Universiteit Leiden
The Netherlands

Bachelor in Data Science and Artificial Intelligence

Topic Modelling Applied to Storytelling Data for Better Patient Care

Teodor-Călin Ionescu

Supervisors:
Dr. Lifeng Han & Prof. Suzan Verberne

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

July 10, 2025

Abstract

This study investigates the use of neural topic modeling to uncover meaningful themes from patient storytelling data, with the goal of offering insights that could contribute to more patient-oriented healthcare practices. BERTopic and Top2Vec are initially compared, for the purpose of individual interview summarization, by using similar preprocessing, chunking, and clustering configurations to ensure a fair baseline. Their outputs for a single interview (I0) are then rated through a small-scale human evaluation, focusing on coherence, clarity, and relevance. Based on the preliminary results and evaluation, BERTopic shows stronger performance and is selected for further experimentation using three clinically oriented embedding models. Results show that domain-specific embeddings improved topic precision and interpretability, with BioClinicalBERT producing the most consistent results across transcripts. The global analysis of the full dataset of 13 interviews, using the BioClinicalBERT embedding model, reveals the most dominant topics throughout all 13 interviews, namely "Medication Management and Symptom Relief in Cancer Care" and "Coordination and Communication in Cancer Care Management", with the use of two different metrics: topic prevalence and approximate distribution. Although the interviews are machine translations from Dutch to English, and clinical professionals are not involved in this evaluation, the findings suggest that neural topic modeling, particularly BERTopic, can help provide useful feedback to clinicians from patient interviews. This pipeline could support more efficient document navigation and strengthen the role of patients' voices in healthcare workflows.

Contents

1 Introduction

Cancer is one of the most challenging global health issues, affecting not only individuals but also entire families and communities. People are subjected to intense physical and mental difficulties to the point where their quality of life changes forever, even after recovering from the disease. In modern healthcare, understanding patient experiences is crucial for improving treatment and care. While clinical research traditionally relies on structured medical data, patient feedback, such as storytelling data, provides valuable insights that should not be overlooked. Healthcare should not only focus on treating the disease, but also on the emotional and psychological needs of patients, recognizing them as individuals in need of comfort and support, rather than just subjects in a medical process.

Natural Language Processing (NLP) is a powerful approach to analyzing patient narratives, enabling the extraction of meaningful topics from a large volume of text. In this research, I aim to use NLP techniques in order to extract relevant topics from cancer patient storytelling data and analyze them systematically in order to see what kind of valuable insights we can offer to healthcare providers in order to make the patients' cancer treatment journeys more bearable. This could give professionals a deeper understanding of patient needs, enabling a more patient-oriented care strategy. I compare two topic modeling algorithms, BERTopic [Gro24] and Top2Vec [Ang24], to determine which one performs better at extracting relevant topics from patient storytelling data. By optimizing these models, we aim to determine which topic modeling approach yields more interpretable and coherent topics within the context of cancer care. Ultimately, this work lays the foundation for a potential feedback tool that allows clinicians to automatically analyze patient files, scanning through large bodies of text easily and focusing on key themes and concerns patients raise.

I proposed two research questions for this paper:

1. **What key topics can current neural topic modeling models extract from patient storytelling data?**
2. **Based on the extracted topics, what feedback can we offer to current healthcare frameworks or procedures to improve patient care?**

This study aims to connect patient feedback with clinical decision-making by addressing these questions. Utilizing NLP, I aim to reveal patients' struggles, emotional states, and concerns. The following sections discuss the background and theory aspects of the thesis, the methodology used to extract and analyze the patient storytelling data, the performance comparison between our chosen techniques, and present our results that could inform improvements in patient care practices.

Lastly, the dataset used in this study is provided by Erasmus Medical Center (Erasmus MC) in Rotterdam, Netherlands [Era25]. It originates from the Metro Mapping Project, which is a design-driven initiative with the goal of supporting cancer patients in navigating their care journey and preparing for substantial medical decisions [GMSS17, TU 21]. By analyzing the cancer patient storytelling data, we aim to contribute to that mission by offering topic modeling insights that reflect the patients' lived experiences and can be used to further improve patient-centered healthcare practices.

2 Background and Related Work

This section explores three key topics that constitute the theoretical foundation of the proposed research. It examines prior work in three critical fields: NLP, Clinical NLP, and Topic Modeling by highlighting prior and potential future developments, existing challenges, and the research gap that motivates the study. These areas are essential to understanding the evolution and application of machine learning techniques in general fields such as journalism, documentation, and healthcare.

This review explores existing literature on these topics, focusing on the accomplishments and challenges that continue to shape their use in healthcare and other fields of use. I highlight where gaps still exist and explore how the proposed study aims to address these issues. The combination of these techniques offers a promising opportunity to enhance the clinical understanding of cancer progression and patient experiences, ultimately aiming to improve diagnosis, prognosis, and care delivery.

Lastly, this section also introduces and discusses two modern topic modeling approaches: **BERTopic** [Gro24] and **Top2Vec** [Ang24], which form the methodological backbone of this research. By exploring their underlying mechanisms and previous applications not only in the clinical domain but also in general, I aim to establish why these approaches are particularly well-suited for uncovering and analyzing topics and themes in patient narratives while also demonstrating their versatility in broader text analysis tasks beyond a clinical context.

2.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is an artificial intelligence field focused on enabling machines to understand, generate, and interpret human language. The roots of NLP could be traced back to the 1950s, with Alan Turing's proposal of the Turing Test [TUR50]. Early NLP systems were rule-based and deterministic. For example, one of the earliest and popular NLP systems was ELIZA, developed by Joseph Weizenbaum in the mid-1960s [Wei66]. ELIZA simulated a psychotherapist using pattern matching and rules to mimic conversation, giving the illusion of a real human understanding of language. Despite its simplicity, ELIZA highlighted both the potential and limitations of early rule-based NLP approaches. For example, the model struggled with ambiguity and context, as well as semantic understanding. These early systems were also heavily constrained by the technological limitations of that time, such as computational power.

While substantial progress was made throughout the years, including in the 1990s with the adoption of statistical and probabilistic models such as n-gram language models with Hidden Markov Models [JM25], these approaches were still limited in their ability to capture deeper semantics and context. The true transformation of NLP as we know it today came in the 2010s with the emergence of neural network-based techniques, particularly word embeddings and transformer architectures, which enabled systems to better grasp meaning, context, and nuance in human language. Models such as Word2Vec [MCCD13] made it possible to represent words in continuous vector spaces, allowing machines to reason about semantic relationships. This shift culminated in the development of transformer-based models, such as BERT [DCLT18], which substantially advanced the field of NLP by introducing deep bidirectional representations, which allow for contextual understanding

of text in a wide range of NLP tasks.

Nowadays, NLP is a critical component of applications ranging from virtual assistants and search engines to automated document summarization. Its ability to process and structure vast amounts of unstructured text makes it particularly viable in domains such as journalism, law, education, and healthcare. By transforming narratives into quantifiable insights, NLP serves as a bridge between human language and machine understanding.

2.2 Topic Modeling

Topic modeling is an unsupervised technique used for automatically identifying themes or topics in a large collection of text documents, and it is the main foundation of this research. One of the "traditional" topic modeling techniques is Latent Dirichlet Allocation (LDA) [BNJ03], which is a Bayesian algorithm for extracting topics from large bodies of text. It can be used for a variety of different tasks related to the observation or extraction of topics, such as the work by Griciūtė et al. [GHN23] which analyses the evolution of topics during the COVID-19 pandemic over a limited period of time in Swedish newspaper articles. As highlighted in the article, however, LDA has its own set of shortcomings, such as limitations in capturing deeper semantic relationships and the irrelevant topic clusters due to LDA's sensitivity to common words and connectors, such as "the" or "or", which can only be avoided by expanding the stop word list. For example, [AYB20] highlights important limitations that LDA suffers from, specifically on short-text data, such as the model's algorithm being non-deterministic, meaning that every single run of the model will result in different outputs, requiring a predefined number of topics (which involves manually determining the right amount of topics needed for each document in order to avoid topic overlap or unnecessarily general topics), issues with data sparsity, and its lack of ability to model relations between topics.

Recent advancements in NLP have led to the development of new topic modeling techniques that overcome many of the limitations of traditional probabilistic models, such as LDA. Among these, **BERTopic** [Gro22] and **Top2Vec** [Ang20] have gained attention due to their ability to generate more coherent and semantically meaningful topics by utilizing dense vector representations of text, referred to as embeddings.

Both BERTopic and Top2Vec move beyond the bag-of-words approach by incorporating modern embedding models, dimensionality reduction, and clustering techniques to extract topics from a corpus. These models do not need intensive preprocessing of the text data, can automatically determine the number of topics without the need for human intervention, and are generally better suited for short or contextually rich texts, such as interviews, where LDA often underperforms. These methods use high-dimensional semantic representations that allow them to capture more subtle relationships between the topic and the document.

Empirical studies have shown that embedding-based models often outperform classical methods in topic coherence and interpretability. For example, Egger and Yu [EY22] compared LDA, BERTopic and Top2Vec on a corpus of COVID-19 Twitter posts data and found that BERTopic produced the most distinct and interpretable topics, while Top2Vec, while producing some overlapping topics, it still managed to outperform LDA and produce unique topics which were not

identified by the other models. Similarly, Wahbeh et al. [WAREg+25] made a comparative study of Top2Vec, BERTopic, and LDA on two datasets, one of short texts and one of longer texts, with performance metrics such as coherence and usability. It was reported that BERTopic achieved the highest coherence scores among all the methods, and the topics produced were the best-defined topics in terms of coherence and human interpretability, while Top2Vec also performed strongly, but ranked slightly below BERTopic. This finding is consistent with other research that sees both Top2Vec and BERTopic outperforming older models, with BERTopic sometimes ranking as the best-performing method.

2.3 Clinical NLP

Clinical Natural Language Processing (NLP) uses NLP techniques, such as topic modeling, to extract and analyze medical text, such as unstructured health data, discharge summaries, patient files, etc. Many possibilities of integrating NLP techniques in the healthcare system have been explored in recent times, particularly in clinical decision support. The methodical review [DFCM09] documents the use of the Linguistic Inquiry and Word Count (LIWC) tool for analyzing the personality of patients through linguistic style, which would be used for various applications, such as predicting the adjustment to cancer, mental and physical improvements after the death of a close one, differentiating between suicidal and non-suicidal patients, etc. These applications emphasize how linguistic feature analysis can contribute to diagnosis and prognosis assessment in healthcare. The paper's focus on NLP for decision support makes it a solid theoretical foundation for our research, as it emphasizes the role of linguistic modeling in improving clinical evaluations and patient outcomes. Clinical NLP, however, has its own set of challenges, some of which are highlighted in [SMD+19], which reviews methods of using NLP for processing clinical notes for various chronic diseases.

Recent advancements in the field of clinical NLP show a shift of focus towards patient narrative data, akin to this study. For example, Yukiko Ohno et al. [OAN+25] published a study on the development of a high-performance NLP tool for monitoring symptoms from patient interviews obtained from a Japanese university hospital. The paper presents the training process of a BERT-based model on patient narrative data, which contains annotated diseases and symptoms. The study proved to be a success, with their newly developed system managing to surpass its predecessors in terms of performance. Some limitations of the work include data scarcity and uncertainty due to the model being trained on data originating from a single clinical facility.

