: Types of qualitative data in educational research.

A defining characteristic of educational research text is its use of domain or discipline-specific language. Students responding to surveys or prompts within an educational setting often use specialized vocabulary and terminology related to their field of study. For instance, engineering students may discuss "problem-solving approaches", "lab experiences", or "interactive methods" in ways uniquely relevant to their educational context. This disciplinary vocabulary requires analytical tools that can capture specialized language while maintaining fidelity to the academic context.

Educational research data also differ from other short-text sources such as tweets, comments on social media platforms, or general online reviews [26]. Unlike tweets or social media comments, which are entirely unstructured and unrestricted, survey and other text-based responses are guided by specific prompts. These prompts focus student responses on specific themes or educational contexts, which results in more coherent and contextually relevant data [27] that supports structured analysis [28]. Moreover, the semi-structured nature of educational research data strikes a balance between structured qualitative formats, such as fixed reflection worksheets, and entirely unstructured qualitative formats, like free-form social media posts [29]. Educational research survey responses provide enough flexibility for genuine, rich expression while still maintaining enough structure to make qualitative analysis both possible and meaningful [27].

In summary, the educational research data emphasized in this dissertation is distinct due to its semi-structured format, discipline-specific language, and guided nature, positioning it uniquely between structured qualitative data (fixed reflection worksheets) and unstructured qualitative data (free-form social media posts).

Given this data regime, the literature leaves several methodological questions open.

1.5 Gap and Thesis Contributions

1.5.1 Research Gap

Prior work in education uses NLP heavily for automated assessment and for sentiment [8, 21, 22], but gives less attention to how topic models can support education research by turning model outputs into human-validated themes from short, prompt-guided student text. Short

responses with domain terms and semi-structured prompts create sparse features, unstable topics, and uneven interpretability [26]. Methods to link topics to themes and to place experts in the loop are not well standardized across classic and neural topic models [27, 30].

1.5.2 Thesis Statement

A structured, human-in-the-loop topic modeling pipeline produces reliable and interpretable themes from short, prompt-guided educational research texts. This thesis is examined through two research questions: RQ1 evaluates which topic modeling approaches perform best on these texts, and RQ2 identifies where and how domain expert input most improves the value and accuracy of the analysis.

1.5.3 Research Questions

This dissertation investigates how Natural Language Processing (NLP) methods can be effectively applied to qualitative educational data, particularly short, semi-structured student responses to open-ended survey prompts. The goal is to assess the suitability, limitations, and potential of a range of NLP tools in this context. The study also evaluates the role of human expertise in complementing automated analysis and considers the methodological implications of using NLP for educational research.

Two primary research questions guide this work:

1. **What are the strengths and weaknesses of five topic modeling approaches for short educational research texts (k-means, LDA, NMF, BERTopic, zero-shot classification)?** This question focuses on comparing the performance of widely used NLP algorithms on short, semi-structured student responses.

2. **How can a domain expert be best integrated into the analysis of short educational research texts to improve the value and accuracy of NLP methods?** This question explores different strategies for combining human expertise with automated models. It asks whether and how domain experts can enhance model performance, either by refining input data, evaluating output topics, or converting topics

into higher-level qualitative themes. This human-in-the-loop approach is evaluated for its effectiveness in improving both interpretability and model outcomes.

These research questions establish the scope of investigation and provide the framework for evaluating different NLP approaches. The following contributions demonstrate how this research addresses each question through systematic analysis and validation.

1.5.4 Contributions

To address these research questions, this dissertation makes three main contributions:

C1: Established performance baselines by comparing five topic modeling methods. This study compared k-means, LDA, NMF, BERTopic, and zero-shot classification on the same datasets using a shared preprocessing pipeline and common evaluation metrics. The results show that no single method works best for all datasets. Instead, the analysis identifies which method fits each dataset type, providing clear guidance for method selection based on dataset characteristics.

C2: Identified where expert input has the highest impact. The study analyzed expert input at different stages: preparation (domain-specific stopwords), prompt design (for zero-shot classification), and theme interpretation (mapping topics to themes). These interventions improved accuracy by 3 to 14 percentage points, showing where expert time produces the most valuable results. This analysis helps researchers prioritize expert involvement to maximize both accuracy and efficiency.

C3: Developed the NLP-Assisted Thematic Analysis framework. The study created a structured, six-stage framework with decision flows for model selection, work flows for method optimization, and documentation of preprocessing rules, theme formation rationales, and parameter selection criteria. These resources support replication and adaptation to new datasets, disciplines, and research contexts beyond the engineering education datasets analyzed here.

1.5.5 Generalizability and Scope

The methods target datasets with three traits: (1) short responses, (2) prompt-guided but open answers, (3) domain-specific vocabulary. Findings are meant to transfer to any dataset that shares these traits, not only the datasets used here.

These contributions provide a systematic framework for addressing the research questions through empirical analysis and validation.

1.6 Chapter Roadmap

The dissertation is organized into eleven chapters.

Chapters 2–3: Background – These establish qualitative research methods (Chapter 2) and the topic modeling literature (Chapter 3) that motivate the study.

Chapters 4–6: Method – These describe the topic modelling pipeline used to address the first research question associated with this dissertation: data collection and preprocessing (Chapter 4), word and document embeddings (Chapter 5), and the five topic modeling approaches along with the evaluation framework used to assess them (Chapter 6).

Chapters 7–8: Results – Chapter 7 presents results from the five discovery-first methods (k-means, LDA, LSA, NMF, and BERTopic) across all three datasets. Chapter 8 presents results from zero-shot classification on the Peer Support dataset.

Chapter 9: Discussion – This chapter interprets results across all five methods, examines cross-method patterns, and provides evidence-based recommendations for practitioners, effectively exploring and answering RQ1.

Chapter 10: Human in the Loop – This chapter explores human-in-the-loop strategies for optimizing the use of topic models in educational research and presents decision flows and work flows that guide the practitioner in how and when to intervene in NLP-based analysis of short texts, thereby addressing RQ2.

Chapter 11: Conclusions – This chapter synthesizes results and discusses implications for both NLP and education research.

Chapter 2

BACKGROUND: QUALITATIVE RESEARCH METHODS

This chapter defines the qualitative foundation for analyzing short, prompt-guided student responses within educational research contexts. It reviews different types of qualitative analysis methods and evaluates their suitability for the semi-structured short text data examined in this dissertation. The focus of this dissertation is on systematic studies in education research that extend beyond single classrooms, using semi-structured student responses. This focus differs from educational assessment, which measures individual outcomes; here, the aim is to identify patterned meanings that inform theory and practice across populations.

Education research seeks to expand knowledge about how people learn and improve teaching practices[29]. It aims to build more inclusive learning environments that accommodate the voices of many different kinds of students[29]. To achieve these goals, researchers must answer questions about what, how much, and how often students are doing, thinking, or feeling certain things, as well as understanding how and why they are doing, thinking, and feeling what they are[29]. The first set of questions is typically addressed by quantitative research methods, while the latter is addressed by qualitative methods[29].

While both quantitative and qualitative approaches are important for advancing knowledge and practice in education, quantitative methods have remained far more prevalent in the peer-reviewed literature than qualitative methods. This is also true for mixed methods, which strategically combine the two types of research methods [31, p. 13]. Nevertheless, while "qualitative research understandably remains a minor pursuit in many broad fields, it is increasingly accepted in all." [31, p. 13].

2.1 Qualitative Research Methods

This chapter considers several qualitative research traditions (Table 2.1) and their applicability to the datasets analyzed in this dissertation:

- **Ethnography:** Ethnography involves long-term immersion and observation to understand cultural practices and meaning systems[29] and can provide rich contextual insights into individual experiences in education. Such observations can produce large quantities of text that must first be broken down by a researcher familiar with the underlying fieldwork into meaningful chunks that are compatible with natural language processing.

- **Phenomenology:** Phenomenological research focuses on lived experience through deep interviews and personal narratives to distill the essence of phenomena[29]. While powerful for understanding individual experiences, it relies on lengthy, unstructured descriptions from few participants and like ethnography, must be broken down strategically into shorter texts to benefit from the NLP techniques explored in this dissertation.

- **Narrative Inquiry:** This approach constructs chronological narratives from individual life stories and experiences[29]. While providing deep, individualized insights, it focuses on few participants and does not typically produce short texts for analysis.

- **Case Study:** Case studies investigate bounded systems in depth using multiple data sources[29]. While excellent for answering "how" or "why" questions in context, they focus on single cases rather than broader analyses across hundreds of responses that stand to benefit from automation through current natural language processing techniques.

- **Grounded Theory:** Grounded theory develops theory through iterative coding and comparison until theoretical saturation[29]. While ensuring rigorous links between

data and theory, it is time-intensive and aims to produce substantive theory rather than descriptive themes, which extends beyond the scope of this dissertation.

- **Content Analysis:** Content analysis systematically codes textual data into categories or counts and provides structured summaries[29]. However, it operates at a micro level and often reduces qualitative data to frequency counts. As a result, it is vulnerable to losing the context and meaning that the interpretive approach adopted in this dissertation seeks to preserve. While NLP tools and topic models may be useful for identifying content relevant to particular research questions, it typically does not require the level of semantic sophistication that many modern NLP topic models offer.

- **Thematic Analysis:** Thematic analysis is a flexible method for identifying and analyzing patterns (themes) within data[27]. It goes beyond content analysis by considering how the use, frequency, and meaning of words map to conceptually grounded themes. Further, unlike methodologies tied to specific theoretical frameworks, thematic analysis can be applied across different research contexts and is well-suited to semi-structured short texts like the survey responses analyzed in this dissertation, providing "enough flexibility for genuine, rich expression while still maintaining enough structure to make thematic analysis both possible and meaningful"[29].

Of these qualitative analysis techniques, thematic analysis is the most closely aligned with evaluating short texts while both considering meaning within those texts and also minimizing the amount of preliminary analysis and processing that must be done to prepare texts for NLP-based analysis.

Thematic analysis is specifically chosen as the guiding qualitative method for this study for the following reasons:

- **Applicability to Education Research Data:** Thematic analysis can handle a large volume of short responses by condensing them into a manageable set of themes. It allows the summarization of key features of the entire dataset while preserving meaning[29]. For instance, instead of reading hundreds of responses in isolation, the

Table 2.1: Suitability of Qualitative Research Methods to This Study

Method	**Primary Goal**	**Typical Output**	**Suitability for This Study**
Ethnography	Cultural immersion and understanding	Narrative of shared practices and norms	Too immersive for survey-based data
Case Study	Deep analysis of a bounded system	Analytic story across multiple data types	Partial overlap, not fully aligned
Phenomenology	Capture of lived experience	Essence of a phenomenon across participants	Ideal for interviews, less so for text
Grounded Theory	Data-driven theory generation	Emergent model explaining social process	Overly generative for this scope
Narrative Inquiry	Study of personal stories and sequencing	Temporal arc and meaning structure	Complementary, but not central here
Content Analysis	Systematic coding into categories or counts	Structured summaries and frequency counts	Reduces meaning to counts; limited interpretive depth
Thematic Analysis	Identification of patterned meaning	Themes with codes, definitions, exemplars	Core method in this dissertation

method can identify that many responses revolve around recurring concerns or experiences across the dataset. This method thus provides an organized overview of student perspectives.

- **Flexibility and Depth:** Unlike a purely counting-based content analysis, thematic analysis allows interpretation beyond word frequencies to the concepts or issues the words represent[29]. This is important in education research, where two students might use different words to express the same underlying concern. Thematic analysis captures such underlying patterns. It also allows mixing inductive coding (letting themes emerge from student voices) with deductive coding (checking against expected issues from literature or theory). This offers a rich analysis that can connect to education research.

- **Human-Centered and Transparent:** Thematic analysis keeps the researcher closely engaged with the data at all times, which aligns with qualitative values of reflexivity and context. Braun and Clarke advocate it as a method that is relatively easy to learn and accessible even to those new to qualitative research[27]. This means it can be clearly explained and audited. The results (themes) are typically straightforward and intuitive, which makes it easier for educators and other stakeholders to understand. The aim is for themes to resonate with educators' intuitive understanding of student issues, thereby amplifying the impact of the research.

- **Efficiency with Rigor:** Thematic analysis can be augmented with computational tools (as will be done with NLP topic modeling) to increase efficiency without sacrificing rigor. It is less rigid than grounded theory because there is no need to follow iterative sampling to build the theory. So, algorithmic support can be incorporated in identifying candidate patterns. Yet, the method remains systematic enough to ensure credibility (through clear coding procedures and reliability checks). Thematic analysis effectively serves as a bridge between qualitative insight and quantitative structure. This is ideal for the mixed-methods aim of this dissertation (addressing both human interpretation and algorithmic analysis)[29].

Thematic analysis is particularly well-suited for combining with NLP-based topic modeling because both approaches share the goal of identifying patterns in text data. While topic models excel at discovering statistical co-occurrence patterns at scale, thematic analysis provides the interpretive framework needed to transform these algorithmic clusters into meaningful themes that address research questions. The next sections detail how this method is combined with NLP topic modeling and how machine-generated topics are distinguished from human-defined themes.

2.2 Thematic Analysis and Topic Modeling

2.2.1 Thematic Analysis Process

Thematic analysis involves a structured yet flexible process of coding qualitative data and identifying themes. The analysis follows the general six-phase guide outlined by Braun and Clarke (2006)[27]:

1. **Familiarization with the data:** Researchers read through all student responses multiple times and transcribe data if needed. The goal is to become deeply familiar with content and note initial impressions.

2. **Generating initial codes:** The analyst works through the data systematically, highlights sections of text, and assigns codes (concise labels) that describe the content or meaning of each segment. A code might be a short phrase like "time management struggles" or "lab equipment access" attached to a sentence where a student mentions that idea. All data relevant to each code are collated for review.

3. **Searching for themes:** The researcher examines codes and groups related ones into candidate themes. For example, codes like "exam anxiety," "fear of failing," "stress before tests" cluster into "Assessment Anxiety." At this stage, the researcher organizes codes into potential themes and gathers all supporting data extracts for each theme.

4. **Reviewing themes:** The research team checks each candidate theme against the full dataset, collapses overlaps, splits overly broad themes, and discards unsupported

ones. The researcher asks: Do these theme candidates make sense given the actual data? Are the data extracts within each theme cohesive? For instance, a broad "Support Needs" theme might be split into "Academic Advising" and "Peer Study Groups" when closer reading reveals distinct types of support. The outcome is a set of tentative themes that reasonably reflect the data. Researchers often visualize this structure as a thematic map linking codes to themes.

5. **Defining and naming themes:** Once the thematic structure is finalized, the research team defines each theme clearly and gives it an informative name. The definition spells out what is unique about that theme and what aspects of the data it covers (and sometimes what is excluded). For instance, "Assessment Anxiety" might be defined as "students expressing fear, stress, or nervousness specifically about quizzes, exams, or graded evaluations." This phase ensures that themes are distinct from one another and tied to specific meanings in the data.

6. **Completing the analysis:** In the final phase, the researcher selects vivid, representative quotes from students for each theme, relates the themes back to research questions and literature, and crafts the narrative that answers those questions. When using automated topic models (NLP-based) to assist in the process, completing the analysis can involve iteration of topic model parameters and (potentially) exploring multiple topic models to optimize performance outcomes. The end deliverable of successful analysis is a scholarly report (such as this dissertation chapter) that presents the themes and interpretations with supporting evidence from the data.

2.2.2 Role of Topic Modeling

Topic modeling refers to a class of machine-learning techniques that identify clusters of words that co-occur within documents. For example, Latent Dirichlet Allocation (LDA) topic models represent each topic as a probability distribution over words and each document as a mixture of topics. An LDA model might output a topic characterized by words like "project, implementation, build, hands-on, prototype" and indicate that a particular student response

is 80% associated with this topic, which could be interpreted as "hands-on projects" based on those keywords. Other topic modeling techniques, such as k-means clustering, Non-Negative Matrix Factorization (NMF), BERTopic, and zero-shot classification, operate on similar principles but differ in their underlying mathematical assumptions. Subsequent chapters discuss these algorithms in detail. Topic models are essentially "statistical methods that analyze the words of the original texts to discover the themes that run through them"[32]. The goal aligns well with that of thematic analysis to find underlying themes in text, but it achieves this through automated statistical pattern-finding rather than deep reading.

Topic modeling is used to support and enhance thematic analysis by providing initial coding structures rather than starting from scratch with open coding. After preprocessing student responses and applying multiple topic modeling algorithms, machine-generated topics act as first-pass categorization, clustering responses that use similar terminology. For instance, one topic might correspond to "mentorship & support" with top words {mentor, advice, guidance, help}. Once topics are generated, thematic analysis builds on those topics rather than starting from scratch in the search for codes and then aggregate groups of codes (themes). Domain experts review all topics by reading sample responses to understand what each cluster represents beyond keywords, since topic models statistically group words rather than interpret meaning. This expert review corresponds to Braun & Clarke's familiarization and coding phases, except that codes were partly suggested by the algorithm.

Domain experts then compare initial topic labels to identify overlaps or splits, merging similar topics (e.g., "difficulty with time management" and "balancing coursework and personal life" into "Time Management Challenges") or splitting overly broad topics into distinct themes. This manual merge/split process ensures that emerging themes are conceptually grounded. This approach tackles the challenge of manual coding large datasets while maintaining interpretive depth essential for qualitative analysis[33].

2.3 Differences Between Topics and Themes

The terms topic and theme can sound similar but they are not the same. They both refer to underlying patterns in the data but have specific meanings and roles. Understanding their differences is key to appreciating why expert interpretation is needed to move from one to

the other[33].

2.3.1 Defining Topics

In topic modeling, a topic is a statistically defined pattern of words. Formally, a topic is a set of terms (or synonyms to those terms) that tend to appear frequently together across the corpus[32]. For example, a topic emerging from a dataset containing student feedback of teaching might be represented by high-probability words like "lecture, slides, notes, textbook." Each topic is typically unlabeled when produced and is essentially a mathematical abstraction (a distribution over the vocabulary) until it is interpreted by researchers.

Documents in a dataset can be associated with multiple topics in varying proportions. This means a single response might be determined to be 60% Topic A and 30% Topic B according to a particular topic model. Topics are thus latent constructs that summarize common co-occurring words, phrases, or segments of text with similar meaning in the data. From a computational perspective, algorithms generate topics without direct human input because they decide groupings based on word patterns and, in more sophisticated models, limited consideration of what those word patterns mean. As a result, topics may sometimes correspond to what a human would consider a coherent subject, but other times they may not.

The main point is that a topic is defined by lexical similarity, not necessarily by semantic coherence. It provides clues to a theme but is not equivalent to one.

2.3.2 Definition of Themes in Qualitative Analysis

A theme in qualitative research is a pattern of meaning or concept that emerges from interpreting the data[33]. According to Braun and Clarke (2006)[27], a theme "captures something important about the data in relation to the research question, representing some level of patterned response or meaning within the dataset." Unlike a topic's list of words, a theme is typically described in natural language as an idea, often a short phrase or sentence, and it is backed by a collection of data excerpts that exemplify it.

For example, a theme might be "Sense of Belonging" if many students describe feeling connected to their peers, instructors, or learning community, even if they express it in different ways. Researchers identify such a theme through careful interpretation, considering context, tone, and meaning. Each theme is explicitly defined and conceptually aligned with the research questions or important aspects of participants' experiences.

2.3.3 Bridging Topics to Themes

The distinction between topics and themes anchors the comparison of computational and qualitative approaches[33]. Structurally, themes and topics often relate in complex ways including but not limited to:

- **One-to-One Alignment (Rare):** In some cases, a topic model might produce a topic with keywords related to a specific concept, such as {"homework, assignments, due, late, deadline"}. Upon expert review, researchers might find that most responses in this topic explicitly mention time management issues with assignments. This topic could map directly to a theme like "Assignment Time Management Challenges," representing a rare case of near-perfect alignment between algorithmic clustering and human interpretation.

- **One Topic to Multiple Themes:** A more common scenario occurs when a topic model generates a topic with mixed keywords, such as {"difficult, professor, accent, understand, class, hard"}. Initial review might suggest this represents "Communication Issues with Instructors." However, detailed examination of the responses could reveal two distinct patterns: some responses mention difficulty understanding due to language barriers, while others reference challenging course content. Researchers might split this single topic into two themes: "Language/Communication Barriers" and "Course Content Difficulty," and they recognize that these represent fundamentally different student support needs.

- **Multiple Topics to One Theme:** Analysis might reveal that related concepts are fragmented across multiple algorithmic topics. For instance, one topic might

contain keywords {"mentor, advisor, guidance, career"} while another contains {"help, support, tutor, office hours"}. While the algorithms treat these as distinct clusters, domain knowledge might recognize both as aspects of academic support systems.

Researchers could merge these into a single theme "Academic Support and Mentorship," and they recognize that students use different vocabulary to describe similar support needs. This decision would be informed by domain knowledge that emphasizes the importance of holistic support systems.

- **Topics That Are Not Themes:** Some topics might be identified as artifacts rather than meaningful themes. For example, a topic dominated by generic words {"student, university, semester, year, college"} might contain responses that provide no actionable insights about the research question. These responses would be excluded from the final theme set. This decision requires human judgment to distinguish between meaningful patterns and survey response artifacts.

An example of how topics map to themes in the context of education research is shown in Figure 2.1. The fundamental differences between topics and themes highlight why expert interpretation is required to create accurate mappings from topics to themes[33].

While topic models excel at scale and identify co-occurrence patterns, produce consistent clusters rapidly, and reveal associations that manual coding might miss (e.g., "difficult" with both "professor" and "class"), they lack contextual understanding and often privilege dominant patterns, overlooking low-frequency but meaningful insights.

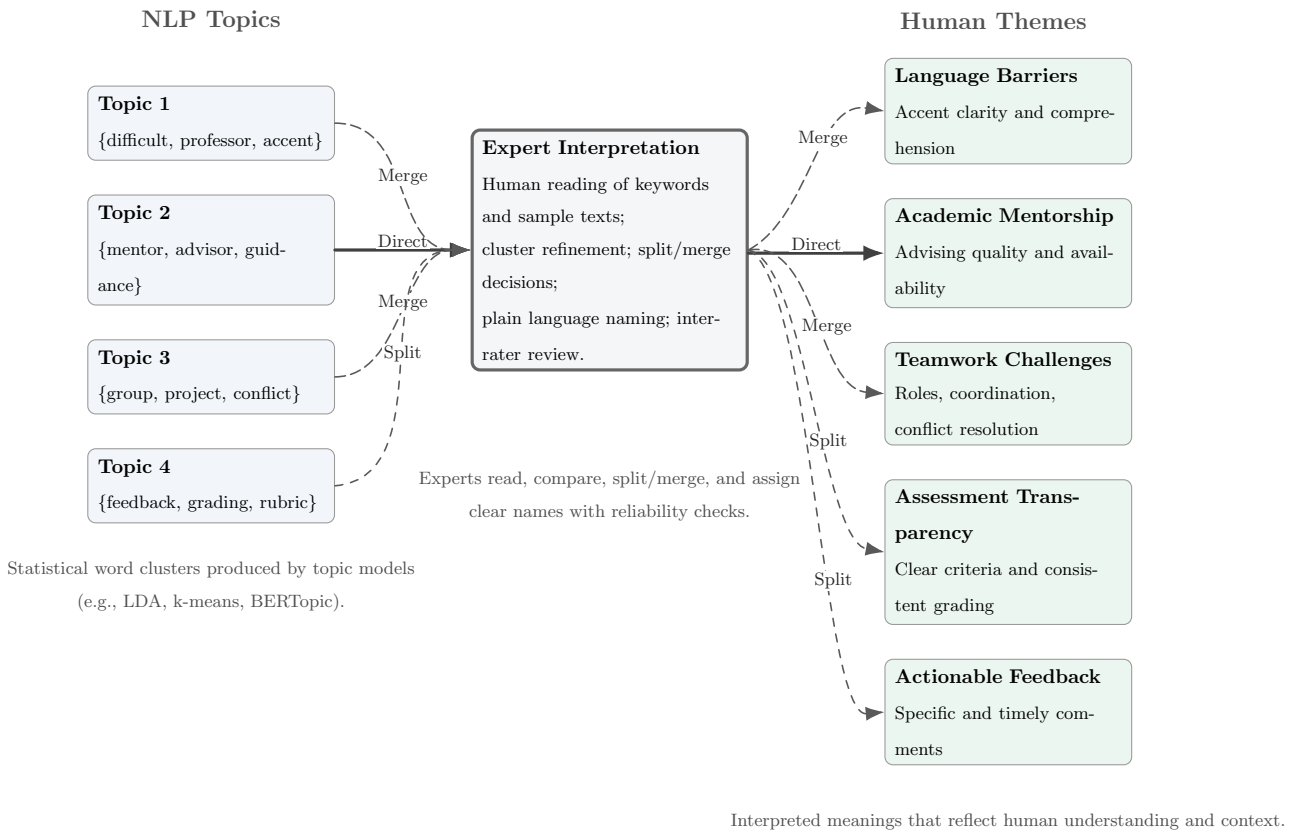


Figure 2.1: Relationship between NLP Topics and Human Themes

Chapter 3

BACKGROUND: TOPIC MODELING

In the previous chapter, the most common qualitative research methods were explored and the rationale for emphasizing thematic analysis as the focus of this dissertation was presented. In this chapter, the NLP-based topic models chosen to explore NLP-driven thematic analysis (with human intervention as appropriate) are examined in the context of the existing literature in order to justify their selection for analyzing short texts in education research.

Artificial intelligence (AI), specifically natural language processing (NLP), offers potential solutions to these challenges. In particular, topic models are unsupervised NLP tools that explicitly seek similarities in the relationships between data points (i.e., words or sequences of words) [34], thereby making these models well suited for identifying patterns, an end goal that it shares with thematic analysis. A wide range of algorithms and approaches are available to capture these patterns and while a full review of topic models is outside the scope of this study, the comparative analysis presented herein examines five fundamentally different approaches to understanding and classifying language based on clustering, probabilistic, algebraic, neural network, and zero-shot classification techniques. Clustering offers a simple approach that groups similar uses of identical words together into individual clusters (i.e., topics), essentially treating words with little regard to context or the possibility that multiple topics may be present in a document. Probabilistic analysis groups documents with the same sets of words (i.e., word co-occurrences) together into topics but still fails to capture word (semantic) meaning even while considering some context of how words are used in combination with one another to express ideas. Matrix factorization (algebraic) methods seek to reduce superfluous information in text (i.e., noise) and actively seek out similarity instead of sameness among language patterns in documents, thereby offering additional insight into semantic meaning. And finally, more complex topic models based on

neural networks offer the opportunity to capture the position of words in documents and their local and global (long-range) context by leveraging their pre-training on large and diverse datasets. Beyond these unsupervised approaches, zero-shot classification offers a fundamentally different paradigm in which experts define themes before classification begins and the model assigns documents to those predefined categories without task-specific training. Representing the straightforward to the complex, these five types of approaches possess different benefits and drawbacks which must be explored and understood for specific types of datasets and specific contexts to develop actionable guidelines for their use.

3.1 Topic Models using Clustering Techniques

Traditional clustering approaches to topic modelling partition documents into groups based on similar word patterns without regard to what a word means (semantics) or the words which surround it (context). Among clustering approaches, k-means is computationally simple and the most widely used clustering technique in scientific and industrial applications including data mining [35]. k-means works by partitioning data into a pre-specified number of non-overlapping clusters by minimizing the distance between each document's (numerical) vector representation and its corresponding cluster centroid. In addition to ignoring context and semantics, k-means is also limited by assigning each document to one and only one cluster and requiring all clusters to be spherical. Thus, poor performance can result when clusters vary significantly by size and density, when multiple topics overlap, or when variability exists in the initial partitioning of the data [36, 37].

Despite these constraints, k-means is useful in short text topic modelling under specific conditions. For example, in a comparative analysis of inherently noisy and short Twitter and Reddit datasets, k-means proved to be the best performing clustering method in two of three datasets when documents were embedded into vectors using word2vec [36], a two-layer, neural network pretrained to learn word associations and semantics using large datasets [38, 39]. In this same study, k-means also outperformed most clustering methods regardless of word embedding technique [36]. The conditions which enabled top level performance included incorporating considerations of word meaning (semantics) through the use of word2vec embedding, confining the data to a relatively narrow set of topics

(discussions of Australian politics), and extracting ground truth from frequently occurring hashtags in documents rather than manual topic assignment [36]. Similar performance levels have been demonstrated for tweets associated with highly specific topics such as the World Cup soccer tournament, where k-means produced comparable results to the more advanced Non-Negative Matrix Factorization (NMF) approach after outlying points were removed to reduce noise [40]. Well-separated topics are also compatible with k-means clustering as illustrated by a study that analyzed health-related tweets and emails and found that k-means performed comparably to probabilistic models [37].

In addition to being useful in and of itself for capturing topics in data corpora under certain conditions, k-means can also serve as an effective baseline for topic modelling by establishing a minimum level of performance for comparison to more sophisticated techniques. Failure to perform at the level of k-means or beyond can suggest an issue with model parameter selection and optimization, a fundamental mismatch between the topic modelling approach and the underlying data corpus, unusual data characteristics that lend themselves to k-means, or a combination thereof. For these reasons as well as the appeal of simplicity and computational efficiency, k-means is included in this comparative analysis.

3.2 Topic Models based on Probability

Like clustering methods, probabilistic topic models do not require any pre-existing information to identify hidden topics in texts. These models capture rudimentary global context by favoring word co-occurrences (i.e., the same words used together) across documents to identify topics. Of the probabilistic models available for topic modelling, Latent Dirichlet Allocation (LDA) is the most common [30] where the Dirichlet is a distribution of probability distributions chosen to model how topics are mixed in documents and how words are mixed in topics in such a way that the resulting mixtures are sparse and realistic. By leveraging the Dirichlet Distribution and evaluating which words and how often words co-occur in different documents, LDA determines how likely (probable) each possible topic is represented in each document, thereby allowing documents to exhibit multiple topics in varying proportions. While LDA ignores word order and local context which limits its capacity for capturing the meaning of words, it has nevertheless been widely used to sort and classify

documents.

The negative impacts of LDA's limited capacity for understanding semantics are minimized when analyzing data corpora containing longer texts with rich vocabulary overlap. Under these conditions, LDA has been able to successfully identify themes that align with human judgement and interest. For example, in the analysis of student feedback, Nanda et al. applied LDA to over 150,000 MOOC (massive open online course) learner comments averaging 16-17 words per document and found that topic assignments corresponded well with subsequent qualitative analysis, with the model successfully distinguishing themes related to course content, instructor quality, and learner satisfaction [41]. In another large study of over 110,000 student evaluations of teaching involving even longer texts, Sun and Yan found that LDA identified eight coherent topics; several of these topics correlated significantly with quantitative survey items, while others revealed themes that the survey questions did not address [42]. Similar alignment of themes with accepted principles of effective teaching practice have been demonstrated in the topic modelling of teacher self-assessment surveys, where LDA-identified topics mapped onto established frameworks for teaching practice [25]. In more applied research, LDA of learning behavior preferences among IT (information technology) students containing a rich vocabulary of 1,956 terms in 163 longer text documents not only revealed twelve distinct topics regarding what students expected of their teachers but clustering of the topics also grouped students according to their level of intellectual development. Doing so enables the development of practical recommendations for not only how instructors can meet student expectations but also how they can adjust teaching practice depending on where students are at [43]. These and related studies not only demonstrate the broad applicability of LDA to different types of data but also suggest that LDA performs particularly well when three conditions are met: documents contain sufficient words to establish co-occurrence patterns, vocabulary overlaps substantially across documents addressing similar themes, and topics are reasonably distinct rather than highly correlated.

Unfortunately, when texts are short, LDA can struggle. Short texts provide limited word co-occurrence information which leads to sparse document representations and unstable topic assignments. This unstable behavior has been consistently demonstrated across a

wide range of short text datasets including tweets, news tidbits, snippets from web-based searches, and texts extracted from question and answer forums [26]. Adding to this instability, the assumption that word order does not matter becomes problematic when word order is meaningful. LDA also assumes that topics are independent with little overlap; this is particularly problematic when analyzing student or teacher experience in education, because what goes on in the relatively small world of the classroom leads to themes or topics that inherently overlap or co-occur. Additionally, LDA's output can be sensitive to hyperparameter settings and initialization, which introduces variability across runs that complicates replication. To address some of these limitations, researchers have developed LDA variants like Correlated Topic Models which relax the independence assumption, Biterm Topic Models that aggregate word pairs to address sparsity, and seeded or guided approaches which incorporate domain knowledge through predefined word lists [44]. These extensions improve performance in specific contexts but add complexity and do not fully resolve the underlying dependence on word co-occurrence. The bottom line is that when documents contain the same word patterns that are meaningful to the goal of a particular topic modelling effort, LDA can be powerful, accurate, and efficient. Its inclusion in this comparative analysis allows direct evaluation of whether the probabilistic framework offers advantages over clustering for student feedback data, and whether its known limitations with short text are offset by its capacity to model mixed-topic documents.

3.3 Topic Models using Matrix Factorization

Matrix factorization breaks down a single matrix into multiple smaller ones which contain hidden (latent) information about features of the original matrix. Two of the most common matrix factorization methods used for topic modelling are Non-Negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA). LSA uses singular value decomposition (SVD) to break the original matrix that represents a document corpus down into three matrices while NMF iteratively adjusts two matrices to minimize the difference between their product and the original matrix. Both methods produce a document-by-topic and a topic-by-word matrix which can be analyzed to identify similarities between documents and similarities between words respectively. The capacity to identify such similarities enables

both methods to detect synonyms and integrate semantic meaning into the topic modelling process. The key difference between the two methods is that LSA allows for negative values in the component matrices which can capture more complex and nuanced language usage while NMF does not allow for such negative values, thereby making the resulting matrices easier to understand and interpret [45, 46].

In education, LSA has been particularly useful in evaluation (i.e., grading) of student work because it is adept at mimicking human judgements of similarities between texts. For example, LaVoie et al. [47] demonstrated a correlation of 0.94 between LSA-driven automated scoring and human scoring of responses from consequence tests which asked students to assess the outcomes of a decision or action. Similar levels of agreement between LSA and human judgement have been demonstrated for language proficiency tests [48], questions which require analytical reasoning of psychology lectures [49], and prompts requiring students to summarize technical texts [50]. In some tests such as those measuring creativity, LSA has even performed better at modelling the meaning and originality of responses than traditional human assessments [51]. Due to its alignment with human judgement, LSA has also been integrated into successful, commercially available products for automated essay scoring [52, 53]. Perhaps more in line with thematic analysis, however, LSA has also been shown to capture key ideas in student feedback. For instance, in a medical school course where students submitted anonymous reflections on gender differences in medicine, LSA was able to identify ten topics that captured over half of the variance in the data [54]. Reasonable topic coherence has also been demonstrated when using LSA to analyze student feedback on teaching quality in multiple courses over an academic year [55].

While LSA is able to capture more complex relationships between texts and thereby produce results that often coincide with human raters and human judgement, non-negative matrix factorization (NMF) offers the distinct potential for providing improved topic coherence (i.e., words within a topic frequently appear together in documents) and interpretability (i.e., words representing a topic are meaningful to human judgement) [56]. Greater coherence and interpretability, in turn, can make it easier to connect the output of NMF to themes that are both formulated by and relevant to human endeavors. The use of NMF for analyzing student work, student feedback, and other textual data from education and edu-

cation research has been very limited compared to studies involving LSA. However, a broad range of studies have demonstrated superior topic coherence with NMF-based topic models outside of education. Compared to LDA, NMF topic models show consistently better statistical measures of topic coherence for analyzing data corpora extracted from Google snippets [57], news headlines [57, 58], news descriptions [57], on-line question and answer forums [57], and restaurant reviews [59]. These benefits of NMF are not limited to statistical coherence measures but extend to both human and large language model measures of coherence as well [58]. Further, NMF has been shown to outperform LDA by producing more distinct topics with lower overlap in the analysis of news articles and Wikipedia pages, suggesting that NMF is better suited to niche or non-mainstream data [60] which are common in education and other specialized domains. As evidence of this, one of the few studies exploring the use of NMF in education showed, in examining student feedback regarding mental health and remote learning during the COVID-19 pandemic, that topics discovered by NMF were similarly coherent and more numerous and granular than those found by LSA and LDA [61]. This improvement in granularity has been duplicated in studies comparing k-means clustering with NMF modeling of World Cup tweets [62] and also of aviation accident reports, although with a corresponding decline in topic coherence compared to LDA [63]. Noisy data such as that obtained from tweets in real time also appears to compromise the ability of NMF to produce interpretable topics compared to other topic modelling techniques [64]. Thus, while the potential of NMF to produce more coherent, more distinct, and more interpretable topics has been clearly demonstrated in multiple previous studies, these benefits are not guaranteed and actual performance is highly dependent on the characteristics of the underlying data. NMF is included in this comparative analysis to evaluate whether matrix factorization offers coherence and interpretability advantages over probabilistic and clustering approaches for short student feedback data, and whether its known sensitivity to data characteristics limits those advantages when vocabulary is domain-constrained and thematically overlapping.

While matrix factorization techniques have proven and distinct advantages over LDA and traditional clustering techniques, they still rely largely on mathematical transformations rather than contextually rich language models. This limitation leads to the next evolution in

topic modeling: neural network approaches which rely on pretrained large language models (LLMs).

3.4 Topic Models based on Neural Networks

Topic models based on neural networks differ fundamentally from traditional approaches to topic modelling in how words are embedded into numerical representations. k-means, LDA, LSA, and NMF typically use bag-of-words embedding which convert words to numbers based on how often they occur in a document (i.e., frequency-embedding) or add a layer of computation to de-emphasize common words while highlighting rare words (e.g., TF-IDF – term-frequency, inverse-document frequency). These methods lack consideration for the order of words, long-range global context in a document, or other clues regarding semantics (meaning). This distinction matters because these methods often fail to recognize such semantic similarities as those between "helpful" and "supportive" if they never appear together in the dataset or are used in different language patterns. Sophisticated neural networks can overcome these barriers to capturing the meaning of words by considering how and where words are used in text relative to other words. Of these neural approaches to word embedding, BERT (Bidirectional Encoder Representations from Transformers) is widely used and represents words based on their surrounding context rather than fixed dictionary definitions [5]. BERTopic then integrates BERT into a topic modelling pipeline where the position of BERT-embedded words is encoded, the dimensions of the embedded matrix are reduced, the reduced data are clustered together into topics, and representative terms (e.g., words) for each topic are strategically extracted [65].

BERTopic performs well when documents express similar ideas with varied vocabulary. For example, in the analysis of MOOC student discussion forums, Khodeir and Elghannam [66] found that BERTopic outperformed both LDA and LSA approaches and was competitive with NMF in terms of both statistical topic coherence and topic diversity scores because it captured semantic content that co-occurrence-based models missed. This advantage reflects BERTopic's ability to recognize that phrases like "need help immediately" and "struggling with deadline" convey similar urgency despite sharing no words. BERTopic modelling in other domains has demonstrated similar performance benefits. Superior statistical topic

coherence of BERTopic has been demonstrated in the analysis of COVID-19 vaccination tweets using multilingual embedding models [67], adolescent health tweets [68], disaster-related tweets [69] and news headlines [70], but when topic coherence was evaluated using human or large language model evaluation, BERTopic consistently outperformed not only LDA but also NMF [71]. And, while BERTopic and NMF often perform similar to one another, BERTopic initially produces more topics which can offer unique insight into data that traditional methods cannot [40]. BERTopic can also complement NMF by providing finer granularity for certain topics while NMF obtains comparable granularity in other topics in the same dataset [61]. BERTopic, however, does not always perform better than traditional topic models. For instance, LDA has been shown to outperform BERTopic when comparing topic diversity from topic models in multiple studies [67].

BERTopic performance can suffer when considering texts in different languages or in specific domains because pretrained embeddings may not represent specialized terminology well. For instance, educational feedback often contains domain-specific terms such as "office hours" and "curve the exam" that general-purpose embeddings may miss. BERTopic also inherits the single cluster limitation of k-means where each document is assigned to exactly one topic, which can misrepresent texts that address multiple topics or belong to different themes. Further, generating embeddings can also be computationally expensive for large datasets, though this cost is incurred once and can be reduced through caching. Thus, when traditional topic models work and work well enough, the additional computational complexity introduced by BERTopic is not warranted. Superior performance by BERTopic, however, can not only be desirable for the task at hand but can also offer insight into the characteristics of the data that make it unsuitable for traditional models and in turn vulnerable to errors using any topic model. BERTopic is included in this comparative analysis to evaluate whether contextual embeddings improve performance on short student feedback texts relative to traditional approaches and whether the performance gains justify the added complexity. Even when BERTopic succeeds, however, its outputs remain statistically derived topics that require post-hoc expert interpretation to connect them to research-relevant themes, a mapping step whose quality depends on how well algorithmic topics align with the conceptual categories that matter to the researcher.

3.5 Zero-Shot Classification

Clustering, algebraic, probabilistic, and neural topic models share a common unsupervised workflow: each first discovers statistical or semantic patterns in a corpus and then requires a human expert to interpret those patterns as meaningful themes. Zero-shot classification (ZSC) inverts this sequence. In ZSC, a domain expert defines theme labels before any automated analysis begins, and the model assigns each document directly to those predefined categories without training on labeled examples from the target dataset [72]. This themes-first approach makes ZSC fundamentally different from unsupervised topic modeling; rather than asking what patterns exist in the data, ZSC asks how well the data fits categories that experts have already determined to be meaningful.

ZSC achieves this classification by leveraging natural language inference (NLI), a task in which a model evaluates whether one piece of text (a premise) logically supports another (a hypothesis). In thematic analysis of student feedback, each student response serves as the premise and each expert-defined theme description serves as the hypothesis. The model estimates how strongly the response supports each theme description and assigns the response to the theme with the highest support. Because this process relies on the semantic relationship between the response and the theme description rather than on word frequency or co-occurrence, ZSC can classify documents that share few surface-level words with their assigned theme, provided that the underlying meaning aligns [72]. This capacity to operate on meaning rather than word overlap gives ZSC a distinct advantage over frequency-based methods when students express similar ideas using diverse vocabulary.

ZSC builds on the broader capabilities of large language models (LLMs) that have transformed text classification in recent years. Transformer-based models such as BERT [5] and BART [73] learn contextual embeddings where the representation of a word depends on the words around it, placing texts with similar meanings close together even when they share few words. This shift from counting co-occurrence to encoding context is particularly valuable for short texts, where limited word overlap makes frequency-based methods unreliable. ZSC extends these capabilities by using a model that has been fine-tuned on NLI tasks to evaluate premise-hypothesis pairs, which enables classification into any set of categories

described in natural language without additional training. This generality distinguishes ZSC from supervised classifiers, which require many labeled examples and may not transfer across courses, institutions, or research contexts.

In education, ZSC has shown promise for scaling qualitative coding while preserving alignment with expert-defined frameworks. Parker et al. [74] used a zero-shot prompting approach with GPT-4 to classify course evaluation comments into predefined categories including pedagogy, curriculum, and sentiment without additional training. Agreement between the model and human annotators was comparable to agreement among human coders, and the outputs entered the analysis already aligned to the researchers' conceptual framework. Katz, Shakir, and Chambers [75] applied a similar prompt-driven workflow to classify peer evaluation comments into deductive categories of teamwork behavior drawn from an expert codebook, with human review to confirm boundary decisions. Fuller [76] had domain experts evaluate themes generated by ChatGPT from course evaluation documents and found substantial overlap with themes identified independently by those experts, though instructors still reviewed the results and cautioned against trusting the model without human oversight. In each of these studies, expert-defined labels shaped the analytic space before or during classification, which reduced the interpretive burden that follows unsupervised topic modeling and made the results immediately usable within established qualitative research practices.

ZSC offers several practical advantages for analyzing short educational texts. It requires no labeled training data from the target domain, which makes it applicable to new datasets without the cost of supervised annotation. Because theme definitions are authored by domain experts, the model's output reflects the conceptual categories that matter to the research rather than statistical artifacts that may or may not correspond to meaningful themes. ZSC also scales efficiently: once prompts are defined, the model can classify thousands of responses in a fraction of the time required for manual coding. These properties are particularly attractive for education research, where datasets are often domain-specific, resources for annotation are limited, and alignment with qualitative frameworks is essential.

However, ZSC also has notable limitations. Performance depends heavily on prompt phrasing because the model matches responses to theme descriptions based on how closely

the description aligns with the model's pretrained understanding of language. Labels that use everyday vocabulary tend to produce more accurate classifications than labels that use specialized terminology, because pretrained models have greater exposure to common language patterns than to domain-specific jargon [72, 74]. This sensitivity to wording means that prompt design itself becomes a source of variability, requiring expert judgment to select phrases that are both conceptually accurate and linguistically accessible to the model. ZSC also cannot discover themes that the expert did not define in advance. If an important pattern exists in the data but was not anticipated, ZSC will not surface it. This contrasts directly with unsupervised methods such as LDA or BERTopic, which can reveal unexpected topics even if those topics require subsequent interpretation. ZSC can also struggle with minority themes that are infrequent in the data because pretrained models have limited exposure to rare concepts and cannot learn dataset-specific patterns without fine-tuning [72]. Finally, when responses address multiple themes simultaneously, ZSC must rely on multi-label classification logic and confidence thresholds, and performance can degrade when secondary themes receive low confidence scores.

Despite these limitations, ZSC addresses a gap that the other four methods in this comparative analysis do not. k-means, LDA, NMF, and BERTopic all require post-hoc expert interpretation to connect model outputs to research-relevant themes, and the quality of that connection depends on how well algorithmic topics align with conceptual themes, an alignment that is not guaranteed and often fails for ambiguous or overlapping categories. ZSC provides a direct path from expert-defined themes to classified data, eliminating the topic-to-theme mapping step entirely and enabling researchers to evaluate model performance against their own conceptual framework from the outset. Its inclusion in this comparative analysis allows direct evaluation of whether a themes-first classification approach can match or exceed the performance of discovery-first topic models for short educational texts, and whether the trade-off between discovery capacity and theme alignment favors one approach over another depending on the characteristics of the data.