Likewise, another study focused on narrative data, this time centered around clinician speech rather than patient speech, has been conducted by Yaniv Alon et al. [ANL+25] in order to study how Hebrew-speaking clinicians make decisions in real clinical scenarios. Word frequency analysis was applied in order to identify dominant concepts in clinician language, after which Large Language Models (LLMs) were called in order to deduce potential cognitive paradigms. Results found that clinicians rely on experience-based heuristics and intuitive reasoning. Some limitations of the research include a small data pool and limited generalizability due to the lack of a multilingual framework (only Hebrew).

The literature reviewed so far shows the promising potential of Clinical NLP to support and

transform healthcare by not only extracting meaningful insights from unstructured patient data, but also from clinician-centered data. However, a lot of the existing approaches still face important limitations. Current classification and information extraction methods often lack depth and semantic understanding, likely due to limited access to high-quality, annotated clinical datasets and the challenges associated with real-world healthcare data, such as privacy concerns, linguistic variability, and contextual dependence. To address some of these issues, this study aims to conduct a more in-depth exploration of neural topic modeling techniques, assessing their ability to extract coherent, patient-centered topics from narrative data. In addition, I make efforts to share anonymized datasets and trained models where possible, in order to contribute to the reproducibility and practical applicability of Clinical NLP research. Finally, this work also briefly explores the possibility for multilingual topic modeling through the potential use of specialized embedding models and translation tools, as the introduction of linguistically inclusive NLP tools that can operate effectively across a diverse linguistic healthcare setting is an important step towards the evolution of the field of Clinical NLP.

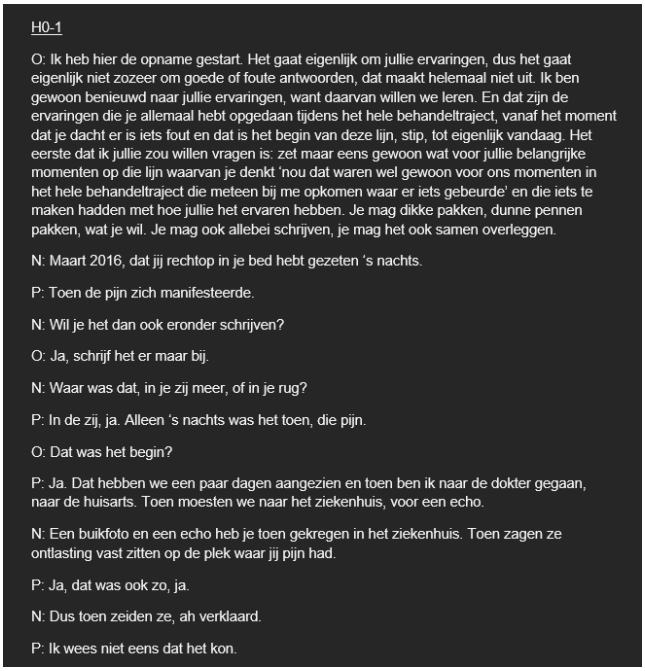
3 The Dataset - Patient Storytelling

The data we use consists of 13 anonymized `.docx` files, each corresponding to a different cancer patient. As mentioned in the introduction section, the dataset is provided by the Erasmus MC [Era25], and it originates from the Metro Mapping Project [TU 21]. They contain multiple interviews in which the patients discuss their experiences with the disease, such as how they were diagnosed, their emotional struggles, coping mechanisms during treatment, and other personal reflections. The length of each document varies, with the shortest document containing 5,596 words (approximately 34,000 characters), and the longest document containing 12,875 words (approximately 58,000 characters). Counting the words for every single interview, from I0 through I12, brings the total collection size in number of words to 132,772 words.

Each document features three different speakers: the patient, the researcher who conducts the interview, and the "naaste", a Dutch word which may be roughly translated in English to "loved one" or "close relation", whose role during the interviews is to offer an outside perspective on the patient's cancer journey. Each line within the document is marked with a capital letter which represents who is speaking: P for the patient, N for the "naaste", and O for the interviewer.

Moreover, all texts preserve marks of orality, such as hesitations, repetitions, and informal speech. While this format reflects the emotional nature of the interviews and provides a better insight into the patient's experiences, it also introduces potential complications in the preprocessing and analysis stages, which are thoroughly discussed and analyzed in the upcoming dedicated section.

Another challenge lies in the language of the documents: all interviews are fully conducted in Dutch. This presents a potential barrier, as results either require manual verification by Dutch-speaking individuals or complete translations into English, which is a complex task due to the nuanced and informal nature of the original language. Additionally, many pre-trained embedding models tend to perform best on English-language data, which may impact the quality and consistency of the results when working with Dutch transcripts.



H0-1

O: Ik heb hier de opname gestart. Het gaat eigenlijk om jullie ervaringen, dus het gaat eigenlijk niet zozeer om goede of foute antwoorden, dat maakt helemaal niet uit. Ik ben gewoon benieuwd naar jullie ervaringen, want daarvan willen we leren. En dat zijn de ervaringen die je allemaal hebt opgedaan tijdens het hele behandeltraject, vanaf het moment dat je dacht er is iets fout en dat is het begin van deze lijn, stip, tot eigenlijk vandaag. Het eerste dat ik jullie zou willen vragen is: zet maar eens gewoon wat voor jullie belangrijke momenten op die lijn waarvan je denkt 'nou dat waren wel gewoon voor ons momenten in het hele behandeltraject die meteen bij me opkomen waar er iets gebeurde' en die iets te maken hadden met hoe jullie het ervaren hebben. Je mag dikke pakken, dunne pennen pakken, wat je wil. Je mag ook allebei schrijven, je mag het ook samen overleggen.

N: Maart 2016, dat jij rechtop in je bed hebt gezeten 's nachts.

P: Toen de pijn zich manifesteerde.

N: Wil je het dan ook eronder schrijven?

O: Ja, schrijf het er maar bij.

N: Waar was dat, in je zij meer, of in je rug?

P: In de zij, ja. Alleen 's nachts was het toen, die pijn.

O: Dat was het begin?

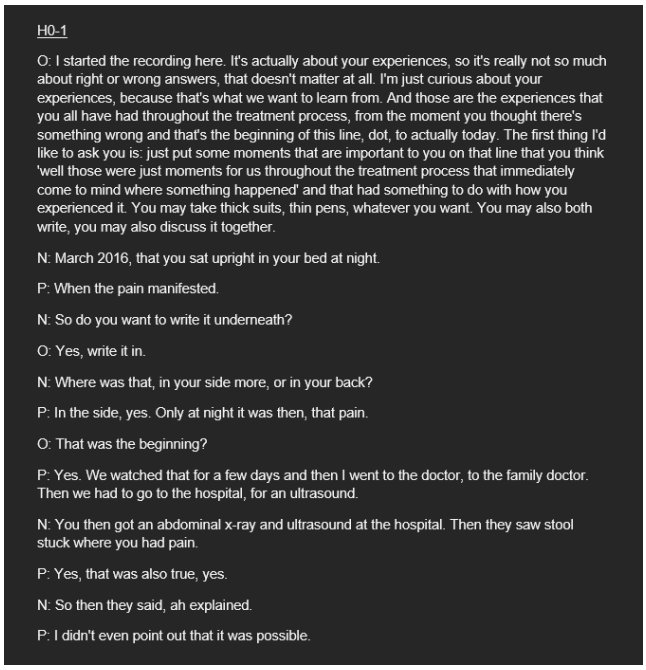
P: Ja. Dat hebben we een paar dagen aangezien en toen ben ik naar de dokter gegaan, naar de huisarts. Toen moesten we naar het ziekenhuis, voor een echo.

N: Een buikfoto en een echo heb je toen gekregen in het ziekenhuis. Toen zagen ze ontlasting vast zitten op de plek waar jij pijn had.

P: Ja, dat was ook zo, ja.

N: Dus toen zeiden ze, ah verklaard.

P: Ik wees niet eens dat het kon.



H0-1

O: I started the recording here. It's actually about your experiences, so it's really not so much about right or wrong answers, that doesn't matter at all. I'm just curious about your experiences, because that's what we want to learn from. And those are the experiences that you all have had throughout the treatment process, from the moment you thought there's something wrong and that's the beginning of this line, dot, to actually today. The first thing I'd like to ask you is: just put some moments that are important to you on that line that you think 'well those were just moments for us throughout the treatment process that immediately come to mind where something happened' and that had something to do with how you experienced it. You may take thick suits, thin pens, whatever you want. You may also both write, you may also discuss it together.

N: March 2016, that you sat upright in your bed at night.

P: When the pain manifested.

N: So do you want to write it underneath?

O: Yes, write it in.

N: Where was that, in your side more, or in your back?

P: In the side, yes. Only at night it was then, that pain.

O: That was the beginning?

P: Yes. We watched that for a few days and then I went to the doctor, to the family doctor. Then we had to go to the hospital, for an ultrasound.

N: You then got an abdominal x-ray and ultrasound at the hospital. Then they saw stool stuck where you had pain.

P: Yes, that was also true, yes.

N: So then they said, ah explained.

P: I didn't even point out that it was possible.

Figure 1: Snippets of the I0 interview in Dutch (left image) and English (right image)

I also examine these structural and linguistic challenges in detail in the upcoming dedicated sections, which explore workarounds and compromises for not only preserving the main essence and important aspects of the original transcripts but also for producing accurate and easily interpretable results by showcasing various experiments and alternative options.

4 Methodology

This section covers the methodological framework implemented to answer the aforementioned research questions. The goal of this study is to identify key topics and themes present in patient storytelling data and examine their relevance within the context of the interview, in order to determine whether state-of-the-art neural topic modeling techniques can be useful for providing patient feedback to medical staff. To achieve this, BERTopic and Top2Vec are applied on a single anonymized cancer patient interview at a time, with the baseline interview being `Interview I0`, and their outputs are evaluated in order to determine their coherence and relevancy in a medical context. This is done in order to test the neural topic modeling techniques' ability to summarize individual interviews without additional context from other interviews.

BERTopic is a relatively recent topic modeling approach introduced by Maarten Grootendorst [Gro22]. It combines transformer-based document embeddings with the HDBSCAN [sci23] clustering algorithm and a class-based TF-IDF procedure to generate coherent topics. BERTopic is designed to improve topic quality by capturing context through BERT and then identifying representative keywords for each cluster. As presented in Section 2.2, this approach has been shown to produce

topics that are more coherent than traditional models, such as LDA, especially on datasets where context matters. BERTopic first generates document embeddings using a pre-trained transformer model. It then reduces the dimensionality of these embeddings, clusters the documents in the embedding space to find potential topics, and lastly extracts the top keywords for each cluster using a specialized TF-IDF. The result is a set of topics, each described by a list of keywords that summarize the themes present in the corpus. BERTopic also offers the option to automatically label the topics through a representative model, although this feature is not used for this study, as I label the topics with my own pipeline.