3.6 Previous Comparative Analyses of Topic Models

This study focuses on comparing different approaches to topic modelling of short texts in the education domain. In this spirit, the comparative analysis investigates baseline models associated with five different approaches (clustering, probabilistic, matrix factorization, neural networks, and zero-shot classification) rather than variants or modifications of these baseline approaches with the goal of developing broad guidelines for how researchers and practitioners in education can strategically use NLP in their work. Each approach has different advantages and disadvantages which makes no single choice the right choice for every dataset. Previous comparative analyses have underscored that there is no one-size-fits-all topic model for short texts (Table 1). Considering 16 comparative analyses conducted between 2019 and 2025, five analyses (31.3%), four analyses (25%), and seven analyses (43.8%) identified NMF, LDA, and neural network-based models (including BERTopic) respectively to be among the top performers. One study also found that LSA performed best, outperforming LDA and several variants of LDA.

Table 3.1: Comparative Analyses of Topic Modeling Studies on Short Text Datasets

Ref	**Algorithms**	**Dataset**	**Avg. Words**	**Corpus Size**	**Performance Measures**	**Top Performers**
Chen et al., 2019 [57]	BTM (standard), DMM (two variants), LDA (standard), LDA (two variants), NMF (standard), NMF (one variant)	Google Snippets	14.3	12,284	Topic Coherence[1]	NMF
		News	11.8	35,592		
		Stack Overflow	7.25	19,783		

(continued on next page)

Table 3.1 – *continued*

Ref	**Algorithms**	**Dataset**	**Avg. Words**	**Corpus Size**	**Performance Measures**	**Top Performers**
		XinlangNews	4.38	8,966		
		TMNtitles	3.46	29,487		
		Tweets (general)	5.81	102k		
Qiang et al., 2022 [26]	BTM (standard), DMM (four variants), LDA (standard), oProb (one variant), SA (two variants)	Pascal Flickr	5.37	4,834	Classification[2], Topic Coherence[1]	DMM
		Stack Overflow	5.03	20K		
		Biomedicine	7.44	19,448		
		Tweets (Queries)	8.55	2,472		
		Google News	6.23	11,109		
		Search Snippets	14.4	12,295		
Albalawi et al., 2020 [71]	LDA (standard), LSA (standard), NMF (standard), DimRed (two variants)	Newsgroup data	28	20K	Topic Coherence[1], Classification[2]	LDA
Doan & Hoang, 2021 [77]	LDA (standard), LDA (two variants), NMF (standard), NN (five variants)	20News	73.52	15,465	Perplexity[3], Classification[2]	NN

(continued on next page)

Table 3.1 – *continued*

Ref	**Algorithms**	**Dataset**	**Avg. Words**	**Corpus Size**	**Performance Measures**	**Top Performers**
		W2E-title	6.9	105,522	Perplexity[3], Classification[2]	NN
		Web Snippets	14.42	12,295	Topic Coherence[1]	NN
Lossio-Ventura et al., 2021 [37]	BTM (standard), DMM (one variant), LDA (standard), LDA (three variants), LSA (standard), k-means	Tweets (Health)	∼9	286,971	Internal Cluster Quality[4]	LDA, oProb
		Emails (Health)	31.78	50,000	External Cluster Quality[5]	LSA, k-means
Egger & Yu, 2022 [40]	LDA (standard), NMF (standard), LLM (one variant), NN (one variant)	Tweets (COVID Travel)	NR[7]	31,800	Qualitative Comparison	NMF, LLM
Murshed et al., 2020 [78]	BTM (standard), LDA (standard), LDA (one variant), NMF (standard), oMF (one variant), oProb (three variants), SA (two variants)	Tweets (COVID)	NR[7]	42k	Topic Coherence[1], External Cluster Quality[5], Classification[2]	oProb
		Tweets (bullying)	NR[7]	20K		

(continued on next page)

Table 3.1 – *continued*

Ref	**Algorithms**	**Dataset**	**Avg. Words**	**Corpus Size**	**Performance Measures**	**Top Performers**
Goyal et al., 2023 [79]	LDA (two variants), NMF (standard)	Image Captions	9.7	10K	Classification[2], Topic Coherence[1]	NMF
					Internal Cluster Quality[4]	LDA
					External Cluster Quality[5]	LDA
Muthusami et al., 2024 [80]	LDA (standard), NMF (standard)	Tweets (stance detection)	NR[7]	2,914	Topic Coherence[1], Internal Cluster Quality[4]	NMF
Doi et al., 2024 [81]	LDA (two variants), NN (four variants)	Tweets	5.47	2k	Topic Coherence[1], Topic Diversity[6]	NN
		GoogleNews	5.25	11k		
		StackOverFlow	4.71	19,000		
Hayat et al., 2024 [61]	NN (one variant), LDA (standard), LSA (standard), NMF (standard)	Student feedback (well-being)	NR[7]	375	Topic Coherence[1], Qualitative Comparison	NN
		Student feedback (remote learning)	NR[7]	321		

(continued on next page)

Table 3.1 – *continued*

Ref	**Algorithms**	**Dataset**	**Avg. Words**	**Corpus Size**	**Performance Measures**	**Top Performers**
Sheils et al., 2024 [82]	LDA (standard), NMF (standard), NN (two variants)	RateMyProfessors	NR[7]	1M	Topic Coherence[1], Internal Cluster Quality[4]	NN
Barker et al., 2024 [44]	LDA (standard), oProb (two variants)	Discussion Forum (Statistics Ed)	NR[7]	946	Qualitative Comparison	oProb
Medvecki et al., 2024 [67]	LDA (standard), NMF (standard), NN (three variants)	Tweets	NR[7]	3,286	Topic Coherence[1]	NN
					Topic Diversity[6]	LDA
Babalola et al., 2024 [58]	LDA, NMF, NN	News Headlines	6.45	149,679	Topic Coherence[1]	NMF, NN
Feng et al., 2025 [68]	LDA, NMF, NN (three variants)	Health Tweets	NR[7]	64,441	Topic Coherence[1], Topic Diversity[6]	NN

[1]A measure of the degree to which words belong to a topic semantically.

[2]A measure of how well topics support predicting document labels (Accuracy and F1-score).

[3]A measure of how well a model predicts unseen text data.

[4]A measure of topic quality based on model structure.

[5]A measure of how well topic clusters align with known categories.

[6]A measure of range and distinctiveness of topics.

[7]Not reported for this dataset, but Tweets are typically limited to 280 characters (46–55 words).

BTM (Biterm); DMM (Dirichlet Multinomial Mixture); LDA (Latent Dirichlet Allocation); oProb (Other Probabilistic topic models); oMF (Other Matrix Factorization topic models); SA (Self Aggregating topic models); LSA (Latent Semantic Analysis); DimRed (Non-LSA Dimension Reduction models); NN (neural network based topic models including BERTopic).

While the comparative analyses outlined in Table 3.1 add further justification for including LDA, NMF, and BERTopic as primary methods in this dissertation, several topic models that performed well in previous comparative analyses were rejected for consideration. Dirichlet Multinomial Models (DMMs) and Biterm topic models (BTM) were not considered primarily because they are designed for sparsity and optimized for ultra-short texts (15 words or less) which were shorter than most documents in the datasets used in this analysis. Self-aggregating (SA) models were also rejected because they merge shorter texts into longer documents prior to topic modelling. Doing so with research data would run the risk of blurring and compromising distinct opinions. Furthermore, NMF variants such as Semantic-assisted NMF (SeaNMF) and Knowledge-guided NMF (KGNMF) were not considered because they are specifically designed to handle word co-occurrence sparsity. This type of sparsity is often characteristic of data with a wide range of language use, such as that found in the public domain, including social media. Given that the data collected for this study was guided by a specific prompt and therefore was much more focused than public domain datasets, these techniques were deemed less appropriate and not considered in this analysis.

Notably absent from the comparative analyses summarized in Table 3.1 is zero-shot classification. All 16 studies compare unsupervised topic models that discover patterns in data without predefined categories, and ZSC does not fit this paradigm because it classifies data into expert-defined categories rather than generating topics from statistical patterns. This absence reflects both the relative novelty of applying ZSC to thematic analysis of short texts and a fundamental difference in approach: the studies in Table 3.1 evaluate how well models discover structure, while ZSC evaluates how well models apply structure that experts have already defined. Recent studies in education have demonstrated the feasibility of ZSC for classifying student feedback [74, 75], but these studies evaluated ZSC in isolation rather than alongside traditional and neural topic models using the same datasets and evaluation metrics. This gap motivates the inclusion of ZSC in this dissertation, which enables direct comparison between discovery-first and themes-first approaches for short educational texts under a shared evaluation framework.

The five approaches reviewed in this chapter each offer different trade-offs for analyzing

short educational texts. Traditional methods (k-means, LDA, NMF) are computationally efficient and well understood but rely on word frequency or co-occurrence and can struggle with the vocabulary diversity and thematic overlap common in student feedback. BERTopic captures semantic meaning through contextual embeddings but inherits the single-topic-per-document limitation and depends on the suitability of pretrained representations for domain-specific language. Zero-shot classification (ZSC) bypasses the topic discovery step entirely by classifying responses into expert-defined themes, but it cannot surface unanticipated patterns and is sensitive to prompt phrasing. No single method addresses all of these challenges, and the previous comparative analyses summarized in Table 3.1 confirm that performance depends heavily on dataset characteristics. The three datasets in this dissertation are short, semi-structured student feedback on instructional support from faculty, teaching assistants, and peers. As such, these data occupy a distinct niche among those studied in previous comparative analyses: focused vocabulary constrained by domain-specific prompts produces both reduced noise and thematic overlap that have not been systematically evaluated across discovery-first and themes-first approaches together.

Chapter 4

METHODS: DATA COLLECTION, CLEANING, AND PREPROCESSING

This chapter describes how the short-text data used in this dissertation were prepared for analysis to support answering education research questions. Each step in the preprocessing pipeline was designed to preserve the substance of the student feedback contained in the datasets while ensuring that the data could be analyzed consistently across datasets.

The cleaning process transformed raw student responses into a format suitable for topic modeling. Guided by expert review, decisions such as stopword selection, removal of non-informative responses, and token-level normalization ensured that differences observed in later analyses reflect real variations in student feedback rather than artifacts of data preprocessing.

4.1 Data Collection Procedures

4.1.1 Datasets and Survey Prompts

The data used in this study were collected as part of a larger investigation into the relationship between instructional support and student engagement within an engineering education context. The study yielded three distinct datasets of student responses, each representing a different dimension of instructional support. One dataset focuses on faculty support, one on teaching assistant (TA) support, and one on peer support.

In each case, students were provided a prompt and provided a free-text answer:

- **Faculty Support:** *"What one action can your professors at <institution name> take to best support you in your classes (please be as specific as possible)?"*
- **TA Support:** *"What one action can your TAs at <institution name> take to best support you in your classes (please be as specific as possible)?"*

- **Peer Support:** *"What one action can students in your <coursename> class take to improve your educational experience (please be as specific as possible)?"*

The responses were short, often a single sentence or a brief phrase, reflecting the concise nature of much educational feedback. The three datasets contained approximately 4,600 responses from engineering students, each set centered on a single type of instructional support. Table 4.1 lists the courses included in the study and summarizes participant demographics.

4.1.2 Collection Procedures and Ethics

IRB (Institutional Review Board) approval (STUDY00000378) was obtained to recruit and survey undergraduate students beginning on October 26, 2016. While the exempt status under which this study was approved did not require continuing review, data collection was discontinued on June 15, 2023. Instructors were asked to offer the survey to their students within two to three weeks of the end of the term in which the course was offered. Instructors offered an incentive to students to complete the survey, with a nominal amount of extra credit being the most popular choice; extra credit has been shown to be a highly effective motivator for college students [83].

For all but one class in the pre-COVID and COVID-era remote instruction (ERT) time periods, the survey was hosted by an institution-specific survey tool (Catalyst WebQ) and students accessed and completed the survey via a link in the learning management system for the course (Canvas) within one to three weeks of the instructors publishing the survey. In the remaining course (a 2016 pre-COVID offering), students completed a paper version of the survey in class. In the post-COVID period, student responses were collected using either Catalyst WebQ (2022) or Google Forms (2023). Instructors were not provided with any survey responses but instead were provided with a list from the researchers of names and percentage of questions completed by each student so that grades could be adjusted according to the incentive offered to students.

All participation was voluntary. Students were offered extra credit regardless of whether they granted consent for their responses to be used in the research, a procedure required by

Table 4.1: Study courses and participant demographics (N = 1,707; 22 unique courses, 43 offerings, 2016–2023). Percentages may not sum to 100 due to non-response.

Category / Course	***N***	**%**
Engineering Courses (22 unique courses, 43 offerings, 2016–2023)		
By Discipline		
Electrical Engineering (EE)	18	81.8
Mechanical/Aerospace Engineering (ME/AA)	4	18.2
By Course Level		
100-level (Introductory)	1	4.5
200-level (Sophomore)	8	36.4
300-level (Junior)	7	31.8
400-level (Senior)	5	22.7
500-level (Graduate)	1	4.5
Participant Demographics[c]		
All participants	1,707	100
Gender (all participants)		
Male	1,260	73.8
Female	422	24.7
Other	12	0.7
Race/Ethnicity (domestic only, n = 1,666)		
Asian	750	45.0
Non-Hispanic White	629	37.8
Mixed Race[a]	95	5.7
Latino/a	72	4.3
Black	37	2.2
Other URM[b]	45	2.7
Other	38	2.3
Country of Origin (all participants)		
Domestic	1,419	83.1
International	263	15.4

[a] Mixed Race includes subtypes with $n \geq 10$: Asian/Non-Hispanic White ($n = 82$) and Latino/Non-Hispanic White ($n = 13$).

[b] Other URM combines groups with $n < 10$ (Native American, Pacific Islander, and other small mixed-race categories).

[c] Each student completed one survey per course offering; the same population is reflected in the Faculty Support, TA Support, and Peer Support datasets.

the institutional IRB to ensure that those who did not provide consent were not excluded from taking the survey. Participants provided informed consent either in written form (for the paper form of the survey in the 2016 pre-COVID course) or in electronic form via a unique link to their student network ID (in all subsequent surveys). No minors participated in the study. Less than 5% of those who completed the survey did not offer consent; their corresponding survey responses were excluded from the final research dataset prior to analysis. Detailed statistics on the resulting datasets, including word counts and character counts, are provided in Chapter 7.

4.2 Data Cleaning Procedures

Data cleaning differs from data preprocessing in that it involves manual or only partially automated processes to ensure that the documents contain language use that is accurate and consistent with accepted use. Regardless of whether thematic analysis is conducted using traditional techniques or with natural language processing tools, data cleaning is an essential first step in preparing data for analysis. The data cleaning steps used on the three datasets in this dissertation are summarized in Table 4.2.

Table 4.2: Data Cleaning Steps

Operation	**Justification**	**Method/Tool**	**Example**
	Phase 1: Data Cleaning		
Response Filtering	Remove non-consenting, blank, or irrelevant responses. <5% removed.	Manual	"N/A" → *Removed*
Anonymization	Remove personal identifiers to protect privacy, while preserving structure.	Manual	"Dr. Smith" → "[Professor]"
Spell Checking	Correct misspelled words and delete repeated words.	Semi-automated with researcher review	"recetation" → "recitation"
Expand Acronyms	Expand domain-specific abbreviations for consistent representation.	Manual	"OH" → "office hours"

Before analyzing text, the dataset was filtered at the response level. Responses that

lacked research consent, contained no substantive content (blank, "N/A," random keystrokes), or were too terse to be informative (e.g., "nothing" with no actual suggestion) were removed. Identifying information such as instructor names was also replaced with generic placeholders to protect privacy [84]. Responses were not removed just for being short if they conveyed a clear idea. For example, "Start class on time" is only four words but is a valid suggestion and was kept. In total, less than 5% of responses were removed through this filtering process.

The language students use in offering feedback is often informal and unstructured, similar to microblog or social media text [33]. Students often used colloquial expressions, incomplete sentences, and discipline-specific jargon and acronyms. For example, a response might say "Prof needs to hold OH" meaning "The professor needs to hold office hours," assuming the reader knows that "OH" stands for office hours. Spelling and grammatical errors were also common, especially when students wrote quickly or on mobile devices [84]. Such misspellings further complicate automated analysis because they fragment the vocabulary. For instance, "recitation" misspelled as "recetation" would be treated as a separate word.

4.3 Data Preprocessing

Data preprocessing was conducted in two phases (Table 4.3): text normalization (lowercasing, tokenization, punctuation removal, and contraction expansion), and linguistic transformations (spelling correction, abbreviation expansion, lemmatization, and stopword removal), verified via researcher review of the preprocessed data. This process used Python with pandas and scikit-learn libraries, combined with researcher oversight at each stage to handle exceptions and judgment calls that automated rules could not cover [84]. Only those decisions that directly affect interpretability are highlighted here.

4.3.1 Stopword Curation (Standard and Domain)

Standard English stopwords like "the" and "and" that appear in many responses simply as a consequence of standard language use were removed from each data corpus because they provided little added value to the meaning of underlying documents.

Table 4.3: Preprocessing Pipeline

Operation	**Purpose**	**Method/Tool**	**Example**
Phase 1: Text Normalization (Surface-Level Standardization)			
Lowercasing & Whitespace	Standardize case and spacing to avoid token fragmentation.	Standard NLP preprocessing	"My TA" → "my ta"
Tokenization	Split text into word tokens at whitespace and punctuation boundaries.	Python string methods	"office hours" → ["office", "hours"]
Punctuation Removal	Remove punctuation, special characters, and numbers.	Regex pattern matching	"don't!" → "don't"
Contraction Expansion	Expand contractions to preserve negation as an explicit token.	Python contractions library	"don't" → "do not"
Phase 2: Linguistic Transformations (Deep Content Processing)			
Spelling Correction	Consolidate vocabulary via spelling correction. Researcher review applied for domain terms.	PySpellchecker [85] + manual review	"offce hrs" → "office hours"
Abbreviation Expansion	Expand domain-specific abbreviations for consistent representation.	Domain dictionary with manual rules	"OH" → "office hours"
Lemmatization	Reduce inflected forms to dictionary base form using POS tags.	spaCy POS-aware lemmatizer	"holding" → "hold"; "classes" → "class"
Standard Stopword Removal	Remove common English function words.	NLTK English stopword list	"and the to" → *Removed*
Domain Stopword Removal	Remove prompt-specific high-frequency words that do not distinguish topics.	Custom list via frequency analysis	"professor class" → *Removed*
Final Validation	Researcher inspection of 10% sample to verify content preservation.	Manual review	Meaning preserved?

Another challenge arose from the way the survey prompt shaped the vocabulary students used. Because all responses addressed a shared question about faculty, TAs, or peers, certain words appeared frequently by design. In the faculty-support dataset, almost every response referenced "professor" or "instructor"; "TA" dominated TA-support responses; and "students" or "peers" appeared throughout peer-support responses. These high-frequency terms were not informative of the type of support being discussed, yet they dominated simple frequency counts and could mislead algorithms into treating them as defining features of a topic when they were essentially background context. Therefore, in addition to standard stopword removal, these domain-specific terms that added little meaning to specific themes or ideas in student responses were also removed.

Two examples showed how domain-specific stopword removal was applied. First, "class" was removed because it appeared in nearly every response as background context. For instance, "give feedback in class" became "give feedback" with no semantic loss, since the faculty dataset assumes classroom context. Verification across many responses showed that removing "class" did not erase content. Second, "help" was kept even though it was very common. Students often wrote things like "help more" or "help faster," and "help" was central to their requests for support [84]. Denny and Spirling [84] show that removing semantically loaded words just because they are frequent can hurt model accuracy.

4.3.2 Token-Level Normalization

Text format was standardized and meaning was preserved [1, 3]. All text was lowercased and extra whitespace was removed to avoid fragmented counts. Punctuation was removed and contractions were expanded to their full form (e.g., "don't" became "do not") so that negation was preserved as an explicit token during analysis. Misspellings were corrected using PySpellchecker with researcher review, so variants like "offce hrs" and "office hourz" all became "office hours" [84, 85]. Common abbreviations were expanded ("OH" became "office hours," "HW" became "homework") for consistency. POS-aware lemmatization using spaCy reduced inflected word forms to their dictionary base form (e.g., "holding" became "hold," "classes" became "class") while using part-of-speech tags to prevent incorrect reductions [3].

Unlike stemming, lemmatization maps each token to a verified dictionary entry rather than truncating word endings, which limits meaning loss in short texts [84].

Example: "My TA should NOT just read off slides, and maybe be more engaging" became "not read slide maybe engage" after lowercasing, expanding contractions, removing punctuation, removing stopwords, removing domain-specific terms, and applying lemmatization. The core suggestion remained clear, and the negation was preserved as an explicit token.

4.4 Additional Considerations

The three support datasets were analyzed both individually and comparatively. It was therefore necessary to ensure that preprocessing and feature extraction made the corpora compatible without accidentally favoring one dataset over another. Multiple additional preprocessing steps were implemented to ensure a balanced analysis.

4.4.1 Separate vs. Combined Analysis

Each dataset was primarily analyzed separately, given that each addressed a distinct support role. Thus, a topic model on faculty responses would yield topics about faculty behaviors, and likewise for TA and peer responses. However, themes across these datasets were also planned to be compared in later chapters. To help such comparison, identical preprocessing was maintained across all three sets (with only minor differences in the domain-specific stopwords appropriate to each dataset), ensuring that, for example, the word "help" or "feedback" was treated the same way whether it appeared in a faculty or peer context. This approach means any differences in discovered topics or patterns are due to differences in the content of the responses, not differences in how they were processed.

4.4.2 Vocabulary Alignment

One benefit of using the same cleaning process on all corpora was that this resulted in compatible vocabularies across datasets. The vocabularies of the three corpora (the set of unique tokens remaining) showed a high degree of overlap, which was expected since students

often brought up similar issues (communication, clarity, etc.) across faculty, TA, and peer support contexts. When a term was truly unique to one corpus (e.g., "grading" might appear mostly in TA and faculty responses but not peer responses, while "participation" might appear in peer but not in TA context), that itself was an interesting content difference. Every corpus was not forced to have exactly the same vocabulary, but when cleaning was performed uniformly, fair comparison across datasets was ensured. If "lecture" had been removed in one corpus but not another, that would have misaligned the features. Such situations were avoided by using a comprehensive approach to stopword removal that applied similarly to all sets. This consistency enables direct comparison of term or topic frequencies across corpora without worry that differences were artifacts of preprocessing.

4.4.3 Corpus Balancing

Since the three datasets were of comparable size (each roughly 1,300 to 1,700 responses), there was not a severe class imbalance issue when analyzing each dataset separately. Each dataset was treated equitably, with no dataset receiving preferential preprocessing or feature handling. This consistent treatment ensures that performance differences observed across datasets in later chapters reflect genuine content variation rather than preprocessing artifacts.

For instance, if one corpus had a lot more unique words (perhaps because students in that group used more varied vocabulary), those were not eliminated to "match" the others; instead, that richness is considered a true characteristic of that corpus. Conversely, if one corpus repeated a particular phrase across many responses (indicating a prevalent concern), it was not down-weighted just to balance term frequencies. This approach preserves the authentic characteristics of each dataset, enabling meaningful comparison of algorithmic performance across different support contexts while maintaining fair preprocessing standards.

4.4.4 Internal Distribution Considerations

Within each corpus, some themes appeared in many responses while others appeared in only a few. This natural imbalance affects clustering algorithms, which may form clusters of very different sizes [84]. Content was not artificially re-balanced through oversampling or undersampling, as this would distort reality. Instead, low-frequency content words were preserved to allow rarer themes to surface and the feature matrix sparsity structure was retained.

The resulting corpora preserved student voice and supported computational analysis. Once cleaned, the textual data were converted into numerical features for modeling; the specific vectorization methods and their evaluation are detailed in Chapter 5, which covers word and document embeddings used in topic modeling. The topic modeling algorithms and performance evaluation framework are described in Chapter 6, and results are presented in Chapters 7 and 8.

4.5 Preprocessing as Methodological Rigor

In summary, preprocessing is not neutral. Small choices change what appears in unsupervised models and what stays hidden. Denny and Spirling [84] show that three key decisions: whether to remove stopwords, whether to stem or lemmatize, and how aggressively to trim rare terms, can shift topic word lists, reorder topic salience, and even change document grouping. They recommend stating transformations explicitly and testing multiple pipelines to see when a preprocessing decision, rather than the data, drives the pattern. This advice was followed in this study by documenting all preprocessing steps explicitly and maintaining consistent procedures across all datasets.

Signal preservation was balanced with noise reduction. Over-filtering discards substance; under-cleaning creates spurious structure. To hold that balance in student feedback, clear rules were applied that could be defended and replicated: (1) remove prompt-anchored role words that mirror the question frame (e.g., “professor,” “TA,” “students”) unless they add meaning beyond the prompt; (2) retain frequent tokens that carry action or sentiment (e.g., “help,” “answer”); (3) expand contractions to their full form so that negation is preserved as

an explicit token ("don't cancel" becomes "do not cancel," retaining "not"); and (4) apply POS-aware lemmatization rather than stemming to reduce inflected forms while limiting meaning loss in short texts [3, 84].

Each preprocessing decision was grounded in methodological considerations relevant to short educational research texts. The standardized pipeline (summarized in Tables 4.2 and 4.3) was applied uniformly across all three datasets to ensure that observed differences in later analyses reflect genuine content variation rather than preprocessing artifacts. This consistent approach enables fair comparison across different topic modeling methods evaluated in Chapters 8 and 9, where vectorization and dimensionality reduction strategies were thoroughly tested while keeping preprocessing constant.

Chapter 5

METHODS: WORD AND DOCUMENT EMBEDDING

Chapter 4 demonstrated how raw survey responses were cleaned and preprocessed into clean, structured text corpora through human-guided preprocessing decisions including stopword curation, response-level filtering, and token-level normalization. This chapter addresses the next step of transforming cleaned educational research data into numerical representations that machine learning algorithms can process. Topic modeling algorithms cannot operate directly on words; they require mathematical representations that capture the essential information in the text. Once expressed numerically, these representations allow algorithms to uncover latent structures in the educational research data.

5.1 Frequency-Based Word Embedding (Bag of Words)

Frequency-based word embedding captures little to no semantic meaning and relies on how often words occur and co-occur in a corpus to capture patterns (topics) in the text. Two approaches to frequency-based word embedding are considered in this dissertation: (frequency) counts and term frequency–inverse document frequency (TF-IDF).

5.1.1 Counts-Based Word Embedding

Counts-based word embedding is the simplest and most common text representation in topic modeling. In a counts-based representation, each document is transformed into a vector of word frequencies (or counts) of a fixed vocabulary. The document is treated as an unlabeled "bag" containing words with certain multiplicities, disregarding word order. For example, consider a student response: "The professor offers extra help during office hours." In a counts-based model with vocabulary that includes ("professor", "offers", "extra", "help", "office", "hours", ...), this document might be represented as something like: $(1, 1, 1, 1, 1, 1, \dots)$ for those six words (each appearing once) and 0 for all other words not

present. Each dimension of the vector corresponds to a word, and its value is the count in the document. Formally, for document d, the counts-based feature vector is defined as:

$$x_d = [f_{d,1}, f_{d,2}, \ldots, f_{d,V}]$$

where $f_{d,i}$ is the frequency of word i in document d and V is the vocabulary size [3].

The main assumption of counts-based representation is that word identity and frequency matter, but syntax and word order do not [3]. This simplification can be powerful for thematic analysis; if two responses use many of the same important words, their counts-based vectors will be similar and may suggest likely shared topics. For domain-specific short texts such as student responses, counts-based representation captures the essential information about content (e.g., how often does a student mention "mentor" or "project" or "stress"?), which is necessary to identify themes.

This variant aligns with the generative assumptions of traditional topic models like LDA (which directly models word counts per topic) and is also the basis for matrix factorization methods like NMF [30].

However, counts-based representation has limitations, especially with short documents. Short responses contain very few words (perhaps one sentence), so a single word's presence or absence can dominate the representation. There is often little redundancy of wording, meaning the raw counts can be very sparse indicators. Also, common function words or prompt words may appear in many responses, contributing counts that do not signal any meaningful topical content (although this can be mitigated with stopword removal, as described in Chapter 4). Another issue is that counts-based representation does not account for word significance; for example, it is possible that a word appearing only once in a corpus could be highly informative but counts-based representation would fail to capture the word's significance.

Despite these issues, counts-based representation is a strong baseline and is used in this dissertation as the foundation for several topic models. However, the limitations of counts-based representation motivate using TF-IDF variant for word embedding.

5.1.2 TF-IDF Weighted Representation

The TF-IDF weighted representation applies term frequency–inverse document frequency (TF–IDF) weighting to counts-based representation, down-weighting common terms and up-weighting rare but potentially significant terms [86]. The TF–IDF weight of word w in document d is usually defined as:

$$\text{tf-idf}(w, d) = \text{tf}(w, d) \times \text{idf}(w),$$

where $\text{tf}(w, d)$ is the term frequency of word w in d (often just the raw count, possibly normalized), and $\text{idf}(w)$ is the inverse document frequency of word w. Inverse document frequency is given by: $\text{idf}(w) = \log \frac{N}{df(w)}$, where N is the total number of documents and $df(w)$ is the number of documents containing word w. Conceptually, $\text{idf}(w)$ is high if w is rare (appears in few documents) and low if w is ubiquitous. Thus, TF–IDF weights words by their relative importance; which means a word that is frequent in this document but rare in the corpus will get a high score, whereas a word that appears in almost every document (like a common stopword or a word from the prompt repeated by everyone) will get a low or even zero score.

Using TF-IDF weighted vectors can improve clustering and matrix factorization performance [87]. In matrix factorization approaches, TF-IDF weighting is first applied to the term counts and then dimensionality reduction techniques are used to reduce the data into a smaller set of dimensions. This step downweights generic words and highlights the words that carry more meaning, which helps uncover broader themes. In clustering methods such as k-means, TF-IDF weighting ensures that distance or similarity measures are not dominated by generic terms [88]. In educational surveys, words like "the", "student", "class" might appear in nearly all responses (especially if the question prompt includes those terms). Counts-based representation would count them, but TF-IDF weighting neutralizes them (if $df(w) \approx N$, $\log(N/df(w)) \approx 0$). On the other hand, a word like "tutoring" that appears only in a subset of responses will have a higher weight, making those responses stand out in similarity comparisons.

For short texts, TF-IDF weighting is particularly helpful because there are so few words per document and each word should be weighted appropriately. If a student writes "I

struggled with calculus", the word "calculus" might appear in only some students' answers; TF-IDF weighting will highlight it, whereas counts-based representation would treat it on par with any other single occurrence. More distinctive terms drive the factorization or clustering rather than common prompt words.

When applying TF-IDF weighting to a small corpus, one must be cautious because rare terms may receive disproportionately high weights. But since documents are usually compared in vector space, this means the document's vector will be oriented along those rare terms, which is acceptable if that truly indicates a unique theme. LDA, being a probabilistic model of raw word counts, typically does not use TF-IDF weighting (it has its own internal weighting via probabilities). For this reason, counts-based representation is used for LDA in this dissertation to adhere to its theoretical foundations, while TF-IDF weighting is used for methods that rely on distance or linear algebra.

Regardless, both counts-based and TF-IDF weighted representations have core limitations. They treat words as independent tokens and cannot capture semantic relationships between words or understand context. For example, "good" and "excellent" would be treated as completely different words, even though they have similar meanings. Similarly, words with multiple meanings would be represented identically regardless of context. This motivates the development of more sophisticated representations that can capture semantic similarity and contextual meaning, which leads to neural embedding-based approaches.

5.2 Neural Word and Document Embeddings

More recently, neural embedding techniques have become popular for representing text, including educational survey responses. Unlike frequency-based embeddings which create sparse vectors (mostly zeros), neural embeddings are dense vector representations learned from large corpora that encode semantic information (Figure 5.1).

5.2.1 Word2Vec Approach to Neural Embedding

Word2Vec, a commonly used neural embedding model, learns word embeddings such that words appearing in similar contexts have vectors that are close to one another in a high-dimensional space [39]. The model learns these representations by analyzing how words

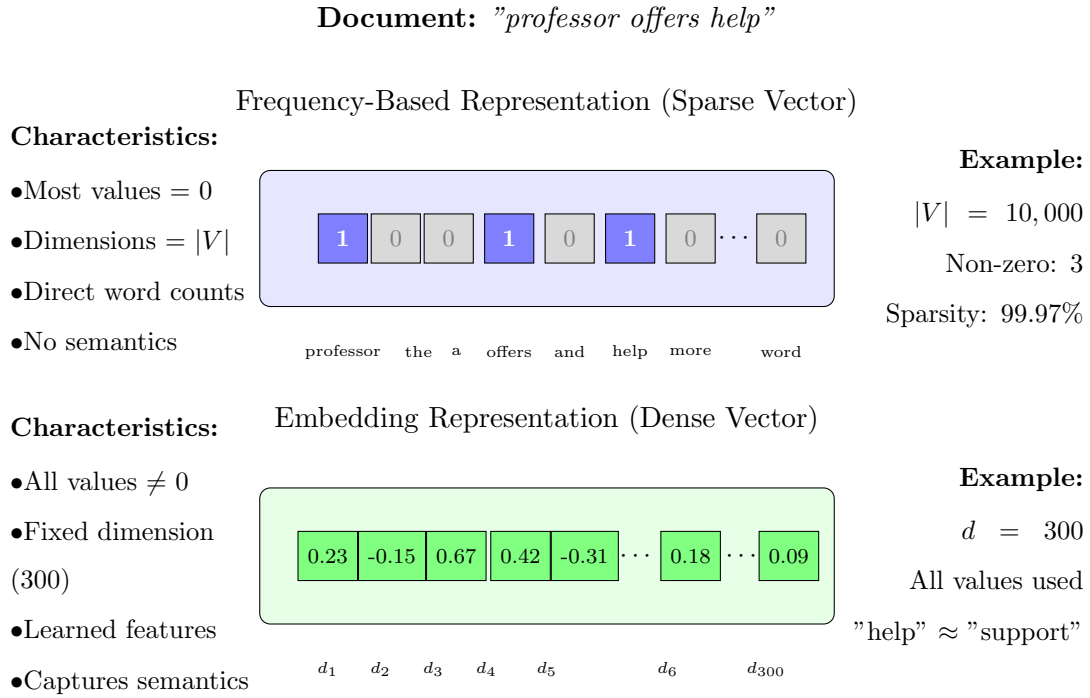


Figure 5.1: Comparison of sparse frequency-based and dense embedding representations. Frequency-based vectors have dimensions equal to vocabulary size with mostly zeros (sparse). Embedding vectors have fixed dimensions (typically 100-300) with all meaningful values (dense).

appear together in large text collections [38, 89]. The core insight is that words appearing in similar contexts likely have similar meanings. For instance, "doctor" and "nurse" might both appear near words like "hospital," "patient," and "medical," so they would get similar vector representations. Similarly, "happy" and "joyful" would have vectors that are close together, while "happy" and "sad" would be far apart.

Mathematically, Word2Vec learns a mapping from each word w to a dense vector $\mathbf{e}_w$ of dimension d (typically 100-300), where words with similar meanings have vectors that point in similar directions. The similarity between two word vectors is measured using cosine similarity:

$$\text{sim}(w_i, w_j) = \frac{\mathbf{e}_{w_i} \cdot \mathbf{e}_{w_j}}{||\mathbf{e}_{w_i}|| \cdot ||\mathbf{e}_{w_j}||}$$

This formula measures the angle between two vectors: similar words have small angles (high similarity scores), while different words have large angles (low similarity scores).

For documents, embeddings can be created by combining the embeddings of individual words. A simple approach is to average the word embeddings in the document:

$$\mathbf{e}_d = \frac{1}{n} \sum_{i=1}^{n} \mathbf{e}_{w_i}$$

where n is the number of words in document d. This document-level embedding represents the overall semantic content of the text, aggregating word-level meanings into a single vector representation that can be used for document comparison and clustering.

5.2.2 BERT Approach to Neural Embedding

Modern approaches like BERT (Bidirectional Encoder Representations from Transformers) [5] create contextualized embeddings that address a key limitation of earlier embedding methods like Word2Vec. While Word2Vec assigns a single vector to each word regardless of its usage, BERT generates different vector representations for the same word based on its surrounding context. For example, "bank" in "river bank" would have a different embedding than "bank" in "financial bank," allowing the model to capture word sense disambiguation.

BERT's architecture, based on the transformer model [90], processes words in relation to all other words in a sentence simultaneously through self-attention mechanisms. This bidirectional processing allows BERT to capture complex semantic relationships and contextual nuances that unidirectional models miss. The model is pre-trained on large text corpora using masked language modeling and next sentence prediction tasks, learning representations that encode language structure, syntax, and semantic meaning [5].

For document-level embeddings, BERT can generate sentence or document representations by pooling the contextualized word embeddings, typically using the [CLS] token embedding or averaging the token embeddings. These document-level embeddings capture the overall meaning of entire texts while preserving the contextual understanding gained from the word-level representations. This makes BERT embeddings particularly suitable for tasks requiring semantic document comparison, such as topic modeling applications where documents with related meanings should be close in embedding space even if they do not share exact words [91]. For example, BERTopic [65] uses pre-trained BERT embeddings to cluster documents, then applies a modified TF–IDF to extract topic-representative words from each cluster, combining the semantic richness of contextualized embeddings with the interpretability of traditional topic modeling.

The main advantage of neural embeddings, whether word-level or document-level, is their ability to capture semantic relationships that frequency-based representations miss. However, they require more computational resources and may be harder to interpret since the numbers in the vectors do not have obvious meanings like word counts do. For topic modeling applications, neural embeddings work particularly well when grouping documents by meaning rather than just by shared vocabulary.

5.3 Vectorization Choice and Evaluation

No matter which representation is chosen from counts-based, TF-IDF weighted, or neural embeddings, the output is a numeric matrix ready for modeling. Each row of this matrix corresponds to a document (e.g., a student's response), and each column corresponds to a feature (a word or a vector dimension). This matrix is the starting point for topic modeling algorithms, which will seek patterns in these numbers. The quality of the topics ultimately

derived is strongly influenced by this representation step.

The choice between frequency-based representations and other embeddings can meaningfully impact model results, so it is important to evaluate and justify the representation used. Simpler representations (like frequency-based methods) preserve direct information about word usage, whereas more complex ones (like embeddings) incorporate levels of abstraction and prior knowledge. The chosen representation must align with the modeling algorithm's assumptions and the nature of the data (short student texts). The choice of vectorization is treated as a hyperparameter that must be empirically determined for each algorithm. Theoretical considerations guide initial expectations: LDA's probabilistic framework favors raw counts [30], while matrix factorization methods like NMF [56] can benefit from either (raw) counts-based or (weighted) TF-IDF representation. Which method is best is determined through thorough comparison of both frequency-based variants across the three datasets using topic model performance metrics (Chapter 6, Section 6.9). This approach maintains that the chosen representation optimizes performance on educational research data rather than relying on general assumptions that may not hold for short educational research texts.

For clustering and matrix factorization methods in this study, both frequency-based variants (counts-based and TF-IDF weighted) are evaluated using the metrics described in Chapter 6 (Section 6.9). Neural-based topic models (BERTopic, zero-shot classification [72]) use different representations (BERT embeddings) and bypass traditional vectorization, though they often reintroduce TF-IDF through techniques like class-based weighting for topic word extraction [65]. In this dissertation, the shared preprocessing pipeline (Chapter 4) provides consistency across methods so that differences in topic quality can be attributed to the algorithms themselves rather than to vectorization choices.

Chapter 6

METHODS: TOPIC MODELS

This chapter describes the five topic modeling methods applied in this dissertation: k-means clustering, Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), BERTopic, and Zero-Shot Classification (ZSC). Each algorithm is presented alongside its assumptions and the implementation choices made for this study. Latent Semantic Analysis (LSA) was also evaluated with default parameters, but because it underperformed compared to its companion matrix factorization technique (NMF) on all three datasets, it is not described in detail here; its results appear alongside the traditional methods in Chapter 7. k-means, LDA, NMF, and BERTopic methods follow a discovery-first approach, where themes emerge from statistical patterns in the text without predefined categories. In contrast, ZSC follows a themes-first approach, where domain experts define themes before any automated labeling occurs.

These five topic modelling methods were selected to represent distinct algorithmic families. k-means provides a non-probabilistic baseline that partitions documents by geometric similarity. LDA introduces a probabilistic framework that models documents as mixtures of topics. NMF offers an additive matrix decomposition that produces non-negative, interpretable topic weights. BERTopic combines pre-trained transformer embeddings with clustering and class-based term weighting. ZSC inverts the workflow entirely by classifying responses into expert-defined categories using natural language inference. Together, these methods span the major approaches available for thematic analysis of short educational text.

All five methods share a common evaluation framework (Section 6.9), the same three datasets (Chapter 4), and the same expert-coded ground truth against which performance is measured. This controlled design isolates the effect of algorithmic differences from parameter selection. The following sections describe the shared dataset context and the selection of the

number of topics that applies to all methods before presenting each method individually.

6.1 Dataset Context and Characteristics

The study participants were undergraduate engineering students from a single institution, with demographic details provided in Chapter 4. The three datasets of student responses varied in size and characteristics, with detailed statistics on word counts, character counts, and preprocessing outcomes summarized in Table 6.1. After preprocessing, the faculty support dataset had a median length of eight words; the TA support dataset had a median length of seven words; and the peer support dataset had a median length of five words. All datasets underwent identical preprocessing as described in Chapter 4 to maintain fair comparison across methods.

Table 6.1: Dataset Characteristics

Dataset	**Docs**	**Word Count**						**Character Count**					
		Raw			**After Preprocessing**			**Raw**			**After Preprocessing**		
		M	Mdn	SD	M	Mdn	SD	M	Mdn	SD	M	Mdn	SD
Faculty	1,667	23.2	17.0	21.7	12.0	9.0	10.2	139.0	101.0	125.0	90.6	69.0	77.3
TA	1,592	19.3	14.0	18.6	9.71	7.0	8.61	111.0	81.0	103.0	69.9	53.0	61.7
Peer	1,376	15.1	11.0	15.6	7.94	6.0	7.36	88.3	61.0	90.3	58.4	42.0	55.6

Note: Docs (Documents), M (Mean), Mdn (Median), SD (Standard Deviation).

6.2 Optimal Topic Number Selection

To support a fair comparison across all topic models, the number of topics chosen to represent each dataset for each topic model was held constant at three. The optimal number of topics was determined using the elbow method by plotting a key metric for each topic model as a function of the number of topics. The point at which the metric begins to level off (that is, the point of diminishing returns) was identified as the optimal number of topics for that model and dataset. Mathematically, this elbow point is found by calculating the point of maximum curvature:

$$\text{Elbow Point} = \arg\max_{T} \left| \frac{d^2 \text{Metric}(T)}{dT^2} \right|$$

The metric used for the elbow plot depends on the model:

For k-means, within-cluster sum of squares (WCSS) measures the total squared distance between each document and its assigned cluster centroid:

$$WCSS(T) = \sum_{i=1}^{T} \sum_{x \in C_i} ||x - \mu_i||^2$$

where C_i is the set of documents assigned to cluster i and μ_i is the centroid of cluster i. As the number of clusters increases, WCSS decreases; the elbow identifies the point at which additional clusters yield diminishing reductions in within-cluster variation. The silhouette score (Section 6.9.1) independently validates the elbow-identified $T = 3$ by confirming that documents are well-separated from adjacent clusters at that value.

For LDA, negative log-likelihood measures how well the model fits the training data given the learned topic distributions. Lower values indicate better fit; as with WCSS, negative log-likelihood decreases as the number of topics increases, and the elbow identifies the point of diminishing improvement.

For NMF, reconstruction error measures how well the matrix factorization approximates the original document-term matrix $\mathbf{X}$. For T topics with learned factor matrices $\mathbf{W}_T$ (document-topic weights) and $\mathbf{H}_T$ (topic-word weights):

$$\text{Reconstruction Error}(T) = ||\mathbf{X} - \mathbf{W}_T \mathbf{H}_T||_F^2$$

where $||\cdot||_F$ denotes the Frobenius norm, which is the square root of the sum of all squared matrix elements. Lower reconstruction error indicates a better approximation of the original data. Unlike a metric that peaks at an optimal value, reconstruction error decreases monotonically as the number of topics increases, because each additional component allows the factorization to capture more variance in the original matrix. The elbow for NMF is therefore identified not at a peak but at the point where the rate of decrease slows substantially, indicating that adding more topics produces only marginal improvement in fit. For all three datasets, this point of diminishing returns occurred at $T = 3$.

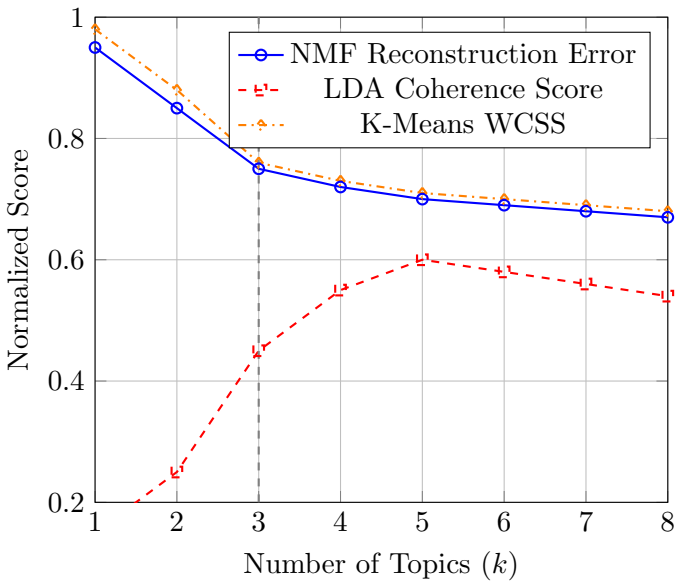


Figure 6.1: Elbow method results for k-means (WCSS), LDA (negative log-likelihood), and NMF (reconstruction error) across all three datasets. The elbow point at $T = 3$ is consistent across these three topic models and datasets.

BERTopic's elbow analysis is described in Section 6.7. ZSC used three expert-defined themes established prior to any model fitting, with the elbow analysis providing heuristic structure cues to inform expert theme definition rather than determining T algorithmically. For all three traditional topic models and across all three datasets, three topics were determined to be optimal or near-optimal. To further validate this selection, domain experts examined clustering results with T values ranging from 2 to 5, reviewing representative sample responses from each cluster. This qualitative validation revealed that $T = 3$ consistently produced the most interpretable and thematically coherent groupings, and the experts' judgment aligned with the quantitative elbow analysis.

6.3 Topic Model Variants

To explore the capability of the four discovery-first topic models, two sets of variants were applied in this comparative analysis. The first (Figure 6.2) evaluated each model using default hyperparameters. Two approaches were used to define ground truth for purposes of computing external performance metrics. Approach 1 used NLP-generated labels, in which the topic model's own cluster assignments served as the reference. Approach 2 used human-coded labels, in which a domain expert manually assigned each response to one of the three themes. Approach 2 provides a stronger external validity check because it is independent of the model output. The first variant compared model performance against both approaches to assess the sensitivity of results to ground truth choice. The second variant (Figure 6.3) used the best ground truth approach identified in the first variant (Approach 2) while also optimizing hyperparameters.

6.4 K-Means Clustering

6.4.1 Algorithm Description

k-means clustering is a centroid-based partitioning algorithm that groups documents into k clusters by minimizing the within-cluster sum of squares [88, 92]. The algorithm operates through three iterative steps:

1. **Initialization**: Randomly place k cluster centroids in the feature space.

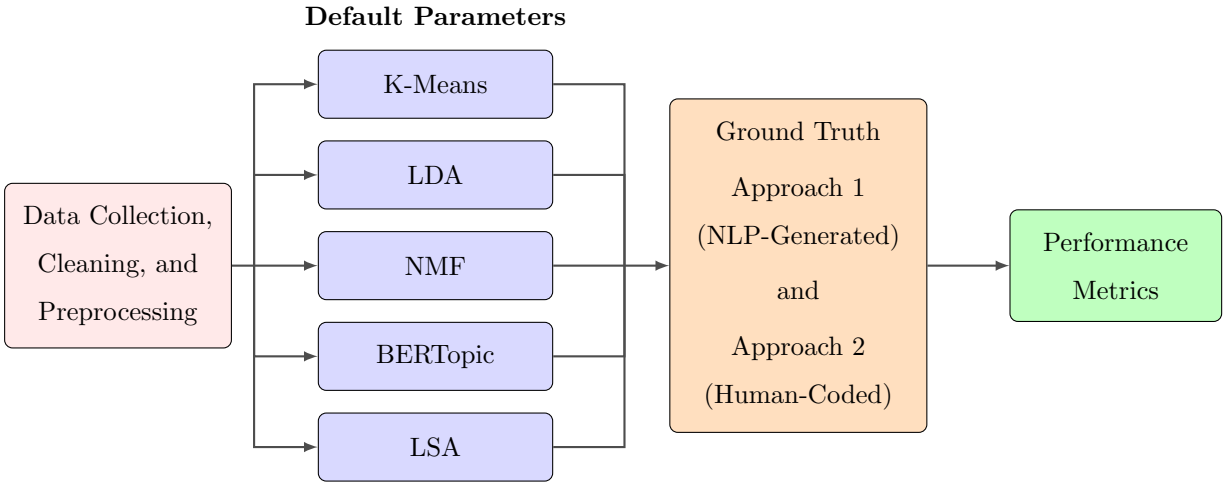


Figure 6.2: Phase 1 analysis process: all five methods evaluated at default parameters, with results compared against both Approach 1 (NLP-generated) and Approach 2 (human-coded) ground truth. Evaluation metrics are defined in Section 6.9.

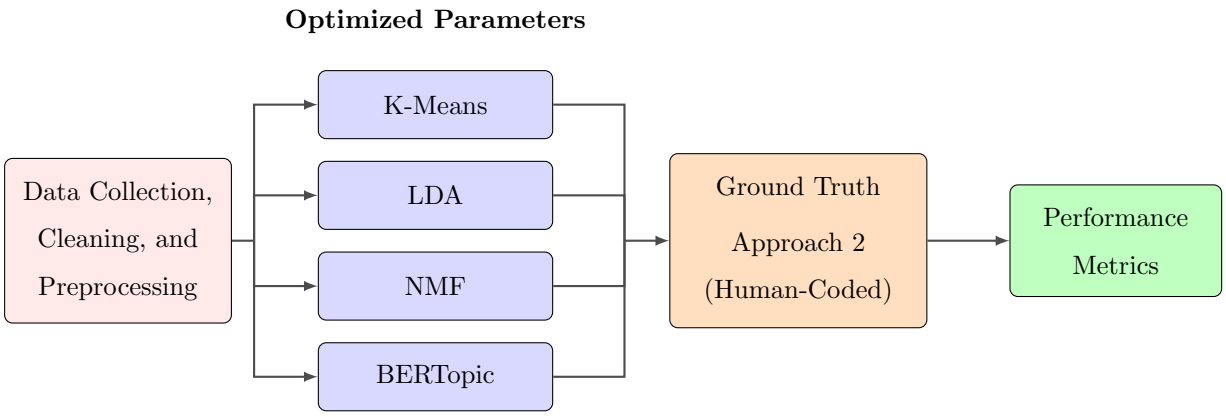


Figure 6.3: Phase 2 analysis process: four methods (LSA excluded) evaluated with dataset-specific optimized parameters, using Approach 2 (human-coded) ground truth exclusively. Evaluation metrics are defined in Section 6.9.

2. **Assignment**: Assign each document to the nearest centroid using Euclidean distance.

3. **Update**: Recalculate centroids as the mean of assigned documents until assignments stabilize.

The algorithm minimizes the within-cluster sum of squares (WCSS) objective function:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

where C_i represents cluster i, μ_i is the centroid of cluster i, and $||x - \mu_i||^2$ is the squared Euclidean distance between data point x and centroid μ_i. The algorithm iteratively updates centroids as:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

where $|C_i|$ is the number of points in cluster i.

k-means assigns each document to exactly one cluster, which makes it a hard clustering method. This property provides clear, non-overlapping assignments but cannot represent documents that express multiple themes.

6.4.2 Implementation Choices

To evaluate different preprocessing approaches, clustering quality was assessed using four performance metrics: silhouette score, Davies-Bouldin index, Calinski-Harabasz index, and within-cluster sum of squares (WCSS) (defined in Section 6.9.1). These metrics capture different aspects of clustering quality, balancing cluster separation with cluster tightness.

Vectorization Strategy: Counts-Based and TF-IDF Weighted Embedding. Both counts-based and TF-IDF vectorization were evaluated. Across all three datasets, counts-based methods outperformed TF-IDF on all clustering quality measures. The counts-based approach achieved higher silhouette scores (0.0471–0.0815) compared to TF-IDF (0.0162–0.0232), lower Davies-Bouldin indices, higher Calinski-Harabasz indices, and lower

Table 6.2: Counts-Based vs TF-IDF Vectorization Comparison

Dataset	**Variant**	**Silhouette (Euclidean) (↑)**	**DB (↓)**	**CH (↑)**	**WCSS (↓)**
Faculty Support	Counts-Based	0.0471	4.19	46.30	19,993
	TF-IDF	0.0162	5.22	25.94	1,571
TA Support	Counts-Based	0.0446	4.18	56.07	14,768
	TF-IDF	0.0225	5.79	32.74	1,481
Peer Support	Counts-Based	0.0815	3.29	69.86	9,278
	TF-IDF	0.0232	5.56	34.45	1,259

Note: Results shown are the best across 5 random states (42, 123, 456, 789, 101). DB = Davies-Bouldin Index; CH = Calinski-Harabasz Index; WCSS = Within-Cluster Sum of Squares.

WCSS values. Thus, the counts-based method was selected as the primary vectorization method for all subsequent k-means clustering analyses.

Dimensionality Reduction and Feature Representation. Following the selection of counts-based vectorization as the primary method, Principal Component Analysis (PCA) was integrated before clustering to address the curse of dimensionality inherent in text vectorization [3]. Three preprocessing variants were evaluated to test the effects of dimensionality reduction and normalization:

- **Counts-Based (L2 Normalized)**: Standardizes document length without dimensionality reduction.
- **Counts-Based + PCA**: Applies PCA without scaling to preserve sparse matrix structure.
- **Counts-Based + PCA (L2 Normalized)**: Combines PCA with L2 normalization to control for document length variation while reducing dimensionality.

L2 normalization scales each document vector to unit length by dividing each element by the vector's Euclidean norm: $\hat{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$. This places all documents on a unit hypersphere so that distances reflect differences in word composition rather than response length.

Without normalization, longer responses produce larger vectors and can dominate distance calculations in k-means clustering.

These variants were evaluated using the four metrics described in Section 6.9. Counts-Based + PCA (L2 Normalized) provided an optimal balance, achieving competitive silhouette scores with the best Calinski-Harabasz indices and lowest WCSS values. In effect, the combination of PCA dimensionality reduction with L2 normalization effectively addresses the curse of dimensionality while maintaining cluster compactness. Counts-Based + PCA (L2 Normalized) was therefore selected as the final preprocessing approach for k-means clustering.

Table 6.3: Preprocessing Variants Comparison: Counts-Based + PCA Approaches

Dataset	**Variant**	**Silhouette (Euclidean) (↑)**	**DB (↓)**	**CH (↑)**	**WCSS (↓)**
Faculty Support	Counts-Based (Normalized)	0.0327	4.45	51.81	1,492
	Counts-Based + PCA	0.1030	4.32	45.15	18,026
	Counts-Based + PCA (Normalized)	0.0382	4.06	54.38	1,340
TA Support	Counts-Based (Normalized)	0.0537	3.85	75.03	1,373
	Counts-Based + PCA	0.0579	3.97	62.72	13,192
	Counts-Based + PCA (Normalized)	0.0611	3.64	84.18	1,223
Peer Support	Counts-Based (Normalized)	0.0701	3.61	85.42	1,127
	Counts-Based + PCA	0.0868	3.38	77.19	8,273
	Counts-Based + PCA (Normalized)	0.0961	3.40	96.16	1,001

Note: Results shown are the best across 5 random states (42, 123, 456, 789, 101).

Seed Selection. For each dataset, clustering was run across five random seeds (42, 123, 456, 789, 101). The seed producing the best combination of Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index was selected for final evaluation.

Computational Requirements. The k-means preprocessing comparison required substantial computational resources to ensure thorough evaluation. Each variant analysis consumed approximately 30–45 minutes per dataset for the full k-means clustering pipeline

including PCA dimensionality reduction, cluster initialization, convergence iterations, and metric calculations. Across three datasets and three preprocessing variants, this analysis represented approximately 4.5 hours of computation time. The thorough comparison was necessary to ensure optimal preprocessing selection, as suboptimal choices could alter clustering behavior and invalidate subsequent comparisons with other methods.

6.5 Latent Dirichlet Allocation (LDA)

6.5.1 Algorithm Description

Latent Dirichlet Allocation models each document as a mixture of topics, where each topic is characterized by a distribution over words [30]. The algorithm assumes a generative process:

1. For each document d, a topic distribution θ_d is drawn from a Dirichlet distribution with parameter α.

2. For each word position in document d, a topic z is sampled from θ_d.

3. A word w is sampled from the topic-specific word distribution ϕ_z with parameter β.

The probability of observing word w in document d is:

$$P(w|d) = \sum_{z=1}^{T} P(w|z)P(z|d) = \sum_{z=1}^{T} \phi_{zw}\theta_{dz}$$

where T is the number of topics, ϕ_{zw} is the probability of word w in topic z, and θ_{dz} is the probability of topic z in document d.

Unlike k-means, LDA provides soft topic assignments. Each document receives a probability distribution over all topics rather than a single cluster label, which better reflects responses that express multiple themes.

6.5.2 Implementation Choices

In the initial analysis of the topic models using default parameters, LDA used counts-based vectorization with standard Dirichlet priors ($\alpha = \beta = 1/T \approx 0.333$ for $T = 3$). The

hyperparameter optimization below identifies the dataset-specific values that replaced these defaults.

Vectorization. Counts-based vectorization was selected over TF-IDF because LDA assumes word frequencies follow a multinomial distribution. TF-IDF converts raw counts into weighted values that no longer satisfy this multinomial count formulation [30].

Hyperparameter Optimization. A two-stage approach identified optimal values for α (document-topic density) and β (topic-word density):

- *Stage 1A (30 experiments per dataset)*: Grid search over six α values [0.05, 0.08, 0.1, 0.2, 0.3, 0.5] with β fixed at 0.01, evaluated across five random seeds.

- *Stage 1B (35 experiments per dataset)*: Fine-tuning of seven β values [0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5] with α fixed at the optimal value from Stage 1A, across five seeds.

This approach resulted in 65 experiments per dataset to provide thorough coverage of the hyperparameter space while maintaining computational feasibility.

Hyperparameter configurations were evaluated using three metrics: log-likelihood, perplexity, and UMass coherence. Log-likelihood measures how well the model explains the observed word counts; perplexity is a monotonically decreasing transformation of log-likelihood defined as

$$\text{Perplexity}(D_{\text{test}}) = \exp\left(-\frac{\sum_d \log p(w_d)}{\sum_d N_d}\right)$$

where D_{test} is a held-out test corpus, w_d are the words in document d, N_d is the word count of document d, and $p(w_d)$ is the probability of the document under the model. Lower perplexity indicates better generalization to held-out data. Although perplexity is a monotonic function of log-likelihood, scikit-learn's LDA implementation reports them through separate API calls (`score()` and `perplexity()`); both were tracked to cross-check numerical consistency across optimization runs. UMass coherence (defined in Section 6.9.1) was included because perplexity alone does not guarantee human-interpretable topics. The optimal values selected were $\alpha = 0.3$ and $\beta = 0.01$ for Faculty Support, $\alpha = 0.5$ and $\beta = 0.05$ for TA Support, and $\alpha = 0.5$ and $\beta = 0.01$ for Peer Support.

Table 6.4: LDA Optimal Hyperparameters and Model Quality Metrics

Dataset	**Best α**	**Best β**	**Coherence (UMass) (↓)**	**Log-Likelihood (↑)**	**Perplexity (↓)**
Faculty Support (n=1,667)	0.3	0.01	-2.125	72.19	706.61
TA Support (n=1,592)	0.5	0.05	-2.348	55.31	458.98
Peer Support (n=1,376)	0.5	0.01	-2.674	41.54	410.30

Note: Results shown are the best across 5 random seeds (42, 43, 44, 45, 46).

Computational Requirements. The two-stage hyperparameter optimization process required approximately 3–4 hours of computation per dataset, with each LDA model taking 2–3 minutes to converge. Across the three datasets, this analysis took approximately 9–12 hours of total computational time. This exploration was important given LDA's sensitivity to hyperparameter selection, as preliminary experiments showed that suboptimal α and β values could degrade topic coherence scores by 20–30% and classification accuracy by 10–15%. The tuning process produced results that represent each method's actual performance rather than artifacts of poor parameter choices.

6.6 Non-Negative Matrix Factorization (NMF)

6.6.1 Algorithm Description

Non-Negative Matrix Factorization decomposes a document-term matrix V (of dimension $m \times n$) into two non-negative matrices [56]:

$$V \approx WH$$

where W (dimension $m \times T$) represents document-topic weights and H (dimension $T \times n$) represents topic-word weights, with T as the number of topics.

The factorization minimizes the reconstruction error between the original matrix V and the approximation WH:

$$\min_{W \geq 0, H \geq 0} ||V - WH||_F^2 + \alpha||W||_F^2 + \alpha||H||_F^2$$

where $||\cdot||_F$ denotes the Frobenius norm, and α controls regularization strength. The

non-negativity constraints ($W \geq 0$, $H \geq 0$) produce an additive parts-based representation where words contribute positively to topics [93].

NMF uses multiplicative update rules that iteratively refine W and H while preserving non-negativity:

$$W_{ij} \leftarrow W_{ij} \frac{(VH^T)_{ij}}{(WHH^T)_{ij}}, \quad H_{ij} \leftarrow H_{ij} \frac{(W^TV)_{ij}}{(W^TWH)_{ij}}$$

This decomposition produces interpretable topics where each document is a non-negative combination of topics, and each topic is a non-negative combination of words.

6.6.2 Implementation Choices

In the default analysis, NMF applied TF-IDF vectorization with no regularization ($\alpha_W = \alpha_H = 0$, $l1\_ratio = 0.5$), following scikit-learn defaults. The optimization below refined these parameters.

Vectorization Strategy: Counts-Based vs TF-IDF. Both counts-based and TF-IDF vectorization were evaluated for NMF, as the method's reconstruction error optimization makes it sensitive to input matrix characteristics. TF-IDF achieved lower reconstruction errors (0.0007–0.0011) compared to counts-based (0.0079–0.0085), showing better model fit across all datasets. TF-IDF was therefore selected as the optimal vectorization approach for NMF.

Table 6.5: NMF Vectorization Strategy Comparison: Counts-Based vs TF-IDF

Dataset	**Vectorizer**	**Reconstruction Error (↓)**	**UMass Coherence (↑)**
Peer Support	Counts-Based	0.0079	-2.6167
	TF-IDF	0.0011	-4.0180
Faculty Support	Counts-Based	0.0085	-2.3456
	TF-IDF	0.0007	-3.6342
TA Support	Counts-Based	0.0084	-3.2451
	TF-IDF	0.0009	-2.6692

Note: Fixed parameters: init='nndsvd', solver='cd', beta_loss='kullback-leibler', alpha=0.0, l1_ratio=0.0

Hyperparameter Optimization Strategy. A two-phase parameter optimization strategy was implemented to identify optimal hyperparameters for NMF across all datasets. This approach follows established practices in sparse NMF regularization where moderate non-zero regularization improves interpretability [46].

Phase 1: Alpha_W and Alpha_H Optimization (80 experiments per dataset)

- α_W values: [0.0, 0.1, 0.3, 0.5] (4 values)
- α_H values: [0.0, 0.1, 0.3, 0.5] (4 values)
- $l1\_ratio$ fixed at 0.5 to balance L1 and L2 regularization
- Seeds: [42, 43, 44, 45, 46] (5 seeds for stability)
- Total experiments: 4 x 4 x 5 = 80 per dataset
- Selection criterion: Highest UMass coherence (Section 6.9.1) with reconstruction error (Section 6.9.1) within 5% of best-fitting model. Among configurations meeting this threshold, the one with lowest reconstruction error was selected. Diagnostic checks (topic entropy and topic balance with more than 60% of documents on a single topic) were used to flag and exclude degenerate configurations.

Phase 2: L1_ratio Fine-Tuning (25 experiments per dataset)

- α_W and α_H fixed at optimal values from Phase 1
- $l1\_ratio$ values: [0.0, 0.25, 0.5, 0.75, 1.0] (5 values)
- Seeds: [42, 43, 44, 45, 46] (5 seeds for stability)
- Total experiments: 5 x 5 = 25 per dataset

- Selection criterion: Highest UMass coherence (Section 6.9.1) with reconstruction error (Section 6.9.1) within 5–8% of Phase 1 best. Configurations where topics became too sparse (fewer than ten words per topic) were excluded by lowering $l1\_ratio$.

For convergence, max_iter was set to 1000 with KL divergence loss and multiplicative update solver (`solver='mu'`), as KL-based NMF with multiplicative updates converges more slowly than Frobenius-based alternatives. Optimization stopped early when tolerance criteria were met to maintain computational efficiency. This approach resulted in 105 experiments per dataset (80 Phase 1 + 25 Phase 2) to provide thorough coverage of the hyperparameter space while maintaining computational feasibility.

Table 6.6: NMF Optimized Hyperparameters (TF-IDF vectorization)

Dataset	**alpha_W**	**alpha_H**	**l1_ratio**	**UMass Coherence (↑)**	**Reconstruction Error (↓)**
Peer Support	0.0	0.1	0.5	-3.9414	0.0011
Faculty Support	0.0	0.3	0.0	-3.6342	0.0007
TA Support	0.0	0.3	0.0	-2.5224	0.0009

Note: Vectorization: TF-IDF (selected from Table 6.5); Fixed parameters: init='nndsvd', solver='mu', beta_loss='kullback-leibler', n_components=3, max_iter=1000

Computational Requirements. The two-phase NMF hyperparameter optimization process required approximately 5.5–7 hours per dataset (Phase 1: 4–5 hours for 80 runs; Phase 2: 1.5–2 hours for 25 runs), with each factorization taking 3–4 minutes to converge. Across all three datasets, this analysis consumed approximately 16.5–21 hours of total computation time. This optimization was necessary as NMF's performance is highly sensitive to regularization parameters: preliminary tests showed that suboptimal α_W, α_H, and $l1\_ratio$ values could degrade F1-scores by 15–25% and increase reconstruction error by 30–40%.

6.7 BERTopic

6.7.1 Algorithm Description

BERTopic is a neural topic modeling method that combines pre-trained transformer embeddings with clustering and interpretable topic representation [65]. It operates through four sequential steps:

Step 1: Document Embedding Generation. Each response is converted into a dense vector representation using a pre-trained transformer model. The transformer assigns numerical weights that capture contextual meaning, producing a high-dimensional vector that encodes semantic relationships.

$$\mathbf{e}_d = \text{Transformer}(d)$$

where $\mathbf{e}_d$ is the embedding vector for document d. Pre-trained on large text corpora, the transformer model places texts with similar meanings near each other in the embedding space regardless of word choice.

Step 2: Dimensionality Reduction. The embedding vectors are reduced to a lower-dimensional space using Uniform Manifold Approximation and Projection (UMAP), which preserves distances between nearby points and the overall topology of the data manifold [94].

$$\mathbf{e}_d^{reduced} = \text{UMAP}(\mathbf{e}_d)$$

Step 3: Clustering. The reduced vectors are grouped using a clustering algorithm. Although BERTopic typically uses HDBSCAN for automatic topic discovery, this study applies k-means with k=3 to provide direct comparison with the traditional methods in Sections 6.4–6.6.

$$\text{clusters} = \text{KMeans}(\mathbf{e}_d^{reduced}, k = 3)$$

HDBSCAN determines the number of topics automatically and treats low-density regions as outliers, which introduces inconsistency that complicates quantitative comparison across methods. k-means enforces a predefined number of topics, providing a fixed structure that

aligns with the three expert-coded themes and isolates the effect of neural embeddings from automatic topic number estimation.

Step 4: Topic Representation via Class-Based TF-IDF. All responses within a cluster are concatenated into a single composite document. Representative topic words are extracted using class-based TF-IDF (c-TF-IDF), which modifies the traditional TF-IDF scheme for topic-level analysis [65]:

$$\text{c-TF-IDF}(w, c) = \text{tf}(w, c) \times \log\left(1 + \frac{A}{\text{tf}(w)}\right)$$

where $\text{tf}(w, c)$ is the frequency of word w within topic c, A is the average number of words per topic, and $\text{tf}(w)$ is the total frequency of w across all topics. Standard TF-IDF operates at the document level, while c-TF-IDF aggregates all documents within a topic and identifies words that are distinctive for each topic compared to all others.

6.7.2 Embedding Model Selection and Comparison

In the default analysis, BERTopic used the all-MiniLM-L6-v2 embedding with UMAP dimensionality reduction and k-means ($k = 3$) clustering. The embedding model comparison below evaluates whether a higher-capacity model improved performance across datasets.

The choice of pre-trained transformer model affects BERTopic's performance because models differ in representational capacity, computational requirements, and training data. This study compares two widely used sentence embedding models from the Sentence-BERT (SBERT) family, which are designed to generate meaningful sentence-level embeddings [95].

all-MiniLM-L6-v2: A compact model with 384 dimensions, 22.7 million parameters, and 80 MB size [96]. This model was trained on over 1 billion sentence pairs and uses a 6-layer MiniLM architecture optimized for efficiency with strong representational quality. Its smaller size makes it computationally efficient and suitable for resource-constrained environments.

all-mpnet-base-v2: A larger model with 768 dimensions, 109 million parameters, and 420 MB size [97]. This model uses the MPNet architecture and was trained on multiple high-quality datasets. Its higher dimensionality and parameter count provide richer repre-

sentations but require more computational resources.

Both models use contrastive learning objectives [95], which place sentences with similar meanings near each other in the embedding space.

Cluster quality metrics (silhouette score, Davies-Bouldin index, and Calinski-Harabasz index, defined in Section 6.9.1) were evaluated for both models across all three datasets using five random seeds (Table 6.7). No consistent advantage emerged for either model: MiniLM achieved higher silhouette scores on Faculty and Peer Support, while MPNet produced higher Calinski-Harabasz values on TA Support. Because neither model dominated across datasets, both were retained for the optimized analysis reported in Chapter 7.

Table 6.7: Cluster quality metrics for embedding models across datasets

Model	**Faculty Support**			**TA Support**			**Peer Support**		
	Sil.	**DB**	**CH**	**Sil.**	**DB**	**CH**	**Sil.**	**DB**	**CH**
all-MiniLM-L6-v2	0.446	0.779	1710.5	0.484	0.761	2019.0	0.446	0.779	1710.5
all-mpnet-base-v2	0.436	0.767	1720.5	0.577	0.638	3904.2	0.433	0.825	1652.1

Note: Sil. = Silhouette Score (higher is better), DB = Davies-Bouldin Index (lower is better), CH = Calinski-Harabasz Index (higher is better). All metrics were computed in the UMAP-reduced space using five random seeds, and the best result was selected.

6.7.3 Fixed Parameters and Implementation Details

Several components of the BERTopic pipeline remained fixed across all experiments to ensure a fair comparison with k-means, LDA, and NMF approaches. This design isolates the effect of embedding model choice and maintains consistency with the traditional methods in Sections 6.4–6.6, as summarized in Table 6.8.

UMAP Parameters. In BERTopic, sentence embeddings were reduced using UMAP with n_neighbors=15, n_components=5, min_dist=0.1, and metric=cosine [94]. UMAP was selected because it preserves local groups of similar texts in embedding space, which supports coherent topic formation. Traditional models in Sections 6.4–6.6 used PCA as their

Table 6.8: Fixed parameters used across all BERTopic experiments

Component	Configuration
UMAP	n_neighbors=15, n_components=5, min_dist=0.1, metric=cosine
Random Seeds	Tested: [0, 24, 42, 77, 99]; best seed selected per dataset/model
Clustering	Algorithm: KMeans, n_clusters=3
Vectorizer	ngram_range=(1,1), stop_words='english'
BERTopic	nr_topics=3, language='english', calculate_probabilities=True

standard reducer. Using the same sentence-transformer embeddings and identical random seeds across models ensures that performance differences reflect clustering behavior rather than the dimensionality-reduction step. For each dataset-model combination, the best seed was selected based on the cluster quality metrics of Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score, as described in Section 6.9.1.

Clustering Parameters. k-means with n_clusters=3 ensured direct comparability with the traditional topic modeling methods. This design disabled BERTopic's automatic topic discovery (HDBSCAN) in favor of a fixed $T = 3$ across all methods. For discovery-mode checks, HDBSCAN was tuned on each dataset with min_cluster_size in {10, 15, 20, 25, 30} and min_samples in {3, 5, 7, 10}. The adopted configuration (min_cluster_size=15, min_samples=5, Euclidean metric) balanced higher UMass coherence with lower outlier rates across datasets; Appendix A.1.2 summarizes the tuning grid and outcomes.

Vectorizer and Tokenization Parameters. After clustering, each topic was combined into one text and tokenized into single words (unigrams) with English stop words removed. Unigrams were used because the responses are short (typically under 20 words) and to keep BERTopic consistent with the traditional models in Sections 6.4–6.6. A test run with ngram_range=(1,2) revealed small subtopics without improving coherence; bigram outputs are shown in Appendix A.1.2.

BERTopic Configuration. The model reduced topics to $T = 3$ (nr_topics=3), computed topic probabilities (calculate_probabilities=True), and processed English-language

responses. A parallel set of discovery runs left the topic count open (nr_topics=None) to check whether automatic clustering diverged from the fixed baseline. Those runs confirmed that the three expert themes remained dominant; metrics therefore report the fixed-T configuration, and the flexible outputs appear in Appendix A.1 for qualitative reference.

Computational Requirements. All BERTopic experiments ran in Google Colab. Each dataset was evaluated across five random seeds for both embedding models (all-MiniLM-L6-v2 and all-mpnet-base-v2). On a GPU-enabled runtime, one full BERTopic run for a dataset of about 1,800 short responses completed within minutes; CPU-only runs required under two hours per dataset. Overall, BERTopic required roughly 1.5–3 times more wall-clock time than traditional methods from Sections 6.4–6.6, but remained feasible for all three datasets. The larger all-mpnet-base-v2 model incurred slightly higher runtime and memory cost than all-MiniLM-L6-v2, but UMAP and k-means dominated total computation, so the difference between models was modest.

6.7.4 Discovery-Mode Sensitivity Check

To verify that fixing $T = 3$ did not hide additional structure, BERTopic was also run in discovery mode using HDBSCAN clustering with a CountVectorizer set to ngram_range=(1,2). These exploratory runs preserved the three main themes while revealing finer subtopics and a small outlier group. The analysis is qualitative and does not report accuracy or F1 because the goal was to inspect topic stability, not to compare metrics. Detailed topic hierarchies and examples appear in the Appendix A.1, based on the higher-capacity MPNet configuration.

6.8 Zero-Shot Classification (ZSC)

Zero-shot classification follows a themes-first approach that differs from the four discovery-first methods described above. In k-means, LDA, NMF, and BERTopic, statistical patterns in the text determine groupings, and expert interpretation occurs after the model has already assigned responses to clusters. ZSC inverts this sequence. Domain experts define themes before any automated labeling occurs, and the model classifies each response directly into those predefined categories. Figure 6.4 illustrates this contrast.

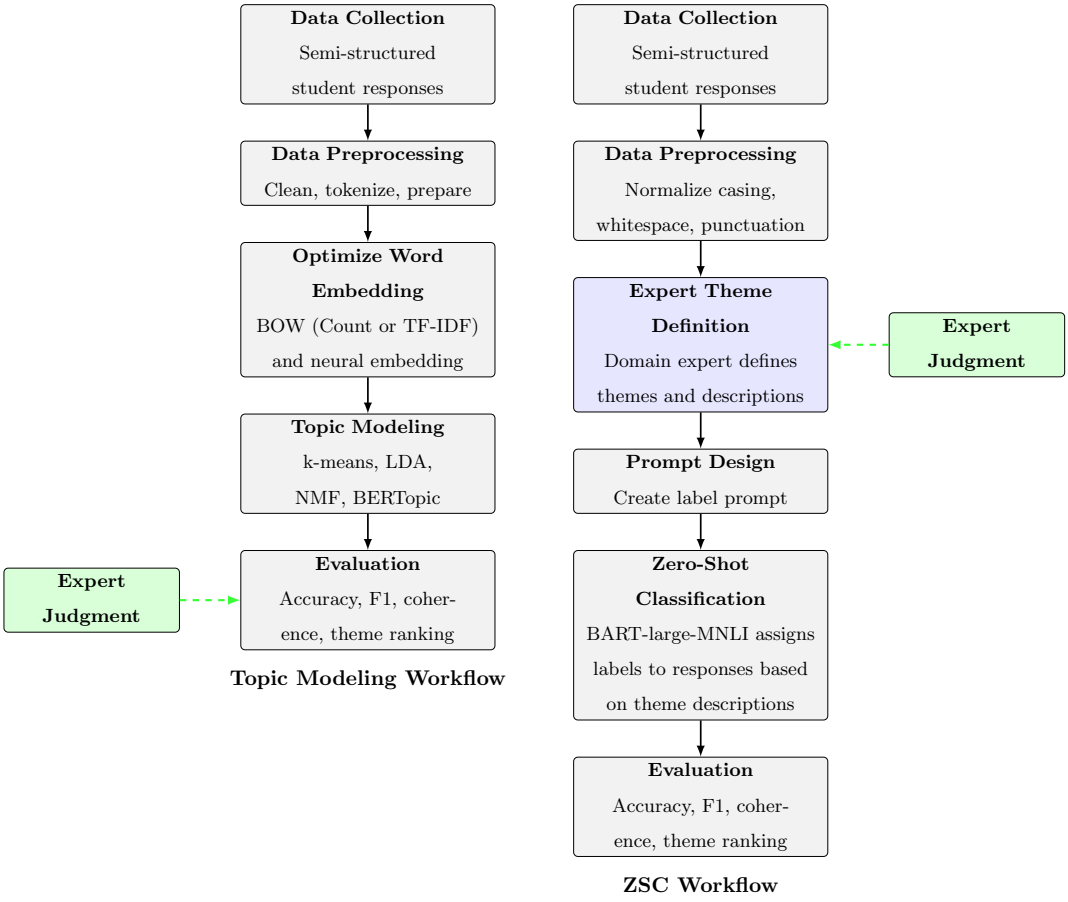


Figure 6.4: Comparison of the discovery-first topic modeling workflow (left) and the themes-first ZSC workflow (right). In the topic modeling workflow, statistical patterns in the data determine groupings before expert judgment is applied. In the ZSC workflow, expert theme definition precedes automated classification.

6.8.1 Algorithm Description

Zero-shot classification assigns text to predefined categories without training on labeled examples from the target dataset. In this study, ZSC is implemented using a natural language inference (NLI) framework. The model evaluates whether a student response supports a given theme description. Each response is paired with each expert-defined theme label, and the model estimates the strength of this relationship.

For each response d and label description h_ℓ, the model outputs a probability vector:

$$\mathbf{p}(d, h_\ell) = f_\theta(d, h_\ell) = \text{softmax}(\mathbf{z}_{\text{ent}}, \mathbf{z}_{\text{neu}}, \mathbf{z}_{\text{cnt}})$$

where $\mathbf{z}_{\text{ent}}$, $\mathbf{z}_{\text{neu}}$, and $\mathbf{z}_{\text{cnt}}$ are the logits for entailment, neutrality, and contradiction; the softmax normalizes these into probabilities that sum to one.

For single-label classification, the model assigns the label with the highest entailment probability:

$$\widehat{y}(d) = \arg\max_{\ell} \; p_{\text{ent}}(d, h_\ell)$$

For multi-label classification, responses are assigned to all labels whose entailment probability exceeds a threshold τ:

$$\widehat{Y}(d) = \{\ell : p_{\text{ent}}(d, h_\ell) \geq \tau\}$$

This entailment-based approach aligns with qualitative thematic analysis. Human coders assess whether a response reflects a theme based on meaning rather than specific words. The model performs a similar comparison by evaluating semantic similarity between responses and theme descriptions. Figure 6.5 illustrates the classification process: a student response and the expert-defined theme labels are provided as inputs, and the model outputs an entailment score for each theme.

6.8.2 Implementation Choices

Model Configuration. The analysis used `facebook/bart-large-mnli`, a transformer with a bidirectional encoder and an autoregressive decoder, fine-tuned on the Multi-NLI

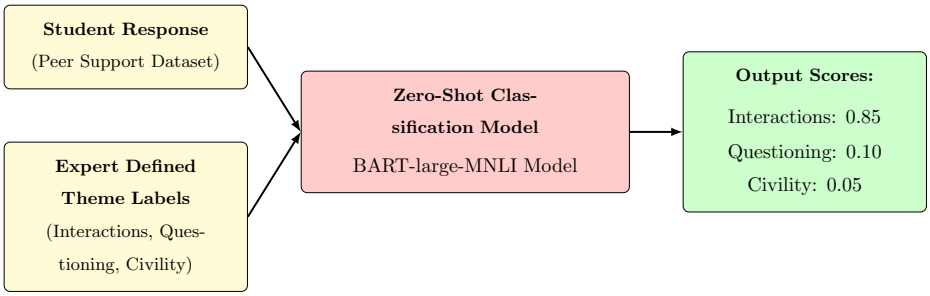


Figure 6.5: ZSC classification process. A student response (Peer Support dataset) and expert-defined theme labels are provided as inputs to the BART-large-MNLI model. The model outputs an entailment score for each theme, indicating how well the response matches each theme description.

corpus [73]. This model treats each student response as the premise and each label description as the hypothesis in an NLI framework and predicts whether the response entails, contradicts, or is neutral to the label. BART-large-MNLI was selected because it has strong reported zero-shot performance in NLI-based classification settings and runs reliably on CPU-only hardware, which avoids the high computational cost of much larger models [73]. The model was implemented using the Hugging Face pipeline interface (Python 3.8+), with data manipulation handled by pandas and numpy and evaluation conducted in scikit-learn.

Threshold. A fixed entailment threshold of $\tau = 0.5$ was applied for multi-label classification. This value was selected as a balanced default that requires strong support before assigning a label while still permitting multi-label assignment when multiple themes are present. This threshold avoids dataset-specific calibration and preserves the zero-shot setting, but future work could explore threshold optimization. If no label exceeded τ, the model assigned the single label with the highest entailment probability to avoid unlabeled responses.

Dataset Focus. The empirical analysis focuses on the Peer Support dataset. This dataset contains three validated themes (Interactions, Questioning, and Civility) with strong inter-rater agreement ($\kappa = 0.75$, Table 7.1) and a low proportion of ambiguous responses (5.89%). These properties reduced the risk that classification errors would reflect theme

ambiguity, making it the most suitable starting point for demonstrating ZSC feasibility. Applying ZSC to all three datasets was beyond the scope of this study, as each dataset would require a separate round of expert theme definition, prompt design, and validation. Extension of ZSC to the Faculty Support and TA Support datasets is identified as a direction for future work in Chapter 11. After de-duplication and cleaning, the dataset included 1,376 responses. To match student language, Interactions and Civility are referred to as Collaboration and Professionalism in the ZSC analysis, while Questioning retains its original label.

6.8.3 Data Analysis Pipeline

Three phases structured the analysis (Figure 6.6).

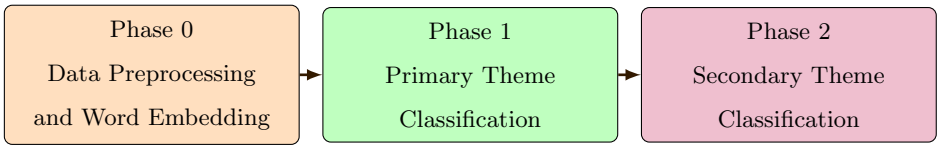


Figure 6.6: Overview of the three-phase ZSC analysis pipeline: Phase 0 (preprocessing and topic estimation), Phase 1 (primary theme classification), and Phase 2 (secondary theme analysis).

Phase 0: Data Preprocessing and Topic Estimation. Unlike the discovery-first methods, ZSC relies on the semantic meaning captured by the pretrained NLI model rather than on word frequency counts. Aggressive stopword removal and lemmatization that benefit frequency-based topic models can strip the natural language context that the NLI model uses for semantic comparison. ZSC therefore uses lighter preprocessing variants, compared in Chapter 8. Three preprocessing steps (basic, moderate, and advanced) were applied to the peer support dataset. Basic preprocessing converted text to lower case, normalized whitespace, and standardized punctuation. Moderate preprocessing also removed common stop words, numbers, and extraneous tokens while lemmatizing words. Advanced preprocessing added light stemming, rare-word filtering, and removal of very short tokens. For

preliminary structure estimation, the preprocessed student responses were embedded with a TF-IDF matrix. Elbow distortion and silhouette scores were evaluated for $k = 2$ to $k = 10$ to provide heuristic guidance on potential theme structure. Phase 0 does not produce final themes. Instead, it provides rough structure cues to support expert theme definition through vocabulary pattern analysis. Although ZSC starts with expert-defined themes, Phase 0 helps the expert understand dominant vocabulary patterns in the corpus before defining themes, ensuring themes reflect both computational patterns and qualitative meaning. Selection among the three preprocessing variants was determined empirically by comparing classification performance against expert-coded labels; results and the selected variant are reported in Chapter 8. The Phase 0 workflow is illustrated in Figure 6.7.

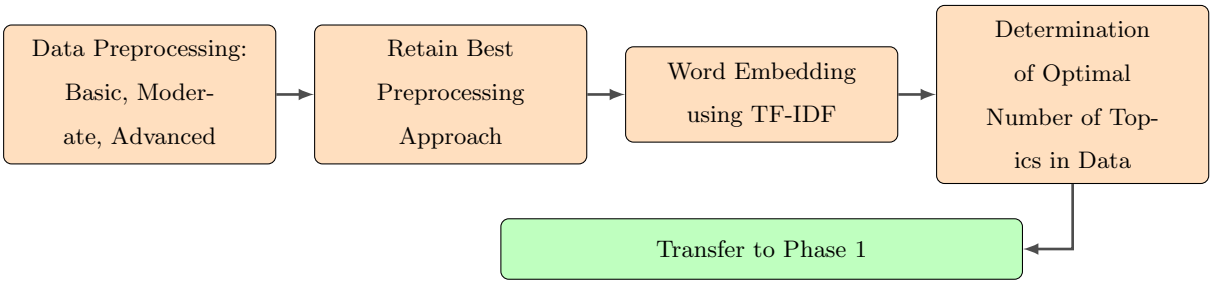


Figure 6.7: Phase 0: Data Preprocessing and Topic Estimation.

Phase 1: Primary Theme Classification. A domain expert reviewed structure cues from Phase 0 along with a 200-response subset to identify and define themes. The expert then labeled each theme using three prompt strategies varying in specificity and language register. The model was configured for multi-label classification so that each response could map to multiple relevant themes. Specific prompt wording is reported in Chapter 8, as these prompts were informed by the topic modeling results in Chapter 7. The Phase 1 workflow is shown in Figure 6.8.

Phase 2: Secondary Theme Analysis. Secondary theme analysis examined primary themes that experts identified as containing subthemes. Since these smaller themes can overlap in meaning, entailment-based prompts were explored to guide the ZSC models.

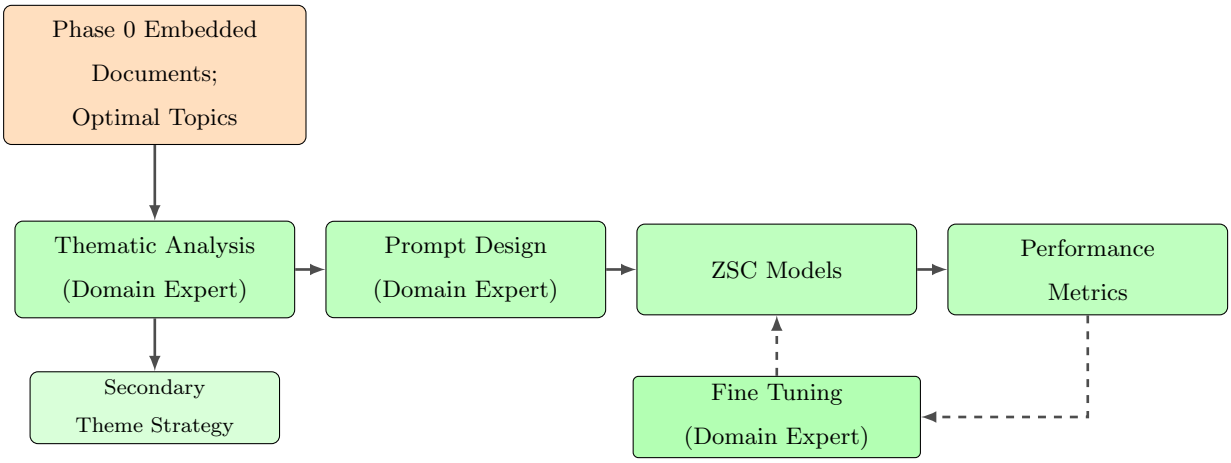


Figure 6.8: Phase 1: Primary Theme Classification.

An entailment prompt is a hypothesis such as "This student response expresses a desire for more collaborative study sessions" that is compared to each document via NLI to determine whether the document entails, contradicts, or is neutral with respect to the hypothesis. Entailment prompts provide additional context relative to labels, which can otherwise restrict models to superficial string matching. Prior work shows that NLI-guided prompting can tease apart closely related topics in zero-shot classification tasks [72]. Specific subtheme prompt wording is reported in Chapter 8. The Phase 2 workflow is shown in Figure 6.9.

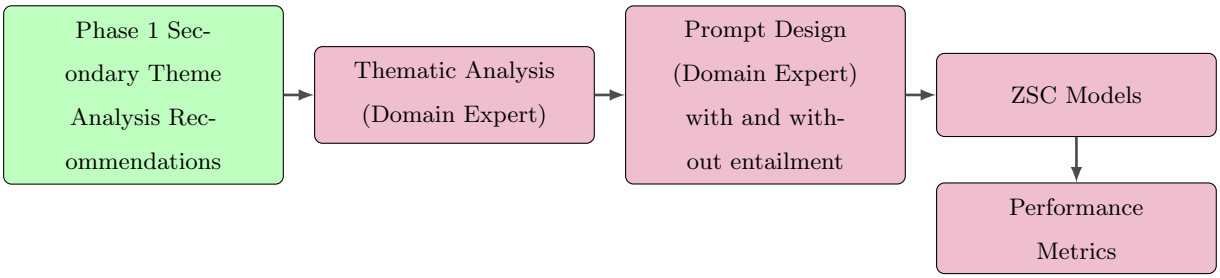


Figure 6.9: Phase 2: Secondary Theme Analysis.

Computational Requirements. Experiments ran on Google Colab with CPU-only runtime. Processing one label set (for example the mainstream label prompts) required about 1.5 hours from prompt execution through metric calculation. Evaluation metrics followed the framework described in Section 6.9 and included accuracy, precision, recall, macro- and sample-weighted F1, Jaccard overlap, and UMass coherence.

6.9 Performance Evaluation

The performance of each topic model was evaluated using a combination of internal and external metrics. Internal metrics assess model quality against the data without requiring labeled examples; external metrics compare model outputs against ground truth labels established through expert coding. Model-specific internal metrics (reconstruction error for NMF, perplexity for LDA, and within-cluster sum of squares for k-means) are defined in the respective method sections above and are not repeated here.

6.9.1 Internal Metrics

Topic Coherence

UMass coherence measures the average log conditional probability of top-word pairs in a topic, computed directly on the training corpus [98]. For a topic with top-N words $W = \{w_1, w_2, \ldots, w_N\}$:

$$C_{\text{UMass}}(W) = \frac{2}{N(N-1)} \sum_{i<j} \log \frac{D(w_i, w_j) + 1}{D(w_j)}$$

where $D(w_i, w_j)$ is the number of documents containing both w_i and w_j, $D(w_j)$ is the number of documents containing w_j, and the $+1$ smoothing avoids zero counts. Unlike PMI-based variants that require an external reference corpus, UMass coherence uses training data directly, making it well-suited for domain-specific corpora such as student survey responses. More negative values indicate lower coherence; values closer to zero reflect better word-pair co-occurrence. In this study, UMass coherence is used to guide LDA hyperparameter tuning, to support NMF regularization selection, and as a cross-method measure of topic quality

reported in results. This metric is particularly useful for short student responses, where sparse co-occurrence patterns can produce misleading topics.