For this particular study, BERTopic is a good choice because of its flexibility in choosing an embedding model, as well as its capability to capture context and synonyms, which is important for medical jargon and marks of orality. For example, the word "patient" has a different meaning when used as a noun, referring to a person receiving medical care, compared to its use as an adjective, where it describes the quality of being tolerant. This distinction is especially important in medical interviews, where context determines whether terms describe people, conditions, or behaviors. Moreover, BERTopic does not require manually setting the number of topics in advance, which is useful if the number of themes that may emerge from a corpus is unknown. This is achieved through its use of the aforementioned clustering algorithm, HDBSCAN, which determines the number of clusters based on the distribution and density of the embedded documents. Lastly, it is highly customizable and offers tools for visualizing topic distances and hierarchies, which is a great bonus as it does not require any further post-processing in order to produce useful and interpretable plots. I apply BERTopic to our specific dataset and then analyze its benefits and weaknesses, as well as compare it to Top2Vec in order to determine which model yields more coherent and useful results in the context of this study, particularly in a clinical setting.

Top2Vec is another unsupervised topic modeling technique introduced by Angelov [Ang20] that takes a different approach in comparison to BERTopic. While both models rely on semantic embeddings and clustering to extract topics, Top2Vec jointly embeds documents and words into the same semantic space. This allows it to directly identify topic keywords by locating words that are semantically close to document clusters, aligning topic discovery with the spatial relationships in the embedding space. Like BERTopic, Top2Vec begins by creating a vector representation of documents using a pre-trained embedding model. However, instead of using a class-based TF-IDF to extract representative terms, Top2Vec finds the nearest word vectors to each document cluster centroid. This results in a highly streamlined pipeline, as it does not require extra steps such as weighting word importance (such as with TF-IDF). Unlike BERTopic, Top2Vec relies entirely on the most semantically similar words to define each topic.

Another key difference lies in the model's handling of context and preprocessing. While BERTopic is designed to use transformer-based models to better capture contextual nuance, Top2Vec originally used Doc2Vec embeddings exclusively, and was later extended to support more powerful models like Universal Sentence Encoder [CYK+18]. However, Top2Vec tends to capture broader or more general topics, as its clustering process often favors semantic density over subtle narrative themes, which may prove to be a detriment in the context of the goal of this research. This stands in contrast to BERTopic's tendency to surface more fine-grained or context-specific topics, particularly when working with chunked input or in fields with complex language such as clinical interviews.

Top2Vec also differs in how it assigns topics. Each document or chunk is associated with a single dominant topic. There is no distribution over multiple topics as in probabilistic models, such as LDA. However, unlike BERTopic, Top2Vec does not offer out-of-the-box tools for topic reduction, visualization, or hierarchical structuring. As such, while Top2Vec's minimalism makes it an easier model to operate, it also offers less control and interpretability.

For the purposes of this study, I used Top2Vec as a baseline comparison to BERTopic. Top2Vec's simplicity, language-model flexibility, and automatic estimation of the number of topics, achieved through its approach of jointly embedding documents and words into the same semantic space using a pre-trained language model, then applying HDBSCAN to the document to detect topics, make it an appealing candidate for the task of extracting topics from patient storytelling data. By comparing the outputs of both models on the same clinical dataset, this research aims to evaluate which technique better captures the nuances of patient storytelling, with a focus on topic coherence, specificity, and clinical relevance.

The methodology consists of several key stages. First, the data set is preprocessed and chunked into segments (the exact number varies per experiment and file) in order to preserve narrative coherence while also ensuring compatibility with the embedding models. Moreover, several different embedding models are tested, but due to the differences between BERTopic and Top2Vec, the focus is placed on a common embedding model, namely **all-mpnet-base-v2** [RG20b], in order to fairly evaluate their performance on as much common ground as possible. Initially, both BERTopic and Top2Vec are tested on the default settings in order to check for preliminary issues and to further experiment with the parameters of the models based on the initial outputs. Next, the parameters of each model, as well as the parameters that control the clustering and dimensionality reduction processes, are experimented with and modified across several runs to determine which settings yield the best results. This involves changing the values of either one parameter at a time or multiple parameters simultaneously and then comparing the resulting output with previous outputs in order to observe the effects of each change.

I manually review all generated outputs to assess the coherence and relevance of the resulting topics by examining both the source documents associated with each topic and their corresponding keyword lists. Depending on the results, I further adjust the parameters by changing their values and comparing the new results in order to achieve more satisfactory outputs. In order to properly compare the performance of BERTopic and Top2Vec on approximately equal ground, I use the same document (Interview I0) for both preliminary sets of tuning experiments. As such, the following results correspond to **Interview I0 only**, unless stated otherwise, and may not apply to the other interviews as well.

Lastly, I select a final list of outputs from both BERTopic and Top2Vec to be manually evaluated by volunteers who are asked to read the original interview and compare it with the extracted topics in order to provide general feedback on the outputs' relevancy and coherence. The best performing neural topic modeling technique is then used for further analysis of the dataset, explicitly a global analysis for identifying recurring themes throughout the entire corpus of 13 interviews, as well as the most dominant themes overall, with the use of two metrics: topic prevalence and approximate distribution.

Table 1: Methodological Framework

Steps	**Purpose**
Initial model run with default settings.	Check for preliminary issues and determine what tweaks need to be made.
Data preprocessing.	Make the data digestible for the embedding model and ensure consistency across the dataset.
Parameter tweaking (on single interview)	Tune the model to yield satisfactory (or as close to satisfactory as possible) results.
Manual verification of output	Check if the resulting topics are coherent and useful.
Tweaked model runs on new data (on unused data)	Verify if the model with adjusted parameters can perform as well on new data.
Volunteer feedback	Obtain unbiased general feedback on the tweaked model outputs.
Global analysis	Identify recurring and dominant themes across the entire dataset to support broader interpretation.

5 Experiments

In this section, I outline the experiments and evaluation procedures used to assess the performance of the two neural topic modeling techniques. I apply each model to the same preprocessed storytelling interview (I0) and test various configurations to explore how different parameter choices and data representations influence the quality and interoperability of the resulting topics. These experiments aim to identify the optimal settings for each model and balance topic granularity, coherence, and clinical relevance.

5.1 Data Preprocessing

In order to prepare the interview data within the `docx.` files for topic modeling, a preprocessing pipeline is applied in order to not only ensure semantic clarity and consistency across the dataset, but also to avoid confusing the embedding models with unnecessary noise or artifacts that could alter the output in substantial ways. As stated in Section 3, the source data consists of 13 anonymized cancer patient interviews which contain speaker labels (P for patient, N for "loved one", and O for the interviewer), structural markings such as headers and internal identifiers, as well as marks of orality, such as stutters in speech and repetitions. Consequently, each of these challenges has to be solved in order to ensure a smooth topic modeling process and a reliable output that can be evaluated and potentially used for medical feedback purposes.

The first, and one of the most crucial steps in the preprocessing pipeline, is the translation of the original documents from Dutch into English. Because I am not fluent in the Dutch language, working with the dataset in a language which I can fluently understand is essential in order to ensure that

Table 2: Preprocessing Pipeline

Preprocessing Step	**Purpose**
Document Translation	Change the documents' language into a readable and verifiable one.
Lowercase Conversion	Redundant. Did not affect topic modeling output and made readability of topic documents more difficult.
Speaker Label Removal	Remove speaker tags (e.g., `P:`, `N:`) that can confuse embedding models.
Section Header Removal	Remove structural markers (e.g., `I0-1`) that do not carry any semantic meaning.
Contraction Expansion	Prevent broken tokens (e.g., *wasn*, *t*) by expanding contractions (e.g., *wasn't* to *was not*).
Custom Stop Words List	Remove uninformative and meaningless words (e.g., "uh", "yeah", "says").

the topic modeling process is accurate and meaningful, as it involves a lot of manual verifications, such as reading the representative documents to ensure that the topics are semantically cohesive and contextually accurate. To address this issue, I use DeepL [Dee24], a widely regarded AI translation tool, which can translate entire documents at once. Although machine translations are known for not being able to capture the complete meaning of a sentence, especially so in the context of casual spoken language, this limitation is not critical within the context of this experiment. As long as the main ideas and overall structure of the dialogues are conveyed in a reasonably accurate and readable manner, the translation suffices for the purpose of topic modeling. From now on, all references to the dataset refer to the English-translated version unless explicitly stated otherwise. A snippet from the first interview from the dataset, I0, is presented in Figure 1, both in Dutch and English.

Because BERTopic and Top2Vec do not natively accept `.docx` files, which is the original format of the interview transcripts, I convert the document to a clean `.txt` file by using the `python-docx` package [Fou24] in order to pass it to the models for fitting. Initially, the preprocessing method responsible for cleaning the text only contained lines for the removal of speaker labels, section headers, as well as a line for turning every single word in lower case in order to keep the text consistent and easy to digest for the embedding model. Initial experimentations show that the lower-case line is seemingly redundant, as it does not affect the output in any way, and also makes it more difficult to verify the documents that formed the topics due to the lack of clarity. Moreover, a critical issue occurs while fitting the model with the preprocessing measures mentioned so far: word contractions, such as "wasn't" or "doesn't", are being processed as separate words. For example, the word "wasn't" is considered "wasn" and "t" as two separate tokens, which introduces noise into the model's input, resulting in low-quality topic keywords and reduced semantic coherence. Upon further investigation, it is revealed that this issue is due to the default regular expression tokenization used in `TfidfVectorizer` from the `scikit-learn` library [PVG+24b], which is used internally by BERTopic for extracting topic keywords via a class-based TF-IDF approach. By

default, `TfidfVectorizer` takes any punctuation as a token separator, resulting in the separation of the aforementioned contraction words. To address this, I integrate contraction expansion to the preprocessing method, which expands every single contraction to its full form (for example, "wasn't" is expanded to "was not" within the cleaned lines of text).

In order to form as many useful and comprehensive topics as possible, I enable stop words within the vectorizer settings in order to filter out unnecessary and uninformative words that carry little semantic meaning. Initially, I only use the default `scikit-learn` English stop word list [PVG+24a], however, running the model with the default stop word list is insufficient, as the list is not comprehensive enough. As a result, I expand the list with my own custom additions composed of words that are observable within the initial model keyword list outputs. The custom list includes profanities, unnecessary words such as "phone" or "rings" (due to the interviewer's phone occasionally ringing during the first interview document), but also words which obstruct the general idea of what is being talked about, such as "said" or "says", which act as noise and often times block the formation of coherent topics by clogging the list of keywords.

The proposed preprocessing pipeline allows for a smooth topic modeling process, which results in the formation of coherent topics without any major complications or noise disturbances. These are discussed in the upcoming sections, which explore the topic outputs and the experimentation process. Although both techniques (Top2Vec and BERTopic) use the same preprocessing pipeline, each model has its own dedicated sections for results and experimentation.

5.2 BERTopic Experimentation

The following subsection details the experimentation phase with BERTopic, with the focus on a single interview (Interview I0). The goal is to identify the optimal setup for capturing relevant, cohesive, and interpretable topics from the storytelling dataset. The results of the experimentation process are located in Section 6.1.

Because the storytelling documents are lengthy, as mentioned in Section 3, I segment the original transcripts into smaller parts depending on certain criteria. This process is called chunking, and it is done in order to avoid exceeding the input size limits of the embedding models, and to ensure that the model analyzes more thoroughly smaller parts of the transcript at once in order to better capture the semantic coherence needed for the formation of a suitable topic. In this study, I employ a sentence-based chunking strategy. After the data processing pipeline completes its course, it produces a single text string, which is then split into individual sentences using regular expression-based detection of punctuation followed by white space. While more advanced sentence tokenization tools such as those provided by `spaCy`[HM17] or `NLTK` [BKL09] exist, I use a regex-based approach due to its simplicity, transparency, and sufficient accuracy given the structured and predictable nature of the interview transcripts. Since the dialogue is generally well punctuated, the added complexity of linguistic rule-based parsing is not necessary for this task. Finally, I group these sentences into chunks of fixed size. I try different chunk sizes to see how they affect the output, more precisely, the number of resulting topics and their coherence. Because of the exploratory nature of topic modeling, it becomes apparent that sequentially tuning the chunking process, as well as any other method, is counterproductive, since there is no way of knowing how each parameter interaction influences the

final output from the start without experimenting with multiple parameters at once. For example, hypothetically, while a smaller chunk size might result in a better-looking output with the default parameters of the model, it may prove to be inefficient once the parameters have been modified. As a result, in order to more accurately capture the effect of chunking on the model's output, I conduct the final round of chunking experiments using the fully tuned version of the model's parameters, rather than the default configuration. This approach ensures that the observed differences in topic generation can be attributed specifically to chunking, rather than to unoptimized model parameters. The tuning process itself is discussed later in this section.

Table 3: Effect of Sentence Chunk Size on Number of Chunks and Topics for Interview I0

Sentences Per Chunk	**Chunks**	**Topics**
5	172	17
6	144	16
7	123	12
8	108	9

Looking at the results in Table 3, it becomes clear that the number of chunks directly influences the number of topics produced. My theory is that this phenomenon occurs because, although BERTopic is capable of identifying topics by taking multiple documents into account, splitting the document into increasingly smaller chunks causes the model to analyze each segment more narrowly. The smaller the chunks, the more the model focuses on localized portions of the corpus, leading to the extraction of a great number of topics, often centered around more subtle or specific ideas, spanning a limited number of sentences. However, this increased granularity comes at the cost of potentially losing broader context and overarching themes present in longer, continuous conversations. Conversely, increasing the number of sentences per chunk results in fewer chunks, and ultimately in fewer, but broader topics. Increasing the number of sentences per chunk too much also comes at the risk of surpassing the token limit of the embedding model, resulting in the chunk being silently truncated. In light of these observations, I decide to ultimately settle on 6 sentences per chunk, as it results in an ideal amount of topics that present a good balance between broad contexts and more subtle or narrow ideas.

The tuning process for BERTopic (and Top2Vec) is inherently exploratory due to the open-ended nature of unsupervised learning and the lack of ground truth labels in the dataset. Instead of altering one parameter at a time, I test multiple configurations at the same time, with iterative refinements based on interpretability, coherence, and relevance of the extracted topics. I select the final model configuration after numerous cycles of manual evaluation and parameter adjustments. I begin the experimentation with BERTopic's default settings, using the **all-mpnet-base-v2** embedding model. This particular model is suitable due to its high average performance and coherence levels. However, I also evaluate alternative sentence transformers, such as **MiniLM-L6-v2** [RG20a], during the early phases, though I ultimately conclude that they are best set aside due to differences in output and token limitations.

The dimensionality reduction component of BERTopic relies on UMAP [MHM25], which projects

high-dimensional sentence embeddings into a lower-dimensional space for clustering. During tuning, I modify several UMAP parameters, most notably `n_neighbors`, `min_dist`, and `n_components` in an effort to balance topic separation with cohesion. These changes aim to influence how closely similar chunks are grouped, thereby affecting the granularity of topic clusters. For the clustering step, HDBSCAN [sci23] is used to group the reduced embeddings into topic clusters. The most impactful parameter here is `min_cluster_size`, which determines the smallest size a group of documents must have to be considered a distinct topic. Smaller values of this parameter tend to yield a higher number of highly specific or noisy topics, while larger values help reduce noise, but consequently result in more generalized topics. In addition, the `min_samples` parameter, which determines the number of points in a neighborhood for a point to be considered a core point, is also modified, along with the `cluster_selection_method`, which controls how the final clusters are selected from the density tree.

In addition, I also optimize the vectorizer parameters within BERTopic by manually modifying them through trial and error by studying the outputs, particularly the n-gram range and the inclusion of stop words (for stop words, see Table 2). The n-gram range is set to `(1, 2)` in order to allow for the emergence of phrase-level keywords made up of a maximum of two words. For example, instead of "pancreatic" and "cancer", the phrase "pancreatic cancer" would be taken as one single keyword if the phrase appears frequently enough within the corpus. This allows for the formation of more coherent topics, as it offers more context for the entire narrative rather than relying solely on isolated, potentially ambiguous single words. Phrase-level keywords are particularly important in clinical narratives, where many key concepts are expressed through multi-word expressions. By capturing these as unified terms, the model can more accurately group semantically related chunks, thereby improving both topic coherence and interpretability.

Lastly, I also adjust the `min_topic_size` parameter value during the tuning process. This parameter sets the minimum number of documents a topic must contain to be considered valid. Smaller values can reveal more fine-grained or niche topics, but risk generating noisy or redundant results. Conversely, larger values suppress the formation of smaller, potentially meaningful clusters. After several iterations, I can conclude that the default (10) value is the best option, as it results in the emergence of both broad and specific themes, while avoiding the creation of overly fragmented or insubstantial topics. Regarding this observation, some interesting results are revealed during the tuning process of this parameter, which are analyzed in detail in Section 6.1.

5.3 Top2Vec Experimentation

The following subsection details the experimentation phase with Top2Vec, with the focus on a single interview (Interview I0). As with BERTopic, the goal is to identify the optimal setup for capturing relevant, cohesive, and interpretable topics from the storytelling dataset. The results of the experimentation process are located in Section 6.2.

In order to ensure a fair comparison between the BERTopic and Top2Vec techniques, I use Top2Vec both in its contextualized form, referred to as C-Top2Vec [AI24], as well as Top2Vec's regular form. C-Top2Vec enhances the original Top2Vec algorithm by incorporating transformer-based sentence embeddings, rather than relying on traditional Doc2Vec representations. As a result, it enables the model to generate more semantically rich and context-aware topics, bringing it closer in methodolog-

ical similarity to BERTopic, which also relies on transformer embeddings. To maintain consistency across both modeling pipelines, I apply the same sentence-based chunking strategy to the dataset prior to fitting both Top2Vec variant models. This ensures that both models receive the same set of input documents: fixed-sized chunks of six sentences extracted from the original transcripts. The goal is to isolate the differences in topic generation purely to the underlying algorithm rather than to variations in data preparation. Although I explore both the original Top2Vec and its contextualized variant during the experimentation phase, I only consider the original Top2Vec implementation in the final comparative analysis. This decision is based on C-Top2Vec's lack of support for viewing the representative documents that formed each topic, a feature essential for interpretability and necessary for aligning the output with that of BERTopic. Initial experimentations with the default configuration of C-Top2Vec using the pre-trained `all-mpnet-base-v2` embedding model result in the detection of only four distinct topics. The limited number of topics is not representative of the diversity present in the interview data. Upon manual inspection, the resulting topics are found to be overly broad, combining semantically distant concepts into a single cluster. This shows that further tuning is required to increase granularity and improve topic coherence, similarly to BERTopic's default configuration trial. The default configuration for regular Top2Vec, however, results in the following error:

"ValueError: need at least one array to concatenate"

This error is probably due to the fact that HDBSCAN did not manage to form any clusters. Tuning the parameters further away from the default configuration values, however, does not result in the error anymore and, instead, successfully forms topics.

Similar to the tuning process employed for BERTopic, Top2Vec is optimized through iterative refinements of key parameters. This time, however, I start from the parameters of the already-tuned BERTopic model from Section 6.1 in order to see how Top2Vec behaves differently under close-to-identical conditions to BERTopic. Most of the tuning process that follows involves altering the parameters passed to the UMAP and HDBSCAN components that are internally used during the dimensionality reduction and clustering. I evaluate the same UMAP parameters explored during BERTopic tuning, namely `n_neighbors` `min_dist` and `metric`, to observe their influence on topic formation. Similarly, I modify the HDBSCAN parameters such as `min_cluster_size` and `min_samples` in order to balance specificity and generalization within the topic clusters. Although Top2Vec does not include a vectorizer component in the same way that BERTopic does, it still produces keywords associated with each topic. These keywords are derived from the joint embedding space of words and documents, which allows the model to identify semantically relevant terms. However, it does not support custom `n-gram` ranges or stop word configurations by default. As a result, fine-tuning keyword granularity proves to be less flexible than BERTopic, but the use of contextual embeddings still provides a solid basis for extracting meaningful topics. One notable limitation noticeable during the use of C-Top2Vec, however, is the inability to retrieve the specific documents that contributed to each topic. Unlike the original Top2Vec implementation, which supports mapping topics directly back to the documents that formed them, similarly to how it works with BERTopic, the contextual version of Top2Vec does not support this feature, as it is still in beta. As a result, while topics and their associated keywords can still be explored, it becomes substantially more difficult to trace these topics back to the chunks or original interview segments that generated them, making manual verification difficult. This limitation also presents a challenge

for interpretability, especially in clinical or qualitative settings where understanding the context of each theme is essential.

Ultimately, after several rounds of parameter adjustment and manual evaluation of topic interpretability, the refined C-Top2Vec model produces a more reasonable set of topics that better reflect the diversity of themes present in the patient interviews. These topics are then used in the comparative analysis with those generated by BERTopic. The specific model output results of the experimentation process are discussed in the 6.2 Section.

5.4 Topic Labeling with LLMs

Because the output topics from both models are essentially lists of keywords, they do not hold an interpretable meaning at first glance. To make sense of them, each topic must be labeled according to its semantic coherence. Traditionally, this involves manually looking at the top keywords for each topic, along with the sample documents that determined the formation of the topic. Labels are usually chosen based on the literal meaning of the keywords, along with the context found in the actual text, in order to make them interpretable and meaningful for human readers. The goal is for the label to be as meaningful as possible, especially in a healthcare setting where clarity and relevance are essential.

In order to achieve this, I take a modern approach to topic labeling by choosing to use a large language model (LLM) to label the topics in order to automate the process. The specific LLM model that I use for this task is OpenAI's **GPT-4o mini** model, because it is a powerful yet cost-efficient and lightweight model that shows strong performance across a range of evaluation metrics [Ope24]. I integrate the model into my pipeline using OpenAI's API, allowing the code to automatically generate descriptive labels for each topic. At first, I only pass the topic keywords to the model to see if the resulting topic labels would be cohesive enough without the need for representative documents, using the following basic prompt:

> "You are an AI that labels discussion topics, from a cancer storytelling interview, for a software that allows doctors to browse through medical files without the need to read them from start to finish. Given the following keywords, provide a clear and specific topic label, and only type the topic label and nothing else:".