Silhouette Score

Silhouette score measures how well each document fits its assigned cluster relative to adjacent clusters [99]. For a document i:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average distance from document i to other documents in the same cluster, and $b(i)$ is the minimum average distance from document i to documents in any other cluster. The score ranges from -1 to 1: values near 1 indicate well-assigned documents, values near 0 indicate documents on cluster boundaries, and values near -1 indicate likely misassigned documents. The overall silhouette score for a given clustering is the mean across all documents:

$$\text{Silhouette}(T) = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

In this study, silhouette score is used to validate the elbow-identified $T = 3$, to compare preprocessing variants during k-means optimization, and to select the best random seed in BERTopic experiments.

Calinski-Harabasz Index

The Calinski-Harabasz index (also called the Variance Ratio Criterion) measures cluster quality by comparing the dispersion between clusters to the dispersion within clusters [100]. For T clusters with N total documents:

$$\text{CH}(T) = \frac{\text{BCSS}/(T-1)}{\text{WCSS}/(N-T)}$$

where BCSS is between-cluster sum of squares and WCSS is within-cluster sum of squares. Higher values indicate more compact and better-separated clusters. In this study,

the Calinski-Harabasz index is used alongside the silhouette score to compare k-means preprocessing variants and to select the best seed for BERTopic.

Davies-Bouldin Index

The Davies-Bouldin index measures average cluster similarity, where similarity is defined as the ratio of within-cluster scatter to between-cluster distance [101]:

$$\mathrm{DB}(T) = \frac{1}{T}\sum_{i=1}^{T}\max_{j\neq i}\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}$$

where σ_i is the average distance from points in cluster i to its centroid c_i, and $d(c_i, c_j)$ is the distance between centroids c_i and c_j. Lower values indicate better-separated, more compact clusters. The Davies-Bouldin index is used alongside the silhouette score and Calinski-Harabasz index for k-means preprocessing comparisons and BERTopic seed selection.

6.9.2 External Metrics

External metrics assess how well discovered topics correspond to expert-coded ground truth themes. Each discovered topic is mapped to the most frequently occurring expert category among the documents it contains; this cluster-to-class mapping then allows standard classification metrics to be applied.

Accuracy, Precision, Recall, and F1-Score

Accuracy is the fraction of documents correctly assigned to their true theme. Precision for a given topic is the proportion of documents assigned to that topic that truly belong to the corresponding expert category. Recall is the proportion of documents in an expert category that the topic captured. The F1-score balances precision and recall:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Macro-averaged F1 is the primary summary performance metric used throughout this study. Macro-averaging weights each theme equally regardless of its frequency in the dataset,

which prevents dominant themes from masking poor performance on less frequent ones. These metrics are standard in classification evaluation [102, 103]. Results are interpreted with the recognition that moderate performance may reflect genuine differences in thematic boundaries between models and expert coders rather than model failure. Related metrics such as Normalized Mutual Information and Adjusted Rand Index are not used here, as classification metrics provide more directly interpretable topic-label alignment information for educational research [3].

Jaccard Index

The Jaccard index measures the overlap between predicted and true label sets and is used for zero-shot classification evaluation in Chapter 8, where each response may be assigned to more than one theme simultaneously [104]. For a document d with predicted label set P_d and true label set T_d:

$$J(P_d, T_d) = \frac{|P_d \cap T_d|}{|P_d \cup T_d|}$$

A value of 1 indicates that predicted and true label sets are identical; 0 indicates no overlap. The mean Jaccard index across the dataset is:

$$J = \frac{1}{N} \sum_{d=1}^{N} J(P_d, T_d)$$

Cohen's Kappa

Cohen's kappa measures inter-rater agreement beyond the level expected by chance [105]:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the proportion of responses on which two raters assigned the same label and p_e is the proportion expected to agree by chance alone. Cohen's kappa values are interpreted using the Landis and Koch scale: values above 0.80 indicate almost perfect agreement; values between 0.61 and 0.80 indicate substantial agreement; values between 0.41 and 0.60 indicate moderate agreement; values between 0.21 and 0.40 indicate fair agreement; and values below

0.20 indicate slight or no agreement beyond chance [106]. In this study, kappa is used to assess inter-rater reliability for expert coding of all three datasets (Table 7.1). The kappa values achieved across all three datasets (0.72–0.75) fall in the substantial agreement range, establishing that the ground truth labels reflect consistent human judgment rather than individual interpretation.

6.9.3 Fine-Tuning the Models

Once the number of topics is fixed at $T = 3$, hyperparameter selection optimizes each model's configuration using the metrics defined above. For LDA, perplexity and UMass coherence jointly guide α and β selection: perplexity assesses how well the model predicts held-out data, while coherence rewards topic-word distributions that form interpretable groupings. This combination guards against selecting parameters that lower perplexity at the cost of thematic clarity. For NMF, reconstruction error and UMass coherence guide regularization selection (α_W, α_H, $l1\_ratio$): reconstruction error measures how accurately the factorization approximates the original document-term matrix, while coherence ensures the resulting topics are semantically meaningful. For k-means, silhouette score and Calinski-Harabasz index are used to identify the preprocessing variant and random seed that produce the most compact and well-separated clusters (Section 6.4). BERTopic does not require hyperparameter tuning in the same sense; instead, the embedding model and random seed that yield the highest silhouette score are selected for each dataset.

6.9.4 Evaluation of Model Performance

The final evaluation stage assesses how well the topics produced by each method correspond to the expert-coded themes. Each topic is mapped to its most frequent expert category, and accuracy, precision, recall, and macro-F1 are computed at both the per-topic and overall levels. Cohen's kappa confirms that the expert-coded themes reflect consistent human judgment across independent coders. For ZSC, the Jaccard index additionally captures how well multi-label predictions match the full set of expert-assigned labels for each response.

Quantitative metrics served a dual role throughout this study: as performance bench-

marks for comparing methods and as diagnostic tools to guide expert review. Low-coherence topics were candidates for splitting, merging, or relabeling based on qualitative reasoning, keeping interpretive control with domain experts rather than with algorithm outputs. When k-means, LDA, NMF, BERTopic, and ZSC converged on the same theme content, this agreement across methods provided additional confidence that the identified themes reflect genuine patterns in student responses rather than method-specific artifacts. These diagnostics served as guidance rather than verdicts, consistent with the human-in-the-loop approach described in the Human in the Loop chapter.

With this evaluation framework established, the dissertation turns to results. Chapter 7 presents results from the five discovery-first topic modeling methods (k-means, LDA, LSA, NMF, and BERTopic) across all three datasets. Chapter 8 presents results from zero-shot classification on the Peer Support dataset.

Chapter 7

RESULTS: *K*-MEANS, LDA, LSA, NMF, AND BERTOPIC

7.1 Introduction

This chapter presents the performance of four discovery-based topic modeling methods (k-means, LDA, NMF, and BERTopic) across three educational support datasets. The analysis proceeded in two phases. The first phase established baseline performance using default parameters across all methods, including Latent Semantic Analysis (LSA). The second phase refined the analysis by optimizing hyperparameters for the remaining four methods after excluding LSA based on Phase 1 findings. This two-phase design addresses RQ1: What are the strengths and weaknesses of topic modeling approaches for thematic analysis of semi-structured short texts?

The chapter begins with the theme identification and validation framework that establishes the ground truth for all evaluations, then presents results from five topic models implemented with default parameters, the rationale for excluding LSA, optimized results for the remaining four methods, and a comparative summary.

7.2 Theme Identification and Validation Framework

Three topics per dataset produced nine total topics. These topics were evaluated for overlap and thematic content and distilled into five overarching themes by domain experts using traditional thematic analysis techniques:

Theme #1: Interactions was present in all datasets and represented students' desire for interactions with others to enhance learning. The specific nature of interactions (e.g., with peers or instructors) and formats (e.g., office hours, forums) varied among the three datasets but shared a consistent underlying focus.

Theme #2: Teaching Practice was emphasized in both faculty and TA support datasets and focused on students' desire for TAs and faculty to be more effective in their

teaching practice, whether through better delivery in the classroom, providing additional resources to support learning, synchronizing faculty-led lectures with TA-led quiz/recitation and lab sections, or similar adjustments.

Theme #3: Examples and Experience was emphasized in both faculty and TA support datasets and called for faculty and TAs to provide ample examples and experiences to reinforce teaching practice (and learning) including appropriate homework, exams, practice problems, active learning/problem solving experiences, and lab support.

Theme #4: Questioning specifically applied to the peer support dataset. Students desired that their peers ask not only more questions but more appropriate questions in class to support a more effective and interactive atmosphere for their learning.

Theme #5: Civility also applied only to the peer support dataset and included a broad range of civil behaviors expected of peers in class including but not limited to refraining from disruptive talking and other distracting activities as well as being on time to class.

These five themes represent the cross-cutting thematic structure that emerged from evaluating the nine computational topics. The expert-coding scheme used as ground truth for performance evaluation mapped these themes to each dataset, with three themes per dataset matching the k=3 validation described in Chapter 6. Two independent domain experts (raters) coded a subset of 100 responses from each of the three datasets using this expert-coding scheme to evaluate interrater reliability. The results showed strong agreement across all datasets: Faculty Support (kappa = 0.74), TA Support (kappa = 0.72), and Peer Support (kappa = 0.75). Once sufficient interrater reliability was established through this validation process, the primary domain expert coded the remaining data. These codes were then used to calculate performance metrics that required a reference (accuracy, precision, recall, F1-score, Cohen's kappa).

Once codes were refined and sufficient reliability established, the primary domain expert coded the remaining data without any further access to the topic model results (Approach 2). For Approach 1, another domain expert used the most relevant and informative of the top-ten words (i.e., keywords) associated with each topic identified by the NLP topic models (Table 7.2) to guide manual coding of the data. These codes, whether they aligned with pedagogically relevant themes or not, became the ground truth for the purposes of

Table 7.1: Interrater Reliability (Cohenś κ).

	Faculty Support	**TA Support**	**Peer Support**
Approach 1	0.64 / 0.77*	0.71	0.73
Approach 2	0.74	0.72	0.75

* A second pass was conducted after revising and clarifying themes.

computing performance metrics.

7.3 Baseline Performance with Default Parameters

The initial comparative analysis applied five topic modeling methods (k-means, LDA, NMF, BERTopic, and LSA) to three datasets using default or standard parameter configurations. Each method assigned student responses to one of three topics (k = 3) corresponding to themes identified through traditional thematic analysis. Performance was evaluated using accuracy, Cohen's kappa, precision, recall, F1-score, and UMass topic coherence. The following sections present results organized by dataset.

7.3.1 Faculty Support Dataset

Quantitative, internal and external performance metrics for the Faculty Support dataset are summarized in Table 7.3. Across all metrics, Approach 2 (A2, expert-led ground truth) performed consistently better than Approach 1 (A1, machine-led ground truth).

In a one-for-one comparison of topic models within Approach 2, k-means produced the highest external performance scores with 74.8% average accuracy and 71.8% macro-F1, followed by NMF (59.1% accuracy, 59.5% macro-F1). BERTopic produced the most consistent performance as measured by the standard deviation of performance across the three topics. LDA reached 51.7% average accuracy and 52.6% macro-F1. LSA performed poorly across all themes, achieving only 45.6% average accuracy and 28.6% macro-F1, with negative Cohen's κ values for two of the three themes.

At the theme level, performance varied substantially. The Teaching Practice theme showed the widest variation, with k-means achieving 79.9% F1 (A2) compared to LSA's

Table 7.2: Keywords for Approach 1 Manual Coding.

Dataset	**Topic**	**Relevant Keywords**
Faculty Support	0	k-means (lecture, material, note, record, clear, recording); LDA (lecture, record, online, flexible); LSA (lecture, record, recording, note, homework); NMF (lecture, record, note, slide); BERTopic (lecture, note, slide, recording)
	1	k-means (example, problem, practice, exam, homework); LDA (problem, example, material, practice, homework); LSA (example, problem, practice, exam); NMF (problem, example, exam, homework, solution); BERTopic (problem, practice, exam, example, homework)
	2	k-means (office, hour, offer, available, email); LDA (hour, office, question, ask, available, help); LSA (office, hour, available, answer); NMF (hour, office, offer, help, available, email); BERTopic (office, hour, help, available, online)
TA Support	0	k-means (lab, problem, section, example, material); LDA (lab, problem, section, example, material, homework); LSA (lab, problem, example, quiz); NMF (lab, problem, example, material, practice); BERTopic (problem, material, example)
	1	k-means (office, hour, offer, available, help); LDA (hour, office, help, offer, email); LSA (office, hour, help, available, answer); NMF (office, hour, offer, available, help); BERTopic (hour, office, email, time, help, available)
	2	k-means (question, answer, ask, available, help, email); LDA (question, answer, ask, help, available); LSA (question, answer, problem, help); NMF (question, answer, available, ask, email, help); BERTopic (lab, help, time, question)
Peer Support	0	k-means (help, group, participate, respectful, study, discussion); LDA (group, study, discussion, breakout); LSA (study, group, help, talk); NMF (study, group, help, willing, discussion); BERTopic (talk, group, help, study, discussion)
	1	k-means (talk, distract); LDA (professor, talk, participate, respectful, distract); LSA (talk, participate, distract, respectful); NMF (talk, participate, distract, respectful, quiet); BERTopic (ask, think, understand)
	2	k-means (question, ask, answer, think); LDA (help, willing, open, understand); LSA (question, ask, answer, participate); NMF (ask, question, answer, think, afraid); BERTopic (question, ask, answer, insightful, valuable)

60.9% F1 (A2). The Examples and Experience theme proved difficult for all methods except k-means, which achieved 78.5% F1 (A2) while the remaining models clustered between 50.3% and 52.6% F1. The Interactions theme produced moderate performance across methods, with LDA reaching the strongest F1 of 67.1% (A2).

In terms of internal performance (measured by U_{mass} topic coherence), NMF produced both the best coherence scores (-2.40) and the smallest variation in coherence across the three topics. BERTopic showed the weakest average coherence (-4.88), driven by poor coherence on the Examples and Experience theme (-10.6).

7.3.2 TA Support Dataset

Quantitative, internal and external performance metrics for the TA Support dataset are summarized in Table 7.4 (Approach 1) and Table 7.5 (Approach 2). The three topics determined by the topic models did not align with the three themes identified independently by the domain expert (Approach 2). Instead, the topic models generated topics associated with TA availability, question-and-answer opportunities (Q&A), and teaching practice, while the domain expert identified themes of examples and experience, interactions, and teaching practice. The availability and Q&A topics were, in combination, aligned with the interactions theme, while the examples and experience theme was absorbed into the teaching practice topic identified by all five topic models.

This misalignment between topics and themes restricted the Approach 2 performance metrics to two topics (Table 7.5) rather than the three topics enabled by Approach 1 (Table 7.4). Because Approach 2 was fundamentally unable to align all themes with topics, it was limited to two-topic evaluation for this dataset.

Approach 1 Results (Three Topics). In a one-for-one comparison of topic models using Approach 1, NMF produced the highest average external performance scores (72.5% accuracy, 71.9% macro-F1). LSA achieved the best per-topic performance on the TA Availability theme (91.0% accuracy, 83.9% F1) but produced 0.0% F1 on Teaching Practice, yielding inconsistent results overall. k-means (67.6% accuracy, 66.4% macro-F1) and LDA (67.3% accuracy, 66.9% macro-F1) produced comparable mid-range performance. LDA pro-

Table 7.3: Faculty Support Performance Metrics (Default Parameters).

		All Metrics are Percentage (%)									
Theme	**Metric**	**K-Means**		**LDA**		**LSA**		**NMF**		**BERTopic**	
		A1	**A2**	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**
Examples & Experience	Accuracy	82.0	92.3	59.4	65.6	70.8	69.8	65.5	67.8	66.4	67.8
	Cohen's κ	0.37	0.74	0.19	0.32	−0.16	−0.16	0.30	0.35	0.31	0.34
	Precision	48.8	82.2	27.2	34.5	0.50	1.00	32.8	36.3	33.3	36.0
	Recall	47.5	75.2	77.6	92.7	0.30	0.60	91.9	95.5	90.5	92.0
	F1-Score	48.1	78.5	40.2	50.3	0.40	0.80	48.4	52.6	48.7	51.7
	U_{mass}	−2.74	−2.74	−4.46	−4.46	−2.07	−2.07	−2.21	−2.21	−10.6	−10.6
Interactions	Accuracy	77.8	81.8	77.0	80.6	71.0	74.8	79.9	83.5	79.2	78.9
	Cohen's κ	0.41	0.47	0.47	0.53	0.16	0.18	0.48	0.55	0.49	0.46
	Precision	91.4	91.8	66.7	66.3	100.0	100.0	86.2	85.6	76.4	67.6
	Recall	36.4	41.3	60.4	67.9	12.1	13.7	46.5	52.2	53.4	53.4
	F1-Score	52.0	57.0	63.4	67.1	21.6	24.1	60.4	64.8	62.8	59.7
	U_{mass}	−2.13	−2.13	−2.09	−2.09	−2.47	−2.47	−2.73	−2.73	−1.75	−1.75
Teaching Practice	Accuracy	61.4	75.5	52.2	57.2	43.2	46.7	67.5	67.0	65.4	61.9
	Cohen's κ	0.23	0.50	0.04	0.16	−0.13	−0.10	0.35	0.35	0.30	0.25
	Precision	57.7	69.8	54.1	73.4	45.6	49.3	75.5	79.0	75.4	74.1
	Recall	81.5	93.5	21.6	27.8	77.9	79.7	50.4	49.9	44.4	41.3
	F1-Score	67.6	79.9	30.9	40.4	57.5	60.9	60.5	61.2	55.9	53.1
	U_{mass}	−2.45	−2.45	−2.13	−2.13	−2.89	−2.89	−2.26	−2.26	−2.25	−2.25
Overall Performance											
Avg. Accuracy		60.6	74.8	44.3	51.7	42.5	45.6	56.4	59.1	55.5	54.4
Avg. Cohen's κ		0.32	0.56	0.22	0.32	−0.05	−0.03	0.37	0.40	0.36	0.34
Avg. U_{mass}		−2.44	−2.44	−2.89	−2.89	−2.48	−2.48	−2.40	−2.40	−4.88	−4.88
Macro F1		55.9	71.8	44.8	52.6	26.5	28.6	56.4	59.5	55.8	54.8
Micro F1		60.6	74.8	44.3	51.7	42.5	45.6	56.4	59.1	55.5	54.4
Wtd. F1		59.0	73.0	43.3	50.0	35.6	38.9	58.3	60.6	56.9	54.8

duced the most consistent performance as measured by standard deviation across topics. BERTopic performed poorly with 41.7% average accuracy and 36.9% macro-F1. In terms of internal performance, k-means produced both the best average coherence scores (-2.50) and the smallest variation in coherence across the three topics.

Approach 2 Results (Two Topics). With the reduced two-topic structure, NMF again led with 69.2% average accuracy and 48.8% macro-F1. LDA followed with 64.3% accuracy and 43.9% macro-F1. k-means achieved 62.3% accuracy and 44.2% macro-F1. BERTopic reached 59.7% accuracy and 41.8% macro-F1. LSA remained the weakest at 47.6% accuracy and 22.0% macro-F1.

7.3.3 Peer Support Dataset

Quantitative, internal and external performance metrics for the Peer Support dataset are summarized in Table 7.6. In terms of the best per-topic metrics, Approach 2 (A2) performed consistently better than Approach 1 (A1). However, when considering the average performance metrics across all topics within a model, results were mixed: A1 performed best for LDA and LSA models while A2 performed best for k-means, NMF, and BERTopic. When comparing the performance of the best approach for each model, the matrix factorization techniques (NMF, LSA) generated the highest scores for both the best per-topic performance measures and the average metrics across all three topics.

NMF achieved the highest overall average performance with 75.6% accuracy and 75.5% macro-F1 (A2). LSA followed closely with 74.4% accuracy and 71.6% macro-F1 (A1) or 71.8% accuracy and 71.3% macro-F1 (A2). BERTopic reached 64.2% accuracy and 61.9% macro-F1 (A2). k-means showed 63.7% accuracy and 57.8% macro-F1 (A2). LDA achieved 60.2% accuracy and 53.7% macro-F1 (A1).

At the theme level, the Questioning theme produced the strongest individual F1-scores across methods, with LSA achieving 86.4% F1 (A2) and NMF reaching 84.5% F1 (A2). The Civility theme also showed strong performance from matrix factorization methods, with NMF at 77.2% F1 (A1) and LSA at 74.2% F1 (A1). The Interactions theme benefited from BERTopic under Approach 2, reaching 78.2% F1.

Table 7.4: TA Support Performance Metrics (Approach 1).

		All Metrics are Percentage (%)				
Theme	**Metric**	**K-Means**	**LDA**	**LSA**	**NMF**	**BERTopic**
Teaching Practice	Accuracy	69.0	78.2	53.0	77.8	66.9
	Cohen's κ	0.42	0.55	−0.15	0.57	0.34
	Precision	56.3	70.8	0.00	65.3	56.1
	Recall	94.7	75.9	0.00	92.8	73.0
	F1-Score	70.6	73.2	0.00	76.6	63.4
	U_{mass}	−2.09	−4.65	−2.79	−4.53	−5.47
TA Availability	Accuracy	89.7	83.6	91.0	89.9	60.0
	Cohen's κ	0.72	0.59	0.78	0.72	−0.04
	Precision	96.7	71.6	84.4	95.2	24.8
	Recall	65.6	68.9	83.4	67.3	20.8
	F1-Score	78.1	70.2	83.9	78.9	22.6
	U_{mass}	−2.41	−2.24	−2.21	−2.17	−2.00
Q&A	Accuracy	76.4	72.9	50.9	77.3	56.5
	Cohen's κ	0.38	0.37	0.11	0.45	−0.05
	Precision	80.5	58.9	37.3	70.5	28.4
	Recall	36.6	55.7	73.8	52.4	22.0
	F1-Score	50.3	57.2	49.5	60.1	24.8
	U_{mass}	−3.01	−2.12	−3.03	−2.53	−2.63
Overall Performance						
Average Accuracy		67.6	67.3	47.5	72.5	41.7
Average Cohen's κ		0.49	0.50	0.23	0.57	0.10
Average U_{mass}		−2.50	−3.00	−2.68	−3.07	−3.36
Macro F1		66.4	66.9	44.5	71.9	36.9
Micro F1		67.6	67.3	47.5	72.5	41.7
Sample Weighted F1		66.1	67.2	39.7	71.9	39.4

Table 7.5: TA Support Performance Metrics (Approach 2).

		All Metrics are Percentage (%)				
Theme	**Metric**	**K-Means**	**LDA**	**LSA**	**NMF**	**BERTopic**
Interactions	Accuracy	73.0	75.2	47.6	79.8	68.9
	Cohen's κ	0.48	0.50	−0.14	0.60	0.38
	Precision	90.9	75.8	51.2	89.0	74.1
	Recall	56.3	80.2	86.6	71.8	66.2
	F1-Score	69.6	78.0	64.4	79.5	70.0
	U_{mass}	−2.71	−2.18	−2.62	−2.35	−2.31
Teaching Practice	Accuracy	62.9	65.0	59.1	70.2	62.1
	Cohen's κ	0.33	0.26	−0.13	0.43	0.25
	Precision	47.6	48.6	4.10	53.6	46.0
	Recall	93.0	60.5	0.90	88.5	69.6
	F1-Score	62.9	53.9	1.50	66.8	55.4
	U_{mass}	−2.09	−4.65	−2.79	−4.53	−5.47
Overall Performance						
Average Accuracy		62.3	64.3	47.6	69.2	59.7
Average Cohen's κ		0.36	0.34	−0.12	0.46	0.28
Average U_{mass}		−2.50	−3.00	−2.68	−3.07	−3.36
Macro F1		44.2	43.9	22.0	48.8	41.8
Micro F1		62.3	64.3	47.6	69.2	59.7
Sample Weighted F1		59.3	60.8	35.7	66.1	57.0

In terms of internal performance (measured by U_{mass} topic coherence), LSA produced the best average coherence scores (-2.78), while NMF showed the smallest variation in coherence across the three topics.

7.3.4 Qualitative Theme Assessment

A qualitative evaluation of topic quality required revisiting the number of topics appropriate for describing the data. While all five topic models independently determined that three topics were optimal, human evaluation through traditional thematic analysis determined that four, three, and five themes for the Faculty, TA, and Peer Support datasets respectively were appropriate. For all three datasets, both the traditional analysis and the topic models identified a theme corresponding to interactions with students and among students. The teaching practice theme also emerged from both traditional analysis and topic models for the Faculty and TA Support datasets.

However, the hospitality theme, prominent in traditional analysis for Faculty Support, failed to emerge as a distinct topic in any of the NLP topic models. The examples and experience theme identified from traditional analysis of the TA Support dataset did not emerge as a distinct theme from the topic models but was instead merged with teaching practice. In its place, the topic models split the interactions theme into two topics corresponding to TA availability and Q&A with students. The topic models absorbed the preparation theme into the civility theme and the engagement theme into the interactions theme in the Peer Support dataset. While these merging events did not compromise the alignment between traditional thematic analysis and topic model analysis for Peer Support, they reduced the granularity of the results.

7.3.5 Theme Rankings

As an additional measure of topic model performance, topics were ranked by frequency for each model and compared to the rankings from traditional thematic analysis. None of the topic models replicated the rankings from traditional thematic analysis for the Faculty Support dataset. Three of the five topic models replicated rankings for the TA Support

Table 7.6: Peer Support Performance Metrics (Default Parameters).

		All Metrics are Percentage (%)									
Theme	**Metric**	**K-Means**		**LDA**		**LSA**		**NMF**		**BERTopic**	
		A1	**A2**	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**
Interactions	Accuracy	51.5	65.2	72.5	61.5	82.2	76.5	79.5	80.9	77.7	83.1
	Cohen's κ	0.20	0.35	0.18	0.13	0.47	0.47	0.51	0.60	0.48	0.64
	Precision	30.3	54.2	36.0	54.8	59.0	89.9	51.6	80.0	49.1	81.6
	Recall	95.6	91.4	34.7	28.3	57.8	47.3	84.7	70.3	84.7	75.2
	F1-Score	46.0	68.0	35.4	37.3	58.4	62.0	64.1	74.8	62.2	78.2
	U_{mass}	−5.43	−5.43	−4.30	−4.30	−2.48	−2.48	−3.80	−3.80	−9.26	−9.26
Civility	Accuracy	76.2	73.6	70.5	67.4	82.7	76.7	86.5	80.4	64.6	67.1
	Cohen's κ	0.23	0.16	0.28	0.21	0.62	0.49	0.68	0.53	0.30	0.35
	Precision	74.0	60.6	46.5	42.4	62.8	56.0	72.1	63.3	42.3	45.4
	Recall	20.6	16.6	49.9	44.5	90.6	79.0	83.1	71.6	79.4	83.4
	F1-Score	32.3	26.0	48.1	43.4	74.2	65.5	77.2	67.2	55.2	58.8
	U_{mass}	−2.54	−2.54	−2.94	−2.94	−3.24	−3.24	−6.16	−6.16	−8.20	−8.20
Questioning	Accuracy	72.6	88.6	77.5	75.3	84.0	90.4	80.1	90.0	59.8	78.1
	Cohen's κ	0.46	0.72	0.55	0.51	0.68	0.79	0.60	0.77	0.21	0.39
	Precision	98.8	91.7	78.5	56.8	94.6	77.9	97.3	82.8	98.0	93.3
	Recall	46.7	70.0	76.7	89.7	72.8	97.0	62.7	86.2	21.4	32.9
	F1-Score	63.5	79.4	77.6	69.6	82.3	86.4	76.2	84.5	35.2	48.6
	U_{mass}	−3.64	−3.64	−6.25	−6.25	−2.63	−2.63	−3.03	−3.03	−6.20	−6.20
Overall Performance											
Avg. Accuracy		50.1	63.7	60.2	52.1	74.4	71.8	73.0	75.6	51.0	64.2
Avg. Cohen's κ		0.30	0.42	0.36	0.29	0.60	0.58	0.60	0.63	0.32	0.46
Avg. U_{mass}		−3.87	−3.87	−4.50	−4.50	−2.78	−2.78	−4.33	−4.33	−7.89	−7.89
Macro F1		47.3	57.8	53.7	50.1	71.6	71.3	72.5	75.5	50.8	61.9
Micro F1		50.1	63.7	60.2	52.1	74.4	71.8	73.0	75.6	51.0	64.2
Wtd. F1		51.1	59.8	60.4	49.2	74.9	70.6	73.9	75.7	46.5	63.5

Table 7.7: NLP Topics vs. Traditional Themes by Dataset.

Dataset	**NLP Topics***	**Themes**	**Theme Descriptions (Traditional Analysis)**
Faculty Support	3	4	*Interactions*: interactions with students, active learning, providing feedback, facilitating collaborations. *Examples & Experience*: plentiful examples, real-world applications, and resources for learning; quality delivery. *Teaching Practice*: reasonable assessment practices, clear expectations, and well-organized materials and assignments. *Hospitality*: kindness, empathy, flexibility, and other elements of care (did not emerge in NLP topic models).
TA Support	3	3	*Interactions*: similar to Faculty Support; split into two topics by NLP topic models (TA Availability and Q&A). *Examples & Experience*: similar to Faculty Support; did not emerge as a distinct topic in NLP models (absorbed into Teaching Practice). *Teaching Practice*: similar to Faculty Support.
Peer Support	3	5	*Questioning*: expectations of other students to ask relevant questions. *Interactions*: social interactions and collaborations with other students (merged with Engagement by NLP topic models). *Civility*: polite, respectful behavior; refraining from disruptive talking; being on time to class (merged with Preparation by NLP topic models). *Engagement*: paying attention and participating in class. *Preparation*: coming to class prepared to be productive.

* Keywords associated with topics identified by NLP topic models are listed in Table 7.2.

dataset (Approach 2), and two of the five topic models did so for the Peer Support dataset (Approach 2).

For Faculty Support, k-means and LSA correctly ranked Teaching Practice first but failed to replicate the correct second- and third-place order (Interactions then Examples and Experience), while LDA, NMF, and BERTopic all placed Examples and Experience first despite its third-place ranking from traditional analysis. For TA Support under Approach 2, k-means, LDA, and BERTopic correctly ranked Interactions first, while LSA and NMF reversed the rankings. For Peer Support under Approach 2, k-means and NMF correctly ranked Interactions first, while the remaining models showed various inversions across the three themes.

Table 7.8: Theme Frequency Rankings by Model.

Dataset	Approach	Theme (Rank)*	k-means	LDA	LSA	NMF	BERTopic
Faculty Support	A1	Examples & Experience (3)	2	1	2	1	1
		Interactions (2)	3	2	3	3	3
		Teaching Practice (1)	1	3	1	2	2
	A2	Examples & Experience (3)	2	1	2	1	1
		Interactions (2)	3	2	3	3	3
		Teaching Practice (1)	1	3	1	2	2
TA Support	A1	Examples & Experience**	—	—	—	—	—
		Interactions (2)	1	1	2	2	1
		Teaching Practice (1)	2	2	1	1	2
	A2	Examples & Experience**	—	—	—	—	—
		Interactions (1)	1	1	2	2	1
		Teaching Practice (2)	2	2	1	1	2
Peer Support	A1	Civility & Preparation (2)	3	2	1	3	1
		Interactions & Engagement (3)	1	3	3	1	2
		Questioning Practice (1)	2	1	2	2	3
	A2	Civility & Preparation (3)	3	2	1	3	1
		Interactions & Engagement (1)	1	3	3	1	2
		Questioning Practice (2)	2	1	2	2	3

* Rank in parentheses indicates ranking from traditional (manual) thematic analysis (1 = most frequent theme).

** Examples & Experience did not emerge as a distinct theme with topic model analysis for the TA Support dataset.

Figures 7.1, 7.2, and 7.3 show the document frequency distributions for Faculty Support, TA Support, and Peer Support, respectively, comparing expert-coded and model-assigned theme counts across all five methods.

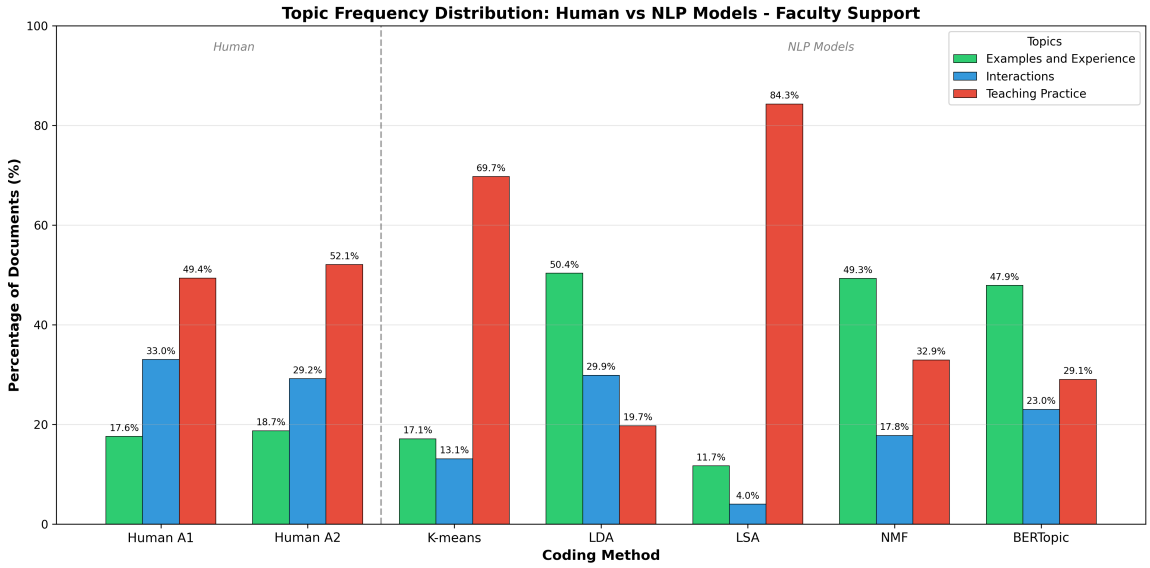


Figure 7.1: Faculty Support Topic Frequency Distributions.

Note on LSA exclusion. In the default phase, LSA applied TF-IDF vectorization with truncated SVD at k = 3 components using scikit-learn defaults. LSA was excluded from the optimized parameter analysis based on Phase 1 results: LSA produced 0.0% F1 for Teaching Practice (TA Support), negative Cohen's κ values in Faculty Support, and was outperformed by NMF on external metrics in two of three datasets. NMF was retained as the representative matrix factorization method.

7.3.6 Interferents in Topic Model Performance

Ambiguity and multiple-topic responses affected performance across all methods and datasets. The TA Support dataset had the highest rates of ambiguous responses (18.6%), and the

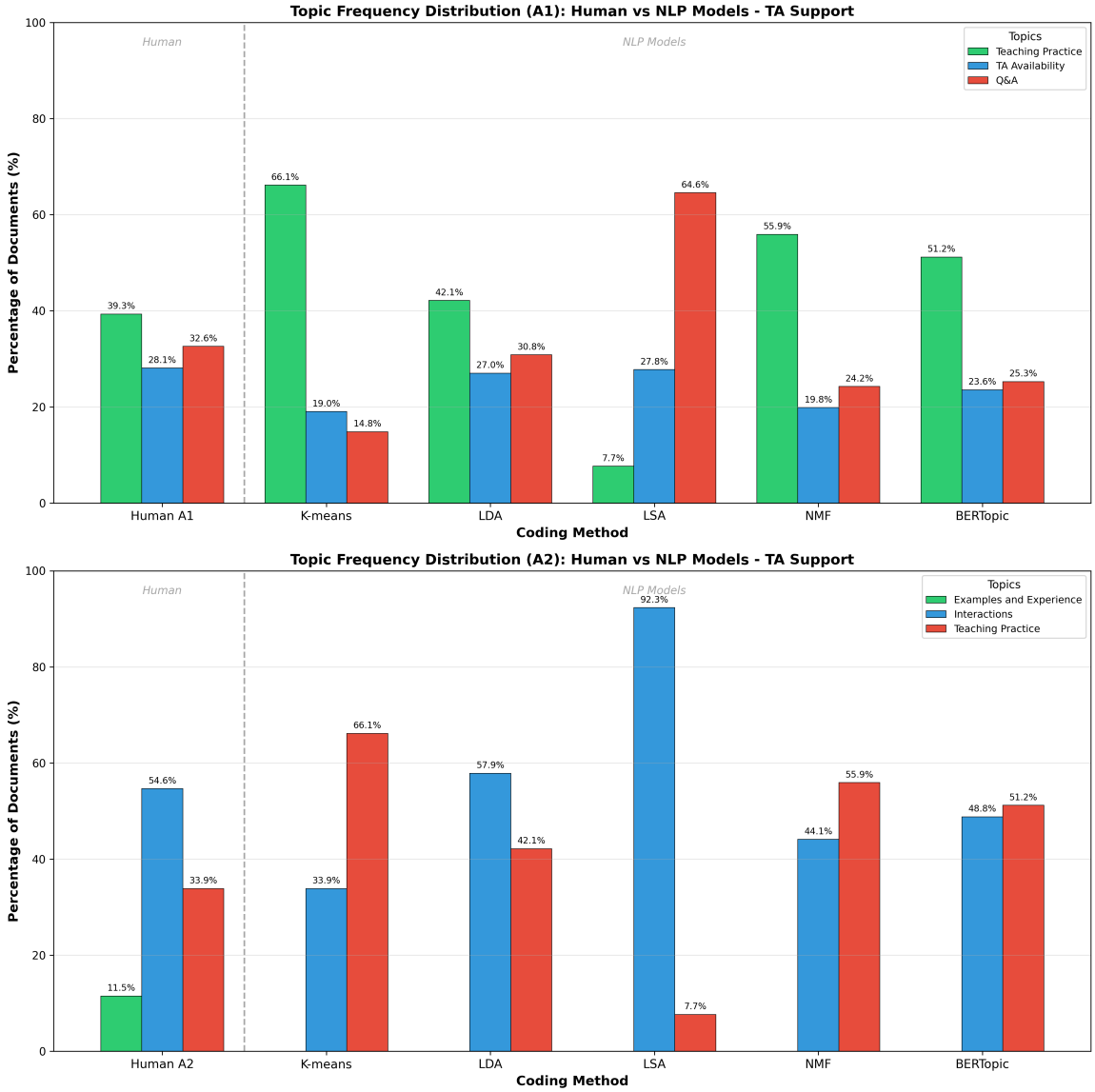


Figure 7.2: TA Support Topic Frequency Distributions.

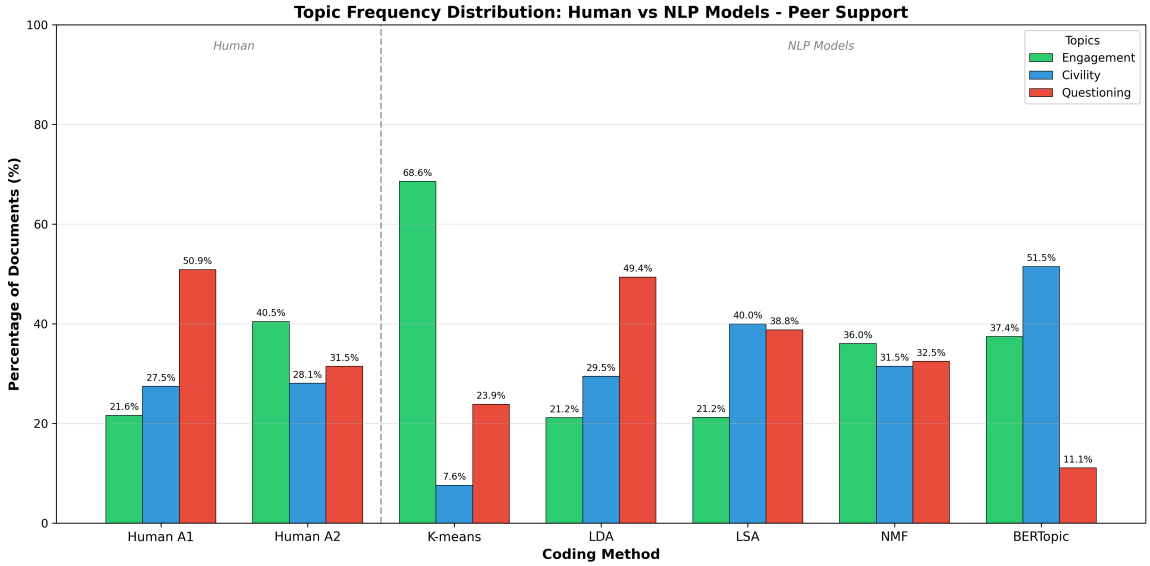


Figure 7.3: Peer Support Topic Frequency Distributions.

Faculty Support dataset had the highest rate of multiple-topic responses (26.5%), while Peer Support demonstrated the lowest rates overall (5.89% ambiguous, 6.86% multiple-topic).

Table 7.9: Ambiguous and Multiple Topic Responses

	Faculty Support	TA Support	Peer Support
Ambiguous Response	13.7%	18.6%	5.89%
Multiple Topic Response	26.5%	11.7%	6.86%

7.4 Optimized Parameter Results

The second phase applied hyperparameter optimization to four methods: k-means, LDA, NMF, and BERTopic. Optimization strategies are detailed in Chapter 6. Since Section 7.3 established Approach 2 as consistently superior to Approach 1 across all methods and datasets, all results in this section use the expert-coded ground truth (Approach 2) exclu-

sively. BERTopic was evaluated with two sentence-embedding models: all-MiniLM-L6-v2 (MiniLM; 384 dimensions, 22.7M parameters) and all-mpnet-base-v2 (MPNet; 768 dimensions, 109M parameters). Both models used UMAP dimensionality reduction and k-means with $k = 3$, maintaining comparability with the traditional methods. Cluster quality metrics showed no consistent advantage for either model across datasets (Chapter 6, Table 6.7); both are therefore included in the per-dataset comparisons below.

7.4.1 Faculty Support Dataset

Quantitative performance metrics for the Faculty Support dataset with optimized parameters are summarized in Table 7.10. NMF and BERTopic MPNet achieved the highest overall accuracy (76.76% and 76.8% respectively). BERTopic MPNet also led on macro-F1 (65.7%), with NMF close behind at 62.82%. k-means and LDA showed comparable accuracy (68.67% and 68.13%) but LDA achieved higher macro-F1 (48.07% vs 45.79%).

Theme-level variation was substantial across all methods. The Examples and Experience theme was the most challenging: k-means reached only 23.64% F1 and LDA 22.01% F1, while NMF improved markedly to 62.81% F1 and BERTopic MPNet to 63.0% F1. Teaching Practice was handled most consistently, with NMF (69.75% F1) and BERTopic MPNet (67.0% F1) leading. For Interactions, LDA achieved the strongest F1 among traditional methods (66.35%), while BERTopic MPNet reached 67.1% F1 with better balanced precision and recall. UMass coherence scores were generally low, with k-means showing the least negative average (-2.10) and NMF the most negative (-3.63).

Frequency distributions (Figure 7.4) showed that k-means and NMF produced nearly identical patterns, both over-representing Teaching Practice (1,003 vs 790 expert-coded) and Examples and Experience (446 vs 307) while under-representing Interactions (218 vs 550). LDA produced a more balanced distribution, approximating expert counts for Interactions (517 vs 550) and Teaching Practice (739 vs 790) while still inflating Examples and Experience (411 vs 307). Both BERTopic models under-represented Teaching Practice by approximately 30% and over-represented Examples and Experience.

Table 7.10: Faculty Support Performance Metrics (Optimized Parameters, Approach 2). All percentage metrics (%); U_{mass} and κ are unitless.

Theme	**Metric**	**K-Means**	**LDA**	**NMF**	**BERT-Mini**	**BERT-MPNet**
Teaching Practice	Accuracy	58.85	59.51	67.01	67.4	68.2
	Precision	54.58	57.78	61.67	72.5	70.8
	Recall	78.48	54.05	80.25	50.3	64.1
	F1-Score	64.38	55.85	69.75	59.3	67.0
	U_{mass}	−1.79	−2.15	−1.91	−2.19	−2.50
Interactions	Accuracy	76.84	78.46	78.70	77.9	84.0
	Precision	88.68	68.47	88.24	70.2	68.4
	Recall	34.18	64.36	40.91	57.3	69.6
	F1-Score	49.34	66.35	55.90	63.1	67.1
	U_{mass}	−2.44	−1.75	−2.40	−2.36	−2.27
Examp. & Exp.	Accuracy	71.33	66.41	84.58	72.6	78.1
	Precision	23.20	19.22	56.51	38.8	52.3
	Recall	24.10	25.73	70.68	84.7	79.2
	F1-Score	23.64	22.01	62.81	53.2	63.0
	U_{mass}	−2.08	−2.48	−1.98	−3.55	−2.69
Overall Performance						
Overall	Avg. Accuracy (%)	68.67	68.13	76.76	72.6	76.8
	Macro F1 (%)	45.79	48.07	62.82	58.6	65.7
	Micro F1 (%)	68.67	68.13	64.94	72.6	76.8
	Wtd. F1 (%)	45.79	54.50	64.55	59.5	68.0
	Cohen's κ	0.207	0.234	0.441	0.390	0.416
	Avg. U_{mass}	−2.10	−2.12	−3.63	−2.49	−2.70

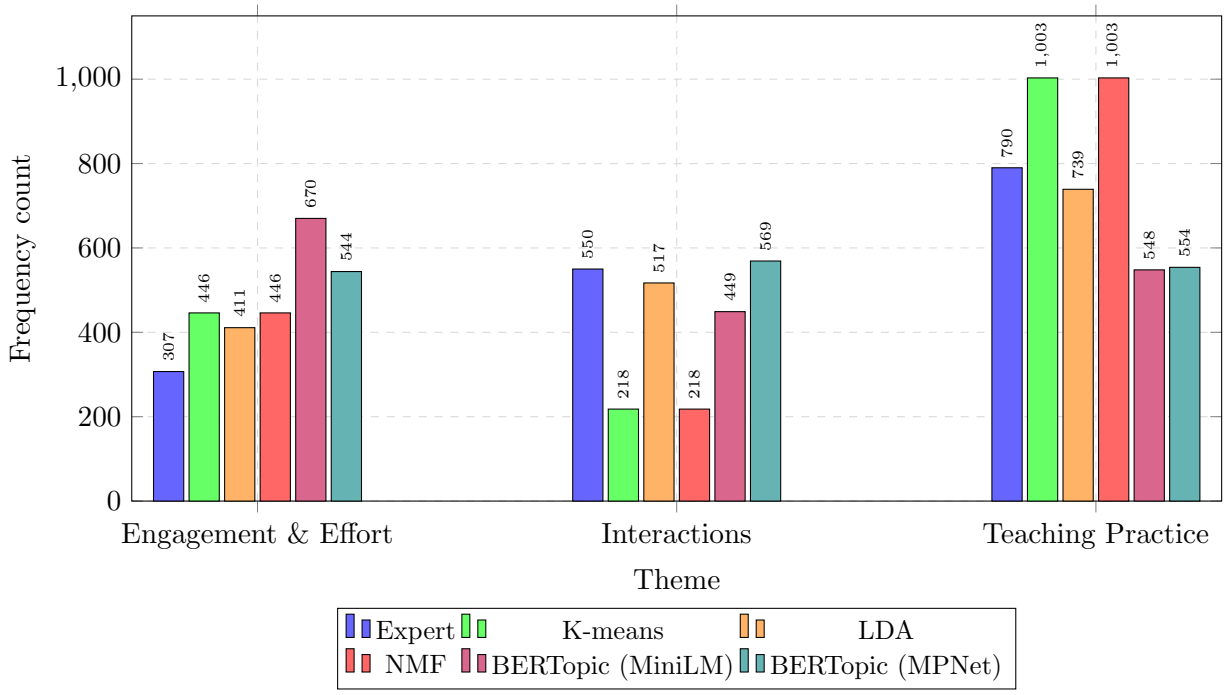


Figure 7.4: Optimized topic frequency distributions for Faculty Support: expert-coded counts versus five NLP methods.