The model generates descriptive labels only based on the topic's keywords, which produces mixed results. While some topic labels are good representations of the actual topics, others are unreliable because they lack the context behind the actual keywords. For example, one of the output topics produced by BERTopic, with some of the top keywords being "size 19, tricky, 25" received the label "Size Discussion in Medical Context", which has nothing to do with the actual context behind the topic. The actual document chunks that form the topic contain a conversation between the patient and the doctor describing needle sizes. It becomes clear that the absence of document context substantially impacts the quality of the generated labels, which is an impediment for individual-interview analysis. As a result, the approach also includes representative documents alongside the keyword list when calling the LLM, which finally leads to consistently more coherent, interpretable, and specific topic labels that better capture the meaning of each topic. The use

of representative documents, however, is not needed for the global analysis phase, which will be explicitly discussed in Section 7.2. The labeling prompt instructs to incorporate the document snippets as supporting context, rather than as the primary focus. This adjustment is made to ensure that the labeling process remains grounded in the keyword list, which represents the core output of the topic modeling process. Overemphasizing the snippets could risk undermining the role of the extracted keywords, effectively turning the labeling task into a generic text summarization exercise. Since keywords are intended to define each topic, they should remain the central focus during the labeling phase, with the document snippets serving as bonus evidence to help interpret them more accurately. The final output is the following:

> "You are an AI that labels discussion topics, from a cancer storytelling interview, for a software that allows doctors to browse through medical files without the need to read them from start to finish. Given the following keywords and sample documents, provide a clear and specific topic label, focusing mainly on the keyword list and using the document snippets as supporting context rather than a baseline. Only type the topic label and nothing else:".

This implementation is only possible using the BERTopic model and the non-contextualized variant of the Top2Vec model because, as mentioned in Section 5.3, C-Top2Vec does not support document mapping features yet, as it is still in beta. As a result, I only use the topic keywords during the labeling process for the C-Top2Vec model. This pipeline is also used for the global analysis section (Section 7.2), with certain modifications, which are explicitly named in the respective section, in order to accommodate the identification and interpretation of general themes across the entire dataset rather than interview-specific topics. Overall, incorporating an LLM for topic labeling not only streamlines the topic labeling process but also ensures that the resulting topics are clinically meaningful and accessible for further analysis.

6 Results

This section presents and compares the outcomes of the topic modeling experiments conducted using BERTopic and Top2Vec on Interview I0. The models are evaluated based on the number, clarity, and clinical relevance of the generated topics. Both quantitative measures, such as the number of topics and chunking effects, and qualitative indicators, such as coherence and interpretability, are taken into account. Each subsection provides an in-depth overview of the experimentation results of each respective model, along with reflections on how well the results align with the overarching goal of the thesis. As previously specified, all of the results originate from a single sample interview: Interview I0. I decide to conduct the experimentation phase in this manner in order to ensure a proper equal baseline for comparison between the two different neural models.

6.1 BERTopic Results

BERTopic is applied to the preprocessed content of interview I0, as outlined in Section 4. The final configuration is selected based on prior experimentation and is designed to optimize topic coherence, relevance, and clinical interpretability. The final model uses the `all-mpnet-base-v2` embedding model for semantic encoding, UMAP for dimensionality reduction, and HDBSCAN

for clustering. The final model is tuned with the parameters showcased in Table 4. Using this configuration, BERTopic generates a total of **17 topics**, each representing a relatively coherent cluster of semantically related segments within the interview. These topics span a wide range of themes, from procedural experiences and emotional reflections to logistical concerns and treatment decision-making, deeming this output suitable for a clinically-oriented software that clinical staff can use to quickly analyze patient data, as proposed in the introduction of this thesis. A comprehensive overview of the topics, along with their automatically generated labels and keywords, is presented in Table 5. The `all-mpnet-base-v2` embedding model manages to produce rich semantic representations that help to distinguish nuanced narratives.

Table 4: Parameters of Fully Tuned BERTopic Model

Parameter	**Setting**
embedding_model	all-mpnet-base-v2
use_embedding_model_tokenizer	True
min_cluster_size (HDBSCAN)	4
min_samples (HDBSCAN)	2
metric (HDBSCAN)	'euclidean'
cluster_selection_method (HDBSCAN)	'leaf'
n_neighbors (UMAP)	8
min_dist (UMAP)	0.0
n_components (UMAP)	10
metric (UMAP)	'cosine'
min_topic_size (BERTopic)	10
min_df (Vectorizer)	2
ngram_range (Vectorizer)	(1, 2)
sublinear_tf (Vectorizer)	True

The model output results in topics that are not only clinically relevant but also reflective of the patient's emotional and experiential journey through cancer treatment. For instance, **Topic 16** describes the patient's experiences during their FOLFIRINOX [Nat23] chemotherapy treatment, highlighting concerns such as neuropathy and treatment planning. Insights like these can serve as valuable feedback for clinical staff, offering a patient-centered perspective on how individuals are coping with the physical and psychological effects of specific therapies. On the other side of the spectrum, **Topic 8** offers the patient's logistical challenges and communication-related experiences, particularly in coordinating appointments and interactions with medical staff at Erasmus Hospital in Rotterdam. These types of topics can offer useful feedback to the hospital itself for improving internal processes, such as appointment coordination, patient communication, and overall administrative support. By surfacing these issues from patient narratives, the model provides actionable insights that can contribute to a more patient-oriented care experience.

From the results, it becomes clear, however, that the BERTopic model in the current state is not perfect by any means. For example, some topic keywords contain duplicate words due to the **ngrams** parameter. For example, in the keywords list of **Topic 7**, the word "port cath" appears as three different entries: "port", "cath", and "port cath". While the inclusion of bigrams increases

contextual richness, it can also lead to partial redundancy when both a phrase and its constituent words appear independently in the keyword list. Although this does not substantially impact interpretability, it can affect visual clarity and compactness of the topic's summary. On the other hand, this overlap may be beneficial when dealing with noisy or varied language, as it may help ensure that key terms are captured even when they appear in different forms across the corpus. This error could potentially be fixed in a post-processing process, or even within the vectorizer settings with extra tuning. Future iterations could involve filtering out redundant unigrams when a high-confidence bigram is present, though this would need to be balanced against the risk of losing relevant variations in phrasing. Another issue is overlapping themes spanning different topics. For example, port-a-cath procedures are discussed in two different topics: **Topic 4** and **Topic 7**. While the lists of keywords are different, meaning that while port-a-cath is discussed, this could reflect a true conversational shift around the same object of interest, they still share a common theme, which could be a sign of over-segmentation driven by overly sensitive clustering parameters. Although this could become an issue, manual document analysis proves that the two topics are distinct enough to be classified in individual clusters. Therefore, this occurrence is not a substantial issue, especially for this preliminary comparison phase, but rather proof that the model is capable of identifying nuanced themes, even from within the same broader topic.

I tune clustering with HDBSCAN using the `min_cluster_size` and `min_samples` parameters to strike a balance between avoiding over-fragmentation and filtering out noise. I employ the use of the `'leaf'` cluster selection method to extract more granular topic clusters. Unlike the default `'eom'` method, which tends to favor broader and more generalized clusters, `'leaf'` prioritizes leaf nodes in the clustering hierarchy, which results in a larger number of specific topics. This approach allows for the identification of narrower themes within the interview transcript, which suits the goal of helping medical staff identify topics within medical documents without having to read the entire files. One of the more interesting finds during the experimentation phase surrounds the `min_topic_size` parameter, which influences the topic count and coherence. As per my observations, lower values lead to the inclusion of small, fragmented clusters that are often redundant or noisy. Interestingly, extremely low values, for example 2, sometimes result in fewer topics, which is counterintuitive, as the parameter dictates the minimum amount of documents that a topic needs in order to be formed, so theoretically, a lower value should allow the formation of more small topics. However, using the default value of 10 not only results in more topics but also generates more cohesive topics than before. My theory is that HDBSCAN filters out the clusters generated at low values as noise, as they lack sufficient density or distinction. This suggests that the parameter interacts closely with the model's noise-handling mechanisms

Table 5: Final Output with the Fully Tuned BERTopic Model (Interview I0)

Topic ID	Topic Label	Top 15 Keywords
0	Experiences and Challenges Navigating Patient Passes at Daniel den Hoed Cancer Center	weird, white, pass, understand, den, hoed, daniel, daniel den, den hoed, patient, outside, room room, team, notice, waiting
1	Experiencing Fear and Anxiety During Medical Examinations Involving Tubes and MRI Scans	period, easy, lying, scary, tubes, throat, mri, tumor marker, marker, examination, prepared, tumor, rest, sorry, ultrasound
2	Radiation Treatment Timeline: Delays, Scheduling, and Scans from June to November	radiotherapist, radiation treatments, june, months, months scan, end november, guus, guus meeuwis, meeuwis, radiation, december, 21, follow, november, radio
3	Timeline of Malignant Diagnosis and Hospital Visits Including MRI and Ultrasound Examinations	malignant, april, write, timeline, place hospital, pretty, mri, walking, town, ultrasound, start, work, recording, raise, hospital hospital
4	Challenges and Experiences with Blood Draws and Port-a-Cath Access in Cancer Treatment	prick, cath, port, port cath, blood, needle, markers placed, hand, day, placed, markers, poked, difficult, puncture, puncture room
5	Discussion on Communication and Decision-Making in Pancreatic Cancer Treatment Conversations	eating, talk, cancer, list, clear, pancreatic cancer, pancreatic, certainly, eventually, conversation, surgeon, helps, time exciting, money, evening
6	Encouraging Patients to Ask Questions and Address Concerns During Appointments	especially, calls, questions, concerns, monday, head, difficult, mean, appointments, ask, involved, asked things, doctor going, reach, regular
7	Challenges and gaps in understanding port-a-cath placement and related surgical procedures.	brand, surgery, kind information, size, port, port cath, cath, example, information, 21, rotterdam, work, happens, eligible, drive
8	Patient's experience coordinating appointments and communication with doctors at Erasmus Hospital in Rotterdam.	forget, wednesday, doctor doctor, hair, date, friday, send, rotterdam, surgeon, erasmus, parking, specialized, hospital came, hospital want, went hospital
9	Discussion of tumor markers and the timeline of metastases detection and monitoring.	metastases, months, heard, tumor marker, marker, lot, year, tumor, spots, normal, lot googled, information, months blood, sampling, blood sampling
10	Impact of 2017 Cancer Cure on Patient's Life and Recovery Experience	2017, cure, year, hands, true course, opinion, gee, intense, 11, husband, home, took long, clear, certainly, rest
11	Bowel Test Results and Persistent Pain Leading to Further Medical Referral	taken, test, feeling, bowel test, piece, bowel, poking, worked, referred, pain, monday, showed, away, read, touch
12	Family Doctor Interactions and Patient Concerns During Cancer Treatment Interviews	family doctor, family, interview, grumpy, doctor family, worry, guys, time time, time doctor, long hair, hair, surgeon, nurse, definitely, puncture
13	Awkward Appointment Experiences and Scan Discussions Over Two and a Half Years	appointments, sigh, showing, half years, appointments new, years, look, scans, came scan, awkward, learned, exactly question, experts, cd, forward
14	Discussion of treatment options and choices made with the doctor during cancer care.	ask doctor, woolly, place place, ask, goodbye, familiar, doctor actually, choice, discussed, options, effective, hospital ask, time heard, kept, want want
15	Discussion on managing interruptions and planning during cancer treatment conversations.	prepared, stop, coming, clearly, time speak, stop moment, plan, mention, fine, going, feel, moment, right, people, happy
16	Concerns about neuropathy during FOLFIRINOX treatment process and course options discussed.	neuropathy, concerns, folfirinox, told, treat, courses, left, process, treatment process, folfiri, disease, huge, want want, write, treatment

In summary, BERTopic produces a diverse and clinically relevant set of topics from interview I0. The final configuration results in coherent and interpretable topic groupings. These findings show BERTopic's effectiveness in analyzing unstructured clinical narrative data and serve as a strong basis for comparison with Top2Vec in the upcoming sections.