7.4.2 TA Support Dataset

Quantitative performance metrics for the TA Support dataset with optimized parameters are summarized in Table 7.11. BERTopic MPNet achieved the highest overall accuracy (70.3%) and macro-F1 (54.2%), followed closely by NMF (69.91% accuracy, 43.74% macro-F1) and k-means (69.87% accuracy, 42.39% macro-F1). LDA showed both the lowest accuracy (58.27%) and the lowest macro-F1 among the three traditional methods (32.38%), despite achieving the strongest Interactions F1 among those three methods (68.29%).

The Examples and Experience theme was the most difficult across all methods. k-means reached only 6.72% F1, NMF 3.35% F1, and both BERTopic models below 28% F1; only LDA partially captured this theme with 33.18% F1 due to high recall (80.0%). Teaching Practice was handled well by NMF (66.67% F1) and BERTopic MPNet (67.1% F1). Interactions showed strong precision across most methods (76.67%–96.20%) but consistently low recall (37.39%–57.07%), indicating that topic models assigned a substantial share of Interactions-labeled responses to other categories. UMass coherence scores were low overall, with the Interactions theme showing the weakest coherence for BERTopic (-4.42 for both models), likely reflecting the thematic breadth of TA-student interaction responses.

Frequency comparisons (Figure 7.5) showed that k-means substantially over-represented Interactions (1,026 vs 813 expert-coded) while under-representing Teaching Practice (250 vs 643). LDA inflated Examples and Experience (516 vs 135 expert-coded) and under-represented Interactions (546 vs 813). Both BERTopic models captured only about half the expert-coded Interactions count (MiniLM: 457; MPNet: 452 vs 813) while inflating Examples and Experience substantially (MiniLM: 580; MPNet: 394 vs 135 expert-coded).

7.4.3 Peer Support Dataset

Quantitative performance metrics for the Peer Support dataset with optimized parameters are summarized in Table 7.12. BERTopic MiniLM achieved the highest performance by a clear margin: 85.0% accuracy and 77.0% macro-F1 with a Cohen's κ of 0.642. LDA followed with 78.46% accuracy and 67.17% macro-F1. k-means (73.33% accuracy, 58.01% macro-F1) and NMF (72.94% accuracy, 59.02% macro-F1) performed comparably. BERTopic

Table 7.11: TA Support Performance Metrics (Optimized Parameters, Approach 2). All percentage metrics (%); U_{mass} and κ are unitless.

Theme	**Metric**	**K-Means**	**LDA**	**NMF**	**BERT-Mini**	**BERT-MPNet**
Teaching Practice	Accuracy	65.01	56.85	71.67	59.3	68.9
	Precision	54.20	45.85	63.52	49.6	66.8
	Recall	86.31	37.79	70.14	42.8	67.5
	F1-Score	66.59	41.43	66.67	45.9	67.1
	U_{mass}	−1.47	−2.56	−1.77	−2.39	−2.05
Interactions	Accuracy	67.27	72.93	67.02	70.5	72.4
	Precision	96.20	84.98	76.67	87.5	91.4
	Recall	37.39	57.07	50.92	49.2	56.8
	F1-Score	53.85	68.29	61.20	63.0	69.6
	U_{mass}	−1.74	−1.93	−1.79	−4.42	−4.42
Examp. & Exp.	Accuracy	77.32	72.68	71.04	67.5	71.2
	Precision	5.16	20.93	2.34	17.1	28.0
	Recall	9.63	80.00	5.93	73.3	24.3
	F1-Score	6.72	33.18	3.35	27.7	25.9
	U_{mass}	−1.57	−2.56	−2.16	−1.93	−2.65
Overall Performance						
Overall	Avg. Accuracy (%)	69.87	58.27	69.91	65.8	70.3
	Macro F1 (%)	42.39	32.38	43.74	45.5	54.2
	Micro F1 (%)	69.87	58.27	54.85	65.8	69.7
	Wtd. F1 (%)	42.39	47.03	54.84	53.1	59.3
	Cohen's κ	0.214	0.256	0.222	0.246	0.295
	Avg. U_{mass}	−1.59	−2.35	−2.52	−3.04	−3.01

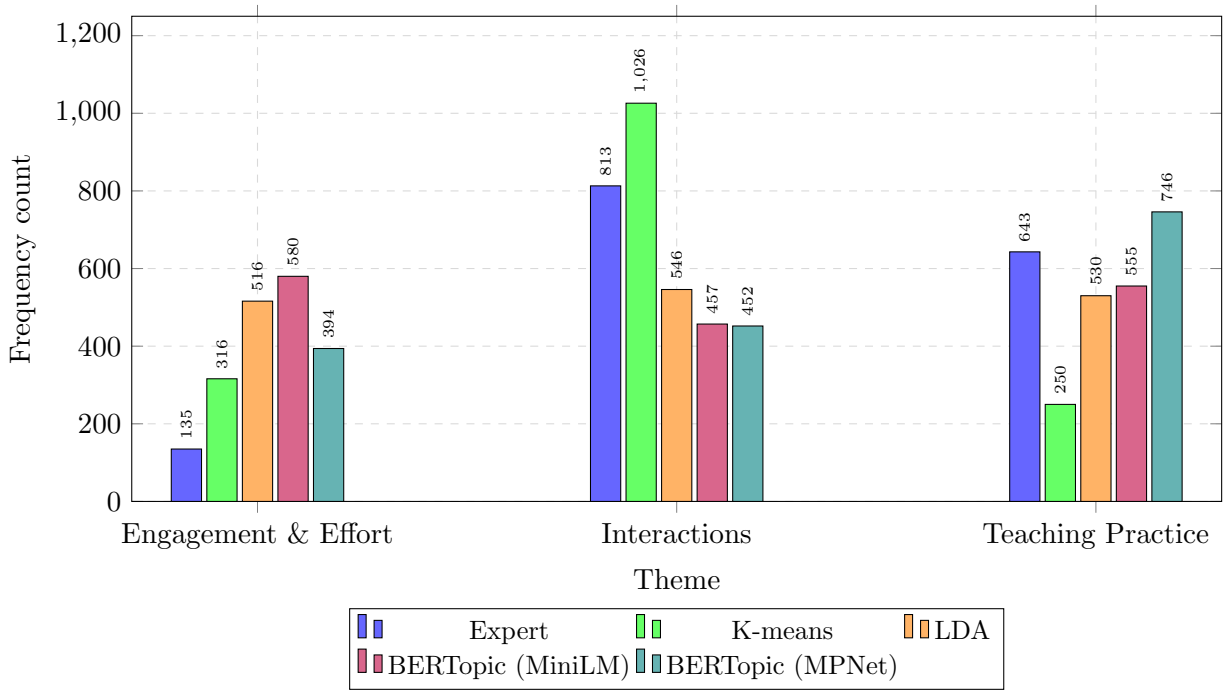


Figure 7.5: Optimized topic frequency distributions for TA Support: expert-coded counts versus four NLP methods (NMF excluded).

MPNet showed notably weaker results (70.7% accuracy, 56.6% macro-F1), representing a 20.4 percentage point macro-F1 gap versus MiniLM on this dataset.

The Questioning theme was handled well by most methods: k-means reached 84.38% F1, BERTopic MiniLM 85.4% F1, and NMF 82.53% F1, while LDA achieved 73.40% F1. BERTopic MPNet was a clear outlier on Questioning (48.8% F1). Civility was the most variable theme: k-means achieved only 31.29% F1 compared to LDA (62.35%), BERTopic MiniLM (71.8%), and BERTopic MPNet (72.9%). Interactions showed the widest spread between BERTopic models: MiniLM reached 73.8% F1 while MPNet achieved only 48.2% F1. BERTopic MiniLM maintained close alignment with expert-coded frequencies across all three themes (within 10%), while MPNet amplified Questioning and under-represented Civility (Figure 7.6).

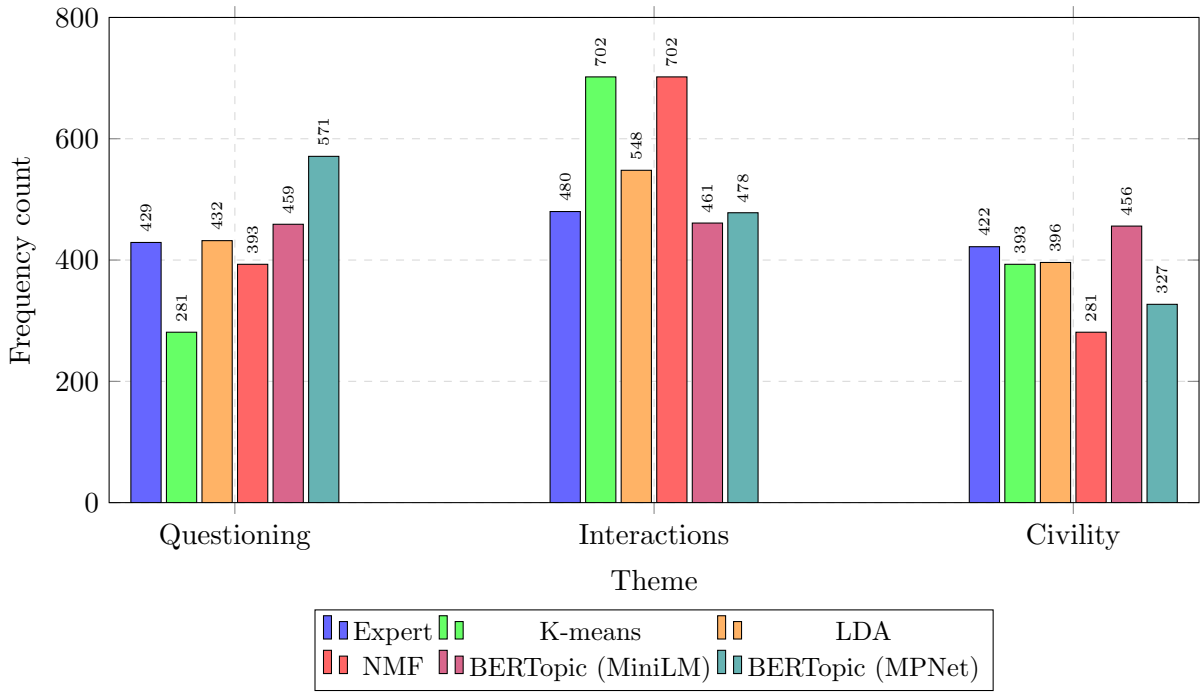


Figure 7.6: Optimized topic frequency distributions for Peer Support: expert-coded counts versus five NLP methods.

Table 7.12: Peer Support Performance Metrics (Optimized Parameters, Approach 2). All percentage metrics (%); U_{mass} and κ are unitless.

Theme	**Metric**	**K-Means**	**LDA**	**NMF**	**BERT-Mini**	**BERT-MPNet**
Questioning	Accuracy	90.55	83.36	88.95	89.7	62.8
	Precision	87.10	73.15	81.41	83.5	42.7
	Recall	81.82	73.66	83.68	87.5	56.9
	F1-Score	84.38	73.40	82.53	85.4	48.8
	U_{mass}	−3.40	−2.91	−1.76	−3.05	−2.10
Interactions	Accuracy	64.53	74.42	64.53	82.1	64.0
	Precision	49.42	61.68	49.20	75.3	48.3
	Recall	71.25	70.42	51.25	72.3	48.1
	F1-Score	58.36	65.76	50.20	73.8	48.2
	U_{mass}	−1.83	−2.66	−2.52	−6.71	−3.19
Civility	Accuracy	64.90	77.62	65.33	82.0	85.3
	Precision	39.15	64.39	43.68	69.1	83.5
	Recall	26.07	60.43	45.02	74.6	64.7
	F1-Score	31.29	62.35	44.34	71.8	72.9
	U_{mass}	−1.63	−2.46	−2.52	−3.54	−4.23
Overall Performance						
Overall	Avg. Accuracy (%)	73.33	78.46	72.94	85.0	70.7
	Macro F1 (%)	58.01	67.17	59.02	77.0	56.6
	Micro F1 (%)	73.33	78.46	58.74	77.1	57.4
	Wtd. F1 (%)	58.01	67.50	57.78	77.2	57.4
	Cohen's κ	0.386	0.511	0.388	0.642	0.348
	Avg. U_{mass}	−2.29	−2.67	−3.94	−3.17	−4.43

7.4.4 BERTopic Performance Summary

BERTopic performance across both embedding models and all three datasets is summarized in Table 7.13. Averaged across datasets, the compact MiniLM model (60.4% macro-F1) slightly outperformed the larger MPNet model (58.8% macro-F1) despite having 4.8 times fewer parameters. Dataset properties determined which model led: MiniLM dominated Peer Support (77.0% vs 56.6% macro-F1), while MPNet achieved higher macro-F1 in Faculty Support (65.7% vs 58.6%) and TA Support (54.2% vs 45.5%). A discovery-mode sensitivity check using HDBSCAN clustering with bigram tokenization confirmed the same three high-level themes in each dataset while revealing finer subtopics. The fixed-k results are retained as the metric baseline, and qualitative discovery outputs appear in Appendix A.1.

Table 7.13: BERTopic Macro-F1 Scores by Embedding Model.

Dataset	**MiniLM (%)**	**MPNet (%)**	**Difference**	**Winner**
Faculty Support	58.6	65.7	+7.1%	**MPNet**
TA Support	45.5	54.2	+8.7%	**MPNet**
Peer Support	77.0	56.6	-20.4%	**MiniLM**
Average	**60.4**	**58.8**	**-1.6%**	**MiniLM**

Chapter 8 presents results from zero-shot classification, a themes-first approach that incorporates expert-defined labels at the start of the analysis rather than discovering topics post hoc.

Chapter 8

RESULTS: ZERO-SHOT CLASSIFICATION

8.1 Introduction

This chapter presents results from the zero-shot classification (ZSC) analysis of the Peer Support dataset. ZSC follows a themes-first approach in which domain experts define themes before any automated labeling occurs. The model then assigns each response directly to these predefined themes using natural language inference. This approach contrasts with the discovery-first methods presented in Chapter 7, where themes emerge from statistical patterns and require post-hoc interpretation.

The analysis focused on the Peer Support dataset (1,376 responses after de-duplication and cleaning), which contains three validated themes with strong inter-rater agreement ($\kappa = 0.75$, Chapter 7) and a low proportion of ambiguous responses (5.89%). These properties make the dataset suitable for evaluating how prompt phrasing affects classification accuracy. To match student language, Interactions and Civility are referred to as Collaboration and Professionalism in this chapter, while Questioning retains its original label.

The ZSC method, model configuration, and three-phase data analysis pipeline are described in Chapter 6. Briefly, the analysis proceeded through Phase 0 (data preprocessing and topic estimation), Phase 1 (primary theme classification with three prompt design strategies), and Phase 2 (secondary theme analysis examining subthemes within primary categories). The following sections present results from each phase. A cross-method comparison with the discovery-first methods from Chapter 7 appears in Chapter 9.

8.2 Data Preprocessing and Topic Estimation (Phase 0)

Phase 0 compared basic, moderate, and advanced preprocessing pipelines on the 1,376 peer-support responses. Moderate and advanced pipelines nearly halved mean word counts, removing short context markers such as negations and pronouns that the NLI classifier uses

for entailment decisions. The analysis retained the basic pipeline, which normalized only casing, whitespace, and punctuation. Table 8.1 summarizes the effect of each preprocessing level on mean word count.

Table 8.1: Peer Support Document Statistics Across Preprocessing Pipelines

Pipeline	Word Count			Character Count		
	Mean	Median	SD	Mean	Median	SD
Raw	16.2	11.0	16.0	94.7	66.0	92.7
Basic	16.2	11.0	16.0	94.5	66.0	92.6
Moderate	7.7	6.0	7.1	58.2	42.0	55.6
Advanced	7.6	6.0	7.0	56.5	41.0	53.9

For preliminary structure estimation, the preprocessed student responses were embedded with a TF-IDF matrix. Elbow distortion and silhouette scores were evaluated for $k = 2$ to $k = 10$ to provide heuristic guidance on potential theme structure. The domain expert reviewed the $k = 3$ elbow point and silhouette scores, along with sample responses from preliminary clusters, to guide theme definition in Phase 1.

The Phase 0 pipeline is illustrated in Figure 6.7 (Chapter 6).

8.3 Primary Theme Classification (Phase 1)

8.3.1 Expert Prompt Design (Phase 1A)

Chapter 7 results confirmed three themes in the Peer Support dataset: Interactions, Questioning, and Civility. These themes guided the domain expert in designing ZSC prompts for Phase 1.

The domain expert reviewed the Phase 0 structure cues along with a 200-response sample (15% of the corpus) and identified five themes describing how students support each other: asking relevant questions, arriving prepared, working collaboratively, maintaining respectful behavior, and engaging during class. When tested on the full corpus, some themes overlapped and others appeared rarely. The expert consolidated these into three

primary themes: Collaboration (mapping to Interactions), Questioning, and Professionalism (mapping to Civility). Engagement and Preparation were treated as minority themes.

The domain expert labeled these themes in three ways:

Domain-Specific Labels (DS-Lab): Prompts used specialized educational terminology, such as "Collegial and Professional Conduct," to mirror researcher vocabulary.

Mainstream Labels (MS-Lab): Prompts adopted accessible, non-specialist language that reflects student phrasing, for example "Respect, Civility, and Professionalism."

Mainstream Labels with Descriptions (MS-LabDesc): Prompts paired mainstream labels with concise descriptions (e.g., "This text addresses professional behaviors in class including not causing disruptions...") to provide additional context.

These prompts were used to configure multi-label classification, allowing each student response to map to multiple relevant themes. Zero-shot outputs were then compared with the domain expert annotations using accuracy, precision, recall, F1-score, and UMass coherence.

The exact wording for each prompt strategy is provided in Table 8.2. The Phase 1 pipeline is illustrated in Figure 6.8 (Chapter 6).

8.3.2 Primary Theme Evaluation (Phase 1B)

Table 8.3 reports performance metrics for all three prompt strategies across primary themes.

Mainstream labels without descriptions (MS-Lab) achieved the highest accuracy (85%) and weighted F1 (60%), matching the best-performing unsupervised method (BERTopic MiniLM) without requiring any training data (Table 8.3). This result indicates that ZSC can achieve competitive performance when prompt language aligns with the model's pre-training distribution. Domain-specific labels scored lower on both metrics (accuracy 82%, weighted F1 56%). Mainstream labels with descriptions achieved higher confidence scores (0.88) and improved recall for some themes, but lower aggregate accuracy (81%) and weighted F1 (50%). The lower Jaccard index (0.35) for this strategy indicates weaker alignment with expert-coded themes when descriptions are added, suggesting that broader descriptions reduce class separation. UMass coherence values were generally low across all three

Table 8.2: ZSC Prompt Designs for Peer Support Themes

Theme	**Domain-Specific**	**Mainstream**	**Mainstream + Description**
		Primary Themes	
Collaboration	Peer-peer Interactions	Talking and Working with Each Other	This text discusses students' desire to interact with each other to be social or to collaborate on class work.
Questioning	Questioning Practice	Asking Appropriate and Relevant Questions	This text discusses students' desire for their peers to adjust the questions they ask in class either by asking more relevant questions or refraining from asking irrelevant or tangential questions.
Professionalism	Collegial and Professional Conduct	Respect, Civility, and Professionalism	This text addresses professional behaviors in class including not causing disruptions in class by talking, playing video games, or other use of technology or being professional and civil to the professor or being friendly, helpful, positive, and polite or avoiding condescending and complaining behaviors.
		Minority Themes	
Engagement	Productive Engagement	Participating	This text addresses paying attention or attending class or being a part of class discussion or similar behaviors.
Preparation	Timely Preparation	Preparing	This text addresses completing assigned reading and videos prior to class or coming to class prepared to ask relevant questions or working on homework before class or similar behaviors.

Table 8.3: Detailed ZSC Model Performance for Primary Themes

Theme	Prompt	n	Acc	Prec	Rec	F1	UMass
Primary Themes							
Collaboration (n=369)*	DS-Lab	192	0.78	0.70	0.40	0.50	-1.55
	MS-Lab	247	0.84	0.81	0.54	0.65	-1.62
	MS-LabDesc	341	0.81	0.66	0.61	0.63	-1.70
Professionalism (n=351)*	DS-Lab	371	0.82	0.66	0.69	0.67	-1.20
	MS-Lab	339	0.86	0.76	0.72	0.74	-1.15
	MS-LabDesc	567	0.78	0.59	0.79	0.68	-1.45
Questioning (n=428)*	DS-Lab	365	0.80	0.72	0.64	0.67	-1.80
	MS-Lab	471	0.83	0.73	0.77	0.74	-1.72
	MS-LabDesc	60	0.72	0.61	0.15	0.24	-1.50
Minority Themes							
Engagement (n=141)*	DS-Lab	415	0.75	0.32	0.65	0.43	-1.65
	MS-Lab	279	0.82	0.36	0.61	0.45	-1.68
	MS-LabDesc	299	0.83	0.39	0.71	0.51	-1.75
Preparation (n=21)*	DS-Lab	33	0.93	0.35	0.40	0.37	-1.25
	MS-Lab	39	0.91	0.30	0.32	0.31	-1.28
	MS-LabDesc	109	0.89	0.22	0.43	0.29	-1.40
No Match (n=66)**							

Note: Abbreviations: Acc=Accuracy, Prec=Precision, Rec=Recall. DS=Domain Specific, MS=Mainstream, Lab=Labels, LabDesc=Label Descriptions. *Expert-coded count. **Unsuitable for any theme.

strategies, consistent with short-text sparsity, and should be interpreted qualitatively. Table 8.4 provides a summary comparison of accuracy and weighted F1 across all three prompt strategies.

Table 8.4: Aggregate ZSC Performance Summary for Primary Themes

Prompt	**Avg Acc**	**Macro F1**	**Weighted F1**	**Avg UMass**	**Jaccard**	**Avg Conf**
DS-Lab	0.82	0.53	0.56	-1.49	0.36	0.82
MS-Lab	0.85	0.58	0.60	-1.41	0.41	0.86
MS-LabDesc	0.81	0.47	0.50	-1.42	0.35	0.88

Note: Abbreviations: Acc=Accuracy, Conf=Confidence. DS=Domain Specific, MS=Mainstream, Lab=Labels, LabDesc=Label Descriptions.

8.4 Secondary Theme Analysis (Phase 2)

Secondary subthemes were analyzed for Questioning ($n = 471$ responses from Phase 1) and Professionalism ($n = 339$ responses from Phase 1) to compare label-based and entailment-based prompting. The exact prompt wording is provided in Tables 8.5 and 8.6. The Phase 2 pipeline is illustrated in Figure 6.9 (Chapter 6).

8.4.1 Questioning Subthemes

Label prompting achieved higher F1 (0.91) than entailment (0.79) for the Productive Questioning subtheme and produced higher average F1 (0.71 vs 0.63) across all Questioning subthemes despite similar accuracy. Table 8.7 presents detailed performance metrics for each Questioning subtheme.

8.4.2 Professionalism Subthemes

Label prompting slightly outperformed entailment prompting (mean F1: 0.68 vs 0.65). Productive Attitude achieved the highest F1 with labels (0.81), while Unproductive Disruptions

Table 8.5: Prompts for Secondary Themes: Questioning Behavior

Theme	**Type**	**Prompt**
Productive Questioning	Label	Asking thoughtful and numerous questions that benefit the entire class
	Entailment	This response describes productive questioning behavior that benefits the entire class
Unproductive Questioning	Label	Asking too many questions or questions only relevant to one or a small group
	Entailment	This response describes unproductive questioning behavior that is disruptive or only benefits few

Table 8.6: Prompts for Secondary Themes: Professionalism Behaviors

Theme	**Type**	**Prompt**
Avoiding Disruptions	Label	Not causing disruptions in class by talking, playing video games, or other technology use
	Entailment	This response describes behavior that avoids classroom disruptions
Professional Conduct	Label	Being professional and civil to the professor
	Entailment	This response describes professional conduct toward professors
Friendly Behavior	Label	Being friendly, helpful, positive, and polite to peers
	Entailment	This response describes friendly and helpful behavior toward peers
Avoiding Negative Behavior	Label	Avoiding condescending and complaining behaviors
	Entailment	This response describes avoiding negative behaviors like condescension or complaining

Table 8.7: ZSC Performance for Secondary Themes: Questioning Behavior

Theme	**Prompt**	**n**	**Acc**	**Prec**	**Rec**	**F1**	**UMass**
Productive (n=350)*	Label	375	0.86	0.93	0.90	0.91	-2.05
	Entailment	350	0.87	0.73	0.86	0.79	-2.40
Unproductive (n=77)*	Label	96	0.85	0.47	0.53	0.50	-1.55
	Entailment	121	0.82	0.43	0.50	0.46	-2.03
Average							
Label		471	0.86	0.70	0.72	0.71	-1.80
Entailment		471	0.85	0.58	0.68	0.63	-2.22
No Match (n=9)**							

Note: Abbreviations: Acc=Accuracy, Prec=Precision, Rec=Recall. *Expert-coded count. **Unsuitable for any theme.

remained the most challenging (label $F1 = 0.42$). UMass coherence values were low across both subsets, reflecting short-text sparsity. Table 8.8 presents detailed performance metrics for each Professionalism subtheme.

Across both Questioning and Professionalism subthemes, label-based prompting consistently matched or outperformed entailment-based prompting, with simpler phrasing producing more reliable classification for short educational text. Interpretation of these results, including comparison with the discovery-first methods from Chapter 7, appears in Chapter 9.

Table 8.8: ZSC Performance for Secondary Themes: Professionalism Behaviors

Theme	**Prompt**	**n**	**Acc**	**Prec**	**Rec**	**F1**	**UMass**
Productive Conduct (n=41)*	Label	61	0.80	0.62	0.66	0.64	-1.62
	Entailment	38	0.84	0.69	0.71	0.70	-1.70
Productive Attitude (n=48)*	Label	48	0.90	0.76	0.86	0.81	-1.40
	Entailment	44	0.88	0.71	0.72	0.72	-1.55
Unproductive Conduct (n=239)*	Label	141	0.86	0.84	0.88	0.86	-1.75
	Entailment	182	0.83	0.80	0.74	0.77	-1.78
Unproductive Disruptions (n=23)*	Label	71	0.80	0.46	0.39	0.42	-1.60
	Entailment	48	0.78	0.50	0.33	0.40	-1.55
Average							
Label		339	0.84	0.67	0.70	0.68	-1.59
Entailment		339	0.83	0.68	0.63	0.65	-1.65
No Match (n=15)**							

Note: Abbreviations: Acc=Accuracy, Prec=Precision, Rec=Recall. *Expert-coded count.

**Unsuitable for any theme.

Chapter 9

DISCUSSION

9.1 Introduction

This study explored how NLP-based topic modeling of short texts can bridge the gap between machine learning techniques and traditional thematic analysis of short texts associated with education and education research. Five topic modeling techniques (k-means clustering, LDA, LSA, NMF, and BERTopic) were applied with default parameters to three datasets of short student responses exploring expectations for faculty, TA, and peer support, and compared to traditional thematic analysis conducted by a domain expert. A second stage optimized hyperparameters for four methods after excluding LSA, and introduced zero-shot classification (ZSC) as a themes-first alternative to unsupervised discovery. Performance was evaluated using a wide range of internal and external metrics.

Overall, no single topic model consistently excelled in extracting topics from the three datasets with high accuracy, reliability (F1-score), topic coherence, and theme ranking (Table 9.1). While no one model fit all performance needs, several trends were observed in performance that provide valuable guidance to practitioners who rely on qualitative data analysis to guide their research, teaching, or other pursuits in education.

This chapter interprets those results. Method-by-method discussion traces each method's behavior under default parameters, explains why each method performed as it did, and then examines what changed after optimization. Subsequent sections address how vectorization choices shaped model behavior, evaluate ground truth considerations, compare discovery-first and themes-first approaches on the peer support dataset, and close with evidence-based recommendations for practitioners.

Table 9.1: Best performing topic models by metric and dataset under default and optimized parameters.

Metric Type	**Metric**	**Faculty Support**	**TA Support**	**Peer Support**
Default Parameters (k-means, LDA, LSA, NMF, BERTopic)				
Internal	Single Coherence (U_{mass})	BERTopic	BERTopic	LSA
	Average Coherence (U_{mass})	NMF	k-means	LSA
	Variation in Coherence (U_{mass})	NMF	LSA	LSA
External	Single Topic Accuracy	k-means	LSA	LSA
	Average Accuracy	k-means	NMF	NMF
	Variation in Accuracy	BERTopic	BERTopic	LSA
	Single Topic F1-Score	k-means	LSA	LSA
	Average F1-Score	k-means	NMF	NMF
	Variation in F1-Score	BERTopic	LDA	NMF
Aggregate	Topic/Theme Ranking	k-means	NMF	NMF
Evaluation Approach		Approach 2	Approach 1	Approach 2
Optimized Parameters (k-means, LDA, NMF, BERTopic; LSA excluded)				
Internal	Single Coherence (U_{mass})	k-means	k-means	k-means
	Average Coherence (U_{mass})	k-means	k-means	k-means
	Variation in Coherence (U_{mass})	BERT-MPNet	k-means	LDA
External	Single Topic Accuracy	NMF	k-means	k-means
	Average Accuracy	BERT-MPNet	BERT-MPNet	BERT-MiniLM
	Variation in Accuracy	BERT-MiniLM	BERT-MPNet	BERT-MiniLM
	Single Topic F1-Score	NMF	BERT-MPNet	BERT-MiniLM
	Average F1-Score	BERT-MPNet	BERT-MPNet	BERT-MiniLM
	Variation in F1-Score	BERT-MPNet	LDA	LDA
Aggregate	Topic/Theme Ranking	k-means	k-means	k-means, NMF
Evaluation Approach		Approach 2	Approach 2	Approach 2

Note: Internal metrics assess cluster or topic structure without reference to ground truth. External metrics measure alignment with expert-coded themes (Approach 1: model-led coding; Approach 2: expert-led coding). Variation = most consistent performance across themes (smallest spread). BERT-MPNet = BERTopic with MPNet embedding; BERT-MiniLM = BERTopic with MiniLM embedding.

9.2 Method-by-Method Interpretation

9.2.1 k-means Clustering

In the analysis of topic models using each model's default parameters, k-means clustering served as the baseline partitioning and topic modeling method because it treated words as words without regard to semantic or contextual relationships. The k-means technique clusters documents based on distance in a vector space, is simple and computationally efficient, but is sensitive to initial centroid placement and assumes spherical cluster shapes of similar sizes. When textual data reflect more complex or indirect ideas such as those involving emotion, context, or multiple overlapping topics, k-means often splits related responses across clusters or groups unrelated responses together, ignoring conceptual meaning in the process.

Given these limitations, it is surprising how well k-means replicated the perspective and assessments of the domain expert (Approach 2) with the Faculty Support dataset. A closer look at the vocabulary associated with each of the three themes in this dataset gives important insight into why. The words students used to describe their preferences for how faculty should support them were very distinct across themes. The words "office hours," "available," and "email" almost exclusively belonged to the Interactions theme, while "example," "real world," and "experience" uniquely signaled Examples and Experience, and "teach," "explain," "lecture," "resources," and "organized" referred to Teaching Practice. These distinct word choices form tight, well-separated clusters in TF-IDF vector space, which is exactly what k-means optimizes. Furthermore, the fact that Faculty Support responses are longer on average (12 words) than those in the TA and Peer Support datasets provided k-means with more lexical information per document, further supporting its performance. Another more subtle performance advantage was k-means' bias toward the Teaching Practice theme, assigning approximately half of student responses to this category. Since Teaching Practice is the majority class, this bias improved correct assignments by chance. Such coincidental alignment cannot be relied upon in every dataset. Nevertheless, the success of k-means on the Faculty Support dataset and its corresponding struggle on the Peer Support dataset illustrate a key point regarding model selection: more sophis-

ticated models are not always better than simpler ones. More complex methods should be used only when topics share substantial vocabulary and require decomposition, probabilistic analysis, or pre-trained language models to separate.

Once topic model parameters were optimized, however, k-means produced mixed results that reveal an important mismatch between internal cluster quality metrics and thematic alignment. The optimized pipeline for k-means used counts-based vectorization (CV) with PCA dimensionality reduction and L2 normalization, a configuration that scored highest on silhouette and Calinski-Harabasz indices during the optimization of preprocessing for the model. TF-IDF was used as the default vectorization in the initial analysis for k-means, NMF, and LSA because it is the scikit-learn default and was described as the standard baseline in Chapter 5. Despite optimization, the accuracy of classifying faculty support documents declined from 74.8% to 68.67% and macro-F1 fell sharply from 71.8% to 45.79%. PCA projects documents onto the directions of maximum variance across the entire corpus, which tends to capture the most frequently occurring vocabulary rather than the rare, discriminative words that cleanly separate themes. In the default model, TF-IDF was used to vectorize words in each document and in so doing, amplified theme-specific terms such as "office hours" for the interactions theme and "real world" for the examples and experience theme. Replacing TF-IDF with counts-based vectorization and PCA discarded this discriminative signal, producing clusters that were geometrically tighter but semantically blurred. The fact that k-means forces each document into a single topic disproportionately hurt the minority themes: once the preference that TF-IDF gave to interactions vocabulary was removed, more responses were absorbed into the majority teaching practice topic. The sharp macro-F1 decline reflects the concentration of errors in the two smaller themes.

In contrast, the accuracy of classifying peer support documents using the optimized k-means model improved from 63.7% to 73.3% and for TA support from 62.3% to 69.87% for the opposite reason. In both datasets, students used informal, overlapping vocabulary where thematically relevant terms appeared frequently across responses. TF-IDF penalized these high-frequency terms, thereby suppressing thematic cues that raw counts preserve. The gain in cluster separation under the counts-based representation was substantial: the Calinski-Harabasz index for peer support nearly doubled from 34.45 (TF-IDF) to 69.86

(counts-based vectorization), and TA support rose from 32.74 to 56.07. Those separation gains translated into accuracy gains because both datasets had low rates of multiple-topic responses (6.86% for peer support, 11.7% for TA support). When most responses address a single theme, better cluster separation converts directly into more correct assignments. Faculty support faced an additional structural constraint: 26.5% of its responses addressed multiple themes simultaneously (Table 7.9, reported in Chapter 7), and the rigid assignment of each document to a single cluster cannot correctly classify a document that genuinely spans two themes regardless of how well the cluster representation separates them. The accuracy ceiling imposed by this incompatibility meant that removing TF-IDF's rare-term amplification produced a net loss rather than a gain.

These contrasting outcomes carry a practical implication: optimizing k-means based on internal cluster metrics can improve performance when the default representation suppresses thematically relevant cues, but gains in cluster geometry cannot compensate for rigid assignment (one topic per document) in datasets with high rates of multiple-topic documents.

Another persistent limitation was consistent in both default and optimized models: the Euclidean distance metric used by k-means to optimize cluster formation is sensitive to differences in magnitude in high-dimensional text data, thereby causing clustering to weight document length alongside semantic content. Since document length is largely irrelevant, this characteristic of k-means can substantially compromise performance in classifying documents by topic/theme.

9.2.2 Latent Dirichlet Allocation (LDA)

LDA, a widely used topic modeling algorithm, represents each document as a mixture of topics and each topic as a mixture of words. To do so, LDA relies heavily on the same words occurring across multiple documents, making short texts prone to inaccurate topic assignments. While this limitation likely affected LDA performance relative to other models, other factors also contributed, since the dataset with the shortest texts (peer support) produced the best LDA results in the default analysis. One possibility is that the improved

performance of LDA on the peer support dataset resulted from the fact that fewer student responses were ambiguous (5.89%) compared to the TA support (18.6%) and faculty support datasets. Since LDA was constrained in this analysis to assign only a single topic to each document and does not consider context, word sequence, or semantic information, topic selection in ambiguous cases is effectively random, leading to errors compared to domain expert assessment. A domain expert is better able to assess subtle differences in language necessary to determine the dominance of a single topic when ambiguous language is in use, an ability LDA lacks. While the lack of robustness in the face of ambiguity is a weakness of LDA, it also provides guidance for model selection: when ambiguity in documents is frequent, LDA should be avoided.

If ambiguity were the only factor contributing to LDA performance variation, a simple decline in scores from the low-ambiguity peer support dataset to the higher-ambiguity faculty and TA support datasets would be expected. Since that monotonic decline was not observed, other factors must have been at play. Prior research has noted that the reliability and validity of LDA-derived topics can be limited when short responses provide sparse word co-occurrence information. In this analysis, word co-occurrence was particularly frequent in the peer support dataset, where the words "ask" and "questions" co-occurred in 60.6% of the documents assigned to the Questioning theme. This strong co-occurrence produced F1-scores of 77.6% and 69.6% and accuracy of 77.5% and 75.3% for Approaches 1 and 2 respectively, underscoring the value of LDA when the same words are frequently used to express similar ideas. The TA support dataset presented a different kind of challenge: the three topics generated by LDA did not align with the three themes identified by the domain expert, constraining Approach 2 evaluation to only two topics in that dataset.

Optimization of LDA showed the widest performance variation across datasets, with accuracy spanning from 58.3% to 78.46%, illustrating a fundamental characteristic of the method: hyperparameter tuning can only amplify what the co-occurrence structure of the corpus already supports. Both alpha and beta were reduced toward sparser distributions during optimization. Alpha controls how many topics dominate each document, and beta controls how many words define each topic. For peer support, this worked well. The three themes (Questioning, Civility, Interactions) are lexically distinct, and most responses

address a single concern in four to six words. Reducing alpha pushed LDA to assign each document strongly to one topic, aligning model output more cleanly with how students actually wrote. Peer support accuracy rose from 60.2% (default Approach 1) to 78.46%, a gain of over 18 percentage points. Faculty support also improved, from 51.7% to 68.13%, because reducing diffuse default distributions sharpened assignments in a dataset where theme vocabularies, while not as distinctly separated as in peer support, are still reasonably distinguishable by word co-occurrence. The improvement was bounded, however, by the 26.5% rate of multiple-topic responses in that dataset. LDA, like k-means, assigns each document to a single topic by selecting the topic with the highest posterior probability; responses that genuinely span two themes will be misclassified regardless of how precisely alpha concentrates topic assignments.

TA support declined from 67.3% to 58.3% despite tuning both alpha and beta. This result reveals the outer limit of what hyperparameter optimization can accomplish. The difficulty in TA support is structural. The corpus's word co-occurrence patterns naturally produce topics corresponding to Q&A, availability, and teaching practice, not the three themes the domain expert identified (interactions, examples and experience, and teaching practice). Reducing alpha made each document's topic assignment more decisive, but it concentrated the model more strongly onto the wrong thematic structure. The frequency analysis confirmed this: LDA assigned 516 responses to Examples and Experience against 135 expert-coded instances, merging documents where students requested "more worked examples in lecture," "a TA available to answer example problems," and "real-world experience in industry." All three share the word "example" and co-occur in the same documents, but each refers to a pedagogically distinct concern. No adjustment of alpha or beta can teach LDA to distinguish these intent differences; that contextual awareness is beyond what counts-based co-occurrence statistics can provide.

The divergence between coherence and accuracy observed in the default phase persisted and sharpened after optimization. Faculty support achieved the strongest coherence (-2.12) yet only moderate accuracy (68.13%), while peer support showed the opposite: weaker coherence (-2.67) but the highest accuracy (78.46%). In TA support, the optimized model produced internally coherent topics that were structurally misaligned with the expert cod-

ing. A high coherence score in this case would mislead a practitioner into accepting outputs organized around the wrong thematic categories. When the intended thematic structure diverges from what the corpus's co-occurrence statistics naturally produce, optimization tends to reinforce that divergence rather than correct it.

9.2.3 Matrix Factorization (LSA and NMF)

Unlike LDA, matrix factorization considers relationships between words by grouping words that often appear together in documents. This approach goes beyond treating documents as sets of words and provides elementary insight into the latent semantic structure of text. Consideration of basic semantics was a major contributor to the fact that 63% of the best performers in terms of internal and external performance metrics were matrix factorization methods (NMF or LSA) in the default analysis.

Latent Semantic Analysis

Despite considering semantic meaning more so than k-means and LDA, LSA still struggles with words that have multiple meanings. This vulnerability was demonstrated in the Faculty Support dataset in the Examples and Experience theme, where interrater reliability (Cohen's κ) was worse than expected by chance (indicated by negative scores). Students often used the word "examples" to express different ideas. Consider the following three documents:

> Go through lots of examples, and when they do, explain your thought process.
>
> Have clear, focused lectures with supplementary materials such as lecture notes and practice problems available.
>
> To give abstract examples for topics before going to examples.

While all three documents refer to opportunities for practice or examples, only the first is pedagogically related to the Examples and Experience theme, where teachers expose their thought processes around problem-solving and gradually reduce scaffolding until students can solve similar problems on their own. The remaining two refer to Teaching Practice with

respect to how and what teachers present in lecture and how they supplement lectures with readily available resources.

LSA also performed particularly poorly in classifying documents as Interactions in the Faculty Support dataset, producing a large number of false negatives (missed responses) as evidenced by recall scores of 12.1% and 13.7%. This poor recall is a direct consequence of how LSA constructs its topic dimensions using singular value decomposition (SVD). SVD requires its components to be mathematically orthogonal, which forces the weight of each word to be distributed across multiple topic dimensions rather than concentrated on a single dimension. The words "office" and "hour," which together appeared in 41.8% of Interactions documents but only 4% of documents belonging to other themes, are the strongest indicators for the Interactions theme. However, these words appear so frequently in the overall corpus that they are captured by the first SVD component, which represents maximum variance and also corresponds to the Teaching Practice topic. As a result, LSA assigns them substantial weight on both the Teaching Practice and Interactions dimensions. This near-equal loading means that documents containing "office hours" are only marginally more likely to be assigned to Interactions than to Teaching Practice, and in practice the vast majority are absorbed into the larger Teaching Practice cluster: of 553 ground truth Interactions documents, LSA misclassified 474 as Teaching Practice.

These limitations of LSA (its vulnerability to polysemy, the orthogonality constraint that distributes high-frequency term weight across multiple dimensions, and its consistently weaker performance on recall and reliability relative to NMF) were the primary basis for excluding LSA from the optimized analysis and retaining NMF as the representative matrix factorization method.

Non-Negative Matrix Factorization

NMF does not experience the multi-dimension loading problem described above because its non-negativity constraint enforces sparsity. Each word concentrates its weight on the single topic where it contributes most, causing "office" and "hour" to load exclusively on the Interactions dimension with zero weight on all other components. This allows NMF to

cleanly separate documents containing these words and produce recall scores substantially higher than LSA for the Interactions theme. Poor recall scores that require such detailed attention are typically confined to a single topic which, while not ideal, may limit the overall impact of this problem.

NMF topic models decompose the document-term matrix into two non-negative matrices, one representing topics and another representing word weights. Previous research has suggested that this decomposition process produces more coherent topics than other techniques. However, in the default analysis, statistical measures of topic coherence (UMass) only partially supported this claim. Average coherence scores were worse than two other models in the Peer Support and TA Support datasets. A qualitative evaluation of topic quality using top keywords contradicted the statistical scores. Keywords such as "ask, question, answer, think, afraid" clearly point to not being afraid to ask and answer thoughtful questions, aligning with the Questioning theme. The same is true for "talk, participate, distract, respectful, quiet" relating to civil behavior and aligning with the Civility theme, and "study, group, help, willing, discussion" relating to collaboration and aligning with the Interactions theme. The fact that qualitative evaluation indicated strong topic coherence while statistical coherence did not strongly suggests that sensitivity to noise words in the documents, not document sparsity or brevity, limited the accuracy of statistical coherence measures. This contradiction underscores the need to include qualitative assessment of topic model performance rather than relying exclusively on quantitative metrics.

Despite the cautionary tales regarding the confounding influence of multiple word meanings (LSA) and noise words (NMF), matrix factorization methods dominated the performance of topic models in the default analysis. While this superior performance can be largely attributed to the consideration of basic semantics among words, it leaves open the question of why BERTopic, which considers semantics at a more sophisticated level, did not produce even better results.

After optimization, NMF produced the largest single improvement observed in the study. Faculty Support accuracy rose from 59.1% (default Approach 2) to 76.76%, a gain of 17.7 percentage points, and macro-F1 improved from 59.5% to 62.82%. NMF's additive structure, which decomposes each document as a weighted sum of topic components, was already

a good match for Faculty Support, where 26.5% of responses addressed multiple themes. Optimization sharpened that match in two ways. First, tuning the regularization parameter made each topic definition more concentrated around its most characteristic words, reducing noise-word contamination. Second, refined preprocessing, including domain-specific stop-word removal, eliminated words that straddled themes without belonging exclusively to either. This allowed the sparsity constraint to work with greater precision: "office" and "hours" loaded even more exclusively onto the Interactions dimension, "example" and "real world" concentrated onto Examples and Experience, and "lecture" and "explain" dominated Teaching Practice. Teaching Practice achieved consistent F1-scores wherever it appeared (69.75% in Faculty Support, 66.67% in TA Support), confirming that this theme's vocabulary is stable across datasets and well-suited to NMF's sparse decomposition. The Questioning theme in Peer Support reached 82.53% F1, similarly reflecting the precision of the sparsity constraint when theme vocabularies are distinct.