6.2 Top2Vec Results

The following subsection presents the results of applying the original Top2Vec model to the storytelling dataset. As stated in Section 4, this analysis is based on the same preprocessed interview (I0) used for BERTopic in order to ensure a fair and consistent comparison. While both C-Top2Vec and the original Top2Vec implementations are explored during the experimentation process, I only use the latter in the final comparative analysis. This decision is due to the fact that the original Top2Vec model supports document-to-topic mapping, which is crucial for both manual implementation and for the overarching goal of creating a clinical support software for navigating patient files with the help of topic modeling. The final configuration of the model, including the UMAP and the HDBSCAN parameters, largely mirrors those used for BERTopic. The only major difference is the use of `min_count` parameter in place of `min_df`, which serves a similar function by controlling the minimum frequency a word must appear in the corpus to be considered for the clustering process. The embedding model used by default in Top2Vec is `all-MiniLM-L6-v2`, as confirmed by the verbose output during the fitting process. The full list of parameter values in the final configuration is shown in Table 6.

Table 6: Parameters of Fully Tuned Top2Vec Model

Parameter	**Setting**
contextual_top2vec	False
embedding_model	all-MiniLM-L6-v2
min_count (Vocabulary Filter)	2
min_cluster_size (HDBSCAN)	4
min_samples (HDBSCAN)	2
metric (HDBSCAN)	'euclidean'
cluster_selection_method (HDBSCAN)	'leaf'
n_neighbors (UMAP)	8
min_dist (UMAP)	0.0
n_components (UMAP)	10
metric (UMAP)	'cosine'

Using this configuration, Top2Vec generates a total of `18 topics`, which is two topics more than the output generated by BERTopic. These topics are generally cohesive and interpretable, and could be reliably mapped back to the original chunks used during the model training. The final model output, along with labels and keyword lists, is presented in Table 7. In several cases, the model produces topic clusters that closely resemble those found in the BERTopic output. For instance, two distinct topics relating to port-a-cath procedures emerge, similar to those identified by BERTopic, with keywords reflecting different aspects of the same clinical theme. However, not all topic clusters capture the broader context with the same accuracy. In one case, a conversation surrounding needle sizes is isolated into a broader, more general topic, namely **Topic 18**, but the model fails to recognize that the discussion is part of a larger conversation about needle types and procedures, as the interview transcripts suggest. As a result, the topic is reduced to a generic theme about "experiences and information", with keywords focusing on different medical and non-medical terminology and procedural words, such as "asked" and "conversation", rather than a cohesive and

concrete theme.

In general, the model produces interpretable results that align reasonably well with human-labeled themes. The ability to trace topic clusters back to specific document chunks makes manual evaluation possible and useful. In addition to the original Top2Vec model, the contextualized variant (C-Top2Vec) is also explored during the experimentation phase. This version allows the use of transformer-based sentence embeddings, such as `all-mpnet-base-v2`, and I initially consider it in order to align the Top2Vec methodology more closely with BERTopic, which also uses contextual embeddings. However, C-Top2Vec currently lacks support for document-to-topic mapping, which is a critical feature for both manual verification and the envisioned clinical support software. Using the same parameters as the final BERTopic model with six-sentence chunks, C-Top2Vec produces up to 86 topics, many of which are overly fragmented or completely redundant. While increasing `min_samples` and `min_cluster_size` parameters does reduce the topic count, to around 30 topics on average at best, the resulting clusters often lack coherence and show heavy overlap in keywords. In addition, without the ability to trace topics back to the original text segments, evaluating the quality of topics becomes nearly impossible. I do not deny the potential of this variant in tasks similar to these ones, as more tuning may improve C-Top2Vec to the point where it could produce results on par with BERTopic, or at least to its non-contextualized counterpart, however, these limitations ultimately result in the exclusion of C-Top2Vec from the final comparative analysis. The Doc2Vec version of Top2Vec is also briefly explored. According to documentation, this variant may perform better on large datasets with high variability, which, in theory, sounds suitable for the storytelling dataset. However, regardless of parameter configuration, the model consistently returns only a single topic, suggesting that either the chunking pipeline is unsuitable or that the preprocessing pipeline removes material needed for Doc2Vec to form meaningful clusters. Consequently, the Doc2Vec-based results are excluded from the final evaluation.

In summary, the original Top2Vec implementation provides a reasonably strong baseline, generating meaningful and traceable topic clusters. While it lacks the topical coherence of BERTopic in some cases, it produces comparable results overall, and its ability to support document mapping makes it viable for qualitative and clinical use cases. The upcoming comparison section further explores the differences between the outputs of Top2Vec and BERTopic in greater detail, and also determines the model to be used for further refinement and global dataset analysis, potentially enabling its use as a foundation within a clinical assistance software tool.

Table 7: Final Output with the Fully Tuned Top2Vec Model (Interview I0)

Topic ID	**Topic Label**	**Top 15 Keywords**
0	Patient Experiences and Emotions During Cancer Appointments and Treatment Environments	suddenly, conversation, room, happened, exciting, talk, appointment, story, hadn, later, move, cry, experiences, felt, was
1	Challenges and Experiences During Cancer Diagnosis and Treatment Journey	mri, surgeon, surgery, appointment, patient, doctor, ultrasound, tumor, hospital, appointments, metastases, scan, cancer, examination, malignant
2	Patient Experiences with Cancer Treatments, Cures, and Recovery Challenges	cure, cures, treatments, patient, treatment, pain, surgery, doctor, metastases, rest, puncture, appointment, no, yes, mri
3	Patient Experiences and Challenges During Cancer Treatment, Appointments, and Surgical Procedures	treatment, treatments, patient, appointment, doctor, surgery, chemo, cure, cures, experiences, metastases, cancer, appointments, concerns, experienced
4	Interactions with Healthcare Professionals During Cancer Diagnosis and Treatment	doctor, surgeon, hospital, patient, appointment, nurse, nurses, surgery, appointments, interview, ultrasound, examination, maybe, secretary, coach
5	Patient Experiences with Doctor Communication and Appointment Management in Cancer Care	appointment, patient, hospital, doctor, appointments, nurse, call, conversation, calls, nurses, surgery, suddenly, treatment, talk, surgeon
6	Timeline of Cancer Diagnosis and Treatment: Metastases, Tumor Markers, and Patient Experience	metastases, tumor, cancer, malignant, chemo, mri, patient, appointment, surgery, treatments, months, march, timeline, treatment, weeks
7	Patient Experiences with Doctors and Hospitals During Cancer Diagnosis and Treatment	doctor, hospital, appointment, patient, surgeon, appointments, nurse, nurses, surgery, ultrasound, mri, tumor, malignant, metastases, examination
8	Patient Experiences and Communication with Medical Professionals at Erasmus Hospital for Cancer Treatment	erasmus, appointment, doctor, surgeon, surgery, hospital, patient, appointments, treatment, examination, tumor, nurse, mri, puncture, treatments
9	Patient Experience with Radiation Treatment and Follow-Up Appointments for Cancer Management	radiation, radiotherapist, radio, patient, treatment, treatments, appointment, surgery, cancer, metastases, tumor, doctor, cure, mri, surgeon
10	Challenges and Experiences with Blood Draws and Port-a-Cath Procedures in Cancer Treatment	needle, puncture, poked, patient, blood, ultrasound, mri, prick, tubes, appointment, scan, port, treatments, surgery, examination
11	Exploring the Role of Turmeric and Patient Questions in Cancer Treatment Decisions	turmeric, cures, treatments, cure, certainly, only, no, concerns, definitely, doctor, maybe, treatment, always, probably, surgeon
12	Navigating Hope and Treatments in Pancreatic Cancer: A Patient's Journey and Conversations	cancer, pancreatic, tumor, chemo, metastases, malignant, patient, cure, treatments, cures, treatment, doctor, hope, surgery, bowel
13	Experiences and Concerns Surrounding Port-a-Cath Placement and Blood Draw Procedures	surgery, port, puncture, patient, needle, cath, surgeon, tubes, nurse, hospital, obviously, fact, appointment, nurses, operation
14	Bowel and Pancreatic Cancer Diagnosis Journey: Pain, Tests, and Treatment Experiences	bowel, pancreatic, patient, puncture, hospital, ultrasound, doctor, pain, appointment, mri, surgery, examination, scan, tumor, surgeon
15	Patient Experiences with Chemotherapy Options and Neuropathy Management in Cancer Treatment	treatments, neuropathy, treatment, cure, doctor, patient, surgery, cures, surgeon, folfirinox, mri, chemo, tumor, cancer, pain
16	Navigating Appointments and Treatment for Pancreatic Cancer at Rotterdam Hospital	rotterdam, hospital, appointment, erasmus, patient, appointments, surgery, doctor, pancreatic, surgeon, tumor, interview, treatment, experiences, examination
17	Patient Experiences and Concerns Regarding MRI and Scan Examinations in Cancer Care	scan, mri, ultrasound, certainly, examination, appointment, yes, patient, radio, no, ask, probably, radiotherapist, well, asked
18	Patient Experiences and Information Gaps During Surgical Appointments and Examinations in Oncology	size, patient, surgery, questions, doctor, information, experiences, examination, nurse, nurses, surgeon, experienced, appointment, conversation, asked

6.3 Model Evaluation and Comparison

This section describes the evaluation process for the two models analyzed in Sections 5.2 and 5.3. At the end of this section, I also provide a personal evaluation of the models and declare my model of choice for further experimentations.

To evaluate the coherence, relevance, and clinical interpretability of the topics produced by BERTopic and Top2Vec, I conducted a small-scale user study involving three volunteer participants. Each participant was provided with the anonymized cancer interview I0, which I used for tuning the parameters of BERTopic and Top2Vec, and asked to read it carefully. They were then asked to answer five questions for each model output. The full survey can be found in Appendix A. The aim of this evaluation is to determine how well each model captures meaningful patterns from the patient interviews that could serve clinical staff in quickly extracting key insights, such as treatment experiences, emotional responses, logistical issues, and decision-making dynamics, without needing to read the entire transcript. While the volunteers are by no means clinical experts, the task of judging the coherence and contextual relevance of the topics does not require specialized expertise and can be meaningfully performed by general readers with enough context from the source material. Participants rated each topic on a scale from 1 to 5 based on its coherence and usefulness (Q1), and separately evaluated the associated top 15 keywords for how well they described the topic (Q3). I chose coherence and usefulness to be evaluated within the same question because, although they are different concepts, they cannot exist within this context without one or the other. A useful topic cannot be incoherent, and an incoherent topic cannot be useful. Consequently, the volunteers were asked to grade the topics on account of both aspects. They were also asked to identify any essential themes that the models may have missed. The outputs that they were provided with were Tables 5 and 7. They did not have access to the representative documents from each topic, but they were allowed to go back to the interview in order to check whether the topic label was, in fact, describing a topic that was talked about during the interview. They did not have access to any other technical information not present in the tables mentioned above or in the interview. The results from survey Question 1 and Question 3 can be found in Tables 8 and 9. These tables present the topic ratings and keyword list ratings, respectively, for both Top2Vec and BERTopic side by side, allowing for a direct and thorough comparison of their perceived quality across individual topics.

Table 8 presents the average topic ratings over usefulness, relevancy, and coherence for BERTopic and Top2Vec across the entire model output. While the topics themselves are not comparable between the two models, a side by side view offers a better perspective for interpretability and comparison. Overall, BERTopic performed strongly, with 12 out of the 17 topics receiving a coherence score of 4.0 or higher, with multiple topics, such as **Topic 1**, **Topic 6**, and **Topic 14**, receiving perfect scores. This points that, according to the participants, BERTopic frequently generated coherent and useful clusters that aligned well with the content of the I0 interview. Keyword ratings were also generally positive, according to Table 9, though slightly more inconsistent than the coherence scores. Most topics scored between 3 and 4 for keyword relevance, with **Topic 4** receiving a high score of 4 and several others falling within a consistent range of 3.3 and 3.6. However, some outliers, such as **Topic 10**, received a high coherence score rating of 4.6 but a relatively low keyword score of 2.6, suggesting that while the theme was meaningful, the associated terms may have been less intuitive or representative for general readers.

Table 8: Comparison of Topic Ratings (Q1) between Top2Vec and BERTopic (Mean Scores, $n = 3$)

Topic ID	Top2Vec Rating	BERTopic Rating
0	4.6	3.6
1	4	5
2	2.6	4
3	4	4.6
4	4.6	4.6
5	4.3	4.6
6	4.6	5
7	3.3	4.6
8	3.3	4.6
9	4.6	4.3
10	4.3	4.6
11	5	4.3
12	4.6	4
13	3.3	3.3
14	3.6	5
15	3.6	4
16	4.3	4.6
17	4	
18	3.6	

This aligns with my theory that, while still helpful in this project's clinical context, keywords are not the main way to make sense of the topics. In more typical topic modeling tasks, keywords are usually the main tool for interpreting what each topic is about. But in this case, what matters more is the content of the representative document segments and the topic labels. These provide clinical staff with a faster and more intuitive way to locate relevant parts of a patient's story. Because of this, minor issues in the keyword lists, such as repeated n-grams or slightly redundant terms, do not substantially detract from the overall usefulness or clarity of the topics. This also suggests that future clinical applications should focus more on improving how topics are labeled and what examples are shown, rather than relying too much on keyword quality. To circle back, upon manual verification, the representative documents do make sense, and the topic label aligns well with the representative document snippets.

To assess the consistency among annotators rating the coherence and relevance of the topics, I compute percent agreement scores for each topic, which are presented in Table 10. Percent agreement is calculated as the proportion of annotators who assigned the most common rating to a given topic, divided by the total number of annotators (three), then expressed as a percentage. In the BERTopic model, several topics such as **Topics 1, 2, 6, 12, and 14** achieved perfect agreement (100%), indicating strong consensus among raters. In contrast, some topics, such as **Topic 15** showed lower agreement (33.3%), reflecting more divergent views. For Top2Vec, agreement is generally slightly lower, with topics such as **Topic 11** showing perfect agreement (100%), but many topics, including Topics **1, 2, and 7** having only 33.3% agreement. Overall, both models demonstrate

Table 9: Comparison of Keywords Ratings (Q3) between Top2Vec and BERTopic (Mean Scores, $n = 3$)

Topic ID	Top2Vec Keywords Rating	BERTopic Keywords Rating
0	3.3	3
1	4.3	3.6
2	4.3	3.3
3	4	3.6
4	4.3	4
5	4.6	3.6
6	4.3	3
7	4.3	3.6
8	4	3
9	4.3	4.3
10	4.3	2.6
11	4.3	3.3
12	5	3.6
13	3.6	3.3
14	4	3
15	4	3.3
16	4	3.6
17	3.6	
18	3.6	

moderate to high inter-rater agreement, with BERTopic slightly outperforming Top2Vec in terms of number of topics with perfect consensus. These findings support the relative robustness of the models, while also highlighting areas where topic definitions may require refinement or further clarifications.

While Top2Vec achieved lower topic scores on average than BERTopic, it performed better in terms of keyword ratings. 15 out of 19 topics received a keyword score of 4 or above, indicating that participants generally found the keywords descriptive and relevant to the topic content. In terms of topic rating, the model showed more variability. Several topics received high ratings, for example, **Topic 0**, **Topic 4**, and **Topic 11**, but others, such as **Topic 2** and **Topic 7**, were seen as less coherent or harder to interpret. This suggests that while Top2Vec was often able to surface meaningful terms, it was somewhat less consistent in grouping them into tightly coherent clusters, as observed in Section 6.2. A common point of feedback from the participants was that some topics appeared to overlap in content, such as between **Topics 4, 5, and 7**, to name a few. This, alongside the more thematically general topic labels, led the volunteers to rank the BERTopic output higher than the Top2Vec model overall, despite the high keyword list scores, further supporting the idea that keywords are less relevant when extracting helpful information from clinical interviews. This shows that while Top2Vec extracted more topics than BERTopic, it often fails to capture nuanced themes, which are crucial for the primary goal of this research.

For Question 2, which asks how well the extracted topics represent the content of the inter-

Topic ID	Top2Vec Agreement (%)	BERTopic Agreement (%)
0	66.7	66.7
1	33.3	100
2	33.3	100
3	66.7	66.7
4	66.7	66.7
5	66.7	66.7
6	66.7	100
7	33.3	66.7
8	66.7	66.7
9	66.7	66.7
10	66.7	66.7
11	100	66.7
12	66.7	100
13	66.7	66.7
14	33.3	100
15	66.7	33.3
16	66.7	66.7
17	66.7	
18	66.7	

Table 10: Annotator agreement percentages per topic for Top2Vec and BERTopic

view overall, the BERTopic model was ranked higher than the Top2Vec model, with the added mention from all three of the evaluation participants that, while both models manage to extract mostly cohesive topics from the I0 interview, the BERTopic model's output is way more precise and contained less overlap. This feedback is confirmation for the observations made in Section 6.2, namely that Top2Vec did not match BERTopic's capabilities of identifying more nuanced topics. It is important to note, however, that this evaluation does not suggest that Top2Vec is inherently worse than BERTopic for topic modeling in general. Rather, it reflects how well each technique performed in the context of this specific task, using this particular dataset, and within the constraints of my own technical expertise and the technical configuration applied in this study. One last observation from the participants, noted under Question 4, is that, while no particular themes are necessarily missing from either of the models' outputs, a clinical feedback tool should ideally list all of the treatment procedures that the patient went through, along with all the medication and prescription mentions from throughout the interview. While I agree with this stance, in practice, it is not easy to produce an output that contains all of this information, as some of these may have only been briefly mentioned in a single sentence, for example. Meaning that the model would most likely filter this information as noise. However, future work should also attempt to focus on this detail as well, as it would bring an extra layer of data for clinical staff to use in the treatment process.

Taking the evaluation process, as well as the experimentation and results, into consideration, I decide to focus the analysis on BERTopic further. Not only did the model demonstrate that it can consistently produce topics of higher overall quality than Top2Vec within the context of

this study, but it also offers greater flexibility and extensibility. A key strength of BERTopic lies in its compatibility with a wide range of publicly available embedding models, which allows for experimentation with domain-specific language models beyond the general-purpose ones, such as `all-mpnet-base-v2`, I used for this preliminary comparison. This flexibility opens up the possibility of integrating clinically-oriented embedding models that may further enhance the model's ability to extract relevant, interpretable, and context-sensitive information from patient narratives. As such, BERTopic not only outperformed Top2Vec in the current setup but also shows more long-term potential as a foundation for developing topic modeling tools tailored to the specific linguistic characteristics of clinical text, as well as to any other specialized domain.

7 Investigating the Generalizability of our Pipeline

7.1 Experimentation with Clinical BERT Models

To further explore the potential of BERTopic in extracting information from patient storytelling data, I decide to experiment with clinically oriented embedding models to investigate how their performance compares to that of general-purpose language models. To do this, I use three of the most commonly used clinically-oriented embedding models: **BioClinicalBERT** [AMB+19], **ClinicalBERT** [Med], and **MSR BiomedBERT**, previously known as **PubMedBert** [GTC+20]. All experiments are initially conducted using the same BERTopic configuration established during initial experimentation with the `all-mpnet-base-v2` model, as detailed in Section 5.2, in order to provide a consistent baseline for comparison. This includes identical chunking logic, vectorization pipeline, and dimensionality reduction components. Because this experimentation stage is carried out at a smaller scale, primarily to explore the suitability and potential of clinically-oriented embedding models for this task, there are no in-depth analysis subsections for each model, but instead, I provide general observations and an overview of the results from all three models. The final outputs of all three embedding models are located in Appendix B.

I begin by testing ClinicalBERT. Using the baseline settings inherited from the earlier experimentation phase, the model produces 17 topics. However, a few of these topics are semantically incoherent. One such example mirrors an issue previously encountered with Top2Vec, in Section 6.2, where conversations about needle sizes are grouped into a general-sounding topic regarding "medical sizes". To reduce noise and improve topic quality, the minimum document frequency (`min_df`) is increased to 3. This change, however, does not yield the desired effect. It merely causes specific discussions, such as the needle size subtopic, to be absorbed into broader, less distinct themes. To provide more contextual information for each chunk, I increase the chunk size from 6 to 7 sentences, which decreases the number of chunks per Table 3. This adjustment results in 15 topics, but these proved quite general or broad. To refine the topic boundaries, the dimensionality reduction setting `n_components` is lowered from 10 to 8, which reduces the dimensionality of the UMAP projection and can help enforce tighter semantic clustering. While this maintains the 15-topic output and results in slightly more coherent themes, the quality of the extracted keywords suffers as a consequence. Redundancies and repeated terms appeared more frequently, making it harder to interpret the topics at a glance. Despite these issues, the output does seem to be an improvement over the initial BERTopic output found in Table 5.

The next model I investigate is BioClinicalBERT. Because of the seemingly better noise filtering with `min_dif = 3`, I retain this setting for the remainder of the experiments. With 6-sentence chunks, the model produces 14 topics. Interestingly, increasing the chunk size to 7 results in 17 topics, which is the opposite of the results seen for ClinicalBERT. In contrast to the other models, the output from BioClinicalBERT is particularly balanced and semantically strong. The topics range from clinical experiences and medical procedures to frustrations regarding logistics, appointments, and hospital communication. These results are both interpretable and clinically meaningful, suggesting that BioClinicalBERT's pretraining on real clinical notes aligns closely with the nature of the storytelling dataset. Further tuning, such as reducing `n_components` to 8, as with the previous model, leads to 16 topics but reduces coherence, indicating that the original setting captures semantic relationships more effectively in this case.

Finally, I examine MSR BiomedBERT. When using 7-sentence chunks, the model produces 14 topics, but these are quite general and lack some of the specificity needed for practical use. Switching back to 6-sentence chunks increases the number of topics to 17 and slightly improves coherence, making it perform similarly to ClinicalBERT or the original `all-mpnet-base-v2` model. Some topics are reasonably focused, but others are too general, and the overall interpretability remains limited compared to BioClinicalBERT. Moreover, MSR BiomedBERT also struggles with topic coherence to a greater extent than the previous models. For example, one of the topics, incorrectly labeled as "Monitoring Eye Health and Treatment Progress in Cancer Care Discussions", includes representative documents that are not only short and uninformative, but one of the documents also includes the expression "to keep an eye on", which is the reason for the misleading topic label. This means that the model struggles to create meaningful clusters, which leads to the LLM mislabeling the topic due to the lack of context and coherence within the representative documents and the list of keywords. This relatively weaker performance may be attributed to the nature of the dataset used to pretrain MSR BiomedBERT, which could be less aligned with narrative-style patient data. Alternatively, it is possible that this model needs further parameter tuning, such as adjustments to clustering or dimensionality settings, to better adapt to the specific structure of this dataset.