The persistent limitation emerged in the Examples and Experience theme, where F1 swung from 62.81% in Faculty Support to 3.35% in TA Support. In Faculty Support, the word "example" predominantly appeared where students requested concrete demonstrations of problem-solving processes, so the non-negativity constraint cleanly concentrated that word on one topic dimension. In TA Support, the same word appeared across all three themes: students asking for "more worked examples in office hours" (Interactions), "real-world examples to illustrate concepts" (Teaching Practice), and "practice example problems" (Examples and Experience). When a word is genuinely multi-thematic in the corpus, the sparsity constraint makes an allocation rather than a contextual judgment. Optimization could not resolve this because the limitation lies in the information that five-to-eight-word responses can provide about student intent, not in the model's parameters.

9.2.4 Neural Networks (BERTopic)

While LSA and NMF rely on patterns determined by word co-occurrence to establish semantic meaning, BERTopic goes a step further by taking into account words surrounding other words (i.e., contextual embeddings) to determine meaning, and uses clustering to

capture these deeper semantic relationships within text. In the default analysis, however, BERTopic failed to deliver competitive results. These performance deficits do not reflect inherent weaknesses of the neural approach. Rather, they underscore the need to select appropriate scaffolding for BERTopic's pipeline. In this analysis, k-means clustering was substituted for HDBSCAN to ensure reproducibility and to fix the number of discovered topics at three. While this choice was necessary for a fair one-for-one comparison among the five topic models, this substitution minimized three essential capabilities available in BERTopic. First, HDBSCAN identifies outliers, allowing documents that do not fit naturally into any cluster to be set aside rather than forced into the nearest cluster regardless of semantic distance. This is important for datasets containing documents with multiple topics or ambiguous assignments, because such documents occupy boundary regions between clusters and introduce noise when forced into a single topic. The Faculty Support and TA Support datasets had much higher proportions of multiple-topic and ambiguous documents, leading directly to degraded performance compared to Peer Support results. Second, k-means assumes clusters are spherical and of similar size, while UMAP's dimensionality reduction step intentionally produces irregular, density-varying cluster shapes that only HDBSCAN is designed to detect. Replacing HDBSCAN with k-means discarded the density structure that UMAP was specifically designed to preserve. Third, fixing $k = 3$ prevented BERTopic from discovering the natural number of topics in the data. Under these constraints, default BERTopic achieved only 54.4% accuracy on Faculty Support (Approach 2), 59.7% on TA Support, and 64.2% on Peer Support.

Optimization introduced two embedding models evaluated within the same UMAP dimensionality reduction and k-means ($k = 3$) clustering framework, differing in both capacity and training objective. All-MiniLM-L6-v2 (384 dimensions, 22.7 million parameters) was pre-trained on shorter paired sentences with a mean-pooling objective that emphasizes overall sentence-level similarity. All-mpnet-base-v2 (768 dimensions, 109 million parameters) was trained on a broader range of tasks including harder paraphrase detection, producing representations that separate similar phrases more finely. Both models used UMAP for dimensionality reduction and k-means with $k = 3$ for clustering. None of the three architectural constraints identified in the default analysis were modified; the performance gains

observed in optimization reflect the contribution of embedding quality within an otherwise unchanged pipeline.

The choice of embedding model proved as consequential as the choice between BERTopic and any traditional method. MiniLM dominated Peer Support (85.0% accuracy, 77.0% macro-F1), where themes are well-separated and responses are short (median 5 words). MiniLM's compact space captured the most salient semantic relationships for these short responses, grouping answers about "office hours," "help sessions," and "available time" into the Interactions theme and linking communication channels such as "discord," "zoom," "email," and "breakout rooms," compensating for the sparse lexical cues that limited counts-based vectorization methods. When themes are more similar or responses longer (Faculty Support averaged 12.0 words per response and TA Support 9.71, compared to 7.94 for Peer Support), MiniLM's pooling compresses away finer distinctions needed for separation. MPNet outperformed MiniLM in Faculty Support (76.8% versus 72.6% accuracy; 65.7% versus 58.6% macro-F1) and TA Support (70.3% versus 65.8% accuracy; 54.2% versus 45.5% macro-F1), where higher-capacity representations preserved those subtler distinctions. Internal cluster quality metrics confirmed this: on TA Support, MPNet produced a Calinski-Harabasz index of 3904.2 versus 2019.0 for MiniLM, reflecting substantially tighter, more separated clusters in the higher-dimensional space. The advantage reversed on Peer Support, where the added capacity encoded more variation within themes than necessary, separating responses that should have been grouped together. The 20.4 percentage-point macro-F1 gap on Peer Support (77.0% for MiniLM versus 56.6% for MPNet) was the largest difference between the two models across any dataset. Averaged across all datasets, MiniLM (60.4% macro-F1) slightly outperformed MPNet (58.8%) despite having 4.8 times fewer parameters.

At the theme level, average UMass coherence scores added nuance to the accuracy results. Both models achieved the strongest average coherence on Faculty Support (MiniLM -2.486, MPNet -2.701), yet Faculty Support produced only moderate macro-F1 scores (58.6% for MiniLM, 65.7% for MPNet). Peer Support showed the opposite pattern: MiniLM's highest accuracy dataset produced weaker average coherence (-3.170). The Interactions topic drove this divergence, recording a UMass score of -6.709, the weakest of any theme

across any dataset in the study. Students addressing peer interaction used a wide range of communication tools and meeting formats, producing a topic with statistically weak word co-occurrence even while the embedding representation still classified those responses correctly. The same coherence-accuracy inversion appeared on TA Support in the opposite direction. The Examples and Experience topic for MiniLM produced the best coherence of any TA Support theme (-1.928), yet F1 was only 27.7%, with precision as low as 17.1% and topic frequencies approximately three times the expert-coded count. Statistical coherence here reflects only that the words co-occur, not that the co-occurrence corresponds to the intended theme.

Two limitations persisted across both embedding models. Retaining k-means with $k = 3$ for cross-method comparability meant HDBSCAN could not identify boundary documents as outliers; responses that did not fit cleanly into any theme were still assigned to the nearest cluster, introducing noise proportional to dataset ambiguity. The Examples and Experience theme in TA Support remained the most resistant case across both phases of analysis. Students requesting more worked examples in office hours, conceptual examples in lecture, and practice example problems all produced semantically similar embeddings, because pre-trained models encode surface meaning and lexical proximity rather than pedagogical intent. No change in embedding model resolved this, because the constraint is in what five-to-eight-word responses can communicate about student intent, not in how the model represents them. A sensitivity check using full HDBSCAN clustering with bigram tokenization confirmed the same three high-level themes in each dataset while revealing finer subtopics, providing qualitative validation that the themes were robust across clustering approaches (see Appendix A.1).

9.2.5 Zero-Shot Classification (ZSC)

Zero-shot classification approaches thematic analysis differently from the discovery-first methods evaluated above. Rather than discovering patterns in the data and then interpreting them, ZSC begins with expert-defined themes and assigns each response directly. The analysis applied ZSC to the Peer Support dataset, selected for its strong inter-rater

agreement (kappa = 0.75) and low ambiguity (5.89%).

The BART-large-MNLI model used for ZSC is pre-trained on general text corpora, so its semantic representations align with everyday language patterns rather than domain-specific terminology. This had direct consequences for prompt design. The model mapped student responses more reliably to labels such as "respect" than to "collegial conduct" because the pre-training distribution contains far more instances of common words than specialized educational terminology. Mainstream language prompts outperformed domain-specific prompts across all themes, with a 3 percentage-point accuracy advantage (85% versus 82%) and a 4-point weighted F1 advantage (60% versus 56%). This finding shows that alignment with the model's pre-training distribution matters more than terminological precision for classification accuracy.

Adding descriptions to mainstream labels improved recall but lowered overall accuracy (81%) and weighted F1 (50%). The lower Jaccard index (0.35) for description-enhanced prompts confirmed weaker alignment with expert-coded themes. Broader descriptions increased multi-label matches but reduced class separation, because extending a label's definition allowed more responses to cross the 0.5 entailment threshold for secondary themes. The simpler label-only approach preserved stricter boundaries between classes.

ZSC achieved 85% accuracy on Peer Support, equal to BERTopic MiniLM on that metric and without requiring any training data. The lower macro-F1 (58.0% versus 77.0% for BERTopic MiniLM) indicates that accuracy alone does not capture theme separation quality. ZSC forms internally consistent label groups, achieving the highest within-label UMass coherence (-1.41) across all methods, but misses the boundaries between themes that BERTopic MiniLM separates more reliably.

At the primary theme level, Questioning (F1 = 74%) and Professionalism (F1 = 74%) outperformed Collaboration (F1 = 65%) under mainstream labels. Professionalism benefited from a behaviorally concrete vocabulary: students consistently used words such as "respect," "attentive," and "distract," which mapped cleanly to the mainstream label and produced precision of 76% with recall of 72%. Questioning achieved similar performance because it was the largest theme in the corpus ($n = 428$) and its vocabulary, centered on "ask" and "questions," was consistent across responses, giving the model strong evidence

even when students phrased their ideas differently. Collaboration produced the lowest recall (54%) and correspondingly lower F1. Students writing about collaborative peer support often combined elements of asking for help and working together, using vocabulary that overlapped with both Questioning and Professionalism. Because ZSC scores each label's entailment probability independently rather than treating labels as competing categories, a response signaling collaborative intent but also mentioning asking questions can cross the 0.5 threshold for Questioning and be assigned there instead of to Collaboration. This independent scoring is the structural reason the 85% overall accuracy did not translate to balanced macro-F1: the majority theme collected many correctly classified responses, while Collaboration's recall deficit and the near-absence of minority theme patterns in the model's pre-training distribution suppressed macro-F1 to 58%. The coherence advantage (average UMass -1.41, best across all methods) reflects the same dynamic. Expert-defined themes drew on more lexically consistent response sets than any unsupervised cluster, because category membership was determined by pedagogical intent rather than word co-occurrence statistics, producing tighter word-topic associations regardless of whether classification was accurate.

A secondary analysis comparing label-based and entailment-based prompting found that simpler prompts worked better. Label prompts outperformed entailment prompts for both subthemes (average F1 of 0.71 versus 0.63 for Questioning; 0.68 versus 0.65 for Professionalism). For short educational text, concise prompts produced more reliable results than elaborated entailment phrasing. Elaborated phrases introduce vocabulary absent from student responses, shifting the model's entailment comparison toward the prompt's own language rather than the response content. In texts of five to eight words, this mismatch reduces the probability assigned to the correct label and increases ambiguity across competing labels.

Three error patterns constrained ZSC performance. Under label-only prompts, multi-label classification failed when entailment probabilities for secondary themes fell below the 0.5 threshold, producing false negatives where responses addressing two themes were assigned to only one. Description-enhanced prompts produced the opposite error: broader definitions allowed more responses to cross the threshold for secondary themes, increasing false positives and reducing class separation, as discussed above. Abstract themes were

harder than concrete ones: the Unproductive Disruptions subtheme reached only 42% F1 despite a clear definition. Minority themes were consistently missed, with the Preparation theme ($n = 21$) achieving only 29–37% F1 across prompt types. Without fine-tuning, the model cannot learn patterns that are rare in its pre-training data. Expert review is still needed for multi-theme responses, abstract concepts, and minority themes, even when a themes-first approach is applied.

9.3 The Role of Vectorization in Topic Discovery

An important consideration that cuts across the model-by-model comparisons above is the role of text vectorization in shaping topic discovery. In this analysis, LDA operated on raw word counts (counts-based vectorization), k-means, LSA, and NMF operated on TF-IDF representations in the default analysis, though k-means optimization subsequently substituted counts-based vectorization, a representation change whose effects on peer and TA support performance are discussed in Section 9.2.1, and BERTopic operated on contextual embeddings from a pre-trained Sentence-BERT model. These are not interchangeable preprocessing steps; each one determines what information the topic model has access to and, consequently, what patterns it can detect. When comparing LDA to NMF or LSA, for example, the observed performance differences reflect not only the mathematical properties of the models themselves but also the advantage that TF-IDF confers over counts-based vectorization in short-text settings. TF-IDF amplifies rare, discriminative words such as "office" and "hours" while suppressing common words such as "class" and "help" that appear across all three themes, giving the model stronger signal from documents that may contain only six to ten words. LDA, by contrast, treats every word occurrence equally, leaving it with less discriminative information in sparse, short documents. This confound is important to acknowledge: the consistently stronger performance of TF-IDF-based models relative to LDA across all three datasets cannot be attributed to the topic model alone.

Contextual embeddings, as used in BERTopic, encode an additional layer of semantic information by representing documents based on how and where words appear relative to one another rather than simply which words are present. This feature is most valuable when themes share substantial vocabulary and can only be distinguished by meaning rather

than by the presence or absence of specific words. However, when the vocabulary associated with each theme is already distinct, as in the Faculty Support dataset, the additional semantic information offers little practical benefit over TF-IDF because the themes are already well-separated in frequency-based vector space. Optimized NMF, operating on TF-IDF representations, reached 76.76% accuracy on Faculty Support, a result nearly identical to MPNet BERTopic at 76.8%, confirming that embedding-based representations provided no measurable advantage when theme vocabularies were already well-separated. For practitioners, these observations suggest that vectorization should be treated as a deliberate design decision rather than a default byproduct of model selection, and that the expected degree of lexical overlap among themes in a given dataset should inform that decision alongside the choice of model itself.

For BERTopic, the choice of embedding model replaces vectorization as the primary representation decision. MiniLM's compact 384-dimensional space worked well for clearly separated themes in Peer Support, while MPNet's richer 768-dimensional space helped distinguish subtle thematic differences in the higher-ambiguity Faculty and TA Support datasets. The choice between MiniLM and MPNet changed which groups of similar texts the model captured and which errors it produced, a finding with direct implications for practitioners selecting BERTopic for short-text analysis.

For zero-shot classification, label phrasing itself serves as the representation. Prompt wording directly affects performance in the same way that TF-IDF term weighting affects traditional methods: by emphasizing the features that best discriminate between categories. Mainstream language that aligned with student vocabulary improved accuracy (85% versus 82% with domain-specific prompts), just as TF-IDF outperforms counts-based vectorization by emphasizing discriminative features over common ones. Expert judgment in prompt design becomes the defining representation choice for ZSC, giving researchers direct control over how the model interprets text.

9.4 Ground Truth Considerations

The study used two approaches for calculating external performance metrics. The first approach (Approach 1) was led by the topic model, where the top words from each topic

were distilled into a list of keywords that enabled a domain expert to code the data manually. The second approach (Approach 2) was human-led, where a domain expert coded the data independently using only the optimal number of topics determined by the topic models as a guide. The findings indicated that Approach 2 led to substantially better performance for most models used to analyze the Faculty and Peer Support datasets. This result suggests that topic models can identify pedagogically relevant themes and, while unexpected, adds to existing literature that advocates for topic models to reduce the tedious and time-consuming work involved in qualitative analysis of text-based data.

However, Approach 2 failed altogether in the analysis of TA Support data, telling a cautionary tale about over-reliance on machine learning for language analysis. In the TA Support data, the topic models produced topics that were misaligned with the themes identified by the domain expert. While the topic models generated topics corresponding to Q&A, availability, and teaching practice (which absorbed Examples and Experience into overall teaching practice), the domain expert identified themes of Interactions (combining the Q&A and availability topics), Examples and Experience, and Teaching Practice. Thus, while topic models can replicate themes assigned by domain experts using traditional analysis methods, topic-to-theme alignment is not guaranteed and must be supervised by a human to avoid the largely erroneous results produced for TA Support.

The optimized analysis used Approach 2 exclusively, providing a consistent reference point but also embedding the domain expert's perspective into all evaluations. This means the evaluation inherits the biases of the expert coding and can blur annotation limitations with model error. Low precision scores in TA Support (as low as 17.1% for BERTopic MiniLM) may reflect boundary issues in the coding scheme as much as model shortcomings. The interrater reliability scores (kappa values of 0.72 to 0.75 across datasets) provide confidence that the expert coding was consistent, but they do not eliminate the possibility that alternative thematic frameworks would produce different performance rankings. All automated methods depend on the quality of their reference categories: when ground truth has clear, well-separated themes, all methods benefit; when it involves themes that overlap or require contextual interpretation, the evaluation itself becomes more uncertain and model rankings should be read with caution.

9.5 Discovery-First versus Themes-First on Peer Support

The Peer Support dataset is the only corpus where all methods, both discovery-first and themes-first, can be compared directly. BERTopic MiniLM and ZSC both achieved 85.0% accuracy on this dataset, but their error profiles differed. BERTopic MiniLM produced balanced theme separation (77.0% macro-F1), while ZSC achieved lower macro-F1 (58.0%) despite matching overall accuracy. The gap between accuracy and macro-F1 indicates that ZSC classified many responses correctly but struggled to distinguish between certain themes. ZSC did produce the best within-label coherence (−1.41) across all methods, confirming that expert-defined themes create more internally consistent groupings than unsupervised discovery.

Table 9.2: Method comparison on peer support dataset: optimized models and ZSC.

Method	**Accuracy (%)**	**Macro-F1 (%)**	**UMass Coherence**	**Theme Ranking**
k-means	73.3	58.0	−2.29	Aligned
LDA	78.5	67.2	−2.67	Aligned
NMF	72.9	59.0	−3.94	Aligned
BERTopic (MiniLM)	85.0	77.0	−3.17	Aligned
BERTopic (MPNet)	70.7	56.6	−4.43	Partially Aligned
Zero-Shot (ZSC)	85.0	58.0	−1.41	Aligned

Note: Accuracy and macro-F1 measure alignment with expert-coded themes (Approach 2). UMass coherence measures internal topic consistency (higher is better). Theme ranking indicates whether method output aligns with expert-defined theme priorities. All topic model results are from the optimized analysis.

These complementary strengths suggest that the two approaches serve different analytical needs. Discovery-first methods are better suited for exploratory analysis when themes are unknown or when the researcher wants the data to reveal unexpected patterns. ZSC is better suited for confirmatory analysis when expert-defined categories are available and the goal is to apply a known coding framework at scale. An important caveat is that ZSC was evaluated only on Peer Support, the dataset with the lowest ambiguity and clearest thematic boundaries. Whether ZSC would maintain its competitive accuracy on the more challenging

Faculty Support (26.5% multi-topic responses) or TA Support (18.6% ambiguity) datasets remains an open question and a direction for future work.

9.6 Implications

The results and discussion to this point have addressed the primary research question of this comparative analysis: how do topic models based on NLP perform when applied to datasets used for education research. The overall goal of this work, however, is to provide actionable recommendations to practitioners (instructors, TAs, researchers, and other stakeholders) for integrating NLP-based topic models into their work at all levels of education. While neither the comparative analysis conducted here nor the prior literature can provide a comprehensive set of recommendations for using NLP effectively, key insights have emerged, summarized in Table 9.3. It is important to note that all topic models in the default analysis were baseline models, constructed in a way that enables one-for-one comparison with minimal bias. The resulting recommendations are therefore starting points for analysis; further optimization of the chosen model will only improve the value and performance of results.

Table 9.3: Recommendations for using topic models in education practice.

I have a dataset (corpus) of documents and...	**Recommendation**
I have no idea what they are telling me.	Process the documents with a simple topic model (k-means, LDA) using the optimal number of topics determined by the elbow method as well as fixed numbers of topics on either side of this optimal number to explore the data using the top n words for each topic or word clouds.

(continued on next page)

Table 9.3 – *continued*

I have a dataset (corpus) of documents and...	**Recommendation**
I want to quickly identify the low-hanging fruit in my dataset so I can spend my time focusing on the documents that are more difficult to interpret.	Process the documents with LDA using the optimal number of topics determined by the elbow method. Generate word clouds of the results and seek out frequent word co-occurrences in the data (indicated by large font in the word clouds). Use these word co-occurrences to fast-track coding of documents that contain them.
I have looked at my data and noticed that students/respondents are using distinct (non-overlapping) vocabularies to express different ideas.	Process the documents with k-means using the optimal number of topics determined by the elbow method as well as fixed numbers of topics on either side of this optimal number to enable identification of dominant topics in the data. Align the topics with themes relevant to the research question or task at hand.
While there seem to be a few distinct patterns in the data, there are subtle differences within patterns that I am interested in diving into.	Use BERTopic with UMAP dimensionality reduction and HDBSCAN clustering to break the data down into as many natural clusters as are relevant. Combine clusters via human evaluation as necessary to find themes relevant to the research question or task at hand. Choose all-MiniLM-L6-v2 for short responses with clearly separated themes; choose all-mpnet-base-v2 when themes overlap or require more contextual differentiation.
Students/respondents are using similar but not the same words to express the same ideas.	Process the data with a matrix factorization method (e.g. NMF) using the optimal number of topics determined by the elbow method to gain a basic idea of semantic similarity across different documents. Advance to BERTopic if granularity is insufficient to address the task at hand.

(continued on next page)

Table 9.3 – *continued*

I have a dataset (corpus) of documents and...	**Recommendation**
Analysis of my data is worthless unless resulting topics or groups of documents have clear semantic meaning.	Process the data using NMF and the optimal number of topics determined by the elbow method as well as fixed numbers of topics on either side of this optimal number. Adjust NMF parameters to optimize both statistical coherence (e.g., U_{mass}) and qualitative coherence (human evaluation of top words corresponding to each topic).
Although my documents tend to use the same sets of words frequently, I need to identify those documents that are ambiguous so I can have multiple human evaluators figure out what's going on with them.	Process the data using LDA and the optimal number of topics determined by the elbow method. Filter out documents where one topic probability stands out (is much higher) than the others. Consider the remaining documents ambiguous.
I know my documents have outliers that are pretty much not interpretable or irrelevant and I want to filter those out before proceeding with the more meaningful documents.	Use BERTopic with UMAP dimensionality reduction and HDBSCAN clustering to identify outliers. Manually evaluate outliers for different types of anomalous behavior and eliminate documents from further analysis as desired.
My documents are very short and use a wide range of words (large vocabulary).	Avoid LDA. Use a matrix factorization method (e.g. NMF) to gain a basic understanding of how semantic meaning differs among different topics. Advance to BERTopic if results do not provide sufficient insight or granularity.

(continued on next page)

Table 9.3 – *continued*

I have a dataset (corpus) of documents and...	**Recommendation**
I already know the themes I am looking for and want to apply them at scale without manual coding.	Use zero-shot classification (ZSC) with a pre-trained entailment model (e.g. BART-large-MNLI). Use plain-language prompts that mirror how students express concepts rather than domain-specific terminology. Keep label descriptions brief to preserve class separation. Plan for expert review of multi-theme responses, abstract concepts, and minority themes, as ZSC performance degrades for these cases without fine-tuning.

Note: this study used k-means within BERTopic to fix $k = 3$ for cross-model comparison under default parameters.

Regardless of topic model type or optimization approach, this study has added to the existing literature in cautioning against fully automated analysis of text-based data when a high degree of rigor or performance is required. In these situations, which are very common in education and education research, manual (human) intervention is not only desirable but required to make sure that topic modelling does not lead the analysis astray and into unacceptably erroneous conclusions. While short-text topic modeling is clearly a useful tool in thematic (and other related qualitative) analyses of text, it is in no way a magic bullet replacement for domain expert assessment. The frameworks for integrating human expertise into automated workflows systematically are developed in Chapter 10.

Chapter 10

HUMAN-IN-THE-LOOP INTERVENTION IN TOPIC MODELS

10.1 Introduction

Expert judgment at defined stages of thematic analysis improved topic model performance on the three datasets examined in this study as evidenced by increases in macro-F1 of up to 17.7 percentage points and increases in accuracy up to 9.3 percentage points. This chapter documents where human-in-the-loop (HITL) judgment entered the workflow, what decisions it drove, and how each decision connects to RQ2: *How can a domain expert be best integrated into the analysis of short educational research texts to improve the value and accuracy of NLP methods?*

Automated topic models find patterns by counting which words appear together most often. In many settings this works well, but the student feedback regarding instructional support used herein (and datasets with similar characteristics) have three problems that word-counting alone cannot solve. First, students describe the same idea using different words which impairs a model that groups responses by word overlap and treats synonymous phrases as separate topics. Second, many responses cover more than one topic or theme in a single answer; models that assign each response to exactly one topic consistently misclassify these. Third, high-frequency domain words embedded in the survey prompts (e.g., students, class, professor) appear in every topic-word list and make different topics look more similar than they are. These problems mean that automated results alone are not sufficient to determine which outputs are trustworthy; a domain expert is needed at targeted stages to review model outputs, identify where they align with underlying, relevant meaning, and correct where they do not.

Results from Chapters 7 and 8 show how much these problems affect model accuracy. The k-means topic model exhibited a macro-F1-score of 84.4% when classifying the questioning theme in the peer support dataset while the NMF topic model demonstrated a

macro-F1-score of only 3.35% when classifying the examples and experience theme in the TA support dataset. Figure 10.1 shows this variation across all methods and datasets.

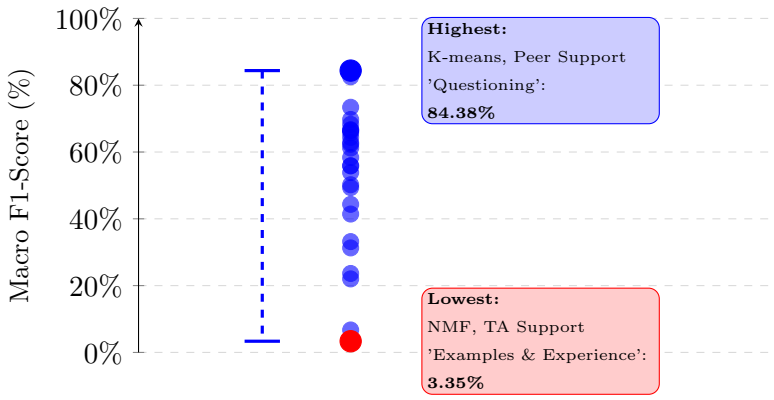


Figure 10.1: Performance Volatility in Traditional Models (macro-F1, %)

A wide range of performance means a model that works well for one theme can fail completely for another, even within the same dataset. Existing approaches address this instability through targeted expert involvement at single points in the workflow.

Existing HITL research in educational feedback analysis clusters human involvement in topic modeling into four patterns: expert review after modeling [107, 108, 109], expert-defined constraints before modeling [110, 111], interactive refinement during modeling [112, 113, 114], and supervised classification that requires predefined categories [115]. Each pattern addresses one point in the workflow but leaves the rest unstructured. None of these approaches distribute expert judgment across multiple stages, and few quantify how much expert involvement improves model performance compared to a fully automated baseline [26]. Table 10.1 summarizes these limitations.

The approach to incorporating expert involvement in the topic modeling process used in this dissertation is fundamentally different than existing approaches in that it does not serve a topic model in isolation but rather uses topic modeling as an integral part of an accepted approach to qualitative data analysis: thematic analysis. The thematic analysis process, adapted with minor changes to the use of topic models, is shown in Figure 10.2. Existing

Table 10.1: Thematic Analysis adapted for Natural Language Processing and HITL

Approach	**Existing Research**	**How this framework addresses it**
Post-model validation [107, 108]	Validates after modeling is complete; cannot correct earlier modeling errors; gains are not quantified	Expert input is present at every stage; performance is measured at each stage so validation informs model choices
Pre-model constraints [110, 111]	Requires predefined categories or seed words; can limit discovery of unexpected themes	Supports open theme discovery; validates impact with macro-F1 and inter-rater reliability at each stage
Interactive refinement [112, 113]	Measures user interaction patterns and perceived usefulness rather than classification accuracy against a validated ground truth; expert decisions are not documented; results are difficult to replicate	Expert decisions are recorded at each stage; accuracy is measured on ground-truth datasets; process is reproducible
Supervised learning [115]	Requires predefined categories; risks missing emergent or unexpected themes	Enables open theme discovery while maintaining validation rigor through inter-rater reliability

HITL efforts tend to be concentrated in the "complete analysis" stage of thematic analysis; in contrast, this study employs HITL at every stage across the full analytic process, rather than concentrating expert involvement at a single correction point.

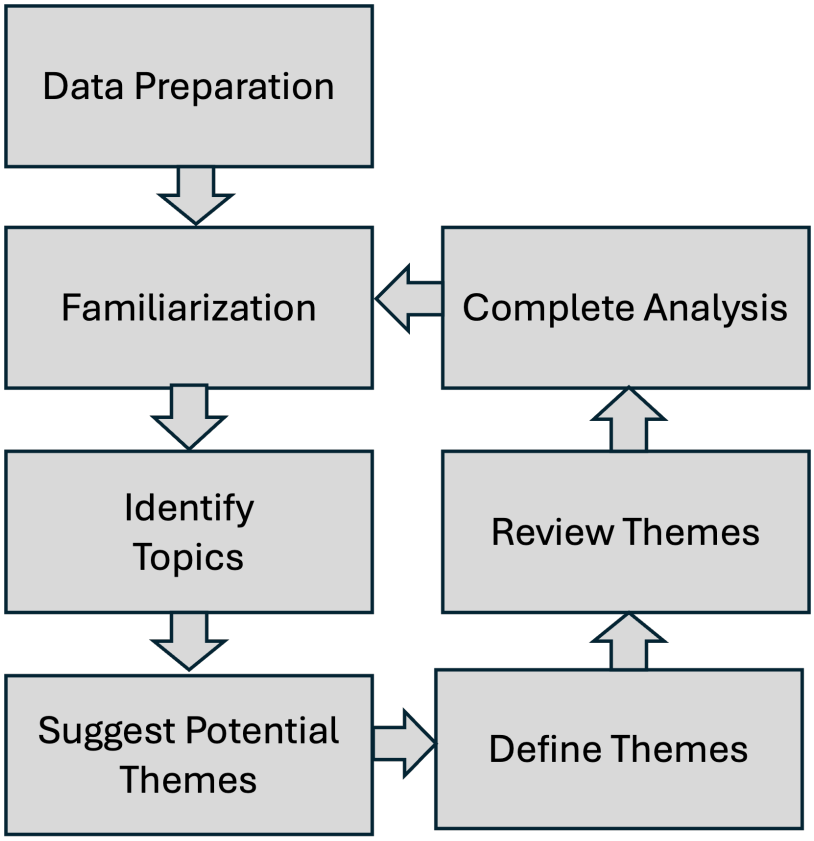


Figure 10.2: Thematic Analysis framework adapted for Natural Language Processing and HITL. The thematic analysis steps produce the verified ground truth codes used in the Complete Analysis, where the full NLP pipeline runs with expert oversight at selected stages.

The thematic analysis framework shown in Figure 10.2 is part of accepted and widely used qualitative research methodology as described by Braun and Clarke [27]. When adapted to the use of topic models to assist in the process, thematic analysis progresses through seven stages, with the option for iteration to improve end results: data prepa-

ration, familiarization, identifying candidate topics, suggesting themes, defining themes, reviewing themes, and completing the analysis by classifying documents in a dataset with the established themes. HITL intervention is valuable in each step to improve both the performance and the rigor of the analysis.

10.2 Data Preparation

The analytic process begins with data preparation, the first of the seven thematic analysis stages shown in Figure 10.2. Short student responses present several characteristics that require deliberate preprocessing before topic modeling. Responses are brief, often a single sentence or brief phrase, which limits the word overlap information that traditional topic models rely on. Students use informal language, colloquial abbreviations (e.g., "OH" for office hours, "HW" for homework), and discipline-specific vocabulary. Some responses address more than one idea in a single sentence. Survey prompts also introduced high-frequency role words ("professor," "TA," "student") that appear in nearly every response but carry no information about the type of support being requested.

The preprocessing pipeline addressed these characteristics through response-level filtering, text normalization, and targeted stopword curation. Prompt-anchored role words were removed; contractions were expanded to preserve negation ("don't" became "do not"); misspellings were corrected; abbreviations were expanded; and lemmatization was applied to reduce inflected word forms. The full preprocessing specification (tools, rules, justifications, and before-and-after examples) is in Chapter 4. These decisions were fixed before any topic model examined the data and applied consistently across all three datasets, ensuring that observed performance differences in the Complete Analysis reflect algorithmic behavior rather than preprocessing variation.

10.3 Familiarization

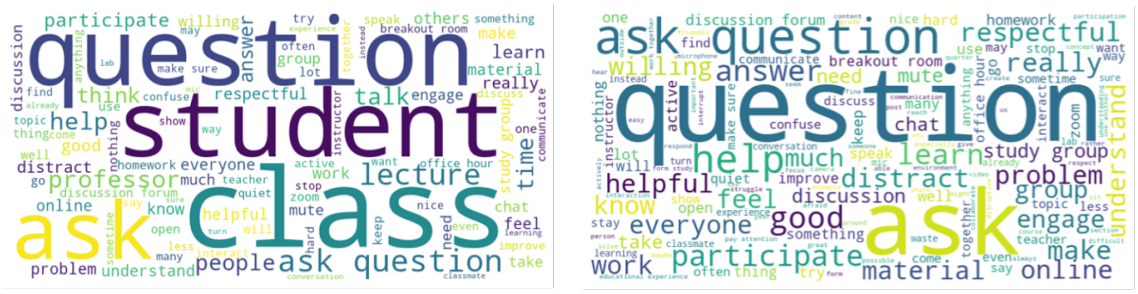
At a conceptual level, familiarization with text-based data focuses on becoming deeply and thoroughly acquainted with the data. In traditional qualitative analysis practice, this step involves a qualified human being (domain expert) reading and re-reading all of the documents within a data corpus as well as taking preliminary notes that include both observa-

tions and impressions/insights. In traditional thematic analysis, familiarization is tedious, time consuming, and vulnerable to bias because human beings invariably bring perspective that is biased by their experience and training. Human beings are also vulnerable to bias introduced by fatigue which invariably comes into play in large datasets. When NLP tools like topic modeling are introduced to the process, however, familiarization can become much more streamlined, thereby reducing fatigue, and in many cases, can reduce bias by reducing the influence of how individual language use patterns may influence the human mind. Select examples of using NLP tools for familiarization are described next.

10.3.1 Word Clouds for Data Visualization

Used frequently in topic models to visualize results, word clouds can also streamline the familiarization process. For instance, a word cloud used to visualize an entire corpus can give valuable insight into additional stopword removal or data preprocessing that can benefit subsequent topic models. For instance, a word cloud of the entire peer support data corpus after initial data cleaning and preprocessing (Figure 10.3a) illustrates that the corpus is dominated by words such as "student," "professor," and "class." While it is not surprising that these words emerge frequently among student responses regarding their expectations for peer support, they are not particularly useful for understanding how student responses differ from one another. Removing these words provides more clarity into these differences (Figure 10.3b) by enabling words like "ask" to join "question" at the forefront of the word cloud and by presenting a larger range of words in a larger font for greater visibility.

Going a step further, a word cloud of an entire data corpus using bi-grams (pairs of words) rather than single words can provide additional clarity about and insight into potential patterns of interest in the dataset. To illustrate this point, a bi-gram word cloud of the peer support dataset is shown in Figure 10.4. Some of the resulting (2,2) n-grams are obvious (e.g., ask_question) while some are not particularly relevant (e.g., would_nice, feel_like) and others provide important context to the (1,1) word clouds (e.g., discussion_board, breakout_room, pay_attention, relevant_question) and support further familiarization with the data.



(a) (b)

Figure 10.3: Peer Support unigram word clouds: (a) after data preparation; (b) after removal of low-value words.

Figure 10.4: Peer Support bigram (2,2) word cloud.

10.3.2 Additional HITL Insights from Familiarization

Word clouds and other data visualization tools borrowed from the natural language processing community can streamline the familiarization step in thematic analysis and eliminate the need for a human being to carefully read all documents in a corpus. For example, the word clouds of the peer support dataset illustrated the dominance of how important it is to students that their peers ask questions in class. Knowing this, a domain expert can justify scanning documents that refer to questions in a single pass, and reserve more in-depth examination and multiple scans for documents whose characteristics are not as readily recognized. In the familiarization phase, however, the interaction of human and NLP tool is bidirectional. NLP visualization tools help reduce the amount of time and effort a human must invest in familiarization while the human also gains insight into how to improve the efficiency of selecting and optimizing subsequent topic models. For example, in the process of reviewing documents, HITL can:

1. Evaluate vocabulary diversity; more diversity (different ways of saying similar things) ultimately points toward using a semantic embedding topic model (e.g., BERTopic, ZSC) to classify the documents in the dataset, while a more limited vocabulary would suggest the more computationally efficient frequency-based methods (LDA, NMF, k-means) may be sufficient.

2. Estimate the prevalence of documents which contain more than one topic. High incidences of multiple topics suggest the need for (a) using a model to complete the classification/theme assignment analysis that makes room for multiple topics (e.g., LDA) or alternatively flags potential multiple topic responses for further analysis and also (b) sets an expectation for reduced topic model accuracy and makes a case for going beyond accuracy in evaluating performance. In this study, more than 20% of TA support responses were flagged as multiple topics, thereby alerting the researcher to expect lower accuracy across all methods in the dataset and informed the decision to report macro-F1 alongside accuracy.

3. Gain an understanding of the proportion of documents in the dataset that are am-

biguous. A high prevalence of ambiguity advises against the use of methods that seed or predetermine the themes in the dataset (e.g., zero shot classification). In other datasets where ambiguity is rare, such as the peer support dataset used in this study, ZSC can be a powerful choice to classify documents (Chapter 8).

In summary, while no topic models per se are likely to be used in the data familiarization step, the data visualization tools used by topic models prove useful not only for gaining initial insight into what the data say but also for optimizing stopword curation, understanding context of frequent words (e.g., via bigrams), and freeing up time and effort for a domain expert to quantify additional characteristics of the data (e.g., vocabulary diversity) that will be useful to complete the analysis in the thematic analysis cycle.

10.4 Identifying Candidate Topics

Preliminary topic models can be effectively used to identify candidate topics, determine the appropriate number of topics for each dataset, and support subsequent HITL definition of relevant and meaningful themes.

10.4.1 Preliminary Topic Models

To illustrate the usefulness of topic models in identifying candidate topics, preliminary LDA models were run at k=2, 3, and 4 topics on the Peer Support dataset. Table 10.2 shows the top-10 words for each topic configuration.

These basic LDA models give insight into relevant topics present in the data including but not limited to questioning (i.e., students desiring that their peers ask questions) and collaboration (i.e., students desiring to work with one another). The model results across different numbers of topics also provide important insight into the stability of topics. The questioning topic emerges clearly regardless of the number of topics used in the model, while other topics are more muddled. Highly stable topics require less HITL involvement in later stages of thematic analysis while less stable topics will require more investment.

The stability of topics can also be explored using multiple topic models on the dataset during the identification of candidate topics. For instance, when implementing preliminary

Table 10.2: Preliminary LDA Topic Model Results: Peer Support Dataset

Configuration	**Topic**	**Top Ten Words**
Two-Topic	Topic 0	question, ask, help, think, time, good, group, people, study, answer
	Topic 1	lecture, talk, discussion, participate, respectful, make, distract, learn, sure, material
Three-Topic	Topic 0	help, discussion, group, people, study, know, really, like, willing, open
	Topic 1	lecture, talk, participate, time, good, respectful, distract, work, engage, helpful
	Topic 2	question, ask, think, answer, make, learn, sure, material, mute, understand
Four-Topic	Topic 0	good, people, distract, really, helpful, open, problem, zoom, attention, pay
	Topic 1	discussion, participate, time, answer, make, sure, chat, mute, forum, lecture
	Topic 2	talk, group, study, learn, work, engage, like, material, feel, homework
	Topic 3	question, ask, help, think, lecture, respectful, know, willing, understand, need

topic models on the peer support dataset using both LDA and NMF, the questioning topic again appears consistently across the two modeling methods, indicating that it is stable and likely dominant in the underlying data. Across the two models, "respectful" is associated with different behaviors related to: (a) talking and study groups according to LDA and (b) talking and distractions in lecture according to NMF. In the first topic/word cloud, the NMF model also points to studying together as a key component of collaboration among peers while the LDA model lacks this clarity.

Applying the same preliminary modeling approach to the Faculty and TA Support datasets produced a different set of candidate topics. For Faculty Support, three topics appeared consistently: one anchored by "office," "hours," "email," and "available" (suggesting themes related to accessibility and direct communication); a second anchored by "explain," "clear," "examples," and "understand" (suggesting themes related to teaching quality); and a third anchored by "practice," "lab," "hands-on," and "project" (suggesting themes related to applied learning). The TA Support dataset produced a closely parallel set of topics, with the same three clusters emerging at k=3, and additional sub-topics appearing at k=4 that split the teaching quality cluster into explanation style and pacing. Together, the three datasets yielded a total of nine candidate topics prior to any expert-driven consolidation.

10.5 Suggesting Potential Themes

Topics do not always align with themes nor do multiple topics aggregate into themes. However, NLP offers some valuable tools for suggesting how alignment or aggregation might proceed in thematic analysis. Based on the word cloud analysis and preliminary topic model outputs, the expert moved from initial patterns to provisional conceptual labels, identifying candidate theme names before any formal coding began.

The preliminary models consistently grouped responses around three clusters per dataset. On Faculty and TA Support, one cluster emphasized communication and availability (words like "office," "hours," "answer," "email"), a second emphasized instructional delivery (words like "explain," "examples," "clear"), and a third emphasized active and applied learning (words like "practice," "hands-on," "lab"). On Peer Support, clusters organized around

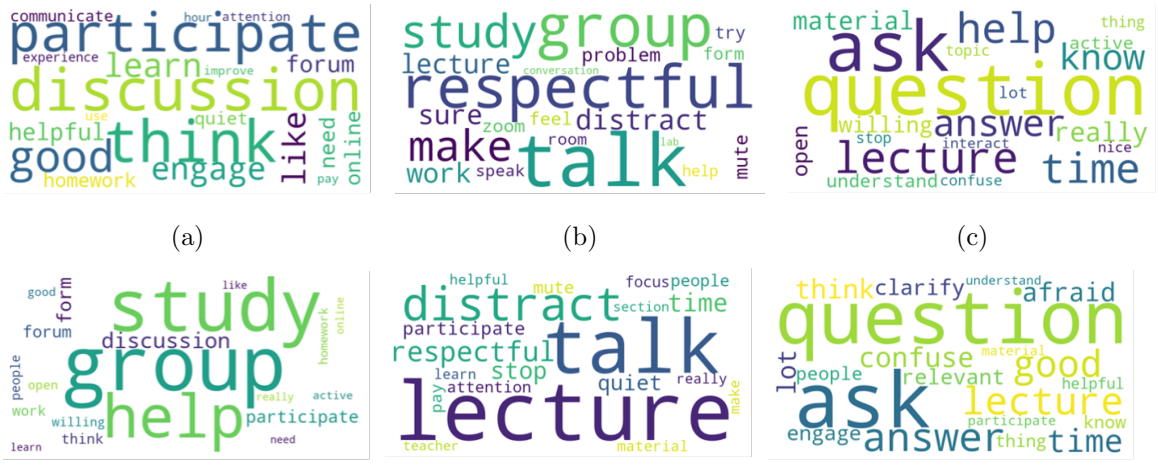


Figure 10.5: Preliminary three-topic model results for the Peer Support dataset: (a)–(c) LDA; (d)–(f) NMF.

participation and questions, respectful behavior, and collaborative support.

These groupings suggested three candidate themes for Faculty and TA Support (Interactions, Teaching Practice, and Examples and Experience) and three for Peer Support (Questioning, Civility, and Peer Support). These labels were provisional at this stage; the next step was to define their boundaries precisely and test whether independent coders applied them consistently.

10.6 Define Themes

Preliminary topic models can be used to provide a range of topics that a domain expert can then (a) use one-to-one as themes; (b) aggregate into themes; or (c) discard as irrelevant to the thematic focus of the overall research design guiding the thematic analysis and use of NLP tools and topic models in the analysis. Domain knowledge is required to distinguish themes that are conceptually different. Experts reviewed candidate themes, assigned precise boundaries, and named each theme to capture what the underlying responses have in common.

10.6.1 Review Themes

In this study, themes were reviewed across three datasets to identify pedagogically similar themes. Across faculty, TA, and peer support datasets, a domain expert reviewed the nine themes resulting from the define themes stage of thematic analysis and consolidated them into five themes conceptually aligned with teaching and learning.

This process requires expert judgment to determine whether terms referred to the same instructional behavior or distinct dimensions of support, a distinction quantitative metrics alone cannot capture. Weak topic associations were also examined to identify ideas that automated clustering had missed. This extra HITL step recovered minority or less frequent themes and prevented dominant patterns from overshadowing smaller but meaningful perspectives. For example, topics that both related to practical examples and hands-on experience were merged into one theme, examples and experience, showing how expert judgment connects fragmented algorithmic outputs to meaningful educational themes.

Table 10.3 shows how dataset-specific topics were consolidated into cross-cutting themes. Faculty Support and TA Support collapsed into the same three themes: Interactions, Teaching Practice, and Examples and Experience. Peer Support added two additional themes, Questioning and Civility, reflecting the distinct nature of peer support experiences.

Table 10.3: Mapping of Algorithmic Topics to Expert-Interpreted Themes

Dataset	**Initial Topic (Algorithmic Output)**	→	**Final Theme (Expert-Interpreted)**
Faculty Support	Topic 1: Office Hours & Communication	→	Interactions
	Topic 2: Lecture Delivery & Organization	→	Teaching Practice
	Topic 3: Examples & Practice Problems	→	Examples & Experience
TA Support	Topic 1: Availability & Help Sessions	→	Interactions
	Topic 2: Mentoring & Lab Guidance	→	Teaching Practice
	Topic 3: Quiz Materials & Examples	→	Examples & Experience
Peer Support	Topic 1: Study Groups & Collaboration	→	Interactions
	Topic 2: Asking Relevant Questions	→	Questioning
	Topic 3: Respectful Classroom Behavior	→	Civility

As a final step in reviewing themes, two independent (human) coders applied the defined themes to 100 responses from each dataset (300 total) to test consistency of interpretation. Cohen's κ was used because it adjusts for agreement expected by chance and is the standard statistic for categorical coding reliability. Agreement levels across datasets were Faculty Support ($\kappa = 0.74$), TA Support ($\kappa = 0.72$), and Peer Support ($\kappa = 0.75$). These values fall within the "substantial agreement" range on the Landis-Koch scale [106], confirming that the expert-defined themes were interpretable and reproducible across coders.

Once reliability was confirmed, the primary expert coded all remaining responses to create the reference dataset used for evaluating algorithmic predictions. Each topic and its transformation were documented in an audit trail that recorded the source cluster, main keywords, provisional labels, merge or split decisions, final theme assignment, and rationale, making every interpretive step traceable and transparent. With $\kappa \geq 0.72$ confirmed across all three datasets, the verified theme codes were ready to serve as the reference for evaluating the five NLP methods in the Complete Analysis.

10.7 Complete Analysis

Unlike most previous studies, this dissertation examined the role of a human being (HITL) at all stages of thematic analysis. The six thematic steps preceding the complete analysis step did not simply precede it but instead, played a substantial role in enabling it. Each step added a layer:

- HITL in data preparation (including collection, cleaning, and preprocessing) established a consistent corpus which highlighted words that had meaning to the task at hand while minimizing or eliminating low-value words.

- Familiarization identified vocabulary patterns, multiple topic prevalence, and ambiguity rates that contributed to shaping method selection during the completion of the analysis. NLP tools used during data familiarization also provided valuable insight into topic stability.

- Identifying candidate topics used preliminary topic models that provided a broad range

of potential codes and topics that could subsequently be aligned with or aggregated into pedagogically (conceptually) meaningful themes.

- Hierarchical analysis tools in preliminary topic models provided important insights into how topics could be combined into relevant themes.

- Defining and reviewing themes via HITL was made more accurate and insightful via NLP tools used in prior steps of thematic analysis.

In short, insights gained by HITL at the first six stages of NLP-based thematic analysis not only produced stable, meaningful themes for classification but also provided important insights into how to configure and optimize the topic models in the final stages of analysis. Before selecting and implementing a model for analysis completion, however, appropriate metrics for capturing topic model performance that are aligned with the task or research question at hand are necessary. These choices are described next.

10.7.1 Selecting Performance Metrics

Performance metrics for topic modeling divide into two categories based on whether a ground truth is required. *Internal metrics* require no prior knowledge of topic labels and can be computed directly from the corpus; they assess whether topics are coherent, stable, and distinguishable. *External metrics* require a known ground truth (either human-coded labels or an established classification) and measure how closely model outputs match that reference.

Within these two categories, three classes correspond to the amount of human effort required. Class A metrics are internal and fully automated: coherence scores (UMass, CV, CNPMI) and topic stability measures are computed after every model run with no human input. Class B metrics require a human to evaluate topic quality without a formal coded dataset: an expert reads model outputs and judges whether topics are interpretable, non-redundant, and correctly ordered. Class C metrics are external and quantitative: accuracy, precision, recall, F1-score, and inter-rater reliability (κ) are computed from expert-coded samples, typically 15–30% of the total response set.

Class C metrics are preferred when topic modeling results will inform educational decisions and a validated ground truth is available. Class A and the automated portion of Class B suit early-stage model selection when no ground truth is available. Class B (expert-driven) suits interpretability validation before committing to a coding scheme.

Table 10.4: Performance Metrics for Topic Modeling: Internal/External Classification and Educational Application

Metric	**Type**	**Class**	**When to Use**	**Relevant Examples in Education**
UMass / CV Coherence	Internal Quant.	A	To: (a) explore data; (b) to guide topic model optimization	Avoids building instructional decisions on incoherent topic structures; surfaces whether student feedback contains distinguishable patterns at all
Topic Stability	Internal Quant.	A	When results do not make sense and major errors or spurious results are suspected	Prevents misleading patterns from driving instructional decisions; confirms the topic structure is consistent enough to interpret
Topic Quality	Internal Qual.	B	Validating that topics are precise, distinguishable, and non-redundant	Ensures model-generated categories correspond to meaningfully distinct student experiences before analysis proceeds
Topic Interpretability	Internal Qual.	B	Confirming that topic words connect to the research goal	Validates that topic-word distributions describe recognizable student experiences, not statistical artifacts

Continued on next page

Table 10.4 continued

Metric	Type	Class	When to Use	Relevant Examples in Education
Topic Diversity	Internal Qual.	B	Assessing whether topics capture the full range of ideas in the corpus	Prevents dominant topics from obscuring minority perspectives; ensures the full range of student concerns is represented
Word Intrusion	Internal Qual.	B	Blind test of topic coherence: evaluator identifies a planted outlier word	Catches topics that look statistically valid but do not describe a coherent student experience, before those topics inform instructional change
Direct Label Comparison	External Qual.	B	Comparing expert-assigned labels to model-assigned labels without numerical scoring	Reveals whether model-assigned categories align with how an education expert would naturally group student responses
LLM Proxy Comparison	External Qual.	B	Using an LLM to generate keywords for sampled documents and comparing to model topics	Rapid diagnostic check of topic alignment when expert time is limited; identifies clusters that do not represent recognizable student concerns
Accuracy	External Quant.	C	Overall classification correctness; cross-method comparison	Supports selection of the best-performing method for classifying student feedback when the goal is broad instructional improvement across all themes

Continued on next page

Table 10.4 continued

Metric	**Type**	**Class**	**When to Use**	**Relevant Examples in Education**
Precision	External Quant.	C	Minimizing false positives when misclassification is costly	Scholarship evaluation: falsely ranking a student as a top candidate denies the award to a more qualified applicant
Recall	External Quant.	C	Minimizing false negatives when missing a case is costly	At-risk detection: failing to flag a student who signals academic difficulty delays intervention
Macro-F1	External Quant.	C	Balanced performance across themes; minority theme coverage essential	Ensures that minority themes in student feedback receive equal weight in evaluation, preventing dominant topics from masking gaps in coverage
Topic Ranking	External Quant.	C	Confirming topic frequency order matches expert or ground-truth prevalence	Confirms that model-identified themes reflect what students actually emphasized most, directly informing where instructional resources should be directed
Inter-Rater Reliability (κ)	External Quant.	C	Confirming that expert codes provide reliable ground truth	Establishes that expert-coded themes are reliably interpretable before using them as ground truth for educational decisions

Defining metrics before modeling begins is itself a HITL decision. A research question that asks "Which method classifies responses most accurately overall?" calls for accuracy. A question that

asks "Which method best captures all themes students express?" calls for F1-score and recall. The same methods rank differently depending on which metric is used, so choosing the wrong metric can lead to the wrong method recommendation.

When to Use Internal Metrics (Class A and Class B)

Class A internal metrics (coherence scores such as UMass and CV, and topic stability measures) require no ground truth and can be computed automatically after every model run. They are the appropriate choice during early-stage exploration when no human-coded reference exists, when comparing alternative parameter settings within a single method, or when assessing whether topics are mathematically coherent before investing expert effort in evaluation. Coherence scores should be treated as diagnostic tools rather than final judgments; a high coherence score does not guarantee that topics are educationally meaningful.

Class B metrics require a domain expert to evaluate topic quality directly, without a formal coded dataset. The expert reads topic-word distributions and assesses whether topics are interpretable, non-redundant, and correctly ordered relative to the research question. Class B is appropriate when a validated ground truth is not yet available, when checking that model outputs make sense before proceeding to inter-rater coding, or when diagnosing whether a parameter change improved topic coherence in ways that quantitative scores do not capture. Class B precedes Class C: expert review of topics should confirm interpretability before coding effort is invested.

Choosing a Ground Truth for External Metrics

When one or more performance metrics is external, the analysis depends on a reference dataset against which model outputs are measured. This reference must be established before any NLP method runs, or the comparison will lack a stable baseline. As established in the Review Themes section, the verified theme codes with $\kappa = 0.72$–0.75 across three datasets serve as the ground truth for this study. When researchers cannot establish a validated ground truth before modeling, evaluation should be restricted to Class A and Class B metrics.

When to Prioritize Accuracy

Accuracy measures the overall proportion of correctly classified responses. It is the appropriate primary metric when the research question focuses on comparative method performance across a dataset, or when all themes carry equal weight in the decision. However, accuracy can mask important differences: two methods can match in overall accuracy while differing substantially in how

well they handle each theme. Theme-level metrics should supplement accuracy whenever research questions address theme-specific performance (Chapter 7).

When to Prioritize Precision

Precision measures the proportion of model-assigned theme labels that are actually correct: of all responses the model assigned to a given theme, how many truly belong there? Precision is the appropriate primary metric when the cost of false positives is high: incorrectly assigning a response to a category leads to inappropriate action, wasted resources, or harm to individuals.

Two educational contexts illustrate high false-positive cost. In merit scholarship evaluation, a model that falsely identifies a student as a top candidate for a competitive award not only misallocates funding but denies the opportunity to a more qualified applicant. In fixed-capacity honors program admissions, a false positive takes a seat from a deserving candidate who is then excluded. In both cases, the harm from the misclassification falls on a third party, and the error cannot be easily corrected after the decision is made. These contexts require precision to be verified before results inform policy.

Per-theme precision, not just aggregate precision, must be verified before model outputs are used to guide educational decisions. Aggregate precision can remain acceptable while a single theme is largely invented by the model, directing resources toward a pattern that does not reflect actual student experience.

When to Prioritize Recall

Recall measures the proportion of actual theme instances that the model correctly identifies. Recall is the appropriate primary metric when missing themes is costly, when research questions focus on coverage, or when identifying all instances of a theme matters more than avoiding false positives. High accuracy with low recall indicates that the model captured dominant themes while systematically missing smaller ones; this pattern appeared in the ZSC results on Peer Support, where strong overall accuracy masked notably low recall on a minority subtheme (Chapter 8).

When to Prioritize F1-Score

F1-score is the harmonic mean of precision and recall, accounting for both false positives and false negatives. It is the appropriate primary metric when research questions target theme-level classification quality, when themes carry different importance, or when using either false positives or

false negatives alone as the criterion would produce a misleading ranking. The comparative analysis showed substantial F1 variation across methods and datasets, even when accuracy was similar (Chapter 7).

When to Prioritize Macro-F1

Macro-F1 averages F1 equally across all themes regardless of theme size, making it more sensitive to minority theme failures than weighted F1 (which favors dominant themes). Macro-F1 is the appropriate primary metric when research questions address theme balance or equity across themes, or when missing a smaller theme is as consequential as misclassifying the dominant one. In educational settings where minority voices (such as students with accessibility needs or students from underrepresented groups) may cluster into smaller themes, macro-F1 ensures those themes are not overlooked by a metric that rewards overall accuracy.

When to Prioritize Topic Ranking

Topic ranking measures whether model outputs align with expert-defined theme priorities or prevalence. It is the appropriate primary metric when research questions ask whether model-identified themes reflect the relative prevalence of issues that respondents actually express. A model that achieves acceptable accuracy may still fail topic ranking by assigning high salience to a minor theme or by suppressing the dominant theme. The educational stakes are direct: if a model ranks a secondary concern as the primary theme, instructors and administrators invest in the wrong area while actual barriers remain unaddressed. The comparative analysis confirmed this pattern: BERTopic MPNet achieved high accuracy on Faculty Support but still failed theme ranking alignment on both Faculty and TA Support, producing statistically coherent clusters that did not match expert-defined theme priorities (Chapter 7). Topic ranking must therefore be validated against expert judgment before results inform instructional change.

When to Use Inter-Rater Reliability

Inter-rater reliability (κ) is the appropriate metric when external (Class C) metrics will be used to evaluate model outputs and a human-coded ground truth has been established. Before any external metric can be meaningfully interpreted, the ground truth itself must be verified as consistently applied. A ground truth coded by a single researcher, or by multiple coders who disagree frequently, introduces noise into every downstream metric. κ must be established before modeling begins; its role in this study is described in the Review Themes section.

Selecting Multiple Metrics: A Balanced Approach

Research questions that address method performance broadly require multiple metrics: overall correctness (accuracy), theme-level balance (F1-score), and alignment with expert priorities (topic ranking). Each dimension captures something the others miss, and using only one metric can lead to false conclusions about which method is best. In this study, two methods matched in accuracy but differed substantially in macro-F1; another matched in accuracy but failed topic ranking; full details are in Chapters 7 and 8. Defining consistent metrics before modeling begins creates a foundation for fair cross-method comparison and connects later performance improvements directly to specific research questions.

10.7.2 Selecting a Topic Model

Method performance varies substantially based on dataset characteristics, and expert review of sample responses identifies the relevant patterns before any model runs. The comparative analysis in Chapters 7 and 8 provides the empirical basis: the same method that performs well on one dataset can perform poorly on another, a gap driven by dataset characteristics rather than method failure.

Five characteristics drive method selection. Reviewing a random sample of 30–50 responses per dataset is sufficient to assess each:

1. **Vocabulary diversity.** Do different respondents use different words to describe the same concept? High diversity (synonymous expressions across responses) favors semantic embedding methods (BERTopic, ZSC) over frequency-based methods (LDA, NMF, k-means), which rely on lexical co-occurrence. Low diversity, where a small shared vocabulary dominates, makes frequency-based methods viable.

2. **Ambiguity rate.** What proportion of responses address more than one theme within a single answer? Rates above 20% indicate that hard-assignment models will misclassify multi-theme responses. Mixed-membership models (LDA, NMF) tolerate ambiguity better than hard-assignment models (k-means). When ambiguity exceeds 30%, no unsupervised method reliably separates themes without post-hoc expert correction.

3. **Mean response length.** Are responses short (under 15 words) or longer? Short responses provide limited word overlap information, which reduces the advantage of frequency-based methods and narrows performance differences across methods. BERTopic MiniLM handles short responses well; BERTopic MPNet benefits from longer context.

4. **Domain terminology density.** How frequently do responses use terms specific to the institutional or disciplinary context (e.g., role labels, course-specific vocabulary) that carry little thematic meaning? High domain terminology density requires targeted stopword curation before any method will produce clean topics. If domain terms are not removed, they dominate topic-word distributions regardless of method choice.

5. **Theme stability across preliminary runs.** Do preliminary LDA topic-word distributions change substantially across runs with the same k? Unstable topics (major keyword shuffles between runs) indicate high ambiguity or insufficient word overlap, and suggest that neural methods or ZSC (which do not depend on co-occurrence frequency) may outperform traditional models.

These observations feed directly into Table 10.5, which maps each characteristic to a recommended method, an empirical basis from this study, and a caution.

For example, when responses are short with clear themes, both frequency-based methods and BERTopic MiniLM perform well. When vocabulary diversity is present, semantic methods are recommended. When multi-theme rates are high, mixed-membership models or human post-processing are needed.

Expert judgment is needed to interpret these characteristics and apply them to new datasets. The table provides guidance, but researchers must assess their own data characteristics through the five-point review above before determining which recommendations apply. The five characteristics are not independent: high vocabulary diversity and high ambiguity rate together are a strong reason to avoid k-means and to weight BERTopic or ZSC higher in the selection. Low vocabulary diversity with short responses and low ambiguity (the Peer Support profile) opens the field to all methods and makes quantitative performance the deciding criterion.

10.7.3 Working with a Topic Model

Text representation selection is the preprocessing decision that directly shapes how each method processes the text. This decision is made before the optimization loops begin and applied consistently across all three datasets. Full descriptions of each algorithm are provided in Chapter 6.

Text Representation

Text representation is the second preprocessing decision shaped by expert input. The choice of representation determines what information is available to the algorithm: frequency counts, term

Table 10.5: Model Selection Guidance Based on Dataset Characteristics

Dataset Characteristic	**Traditional Methods**	**Neural Methods**	**Zero-Shot Classification**
Short text, clear themes (5–8 words, distinct boundaries)	**LDA recommended** (78.5% accuracy) Cannot group responses that use different words but mean the same thing	**BERTopic MiniLM recommended** (85.0% accuracy) May combine themes that sound similar but are different	**Use when:** Themes already defined Requires testing and improving prompts
Multi-topic responses (>20% multi-topic)	**NMF with TF-IDF** (76.8% accuracy) Can assign responses to multiple themes	**BERTopic MPNet** (70.3% accuracy) Assigns each response to only one theme	**Limited:** Assigns one label per response Requires design to handle multiple themes
High ambiguity (Abstract themes, context-dependent)	**Baseline only** (58–70% accuracy) Word meaning depends on context, which limits these methods	**Use with caution** (54–70% accuracy) Requires expert review after analysis	**Use when:** Theme framework is stable Expert review is essential
Predefined themes available	**Not ideal:** Discovers topics first, requires mapping to themes after analysis	**Not ideal:** Discovers topics first, requires mapping to themes after analysis	**Recommended:** Classifies directly into themes (85% accuracy) Matches existing theme categories

Note: Based on empirical results from this study's datasets. Performance may vary with different data characteristics.

weights, or semantic vectors. For the three datasets in this study, the expert evaluated three criteria before selecting a representation method: (1) whether the algorithm requires sparse or dense input, (2) whether vocabulary diversity in the data would be captured by term frequency or requires semantic encoding, and (3) whether consistent representation across all datasets was feasible.

LDA received counts-based vectorization because its multinomial formulation expects integer word counts. NMF and k-means used TF–IDF as the default representation, since TF–IDF down-weights terms that appear in nearly every response and emphasizes discriminative terms, reducing noise from common support-role vocabulary. Counts-based vectorization would have treated high-frequency common terms as equally informative, hiding the words that actually separate themes. During k-means optimization, the expert evaluated counts-based vectorization against TF–IDF and selected the representation that produced better cluster separation for each dataset, confirming that the representation choice has measurable impact on model performance (see Chapter 7).

BERTopic received sentence-level embeddings from sentence-BERT [95]. Sentence embeddings encode semantic relationships across the full response rather than treating each term independently, which directly addresses the vocabulary diversity problem observed in Faculty and TA Support responses. The expert selected MiniLM as the default embedding model for initial runs (lower computational cost, better performance on short responses) and MPNet as the alternative for datasets with longer or more varied sentence structures. The embedding model choice is itself an expert-guided decision, documented in the BERTopic optimization workflow below.

Zero-shot classification uses raw text passed directly to the natural language inference model (BART-large-MNLI). No preprocessing beyond tokenization is applied because the model was already trained on vocabulary and context; additional vectorization would strip the information the model needs to match responses to labels. The representation choice for ZSC is therefore fixed by the model architecture, and the expert's input focuses instead on label phrasing rather than vectorization.

These selections were applied consistently across all three datasets to ensure that representation differences did not confound cross-method comparisons in Chapters 7 and 8.

k-means Vectorization and Configuration Optimization

k-means optimization targets three configuration choices: vectorization method, dimensionality reduction, and centroid normalization. Per-theme precision is the primary indicator alongside accuracy and macro-F1, because aggregate accuracy can remain stable while a single theme collapses. The HITL role is to identify when per-theme precision fails even as aggregate accuracy looks acceptable, and to select the configuration that best aligns with the dataset's vocabulary structure.

[HITL] = Human-in-the-Loop decision point **[NLP]** = Automated step

Pre-Loop Setup

[HITL] Step 1: Assess Data Characteristics

↓

[HITL] Step 2: Select Priority Performance Metrics

Select 1–3 metrics before modeling begins; see the Selecting Performance Metrics section. Define metrics based on the research question, not after results are known.

Optimization Loop (repeat until accepted)

[NLP] Step 3: Run K-means with Default Configuration

TF-IDF vectorization, k-means++ initialization, $k = 3$; no dimensionality reduction; no L2 normalization

↓

[NLP] Step 4: Calculate Per-Theme Performance Metrics

Calculate the metrics selected in Step 2; record baseline values.

↓

[HITL] Step 5: Review Vectorization Method

Examine: (1) per-theme precision by theme; (2) feature space characteristics (term frequency distribution, IDF weight distribution); (3) sample responses from the lowest-precision theme

Decision	Observation	Action / Evidence
Switch to counts-based vectorization	Per-theme precision: lowest theme below your target threshold. Feature space: TF-IDF IDF weights heavily penalize terms that appear frequently within one theme but across many responses corpus-wide. Sample responses: thematic keywords for the low-precision theme are underrepresented in the feature space.	Switch to counts-based vectorization. This reduces IDF-driven suppression of thematic vocabulary when themes share high-frequency terms.
Keep TF-IDF	Per-theme precision meets target for all themes; vocabulary is discriminative across themes and response lengths are consistent.	Keep TF-IDF.

↓

[HITL] Step 6: Review Dimensionality Reduction

Examine: (1) macro-F1 stability across repeated runs (high variance indicates noisy feature space); (2) vocabulary size after stopword removal; (3) centroid separation across themes

Decision	Observation	Action / Evidence
Apply PCA	Macro-F1 varies noticeably across repeated runs. Feature space is high-dimensional and sparse after stopword removal; centroid positions shift between runs.	Apply PCA to reduce feature space noise before centroid computation.
No PCA	Macro-F1 stable across runs; vocabulary space is already compact after preprocessing.	No PCA.

↓

[HITL] Step 7: Review Centroid Normalization

Examine: (1) per-theme precision; (2) theme cluster sizes (whether one cluster dominates); (3) high-frequency vocabulary in centroid word lists

Decision	Observation	Action / Evidence
Apply L2 normalization	Per-theme precision: one theme still below target despite vectorization and PCA adjustments. Theme cluster sizes: one cluster absorbs disproportionately many responses, pulling the centroid toward high-frequency generic vocabulary.	Apply L2 normalization (scales each response vector to unit length before cluster assignment; prevents large clusters from dominating centroid positions).
No L2 normalization	Per-theme precision acceptable; cluster sizes balanced across themes.	No L2 normalization.

↓

[NLP] Step 8: Run K-means with Revised Configuration

Apply revised vectorization, dimensionality reduction, and normalization decisions; keep $k = 3$

↓

[NLP] Step 9: Calculate Performance Metrics

Calculate selected metrics; compare to baseline from Step 4.

♦ EVALUATE: Metrics improved AND topics remain qualitatively interpretable? ♦

YES

Accept optimized configuration. Proceed to Step 10: Select Optimal K-means Configuration.

NO: Return to Step 3

Revise one configuration decision at a time and re-run. If per-theme precision consistently fails to improve for a given theme, check the ambiguity rate from Step 1: k-means assigns each response to exactly one cluster, and no configuration resolves genuinely multi-theme responses. Document as a data structure limitation and proceed to Step 10.

Post-Loop

[HITL] Step 10: Select Optimal K-means Configuration

Record final vectorization method, dimensionality reduction decision, normalization decision, and the per-theme precision that drove each choice

The expert evaluates vectorization choices (counts-based vs. TF–IDF) based on the vocabulary diversity pattern observed during familiarization. When themes are lexically distinct and word overlap between themes is low, counts-based vectorization with PCA and L2 normalization tends to produce better cluster separation. When discriminative terms are rare but meaningful, TF–IDF amplifies those terms and performs better. Because no single configuration generalizes across datasets, the expert reviews per-theme metrics after each run to confirm which configuration to keep. Ambiguity rates above 20% are documented as a data structure limit: k-means assigns each response to exactly one cluster, which cannot correctly handle responses that genuinely cover two themes, and no configuration change resolves that structural constraint.

LDA Hyperparameter Optimization

Once k is fixed at 3 (validated in Section 10.7), LDA optimization focuses on two hyperparameters: α (document-topic concentration) and β (topic-word concentration). Alpha is adjusted before beta; adjusting them simultaneously obscures the source of any performance change. UMass coherence serves as the internal check because it uses document-level co-occurrence probabilities that match the TF-IDF representation and can be computed after every run without additional human input.

[HITL] = Human-in-the-Loop decision point **[NLP]** = Automated step

Pre-Loop Setup

[HITL] Step 1: Assess Data Characteristics

↓

[HITL] Step 2: Identify Priority Performance Metrics

Select 1–3 metrics before modeling begins; see the Selecting Performance Metrics section. Define metrics based on the research question, not after results are known.

Optimization Loop (repeat until accepted)

[NLP] Step 3: Run LDA with Default Parameters

Symmetric $\alpha = 1/k \approx 0.33, \quad \beta = 0.1, \quad k = 3$

↓

[NLP] Step 4: Calculate Performance Metrics

Calculate the metrics selected in Step 2; record baseline values.

↓

[HITL] Step 5: Alpha (α) Tuning

Examine: (1) performance metrics (accuracy, macro-F1); (2) topic-document distribution (how evenly topics are assigned across responses); (3) sample responses per topic (read 10–15 responses per topic to check theme alignment). Use ambiguity rate from Step 1 to determine which row below applies before examining the distribution.

Decision	Observation	Action / Evidence
Decrease α	*Performance metrics:* macro-F1 below target; per-theme F1 varies widely across themes. *Topic-document distribution:* diffuse: most responses show near-equal probability across all topics with no clear dominant assignment. *Sample responses:* responses assigned to one topic contain vocabulary characteristic of two or more themes. Data characteristics (Step 1): ambiguity rate below 15%; vocabulary diversity low or medium.	Set $\alpha \rightarrow$ 0.01–0.10 (sparse); tighter per-document distributions help separate themes when vocabulary overlap is low and most responses address a single theme.
Increase α	*Performance metrics:* minority theme recall low; macro-F1 low and per-theme F1 varies widely. *Topic-document distribution:* minority theme cluster undersized relative to expected proportion; dominant theme oversized. *Sample responses:* responses about minority themes assigned to the wrong topic despite using vocabulary that clearly belongs to that theme.	Set $\alpha \rightarrow$ 0.5–1.0 (allows each response to span multiple topics). Not applicable to the three datasets in this study.
Keep α (do not tune)	*Data characteristics (Step 1):* ambiguity rate above 15%, indicating genuine multi-theme responses in the corpus. *Topic-document distribution:* distribution balance does not improve with tuning; per-theme accuracy shifts without a net gain across iterations. *Sample responses:* multi-theme responses confirmed by reading; theme vocabulary overlap is genuine, not a modeling artifact.	Keep α at default ($1/k$); document as a data structure limitation. Forcing sparse assignments onto genuinely multi-theme responses reduces accuracy without resolving the underlying overlap.

↓

[HITL] Step 6: Beta (β) Tuning

Examine: (1) performance metrics relative to baseline; (2) topic-word overlap (check whether top-10 words repeat across topics); (3) sample responses (confirm top words match how students actually express each theme)

Decision	Observation	Action / Evidence
Decrease β	*Performance metrics:* accuracy low despite acceptable UMass coherence. *Topic-word overlap:* top-10 words share more than 50% of terms across two or more topics. *Sample responses:* top words per topic do not align with the vocabulary students use to describe that theme.	Set $\beta \rightarrow$ 0.01–0.10 (sparse); tighter per-topic word distributions reduce word overlap between topics when themes have distinct vocabulary.
Increase β	*Performance metrics:* UMass coherence low; per-theme F1 varies widely. *Topic-word distribution:* topics fragmented; each topic contains only one or two high-frequency keywords, too sparse for stable word assignment. *Sample responses:* top words per topic do not form a coherent theme readable by the expert.	Set $\beta \rightarrow$ 0.5–1.0 (allows words to spread across topics).
Keep β **(do not tune)**	Same high-ambiguity condition as α (Step 5): ambiguity rate above 15% from Step 1; or the α change alone closed the accuracy gap and topic-word overlap is already below 30%. Changing β as well would obscure which parameter drove the improvement.	Adjust α first; change β only if α adjustment is insufficient.

↓

[NLP] Step 7: Run LDA with New Parameters

Apply revised α and/or β; keep $k = 3$ (validated in the Complete Analysis section)

↓

[NLP] Step 8: Calculate Performance Metrics

Calculate selected metrics; compare to baseline from Step 4.

♦ EVALUATE: Metrics improved AND topics remain qualitatively interpretable? ♦

YES

Accept optimized parameters. Proceed to Step 9: Select Optimal LDA Model.

NO: Return to Step 3

Revise parameters and re-run. If top priority performance metrics consistently fail to improve after three iterations, the data characteristics identified in Step 1 (high ambiguity rate, genuine vocabulary overlap between themes) are the binding constraint. Document the limitation and proceed to Step 9.

Post-Loop

[HITL] Step 9: Select Optimal LDA Model

Record final α, β, k values and the rationale for each parameter choice; note which data characteristic from Step 1 drove each tuning decision

The expert adjusts alpha before beta to isolate the effect of each change. Sparse alpha concentrates each document's topic distribution more tightly, which helps when themes are lexically distinct but can hurt when responses draw vocabulary from multiple themes. The expert diagnoses which situation applies by reading topic-word distributions and checking the ambiguity rate established during familiarization. When the ambiguity rate is low and themes are well-separated, sparse alpha improves performance; when responses genuinely span multiple themes, sparse alpha forces single-topic assignments that the data does not support. This diagnosis–based on familiarization data rather than the metric alone–determines whether alpha should be lowered or held at default.

NMF Regularization Optimization

NMF optimization targets two decisions: regularization strength (α_W, α_H) and domain stopword curation. Default NMF applies no regularization, which allows word weights to spread freely across topics and often produces topics with high word overlap. Expert review distinguishes whether

overlap reflects genuine theme similarity or parameter underfitting before removing any term from the vocabulary.

[HITL] = Human-in-the-Loop decision point **[NLP]** = Automated step

Pre-Loop Setup

[HITL] Step 1: Assess Data Characteristics

↓

[HITL] Step 2: Select Priority Performance Metrics

Select 1–3 metrics before modeling begins; see the Selecting Performance Metrics section. Define metrics based on the research question, not after results are known.

Optimization Loop (repeat until accepted)

[NLP] Step 3: Run NMF with Default Parameters

$\alpha_W = 0, \quad \alpha_H = 0, \quad k = 3$*; no regularization*

↓

[NLP] Step 4: Calculate Performance Metrics and Word Overlap

Calculate the metrics selected in Step 2; record top-10 word overlap across topic pairs as an indicator of theme separation; record baseline values.

↓

[HITL] Step 5: Review Regularization Strength

Examine: (1) top-10 word overlap percentage across all topic pairs; (2) performance metrics vs. baseline; (3) sample responses from each topic to check whether themes are distinguishable

Decision	Observation	Action / Evidence
Increase regularization (α_W, α_H)	*Word overlap:* top-10 words share more than 50% of terms across two or more topic pairs. *Performance metrics:* macro-F1 low; accuracy below baseline. *Sample responses:* responses from two different expert-coded themes appear in the same NMF topic.	Increase α_W and α_H (tested values: 0.1, 0.5); narrowing word weight distributions reduces artificial overlap between themes.
Decrease regularization	*Word overlap:* acceptable; but topics fragmented into narrow single-keyword clusters. *Performance metrics:* per-theme F1 collapses for minority theme; coherence low. *Sample responses:* each topic contains only one or two high-frequency words; themes not readable.	Decrease α_W and α_H toward 0; allows word weights to spread more freely across topics.
Keep defaults	*Word overlap:* below 30% across all topic pairs. *Performance metrics:* accuracy and macro-F1 meet the target threshold.	Keep $\alpha_W = 0$, $\alpha_H = 0$.

↓

[HITL] Step 6: Review Domain Stopwords

Examine: (1) top-5 words per topic after regularization; (2) sample responses for any high-frequency term appearing in multiple topics; determine whether that term is background vocabulary or a genuine thematic signal

Decision	Observation	Action / Evidence
Add term to stopword list	A domain term appears in the top-5 words of two or more topics despite regularization. Sample responses: term appears regardless of which expert-coded theme the response belongs to, confirming it is background vocabulary.	Add to stopword list; re-run full optimization loop from Step 3. **Caution:** verify the term is background vocabulary before removing. If removing it reduces performance, the term is genuinely multi-thematic; document as a data structure limitation.
Keep term	Term appears in top-5 of one topic only; sample responses confirm it marks a specific theme and is not background noise.	Keep term; it carries signal for at least one theme.

↓

[NLP] Step 7: Run NMF with New Parameters

Apply revised α_W, α_H, and updated stopword list; keep $k = 3$

↓

[NLP] Step 8: Calculate Performance Metrics

Calculate selected metrics; compare to baseline from Step 4.

♦ EVALUATE: Metrics improved AND topics remain qualitatively interpretable? ♦

YES

Accept optimized configuration. Proceed to Step 9: Select Optimal NMF Configuration.

NO: Return to Step 3

Revise regularization or stopwords and rerun. If overlap persists despite regularization, check whether the overlapping terms are genuinely multi-thematic (as with "example" in TA Support); if so, document as a data structure limitation rather than a modeling failure, and proceed to Step 9.

Post-Loop

[HITL] Step 9: Select Optimal NMF Configuration

Record final α_W, α_H, stopword additions, k, and the word overlap observation that drove each decision

The expert reviews topic-word overlap after each regularization change to determine whether overlap reflects genuine theme similarity or parameter underfitting. Increasing regularization narrows the word profiles of each topic, reducing overlap between themes; when that overlap was artificial, performance improves. However, when a word appears across multiple topics because it is genuinely multi-thematic in the corpus–not because of underfitting–the expert must recognize this as a data structure limit and stop iterating. For example, a word that appears consistently in the top words of every topic likely reflects how students write rather than a modeling error; removing it degrades performance because it was semantically central, not peripheral. The expert's judgment about whether to remove a term or accept it as multi-thematic is the decisive HITL input in NMF optimization.

BERTopic Embedding and Cluster Parameter Optimization

BERTopic optimization begins with an embedding model selection that precedes the parameter tuning loop, because the embedding model determines how responses are compared numerically and this choice interacts with vocabulary diversity and response length in ways that HDBSCAN adjustments alone cannot fix.

[HITL] = Human-in-the-Loop decision point **[NLP]** = Automated step

Pre-Loop Setup

[HITL] Step 1: Assess Data Characteristics

↓

[HITL] Step 2: Select Embedding Model

Review data characteristics from Step 1 to select MiniLM or MPNet (both sentence-BERT variants [95]); when characteristics are ambiguous, run both and compare

Decision	Observation	Action / Evidence
Use MiniLM	Data characteristics (Step 1): mean response length below 12 tokens; ambiguity rate below 10%; vocabulary diversity low. Sample responses: themes are lexically distinct; students use consistent vocabulary.	Use MiniLM (efficient sentence-BERT; captures sentence meaning for short, syntactically simple, lexically consistent inputs).
Use MPNet	Data characteristics (Step 1): mean response length above 15 tokens; vocabulary diversity high; ambiguity rate moderate (10–20%). Sample responses: responses use varied phrasing for the same theme; lexical overlap between themes is high.	Use MPNet (larger architecture; captures nuanced semantic distinctions across varied phrasing when vocabulary diversity is high).
Run both and compare	Data characteristics assessment does not clearly indicate either model; ambiguity rate between 8–12% and vocabulary diversity medium.	Run both MiniLM and MPNet; compare accuracy and macro-F1; select the higher-performing model before entering the optimization loop below.

Optimization Loop (repeat until accepted)

[NLP] Step 3: Run BERTopic with Selected Embedding

Use embedding from Step 2; HDBSCAN default parameters; k target = 3

↓

[NLP] Step 4: Calculate Performance Metrics and Topic Count

Calculate the metrics selected above; confirm HDBSCAN topic count matches the target k.

↓

[HITL] Step 5: Review HDBSCAN Minimum Cluster Size

Examine: (1) topic count from Step 4; (2) per-theme macro-F1; (3) sample responses from any outlier cluster (responses HDBSCAN could not assign to a topic)

Decision	Observation	Action / Evidence
Increase `min_cluster_size`	Topic count exceeds $k = 3$; small spurious clusters form around outlier responses. Per-theme macro-F1: one or more micro-clusters have very low F1.	Increase `min_cluster_size` until topic count matches $k = 3$ from expert review in the Complete Analysis section.
Decrease `min_cluster_size`	Topic count below $k = 3$; responses cluster too broadly, merging themes that expert review identified as distinct.	Decrease `min_cluster_size`.
Keep default	Topic count equals $k = 3$ on first run; accuracy and macro-F1 within expected range for the dataset profile.	Accept default; proceed to evaluation.

↓

[NLP] Step 6: Run BERTopic with Adjusted Parameters

Apply revised `min_cluster_size`; keep selected embedding model and $k = 3$ target

↓

[NLP] Step 7: Calculate Performance Metrics; Confirm Topic Count

Calculate selected metrics; confirm topic count matches target k before accepting configuration.

♦ EVALUATE: Metrics improved AND topics remain qualitatively interpretable? ♦

YES

Accept optimized configuration. Proceed to Step 8: Select Optimal BERTopic Configuration.

NO: Return to Step 3

Revise `min_cluster_size` or switch embedding model and re-run. If embedding overlap between themes persists across both MiniLM and MPNet, the overlap is a property of how students write, not a modeling error; document as a data structure limitation and proceed to Step 8.

Post-Loop

[HITL] Step 8: Select Optimal BERTopic Configuration

Record final embedding model, `min_cluster_size`, and the data characteristic observation from Step 1 that drove the embedding selection

The expert selects the embedding model based on vocabulary diversity and response length observed during familiarization. Short, lexically consistent responses with well-separated themes favor compact embedding models, while datasets with higher vocabulary diversity and longer responses benefit from higher-capacity models that preserve finer semantic distinctions. After embedding model selection, the expert runs an HDBSCAN sensitivity check to confirm that the expected number of coherent themes is produced. When theme overlap persists across multiple embedding configurations, the expert documents this as a data structure limit–the overlap reflects how students write, not a modeling error–and proceeds to the evaluation stage rather than continuing to adjust parameters.

Zero-Shot Classification: Prompt Design Cycle

ZSC optimization targets label phrasing rather than model hyperparameters: the model weights are fixed (BART-large-MNLI, pre-trained on natural language inference tasks); the only expert-controlled input is the candidate label set. The prompt design cycle iterates over text formulations, not numeric parameters, making it structurally distinct from the hyperparameter loops described above.

[HITL] = Human-in-the-Loop decision point **[NLP]** = Automated step

Pre-Loop Setup

[HITL] Step 1: Review Corpus Vocabulary

Randomly select $n = 200$ responses; read and record how students phrase each theme in their own words; note the most common expressions before drafting any label

↓

[HITL] Step 2: Draft Initial Label Set

Write candidate labels in mainstream everyday language; avoid domain-specific or technical terminology; BART-large-MNLI was trained on general-purpose text and labels in everyday language match its pre-training vocabulary

Prompt Design Cycle (repeat until accepted)

[NLP] Step 3: Run ZSC with Initial Labels

Fixed model: BART-large-MNLI; input: candidate label set from Step 2

↓

[NLP] Step 4: Calculate Per-Label F1, Accuracy, Macro-F1

Calculate the metrics selected above; include per-label F1 to identify which specific labels underperform.

↓

[HITL] Step 5: Diagnose Low-F1 Labels

For each label below your target F1 threshold: (1) check label phrasing (technical vs. plain language); (2) read 15–20 misclassified responses; (3) check theme sample size (n); diagnose whether failure reflects phrasing mismatch, threshold miscalibration, or a structural data constraint

Decision	Observation	Action / Evidence
Revise to plain language	Per-label F1 below target. Misclassified responses: label uses terminology not present in student responses; students phrase the theme in everyday words. Phrasing: technical or domain-specific.	Rephrase in mainstream everyday language; match the vocabulary students use in their responses rather than formal academic or domain-specific terminology.
Test label with description	Per-label F1 low despite plain language phrasing. Misclassified responses: plain label alone is too broad; model assigns it to responses it should not.	Add a 1-sentence description to the label; test against the no-description baseline using per-label F1, not aggregate accuracy, because description effects differ by theme. Adding descriptions can make the model overly selective; test against the no-description baseline using per-label F1 before keeping any description.
Adjust confidence threshold	Jaccard index low (model assigns fewer labels per response than expert coding indicates; multi-label responses are under-captured).	Lower the confidence threshold; test effect on per-label precision-recall balance.
Document structural limit	Per-label F1 below target despite plain language, description, and threshold testing. Misclassified responses: theme describes an absence of behavior students rarely articulate explicitly; or minority theme with very few examples.	Document the structural limit; apply manual override for that label. Do not continue iterating: the failure reflects how the data is structured, not a label engineering problem.

↓

[HITL] Step 6: Revise Label Phrasing or Threshold

Apply the diagnosis from Step 5 to the affected labels only; keep labels that met the F1 threshold unchanged

↓

[NLP] Step 7: Run ZSC with Revised Labels

Fixed model: BART-large-MNLI; input: revised label set

↓

[NLP] Step 8: Calculate Performance Metrics

Calculate selected metrics; compare to baseline from Step 4.

♦ EVALUATE: Metrics improved AND topics remain qualitatively interpretable? ♦

YES

Accept final label set. Proceed to Step 9: Select Final Label Set.

NO: Return to Step 5

Return to Step 5: re-diagnose the remaining low-F1 labels. If F1 fails to improve after three revision cycles despite plain language and threshold adjustment, apply the structural limit decision from the table above; continuing to iterate on a structurally constrained label does not improve results.

Post-Loop

[HITL] Step 9: Select Final Label Set

Record final label text for each theme, the diagnosis from Step 5 that drove each revision, and any structural limits documented during the cycle

The alignment between student language and model pre-training is the primary performance driver. The difference becomes concrete when comparing label formulations. For the Questioning theme in Peer Support, the domain-specific label "Questioning Practice" failed because formal academic phrasing does not match the model's pre-training distribution. A representative student response reads: "I like when people ask questions in lecture, most times I have the same question and it makes it easier to ask a question if someone else does it first." The mainstream label "Asking Appropriate and Relevant Questions" matched this response directly: "ask questions" appears three

times and maps to "Asking...Questions;" the lecture context maps to "Appropriate and Relevant." The domain-specific label offered no such word-level match.

The expert iterates over label formulations, testing whether mainstream language that aligns with how students write outperforms formal academic phrasing. Concise labels that match student vocabulary generally perform better than extended descriptions, which can make the model too selective. The expert stops iterating when further label changes produce no meaningful improvement or when the error source is structural rather than linguistic–for instance, when a theme describes an absence of behavior that students rarely articulate explicitly, or when a minority theme has too few instances for label engineering to have a reliable effect. In these cases, the expert documents the structural limit and applies manual override instead of continuing to redesign labels.

Balancing Quantitative Metrics with Qualitative Meaning

Quantitative metrics identify when model outputs may need adjustment but cannot determine whether a parameter change produces topics that are educationally interpretable. Expert review at the end of each optimization iteration evaluates both dimensions: did the metrics improve, and do the topics still make sense?

Three cases illustrate where quantitative metrics and qualitative review point in different directions. First, coherence can improve while accuracy drops: a parameter setting that produces statistically tight topics may exclude the range of vocabulary students actually use, and the expert rejects it despite the coherence gain. Second, accuracy can improve while topics fragment: a regularization change that raises macro-F1 may split a theme into two overlapping subtopics that require manual merging before theme assignment can proceed. Third, a parameter change can produce metrics near-identical to baseline while qualitatively improving theme separation; in these cases, the expert's reading of the top-10 keywords per topic is the deciding criterion, not the quantitative comparison.

The optimization loop therefore has two exit conditions: quantitative improvement sufficient to justify the parameter change, and expert confirmation that topic-word distributions remain interpretable and non-redundant. When these two conditions conflict, the expert's qualitative judgment takes precedence. This precedence is intentional: the purpose of the optimization stage is to produce topics that are useful for educational analysis, not to maximize any single statistical metric.

10.7.4 Evaluation

Once the optimization loops exited with both quantitative metrics and qualitative review satisfied, expert input produced measurable gains in accuracy, coherence, and reliability across all three datasets. Preprocessing, tokenization, and metric calculation were automated; expert effort was concentrated where domain reasoning altered outcomes.

Identifying Model Misalignment Through Expert Evaluation

BERTopic failed to match the expert-coded theme ranking on the TA Support dataset. The model grouped responses by statistical similarity but failed on multi-idea responses, cross-over themes, and domain-specific terminology. Correctly classifying responses about assessment scheduling, laboratory support, and peer interaction required knowledge of teaching norms that unsupervised models do not encode. Each failure mode became visible only during expert evaluation; a targeted review addressed them directly.

Targeted HITL Process for Error Correction

Figure 10.6 illustrates the four-step targeted human-in-the-loop process that was used to address model misalignment. In the TA Support dataset, BERTopic's HDBSCAN identified 232 outliers (responses assigned to the -1 cluster that the model could not place confidently), representing about 14% of the total dataset of 1,592 responses.

Expert review of these 232 outliers identified three problem areas: assessment-related comments (approximately one-third of outliers), delivery-related comments (approximately one-fourth), and laboratory support language, a domain-specific concept the model did not encode. All 232 outliers were reclassified by the expert; performance metrics were recalculated and theme ranking was verified against expert-coded labels.

Table 10.7 shows the performance improvements after expert correction. Accuracy increased from 70.3% to 80.1%, an improvement of 9.30 percentage points. F1-score increased from 54.2% to 69.3%, an improvement of 14.10 percentage points. These gains demonstrate that strategic expert input transformed the hardest cases into measurable performance improvements.

Figure 10.7 shows that after reclassification, the theme ranking aligned with expert coding, confirming that expert correction improved not only quantitative metrics but also the conceptual validity of model outputs.

HDBSCAN outliers are a clear indicator: responses the model could not place confidently are precisely the cases where statistical similarity fails to capture educational meaning. Expert review of

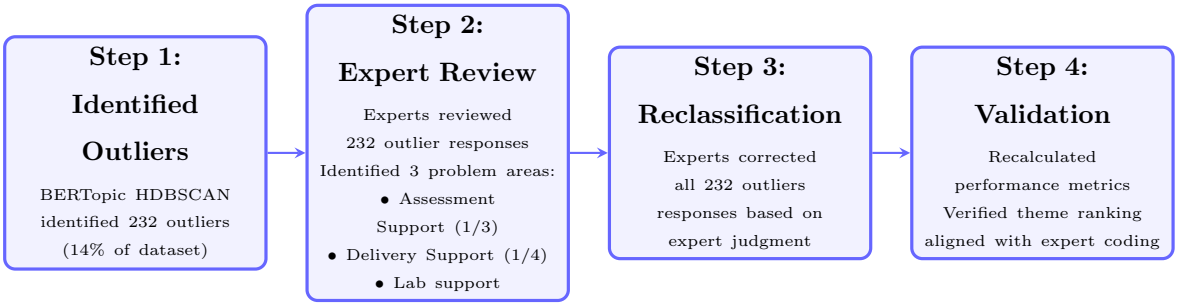


Figure 10.6: Targeted Human-in-the-Loop Process for Error Correction

Table 10.7: TA Support Dataset: Performance Metrics After HITL Outlier Correction

Metric	Without HITL BERTopic (MPNet)	With HITL BERTopic (MPNet)	Improvement
Accuracy	70.3%	80.1%	+9.30%
F1-Score	54.2%	69.3%	+14.10%

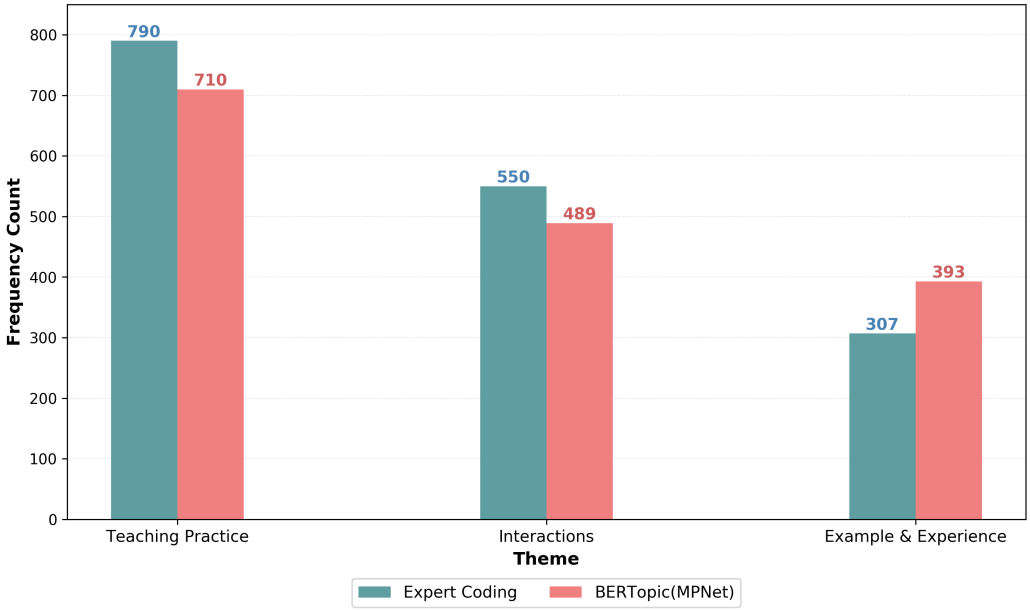


Figure 10.7: Theme Ranking Comparison: Expert Coding vs. BERTopic (MPNet) After HITL Correction

that 14% subset uncovered domain-specific patterns, including assessment-related and laboratory-support language, that unsupervised algorithms cannot resolve.

10.7.5 Next Steps: Iteration

When the five methods evaluated here achieve high accuracy and macro-F1 on expert-coded labels and theme ranking aligns with expert priorities, the analysis is complete. When performance gaps persist despite optimization, the data characteristics identified during familiarization and data preparation point toward what a more capable method would need to address.

Three patterns in the results indicate that the current methods have reached their ceiling on a given dataset. First, when ambiguity rate exceeds 20% and no method exceeds 70% macro-F1, the single-topic-per-response constraint shared by k-means and BERTopic is the main constraint, and multi-label deep learning models that allow each response to belong to multiple themes would address this directly. Second, when ZSC substantially outperforms all unsupervised methods on F1, the data structure favors expert-defined labels over statistical discovery, and fine-tuned classification models trained on the IRR-validated codes would likely improve further. Third, when embedding overlap between themes persists across both MiniLM and MPNet configurations (as observed in TA Support), domain-specific pre-training or adapter-based fine-tuning on educational text would give the embedding model exposure to the vocabulary patterns that general-purpose pre-training does not capture.

Moving to more sophisticated models is appropriate when these patterns are confirmed and when the verified ground truth is large enough to support fine-tuning (typically 500 or more labeled examples per theme). The IRR-validated codes produced through the thematic analysis steps provide exactly this kind of reference dataset, making the ground-truth work not only the foundation for the current evaluation but also the preparation for the next iteration. Recommendations for future directions are discussed in Chapter 11.

10.8 Practical Implications and Recommendations

Interpretation and validation stages produced the greatest impact; routine computational steps achieved strong results through automation alone. The findings provide clear direction for allocating expert effort.

10.8.1 Implementation Guidance

Focus expert time on interpretation and validation. Topic-to-theme bridging and modeling optimization produced the two largest performance gains in this study. Preprocessing and metric calculation are automated tasks; expert effort should concentrate where conceptual judgment, domain knowledge, and interpretive decisions alter outcomes.

Use quantitative metrics as guidance, not as final judgments. Coherence can improve while accuracy drops (LDA sparse alpha on TA Support), and accuracy can improve while topics fragment (NMF regularization on Faculty Support). Metrics identify where model outputs may need review; expert inspection of topic-word distributions is the deciding criterion when quantitative and qualitative results point in different directions.

Maintain transparent decision records. Document every expert action: merge and split decisions, rationale for parameter choices, and override justifications. The audit trail supports reproducibility and makes it possible for other researchers to apply the same methodology to new datasets.

Validate interpretations with independent coders before treating model outputs as ground truth. Inter-rater reliability (κ = 0.72–0.75 across three datasets) confirmed that the expert-defined themes were interpretable and reproducible. Without this validation step, expert-coded labels reflect a single researcher's interpretation, not a verified coding scheme.

Alternate between quantitative and qualitative review within each optimization loop. No single metric captures all relevant information. Aggregate accuracy can mask a substantial per-theme precision failure in a single method. Per-theme F1 and expert review of topic-word distributions together give a clearer picture of when to continue iterating and when to accept a configuration.

10.9 Conclusion

Distributing expert judgment across all seven stages of thematic analysis produced a verified ground truth ($\kappa = 0.72$–0.75 across three datasets) and measurable performance gains: macro-F1 improved by up to 17.7 percentage points and accuracy by up to 9.3 percentage points.

The largest gains occurred during topic-to-theme bridging and model optimization. Expert review selected optimal topic counts, adjusted label prompts and hyperparameters, and converted statistical clusters into themes aligned with educational constructs. Each decision drew on knowledge of domain vocabulary, response ambiguity rates, and educational constructs that algorithms alone could not capture. The resulting themes were empirically grounded and conceptually transparent,

which ensured that models reflected educational processes rather than word-level patterns.

Metric selection (Section 10.7.1) set consistent evaluation standards that supported comparison across methods. Quantitative metrics informed expert interpretation, and expert judgment guided parameter decisions in turn. This structure defines human-in-the-loop analysis not as constant intervention but as the selective use of expertise where it alters outcomes in measurable ways.

Prior work in human-in-the-loop educational feedback analysis typically restricts expert involvement to a single stage: either post-model validation [107, 108] or pre-model constraints [110, 111]. Interactive refinement systems that support iterative adjustment [112, 113] do not quantify performance gains on validated educational datasets. This framework differs by distributing expert judgment across all stages of the workflow and measuring the improvement each intervention produces. It also supports theme discovery rather than relying on predefined categories [115], while maintaining the validation rigor that educational applications require.

These results confirm that integrating expert judgment across all seven thematic analysis stages produces results that are both statistically sound and educationally meaningful. Documenting expert decisions and measuring outcomes at each stage provides a replicable methodology that others can apply to educational text analysis. Chapter 11 synthesizes these findings with the limitations of the current methods and outlines the conditions under which multi-label and fine-tuned models would be the appropriate next step.

Chapter 11

CONCLUSIONS

This dissertation examined how natural language processing can support qualitative educational research when student responses are short, prompt-guided, and rich with discipline-specific language. The study drew on three datasets of undergraduate engineering student survey responses about faculty support, TA support, and peer support expectations, compared five topic modeling methods across all three datasets, and evaluated how domain expert judgment, integrated throughout a structured NLP-Assisted Thematic Analysis framework, shaped model performance and interpretive quality. Two research questions guided the work. RQ1 asked which topic modeling approaches perform best on these texts, and RQ2 asked how domain expert input can be best integrated to improve accuracy and analytical value. This chapter synthesizes those findings, articulates contributions to educational research and NLP, acknowledges limitations, and proposes directions for future study.

11.1 Summary of Findings

RQ1: Comparative Performance of Topic Modeling Approaches

The cross-method evaluation demonstrated that method suitability depends on dataset characteristics rather than model sophistication. Across all three datasets, the analysis identified five cross-cutting themes (Interactions, Teaching Practice, Examples and Experience, Questioning, and Civility). All five methods were evaluated against these themes on student responses averaging five to twelve words per document.

Traditional models provided complementary baselines whose performance reflected specific dataset properties. k-means worked well when theme vocabularies were lexically distinct, as in Faculty Support, where office-hours vocabulary cleanly separated the Interactions theme from Teaching Practice, but it struggled where responses were short or addressed multiple themes simultaneously. LDA performed best on Peer Support (78% accuracy after optimization), where consistent vocabulary and strong word co-occurrence patterns compensated for its reliance on bag-of-words statistics, but degraded on TA Support because the corpus co-occurrence structure did not align with the themes the domain expert identified. NMF produced the largest single accuracy gain in the study, with Fac-

ulty Support improving from 59.1% to 76.7% after optimization because its non-negative sparsity constraint concentrated theme-specific vocabulary precisely and handled multi-topic responses that hard-assignment models cannot accommodate.

The TA Support dataset demonstrates a risk of misalignment between statistically valid topics and researcher-defined themes that practitioners should account for before applying discovery-first methods to a new corpus. k-means, LDA, NMF, and BERTopic each produced statistically coherent topics on TA Support that organized responses around availability, question-and-answer structure, and instructional delivery. These topics, however, did not correspond to the five themes the domain expert identified as educationally meaningful. Strong coherence scores accompanied this misalignment, indicating that internal metrics alone are insufficient to confirm that model output serves a study's analytical goals. Expert validation of topics against researcher-defined themes is therefore not a supplementary quality check for discovery-first methods; it is the step that determines whether statistically valid topics translate into findings that are interpretable and relevant to the research purpose.

For both traditional and neural methods, input representation choices proved as consequential as algorithm selection. For traditional methods, applying TF-IDF weighting rather than raw frequency counts produced inconsistent performance changes across datasets, indicating that the optimal vectorization strategy depends on dataset characteristics rather than method type. For BERTopic, embedding model selection was equally consequential. MiniLM dominated Peer Support (85% accuracy, 77% macro-F1), where responses were short and themes were well-separated. MPNet outperformed MiniLM on Faculty and TA Support, where richer semantic representations preserved finer thematic distinctions in longer, more ambiguous responses. These findings indicate that practitioners who compare algorithms on a single fixed preprocessing configuration may miss the best-performing combination for their data, and that evaluating vectorization and embedding choices alongside algorithm selection is necessary for a fair and informative comparison.

Zero-shot classification applied expert-defined labels without training data and reached 85% accuracy on Peer Support, matching BERTopic MiniLM on that metric. The lower macro-F1 (58% versus 77%) indicates that ZSC formed internally consistent label groups but did not separate themes as reliably as the neural method. Prompt phrasing governed ZSC performance. Mainstream labels, which used student vocabulary instead of specialist terminology, improved accuracy by 3 percentage points over domain-specific labels. This result parallels the advantage that TF-IDF weighting provides in frequency-based methods, where both approaches succeed by emphasizing the terms that discriminate most clearly between categories.

One finding applies across all five methods. Discovery-first methods (k-means, LDA, NMF, BERTopic) suit exploratory contexts where themes are unknown and the data must reveal them; ZSC suits confirmatory contexts where an established coding framework must be applied at scale. These two approaches are complementary, and the choice between them depends on whether the research goal is exploratory or confirmatory rather than on which method is technically superior.

Across all three datasets, the proportion of multi-topic responses emerged as a consistent predictor of the upper bound of achievable accuracy, independent of algorithm selection. Peer Support, where only 5.89% of responses addressed more than one theme simultaneously, supported the highest accuracy levels across all five methods. Faculty Support, where 26.5% of responses addressed multiple themes, constrained performance across all methods regardless of model complexity. TA Support, with an intermediate ambiguity rate of 18.6%, produced intermediate performance. This pattern indicates that dataset ambiguity rate predicts the performance ceiling more reliably than algorithm choice. Practitioners should therefore assess ambiguity rate before selecting a method to set realistic performance expectations and to determine whether the data, rather than the algorithm, is the primary limiting factor.

These results answer RQ1: no single approach dominates across datasets, and dataset properties predict performance more reliably than algorithm sophistication. Among those properties, ambiguity rate sets the ceiling on what any method can achieve, while thematic clarity, vocabulary overlap, document length, and the availability of expert labels determine which method is most likely to approach that ceiling.

RQ2: Roles for Human Expertise

The human-in-the-loop analysis showed that expert involvement shapes outcomes at every stage of the analysis process, not only at the end. The NLP-Assisted Thematic Analysis framework developed in Chapter 10 structures expert judgment across six stages: data preparation, corpus familiarization, candidate topic identification, theme definition, inter-rater review, and the complete analysis with expert oversight at selected points.

In the data preparation stage, domain-specific stopword curation removed prompt-anchored role words that appeared in nearly every response but conveyed no information about what students were requesting. In the theme definition stage, the topic-to-theme bridge converted nine algorithmic topics into five validated themes, a step that required domain knowledge to distinguish patterns that automated models treated as similar but that differed in pedagogical meaning. These five themes achieved strong inter-rater reliability (Cohen's κ between 0.72 and 0.75 across all three datasets),

establishing the verified ground truth used in all subsequent evaluations. The shared topic count of three per dataset was confirmed by expert review of preliminary model runs, not by internal metrics alone.

In the optimization stage, expert involvement produced the most measurable gains. BERTopic outlier correction, in which the domain expert reviewed 232 responses the model could not assign confidently, improved TA Support accuracy by 9.3 percentage points (70.3% to 80.1%) and macro-F1 by 14.1 percentage points (54.2% to 69.3%). Prompt engineering for zero-shot classification showed that expert phrasing raises weighted F1 by 4 to 10 percentage points compared to technical labels.

The framework documents these interventions through decision flows and work flows that guide practitioners in model selection and optimization. Five dataset characteristics (vocabulary diversity, ambiguity rate, mean response length, domain terminology density, and theme stability) determine which method is most likely to perform well, and the documented work flows show how to optimize the chosen method with expert oversight at each stage. These structured guides translate the findings of this study into steps that other researchers can apply directly to similar datasets.

Across both research questions, the findings form the basis for three contributions to educational research and NLP practice.

11.2 Contributions and Implications

This dissertation delivered three main contributions that collectively link NLP practice with qualitative research needs:

C1: Established performance baselines by comparing five topic modeling methods. The study compared k-means, LDA, NMF, BERTopic, and zero-shot classification on the same datasets using a shared preprocessing pipeline and common evaluation metrics. The results show that no single method works best for all datasets. Instead, the analysis identifies which method fits each dataset type, providing clear guidance for method selection based on dataset characteristics.

C2: Identified where expert input has the highest impact. The study analyzed expert input at different stages: preparation (domain-specific stopwords), prompt design (for zero-shot classification), and theme interpretation (mapping topics to themes). These interventions improved accuracy by 3 to 9 percentage points and macro-F1 by up to 14 percentage points, showing where expert time produces the most valuable results. This analysis helps researchers prioritize expert involvement to maximize both accuracy and efficiency.

C3: Developed the NLP-Assisted Thematic Analysis framework. The study created a structured, six-stage framework (Chapter 10) that integrates expert judgment throughout the

analysis process instead of reserving it for a final review. The framework includes decision flows for model selection based on dataset characteristics, work flows for running and optimizing each method with expert oversight, and documentation of preprocessing rules, theme formation rationales, and parameter selection criteria. These resources support replication and adaptation to new datasets, disciplines, and research contexts beyond the engineering education datasets analyzed here.

These contributions have three implications for educational research practice. First, analysts should select modeling approaches based on thematic clarity and data characteristics, not by defaulting to the most sophisticated algorithm, because simpler methods can outperform complex ones when data conditions favor them. Second, expert engagement is indispensable when converting model output into findings that inform pedagogy or policy, and the study identifies where to focus that engagement to achieve the highest return. Third, the NLP-Assisted Thematic Analysis framework offers a reusable starting point for researchers who want to apply topic modeling to short educational texts but lack a structured guide for integrating computational methods with domain knowledge. The shared evaluation framework, including the performance baselines and metric selection guidance, also provides a template for benchmarking future models on short educational research texts.

11.3 Limitations

While this study provides a comparative analysis of topic modeling techniques across multiple methods and datasets, it is not without limitations. Most importantly, the scope of the datasets was confined to a single institution, a specific type of educational data (student feedback), and a particular context (engineering courses). These constraints limit the generalizability of the findings to other data, settings, and contexts. To broaden the relevance of the findings, the guidelines in Chapter 10 rely on data characteristics (text length, ambiguity, vocabulary overlap) that are largely independent of the domain and context in which data are collected.

Furthermore, the datasets, while sizable (between one and two thousand responses per dataset after preprocessing), are moderate by large-scale data standards. A larger dataset would improve performance but is unlikely to change the nature of the results, including the number of themes, the meaning of themes, or the relative performance patterns among methods. Theme assignment in the default analysis was also limited to a single theme per document, which ignored the possibility that multiple themes were present. Multiple-topic documents can deflate topic coherence scores by introducing words from secondary topics into the assigned topic. These deflationary effects should be consistent across models, minimizing their impact on relative performance comparisons.

A potential source of bias is the assumption that theme assignments by domain experts represent the gold standard. This potential bias was mitigated by requiring multiple domain experts to reach the same conclusions on theme assignments, as indicated by adequate Cohen's kappa (0.72 to 0.75 across datasets). Model comparisons in this study relied on predominantly quantitative performance metrics, which may introduce bias toward methods that score well numerically but produce less interpretable themes. Future work should consider an expanded and complementary study that focuses on qualitative assessments of model performance, including evaluations of topic quality, topic diversity, and topic interpretability.

In terms of model selection, this analysis was limited to five methods and did not examine the wide range of variants developed for each. Some of these variants may be better aligned with the type of data in this study and could improve performance. However, since the goal was to develop broad guidelines that provide a starting point for practitioners, further examination of variants or advanced NLP models fell outside the scope of this study.

The ZSC analysis introduced additional constraints. Only a single model (BART-large-MNLI) was evaluated, and only on a single dataset (Peer Support) with manually crafted prompts that introduce subjectivity. Different transformer architectures, datasets, or prompt designs might yield different results. Without fine-tuning, ZSC cannot learn rare patterns, which limits its performance on minority themes such as Unproductive Disruptions (42% F1) and Preparation (37–40% F1). Whether ZSC maintains competitive accuracy on the more complex Faculty Support and TA Support datasets is an open question that warrants investigation.

Finally, UMass coherence scores were generally low across all models, reflecting the short-text sparsity that limits word co-occurrence information. Quantitative coherence alone is unreliable for very short texts, and this study's NMF results (where qualitative evaluation of keywords contradicted statistical coherence scores) illustrate that point directly. Coherence values throughout this study should be interpreted qualitatively and with awareness that they measure statistical properties of word co-occurrence rather than conceptual quality.

The NLP-Assisted Thematic Analysis framework was developed and validated using data from a single institution and a single disciplinary context. Whether the six-stage structure and the decision flows for model selection generalize to other institutions, disciplines, or survey instruments is an open question that future work should address.

11.4 Future Directions

Four open questions identified during the analysis point toward the most productive extensions of this work.

The most direct extension is to evaluate zero-shot classification on the Faculty Support and TA Support datasets. ZSC was applied only to Peer Support in this study, which had the lowest ambiguity (5.89%) and the clearest theme boundaries. Whether ZSC maintains competitive performance on Faculty Support, where 26.5% of responses address multiple themes simultaneously, or on TA Support, where word co-occurrence patterns naturally align with different topics than those the domain expert identified, is an empirical question that the current study cannot answer.

A second direction is cross-institutional and cross-disciplinary testing of the NLP-Assisted Thematic Analysis framework. The five dataset characteristics used in the decision framework (vocabulary diversity, ambiguity rate, response length, domain terminology density, and theme stability) were developed from engineering education data at one institution. Applying the framework to social science, health professions, or multilingual educational contexts would clarify which elements of the decision flows are broadly applicable and which require domain-specific adaptation.

Third, the analysis identified three conditions where the methods evaluated here have reached a performance ceiling: ambiguity rates above 20% combined with macro-F1 below 70%, cases where ZSC substantially outperforms all unsupervised methods, and persistent embedding overlap across MiniLM and MPNet. Each condition points toward a next step: multi-label deep learning models for high-ambiguity data, fine-tuned classification models where ZSC demonstrates strong superiority, and domain-specific pre-training where general embeddings fail to separate themes. Testing whether these targeted interventions resolve the performance limits observed here would advance the practical guidance the framework can offer.

Finally, longitudinal deployments that integrate the pipeline into institutional decision cycles would reveal how iterative expert feedback and model updates affect accuracy over time, and whether the inter-rater reliability thresholds established here hold as datasets and research teams change.

The dissertation demonstrates that topic modeling guided by domain expert judgment produces reliable and interpretable themes from short educational research texts when model selection and expert integration follow a structured, validated process. The NLP-Assisted Thematic Analysis framework developed here offers a practical starting point for researchers who want to apply these methods to their own data without repeating the full comparative analysis reported in this work.

Appendix A

APPENDIX A

A.1 BERTopic Discovery-Mode Outputs

This appendix documents the exploratory BERTopic runs that retained the model's native discovery mode. The primary analysis in Chapter 8 used k-means with $k = 3$ to enable direct comparison with traditional topic modeling methods. To confirm that this fixed-k design did not mask additional structure, BERTopic was re-run with HDBSCAN clustering and bigram tokenization. The three high-level themes remained intact across datasets, while discovery mode surfaced finer-grained subtopics and a small proportion of outlier responses.

A.1.1 Configuration Summary

The discovery-mode runs used the following configuration; Appendix figures and tables report the outputs from the higher-capacity MPNet checkpoints (MiniLM produced the same theme structure and is omitted for brevity).

- **Embeddings:** SentenceTransformer(`"all-mpnet-base-v2"`); `"all-MiniLM-L6-v2"` was also tested for runtime comparison.
- **Vectorizer:** CountVectorizer with `ngram_range=(1,2)`, English stop words, and `min_df=2` to retain informative bigrams.
- **UMAP:** `n_neighbors=15`, `n_components=5`, `min_dist=0.1`, cosine metric, `random_state=42`.
- **HDBSCAN:** `min_cluster_size=25`, `min_samples=5`, Euclidean metric, EOM cluster selection, `prediction_data=True`.
- **BERTopic:** `nr_topics="auto"`, `calculate_probabilities=True`, low-memory disabled.

All descriptors were lemmatized to avoid duplicate stems (for example, *lecture* versus *lectures*).

A.1.2 Hyperparameter Checks

Exploratory sweeps preceded the discovery-mode configuration. HDBSCAN min_cluster_size values between 10 and 30 and min_samples values between 3 and 10 were evaluated on the Faculty and

Peer Support datasets. Settings below 15 merged the Interactions and Examples & Experience themes, while settings above 20 increased outlier rates beyond 12%. The adopted combination of min_cluster_size=15 and min_samples=5 provided the highest UMass coherence without inflating the unassigned cluster. Vectorizer trials compared unigram and $(1, 2)$ n-gram ranges; bigrams did not raise coherence and occasionally surfaced duplicate descriptors, so the unigram configuration was retained for the fixed-k evaluation.

A.1.3 Peer Support

Discovery mode maintained the core themes (Interactions, Questioning, Civility) and separated recurring subtopics such as Zoom etiquette and study groups. An outlier cluster captured generic "ask more questions" responses that repeated across themes.

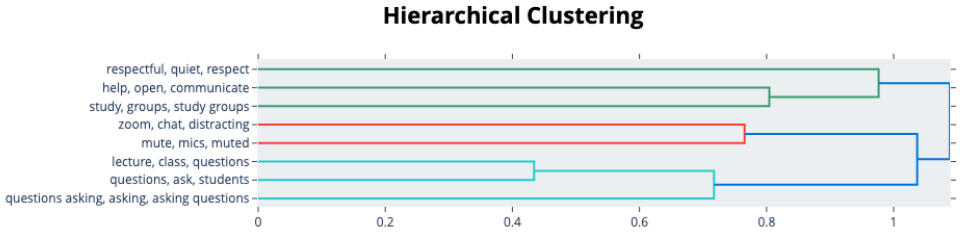


Figure A.1: Hierarchical clustering of discovery-mode BERTopic topics for Peer Support.

A.1.4 Faculty Support

The hierarchy preserved the three headline themes (Teaching Practice, Interactions, Examples & Experience) while isolating frequent requests around lecture recordings, Canvas organization, and deadline flexibility.

Table A.1: Discovery-mode BERTopic topics for Peer Support (HDBSCAN, MPNet)

Topic ID	**Count**	**Top terms**	**Representative action**
−1	566	questions, sessions, peers, ask	General "ask more questions" feedback clustered as outliers due to frequent phrasing.
0	265	questions, ask, peers, sessions, discord	Encourage peers to ask questions and use shared forums to resolve confusion.
1	264	lecture, sessions, questions, guidance, talking	Remind students to ask questions during lecture and limit side conversations.
2	87	help, open, communicate, collaborate, interact	Promote openness to helping classmates and communicating during group work.
3	60	study, groups, form, homework, help	Form and maintain study groups for collaborative problem solving.
4	44	mute, mics, muted, speakers	Mute microphones when not speaking to reduce classroom distractions.
5	34	zoom, chat, distracting, etiquette	Maintain professional behaviour in Zoom chat; avoid distracting messages.
6	31	respectful, quiet, courteous	Encourage respectful classroom conduct and quiet attention.
7	25	asking questions, clarification	Reinforce repeated requests to keep asking clarifying questions.

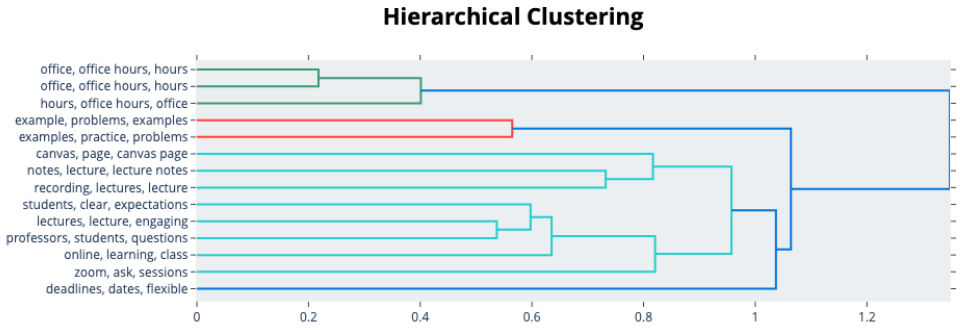


Figure A.2: Hierarchical clustering of discovery-mode BERTopic topics for Faculty Support.

A.1.5 TA Support

For TA Support, discovery mode split the broad themes into actionable subtopics, highlighting lab logistics, office-hour scheduling, quiz-section alignment, and communication expectations.

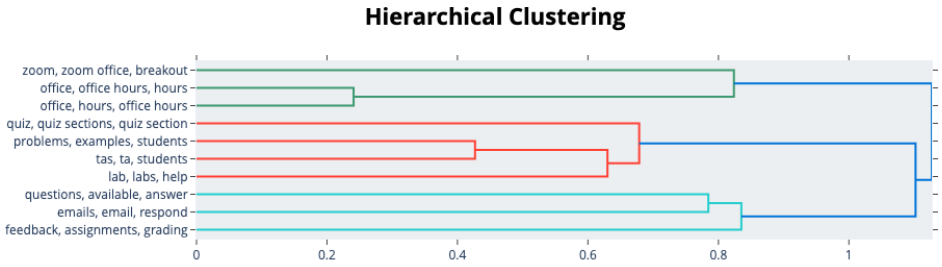


Figure A.3: Hierarchical clustering of discovery-mode BERTopic topics for TA Support.

These discovery-mode outputs corroborate the fixed-k findings in Chapter 8, providing qualitative detail that can support future exploratory work while keeping the primary evaluation framework consistent across methods.

Table A.2: Discovery-mode BERTopic topics for Faculty Support (HDBSCAN, MPNet)

Topic ID	**Count**	**Top terms**	**Representative action**
−1	721	sessions, learners, lecture, time, material	Broad lecture-quality comments: pace content appropriately and provide helpful materials.
0	186	examples, practice, problems, exams	Provide additional worked examples and practice problems before exams.
1	101	office, office hours, available, email	Offer more flexible office hours and timely email responses.
2	91	facilitators, learners, questions, support	Set clear schedules and remain accessible for student questions.
3	90	recording, lectures, upload, watch	Record and share lectures so students can review difficult material.
4	87	learners, clear, expectations, communication	Communicate expectations, grading policies, and deadlines transparently.
5	84	lectures, engaging, understand, material	Make lectures engaging and ensure concept mastery before moving on.
6	72	office hours, hour, extra	Add extra or smaller office-hour sessions around exams.
7	55	online, learning, sessions, remote	Adjust workload and support for remote learning conditions.
8	37	hours, provide, resources, available	Supply plentiful study resources alongside ample office hours.
9	35	canvas, page, assignments, announcements	Keep the Canvas site organized with clear assignment postings.
10	30	deadlines, dates, flexible, assignments	Provide flexibility with deadlines and communicate timing clearly.
11	28	example, problems, real world	Tie course concepts to real-world examples to motivate learning.
12	25	notes, lecture notes, detailed	Publish detailed lecture notes to supplement in-class explanations.
13	25	zoom, ask, sessions, questions	Use Zoom sessions for interactive Q&A when students are confused.

Table A.3: Discovery-mode BERTopic topics for TA Support (HDBSCAN, MPNet)

Topic ID	**Count**	**Top terms**	**Representative action**
−1	232	homework, learners, problems, questions	General homework-help requests that fell outside specific clusters.
0	369	mentors, learners, sessions, helpful, office	Make TAs accessible and proactive in supporting current coursework.
1	309	problems, examples, lecture, practice	Work through practice problems drawn from lecture and upcoming assessments.
2	181	lab, section, help, time	Provide clearer lab guidance and restructured lab sections.
3	167	office, office hours, holding	Expand office-hour availability and vary meeting formats.
4	77	quiz, sections, material, exams	Align quiz sections with lecture coverage and record sessions.
5	74	questions, available, answer, possible	Remain available to answer individual questions promptly.
6	57	feedback, assignments, grading, canvas	Deliver faster feedback and clearer grading commentary.
7	56	office hours, scheduling, learners	Offer flexible office-hour scheduling for students with conflicts.
8	40	emails, respond, timely	Respond to emails quickly with actionable guidance.
9	30	zoom, breakout, office hours	Host Zoom-based office hours and breakout rooms for small-group help.

BIBLIOGRAPHY

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2009.

[2] E.D. Liddy. Natural language processing. In *Encyclopedia of Library and Information Science*. Marcel Dekker, Inc., NY, 2nd edition, 2001.

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[7] Educational Testing Service (ETS). About e-rater, 2022.

[8] Ahmed Shehab, Mohamed Elhoseny, and Aboul Ella Hassanien. A hybrid scheme for automated essay grading based on lvq and nlp. In *2016 12th International Computer Engineering Conference (ICENCO)*, pages 201–206, Cairo, Egypt, 2016. IEEE.

[9] Laura K. Allen and Danielle S. McNamara. You are your words: Modeling students' vocabulary knowledge with nlp tools. In *Proceedings of the 8th International Conference on Educational Data Mining*, pages 298–305, Madrid, Spain, 2015. International Educational Data Mining Society.

[10] V. Melissa Holland and Jonathan D. Kaplan. Nlp techniques in computer-assisted language learning. *Instructional Science*, 23(5–6):351–380, 1995.

[11] Soma Ghosh. Online automated essay grading as a web-based learning tool in engineering education. In *Web-based Engineering Education*, pages 1–15. IGI Global, Hershey, PA, USA, 2010.

[12] Joseph Wilson, Benjamin Pollard, John M. Aiken, Marcos D. Caballero, and H. J. Lewandowski. Classification of open-ended responses to a research-based assessment using natural language processing. *Phys. Rev. Phys. Educ. Res.*, 18:010141, Jun 2022.

[13] Matthew A. Verleger. Using nlp to classify responses to open-ended engineering problems. In *2014 ASEE Annual Conference & Exposition*, pages 24.1–24.15, Indianapolis, IN, USA, 2014. American Society for Engineering Education.

[14] C.L. Dym, Alice Agogino, Ozgur Eris, Daniel Frey, and Larry Leifer. Engineering design thinking, teaching, and learning. *IEEE Engineering Management Review*, 34(1):65–65, Sep 2006.

[15] Scott A. Crossley, David R. Russell, Kristopher Kyle, and Ute Römer. Applying nlp to student writing across science and engineering. *Journal of Writing Analytics*, 1(1):48–81, 2017.

[16] Sunghoon Lim, Conrad Tucker, Kathryn Jablokow, and Barton Pursel. Quantifying the mismatch between course content and students' dialogue in online learning environments. In *ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, page V003T04A016, Cleveland, OH, USA, Aug 2017. American Society of Mechanical Engineers.

[17] Caitlin Robinson, Michael Yeomans, Justin Reich, Chris Hulleman, and Hunter Gehlbach. Forecasting student achievement in moocs with nlp. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 383–387, Edinburgh, United Kingdom, 2016. Association for Computing Machinery.

[18] Scott Crossley, Jaclyn Ocumpaugh, Matthew Labrum, Franklin Bradfield, Mihai Dascalu, and Ryan S Baker. Modeling math identity and math success through sentiment analysis and linguistic features. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*, pages 11–20, Buffalo, NY, USA, 2018.

[19] Damji Heo Stratton, Saira Anwar, and Muhsin Menekse. How do engineering students' achievement goals relate to their reflection behaviors and learning outcomes? In *2017 ASEE Annual Conference & Exposition*, 2017.

[20] Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S McNamara, and Ryan S Baker. Combining click-stream data with nlp tools to better understand mooc completion. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 6–14, 2016.

[21] Michelle Soledad, Jacob Grohs, Sreyoshi Bhaduri, Jennifer Doggett, Jaime Williams, and Steven Culver. Leveraging institutional data to understand student perceptions of teaching in large engineering classes. In *2017 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE, 2017.

[22] Venugopal Dhanalakshmi, Dhivya Bino, and Abinaya M Saravanan. Opinion mining from student feedback data using supervised learning algorithms. In *2016 3rd MEC international conference on big data and smart city (ICBDSC)*, pages 1–5. IEEE, 2016.

[23] Ashwin Satyanarayana, Karen Goodlad, Jennifer Sears, Philip Kreniske, Mery F Diaz, and Sandra Cheng. Using natural language processing tools on individual stories from first-year students to summarize emotions, sentiments, and concerns of transition from high school to college. In *2019 ASEE Annual Conference & Exposition*, 2019.

[24] Andrew Katz, Matthew Norris, Abdulrahman M Alsharif, Michelle D Klopfer, David B Knight, and Jacob R Grohs. Using natural language processing to facilitate student feedback analysis. In *2021 ASEE virtual annual conference content access*, 2021.

[25] Diego Buenano-Fernandez, Mario González, David Gil, and Sergio Luján-Mora. Text mining of open-ended questions in self-assessment of university teachers: An lda topic modeling approach. *Ieee Access*, 8:35318–35330, 2020.

[26] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1427–1445, 2022.

[27] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[28] Steven Tenny, Janelle M. Brannan, and Grace D. Brannan. Qualitative study, 2022. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.

[29] Michael Quinn Patton. *Qualitative research & evaluation methods: Integrating theory and practice*. SAGE Publications, Thousand Oaks, CA, 4th edition, 2015.

[30] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[31] Mike Thelwall and Tamara Nevill. Is research with qualitative data more prevalent and impactful now? interviews, case studies, focus groups and ethnographies. *Library & Information Science Research*, 43(2):101094, 2021.

[32] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[33] Paul DiMaggio, Manish Nag, and David Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606, 2013.

[34] Glenn G. Smith, Robert Haworth, and Slavko Žitnik. Computer science meets education: Natural language processing for automatic grading of open-ended questions in eBooks. *Journal of Educational Computing Research*, 58(7):1227–1255, 2020.

[35] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping Multidimensional Data*, pages 25–71. Springer, 2006.

[36] Stephan A. Curiskis, Barry Drake, Thomas R. Osborn, and Paul J. Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*, 57(2):102034, 2020.

[37] Juan Antonio Lossio-Ventura, Sergio Gonzales, Jorge Morzan, Hugo Alatrista-Salas, Tina Hernandez-Boussard, and Jiang Bian. Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial Intelligence in Medicine*, 117:102096, 2021.

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

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, volume 26. Curran Associates, Inc., 2013.

[40] Roman Egger and Joanne Yu. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7:886498, 2022.

[41] Gaurav Nanda, Kerrie A. Douglas, Danielle R. Waller, Hanna E. Merzdorf, and Dan Goldwasser. Analyzing large collections of open-ended feedback from MOOC learners using LDA topic modeling and qualitative analysis. *IEEE Transactions on Learning Technologies*, 14(2):146–160, 2021.

[42] Jerry Sun and Liang Yan. Using topic modeling to understand comments in student evaluations of teaching. *Discover Education*, 2(1):25, 2023.

[43] Dan Shi, Jiani Zhou, Fan Wu, Dui Wang, Di Yang, and Qi Pan. Characteristics of students' learning behavior preferences—an analysis of self-commentary data based on the LDA model. *Journal of Intelligent & Fuzzy Systems*, 46(2):4495–4509, 2024.

[44] Hollylynne A. Barker, Hye-Sun Lee, Shaun Kellogg, and Rob Anderson. The viability of topic modeling to identify participant motivations for enrolling in online professional development. *Online Learning Journal*, 28(1):3571, 2024.

[45] Thomas K Landauer, Danielle S McNamara, Simon Dennis, and Walter Kintsch. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, NJ, 2007.

[46] Nicolas Gillis. Nonnegative matrix factorization: Complexity, algorithms and applications. *Foundations and Trends in Machine Learning*, 13(2-3):203–365, 2020.

[47] Noelle LaVoie, Jacqueline Parker, Peter J. Legree, Steven Ardison, and Robert N. Kilcullen. Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement*, 80(2):399–414, 2020.

[48] Mingqing Zhang, Shudong Hao, Yanyan Xu, Dengfeng Ke, and Hengli Peng. Automated essay scoring using incremental latent semantic analysis. *Journal of Software*, 9(2):429–436, 2014.

[49] Eva Seifried, Wolfgang Lenhard, Harald Baier, and Birgit Spinath. On the reliability and validity of human and LSA-based evaluations of complex student-authored texts. *Journal of Educational Computing Research*, 47(1):67–92, 2012.

[50] Jad Hoblos. Experimenting with latent semantic analysis and latent Dirichlet allocation on automated essay grading. In *Proceedings of the 7th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–7, 2020.

[51] Elisabeth Forster and Kevin Dunbar. Creativity evaluation through latent semantic analysis. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2728–2733, 2009.

[52] Peter W Foltz, Darrell Laham, and Thomas K Landauer. Automated essay scoring: applications to educational technology. In *Proceedings of ED-MEDIA 99—World Conference on Educational Multimedia, Hypermedia & Telecommunications*, pages 939–944, Seattle, WA, USA, June 19–24 1999. Association for the Advancement of Computing in Education (AACE).

[53] Peter W. Foltz, Lynn A. Streeter, Karen E. Lochbaum, and Thomas K. Landauer. Implementation and applications of the Intelligent Essay Assessor. In Mark D. Shermis and Jill Burstein, editors, *Handbook of Automated Essay Evaluation*, pages 68–88. Routledge, 2013.

[54] Rola Khamisy-Farah, Raymond Farah, Haneen Jabaly-Habib, Yara Nakhleh Francis, and Nicola Luigi Bragazzi. Exploring gender perspectives in medical education: Latent semantic analysis of Israeli first-year medical students' reflections. *JMIR Medical Education*, 11(1):e78371, 2025.

[55] Riki Kurniawan and Zikri Indra. Analyzing student perspectives on learning experience using latent semantic indexing algorithm. In *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, pages 287–291, 2023.

[56] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct 1999.

[57] Yulong Chen, Hongzhi Zhang, Rui Liu, Zhaonian Ye, and Jianliang Lin. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, 163:1–13, 2019.

[58] Olusola Babalola, Bolanle Ojokoh, and Olutayo Boyinbode. Comprehensive evaluation of LDA, NMF, and BERTopic's performance on news headline topic modeling. *Journal of Computing Theories and Applications*, 2(2):268–289, 2024.

[59] Sanjo George and Sathiya Vasudevan. Comparison of LDA and NMF topic modeling techniques for restaurant reviews. *Indian Journal of Natural Sciences*, 10(62):28210–28216, 2020.

[60] Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, 2015.

[61] Fakhar Imam Hayat, Sara Shatnawi, and Edward Haig. Comparative analysis of topic modelling approaches on student feedback. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024)*, pages 226–233, 2024.

[62] Daniel Godfrey, Caley Johns, Carl Meyer, Shaina Race, and Carol Sadek. A case study in text mining: Interpreting Twitter data from world cup tweets, 2014.

[63] Agnes Nanyonga, Henry Wasswa, and Graham Wild. Topic modeling analysis of aviation accident reports: A comparative study between LDA and NMF models. In *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pages 1–2, 2023.

[64] Prerna Suri and Namita R. Roy. Comparison between LDA & NMF for event-detection from large text stream data. In *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, pages 1–5, 2017.

[65] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[66] Nayera Khodeir and Fatma Elghannam. Efficient topic identification for urgent MOOC forum posts using BERTopic and traditional topic modeling techniques. *Education and Information Technologies*, 30(5):5501–5527, 2024.

[67] Dragan Milošević, Bojana Bašaragin, Adela Ljajić, and Nikola Milosevic. Multilingual transformer and BERTopic for short text topic modeling: The case of Serbian. In *Disruptive Information Technologies for a Smart Society. ICIST 2023. Lecture Notes in Networks and Systems*, volume 872, pages 161–173. Springer, 2024.

[68] Yue Feng, Yuxuan Wu, Yilin Wang, Jiaxin Zhang, Yuxin Zhao, Ke Yang, et al. BERTopic_teen: A multi-module optimization approach for short text topic modeling in adolescent health. *Frontiers in Public Health*, 13:1608241, 2025.

[69] R. J. Julanta Leela, A. Bhuvaneswari, and M. Kumudha. Topic modeling based clustering of disaster tweets using BERTopic. In *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)*, pages 1–6, 2024.

[70] Amith Udupa, K. N. Adarsh, A. Aravinda, Nirmala H. Godihal, and N. Kayarvizhy. An exploratory analysis of GSDMM and BERTopic on short text topic modelling. In *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*, pages 1–9, 2022.

[71] Reham Albalawi, Tet Hin Yeap, and Morad Benyoucef. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3:42, 2020.

[72] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923. Association for Computational Linguistics, 2019.

[73] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics, 2020.

[74] MJ Parker, C Anderson, C Stone, and Y Oh. A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education*, 35:444–481, 2024.

[75] A Katz, S Wei, G Nanda, C Brinton, and M Ohland. Exploring the efficacy of ChatGPT in analyzing student teamwork feedback with an existing taxonomy, 2023.

[76] Kathryn A. Fuller, Kathryn A. Morbitzer, Jacqueline M. Zeeman, Adam M. Persky, Amanda C. Savage, and Jacqueline E. McLaughlin. Exploring the use of ChatGPT to analyze student course evaluation comments. *BMC Medical Education*, 24:423, 2024.

[77] Thanh Nam Doan and Tuan-Anh Hoang. Benchmarking neural topic models: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4363–4368, 2021.

[78] Basheer Abdullah Hezam Murshed, Hashem Daowd Esmail Al-ariki, and Suresha Mallappa. Semantic analysis techniques using Twitter datasets on big data: Comparative analysis study. *Computer Systems Science and Engineering*, 35(6):495–512, 2020.

[79] Aarti Goyal and Indu Kashyap. Comprehensive analysis of topic models for short and long text data. *International Journal of Advanced Computer Science and Applications*, 14(12), 2023.

[80] R. Muthusami, N. Mani Kandan, K. Saritha, B. Narenthiran, N. Nagaprasad, and K. Ramaswamy. Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. *Scientific Reports*, 14(1), 2024.

[81] Toshiki Doi, Masaru Isonuma, and Hitomi Yanaka. Topic modeling for short texts with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 21–33, 2024.

[82] Joseph C Sheils, David A Dampier, and Hassan Malik. A comparative study of topic models for student evaluations. In *Proceedings of the ASEE Annual Conference & Exposition*, 2024.

[83] Stephen R Porter, Michael E Whitcomb, and William H Weitzer. Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004(121):63–73, 2004.

[84] Matthew J. Denny and Arthur Spirling. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189, 2018.

[85] Barr, Richard. Pyspellchecker. Python Package Index (PyPI), 2023. Version 0.7.0 (as of 2023).

[86] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[87] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[88] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[89] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[90] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[91] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[92] S Lloyd. Least squares quantization in pcm. *IEEE Trans. on Information Theory*, 28(2):129–137, 1982.

[93] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556–562. MIT Press, 2001.

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

[95] Nils Reimers and Iryna Gurevych. Sentence-bert: sentence embeddings using siamese bert networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics, 2019.

[96] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[97] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[98] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[99] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[100] Tadeusz Calinski and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[101] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 1(2):224–227, 1979.

[102] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.

[103] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, New York, 1999.

[104] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, Boston, 2006.

[105] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[106] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[107] Maija Hujala, Antti Knutas, Timo Hynninen, and Heli Arminen. Improving the quality of teaching by utilising written student feedback: A streamlined process. *Computers & Education*, 157:103965, 2020.

[108] Abbirah Ahmed, Martin J. Hayes, and Arash Joorabchi. Assessing student engagement: A machine learning approach to qualitative analysis of institutional effectiveness. *Future Internet*, 17(10):453, 2025.

[109] Shubham Pyasi, Swapna Gottipati, and Venky Shankararaman. Sufat: An analytics tool for gaining insights from student feedback comments. In *Proceedings of the 26th International Conference on Computers in Education*, pages 365–370. Asia-Pacific Society for Computers in Education, 2018.

[110] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé III, 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*, pages 28–33, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[111] Eswari Santhanam, Bernadette Lynch, and Peter Jones. Making sense of student feedback using text analysis – adapting and expanding a common lexicon. *Quality Assurance in Education*, 26(1):60–69, 2018.

[112] Enamul Hoque and Giuseppe Carenini. ConVisIT: Interactive topic modeling for exploring asynchronous online conversations. *ACM Transactions on Interactive Intelligent Systems*, 6(1):7:1–7:24, 2016.

[113] Alison Smith-Renner, Varun Kumar, Jordan L. Boyd-Graber, Kevin Seppi, and Leah Findlater. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pages 293–304. ACM, 2018.

[114] Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and Rob Procter. A user-centered, interactive, human-in-the-loop topic modelling system. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 505–522. Association for Computational Linguistics, 2023.

[115] Jenny McDonald, Adon Christian Michael Moskal, Allen Goodchild, Sarah Stein, and Stuart Terry. Advancing text-analysis to tap into the student voice: a proof-of-concept study. *Assessment & Evaluation in Higher Education*, 45(1):154–164, 2020.