Across all these models, it becomes clear that the choice of embedding model plays a critical role in determining the coherence, interpretability, and clinical relevance of the generated topics. While ClinicalBERT and MSR BiomedBERT show occasional improvement over the baseline, their outputs often lack the specificity and clarity necessary to reliably support clinical interpretation, or at least did not show a noticeable improvement from the baseline. In contrast, BioClinicalBERT consistently produces the most balanced and contextually appropriate topics, capturing both technical medical experiences and the patient's emotional experiences with greater nuance. These findings suggest that pretraining on a combination of biomedical literature and real-world clinical notes, as is the case with BioClinicalBERT, yields effective semantic representations for narrative healthcare data.

Due to the seemingly considerable performance improvement, I further examine the consistency and adaptability of BioClinicalBERT. I apply the model to an additional interview (I2), which is the shortest in the dataset at approximately 5596 words. When using 7-sentence chunking, the model produces only 8 topics. While these remain coherent, the output lacks granularity. Reducing

the chunk size to 6 sentences increases the number of topics to 12, with a noticeable improvement in nuance and precision. The resulting topics capture more specific clinical moments and emotionally substantial statements, enhancing the interpretability and relevance of the output. Testing this idea on several additional interviews that are slightly longer than I2 but still shorter than I0 (the longest interview in the dataset) confirms the emerging pattern. Shorter documents tend to benefit from smaller chunk sizes. In these cases, reducing the chunk length leads to the generation of more nuanced and clinically relevant topics, whereas longer chunks often result in overly broad themes that fail to capture important details. This further supports the hypothesis that a dynamic chunking strategy, adapted based on document length, may be necessary to optimize topic modeling performance across datasets containing interviews of varying lengths and complexities.

7.2 Global Analysis

7.2.1 Model Setup

In order to gain deeper insights into the entire dataset of 13 interviews, I expand the scope of the analysis beyond individual interviews and apply topic modeling to the entire corpus. This global analysis aims to uncover overarching themes and recurring patterns that may not be apparent when examining the interviews in isolation. By treating the full dataset as a unified narrative space, it becomes possible to identify more general topics that reflect shared experiences and systemic challenges across patients. For this experiment, I employ the BioClinicalBERT embedding model, which was previously identified as the most effective in producing coherent and clinically meaningful topics in Section 7.1. Its domain-specific training on biomedical and clinical text makes it particularly well-suited for capturing general themes within the context of the dataset.

I begin with a similar approach to the one used in Section 5.2, but this time, rather than aiming for fine-grained and nuanced topics meant to thoroughly summarize a single interview, I aim to extract broader, more general themes that reflect overlapping ideas from all 13 interviews combined. While the chunking strategy remains sentence-based, I adjust the number of sentences per chunk from 6 to 7 for this phase. As observed in Table 3 during the chunking experimentation, smaller chunks tend to increase topic granularity by focusing on localized content, often resulting in a greater number of narrower topics. In contrast, larger chunks provide the model with more context per document, leading to fewer but broader and more generalized topics. Since the goal of the global analysis is to capture overarching themes rather than fine-grained ones, using 7-sentence chunks is an appropriate balance by encouraging generality without exceeding the token limitations of the embedding model.

I then tune the rest of the BERTopic pipeline for global modeling, in a similar manner to Section 5.2, through trial and error, experimenting with different values of the parameters and further refining them by observing the resulting outputs. The full list of parameter values for the global analysis model can be found in Table 11. To accomplish this, I first modify the dimensionality reduction step by adjusting the UMAP parameters. Unlike the configuration used in the per-interview analysis, found in Table 4, which favors granularity and sensitivity to local patterns by using smaller neighborhoods and minimum distance between points, I increase the values in order to promote broader semantic clustering. In particular, I raise the `n_neighbors` parameter from 8 to 16 and the

`min_dist` from 0.0 to 0.2. These changes encourage UMAP to consider a wider semantic context when placing points in the reduced embedding space and to maintain more separation between clusters, which helps generalize across individual interview narratives. I also reduce the `n_components` from 10 to 4 to simplify the structure and avoid excessive fragmentation in the later clustering stage.

For the clustering step, I adjust the HDBSCAN parameters to further support the model's goal of extracting generalized topics. In contrast to the per-interview configuration, where `min_cluster_size` and `min_samples` were set to very low values, 4 and 2 respectively, to allow the formation of small, highly specific clusters, I increase `min_cluster_size` to 11 for the global model. This ensures that only patterns recurring across a larger number of chunks are promoted to full topics. I also use `eom` cluster selection method instead of `leaf`, as it favors more stable and prominent clusters, reducing the likelihood of generating noisy or overly fragmented topics. These changes collectively bias the model toward extracting higher-level themes that appear repeatedly across the dataset, aligning with the goal of capturing broad ideas rather than isolated experiences.

Other notable changes and additions are, first of all, the slightly different labeling process, initially presented in Section 5.4. Previously, the labeling approach involved considering both the keywords and representative documents from the interviews to assign topic labels. However, the new approach uses a different LLM prompt that focuses exclusively on the list of keywords for each topic, excluding the contextual information from individual documents. This change makes the labeling pipeline more suitable for labeling overarching themes, as by emphasizing keywords alone, the labeling becomes more objective and consistent, enabling for clearer and more precise topic descriptions that are interpretable without relying on specific document transcripts. In contrast, the previous labeling strategy allows for specific labels suited for summarizing individual interviews. The revised prompt is the following:

> "You are an AI that labels discussion topics, from a collection of cancer storytelling interviews, for a software that allows doctors to browse through medical files without the need to read them from start "to finish. Given the following keywords, provide a clear and specific topic label, with enough context to be interpretable, focusing on the keyword list.Be general. Keep it short. Only type the topic label and nothing else."

Lastly, I add new words to the stop word list in order to cover for additional noise observable in the global output. Because of machine translation issues throughout all 13 interviews, however, a perfect stop list is very difficult to obtain, as it would require manually going through every single interview in order to identify mistakes, such as words which are linked together, spelling errors, or mistranslated words.

Table 11: Parameters of BioClinicalBERT-Tuned BERTopic Global Model

Parameter	**Setting**
min_topic_size (BERTopic)	10
top_n_words (BERTopic)	15
min_cluster_size (HDBSCAN)	11
prediction_data (HDBSCAN)	True
cluster_selection_method (HDBSCAN)	'eom'
n_neighbors (UMAP)	16
min_dist (UMAP)	0.2
n_components (UMAP)	4
metric (UMAP)	'cosine'
random_state (UMAP)	42
min_df (Vectorizer)	3
ngram_range (Vectorizer)	(1, 2)
sublinear_tf (Vectorizer)	True

7.2.2 Topic Analysis

Using the above configuration, the topic model is fitted on the entire corpus of 13 interviews. The resulting output is presented in Table 17, in Appendix C. The extracted topics reflect overarching themes that span across every interview, offering a wide view of the cancer treatment experience. This broad perspective is particularly useful for researchers aiming to understand patterns in patient narratives without being limited to one-on-one analysis. Unlike the more granular per-interview models, which may capture individual nuances or interview-specific details, the global model reveals recurrent concerns and shared points across a diverse set of patients. This allows for the identification of systemic issues and common emotional or physical pain points in the cancer care process. For instance, **Topic 0** includes keywords such as "oxycontin", "medications", and "nausea", which highlight a widespread struggle among patients with managing medication side effects. Other topics identify specific procedures and systems that reoccur frequently. **Topic 2** provides insight into how patients describe and remember surgical interventions, specifically keyhole surgery. Similarly, **Topic 7** clusters together references to timelines and appointments, indicating that patients recall treatments not just medically, but temporally. This could support the development of visual treatment timelines, helping patients contextualize their cancer journey. **Topic 4** (Sleep Patterns and Nighttime Activities) focuses on the sleeping patterns of patients, indicating that sleep habits are a recurring and important theme across all 13 interviews. This suggests that sleep is an important aspect of the cancer journey that patients frequently feel the need to discuss. A deeper analysis of this topic could uncover common sleep issues or patterns experienced by patients, providing valuable insights into how cancer and its treatment affect rest and overall well-being.

Some topics capture institution-specific insights. **Topic 9** focuses on "Support and Resources for Cancer Care at Erasmus MC", referencing amenities like food, lighting, and buildings. While such content might seem irrelevant from a clinical standpoint, it underlines the importance of the hospital environment in shaping a patient's overall experience and could reveal common complaints or feedback points regarding the facilities. Using such feedback in order to improve hospital facilities could significantly affect perceived care quality. Importantly, several topics highlight the emotional

responses of patients. **Topic 8**, labeled "Coping with Treatment Setbacks and Emotional Reactions", groups together keywords related to emotional processing. These expressions reveal how patients narrate their coping mechanisms. Identifying these themes can support the integration of mental health professionals into oncology teams, ensuring that patients have psychological support as they process complex outcomes.

Lastly, **Topic 12** (Navigating Treatment Decisions with Specialist Nurses) emphasizes the critical role that specialist nurses play in the cancer treatment journey. The keywords reflect the patients' reliance on professionals not only for clinical information, but also for emotional reassurance and help in understanding complex medical choices. Rather than being passive recipients of care, patients appear to engage actively in their treatment planning, often assisted by nurses who help them comprehend their medical experiences better. This highlights the support system provided by nurses during key decision-making moments, reflecting narratives of trust, reassurance, and human connection.

Overall, the topics extracted by the globally tuned model are balanced between medical procedures, emotional resilience, logistical coordination, and support systems, demonstrating that this approach successfully captures high-level patterns across patient journeys. While some issues, previously encountered during the per-interview fitting of the model, such as some keywords being repetitive, for example "day day" in **Topic 0** (Medication Management and Symptom Relief in Cancer Care), these problems are less evident in the global model's output. This broader perspective complements the per-interview analyses by revealing common thematic ground shared between individuals, despite the unique elements of each narrative.

7.2.3 Topic Prevalence and Approximate Distribution

To better understand how different themes are distributed across individual patient interviews, I compute two different metrics: **Topic Prevalence** and **Approximate Distribution**. These metrics provide distinct perspectives on how topics are distribute within each interview. Topic prevalence reflects how frequently a topic is assigned to text segments within an interview, while approximate distribution captures the average probability of each topic being present throughout the interview. Together, they allow for both a discrete and probabilistic interpretation of themes.

Topic prevalence refers to how frequently a given topic appears within the content of an interview, based on the model's most confident assignment for each chunk. Each chunk is associated with exactly one topic, the one deemed most representative of its content. By counting how many chunks within a single interview are assigned to each topic, and then normalizing these counts by the total number of chunks in that interview, the result is a proportion between 0 and 1 that represents the relative dominance of each topic in that specific interview. This approach gives a discrete view of what patients focus on, making it useful for identifying predominant themes in each interview. To calculate topic prevalence, after the chunking of the corpus, every chunk was tagged with an identifier of the interview it originated from. This tagging ensures that all chunks can be traced back to their respective interviews. Two lists were maintained in parallel: one containing the chunks, and the other containing their corresponding interview identifiers. Once the topic modeling had been applied to the chunks, each one received a topic assignment. These topic labels are then combined

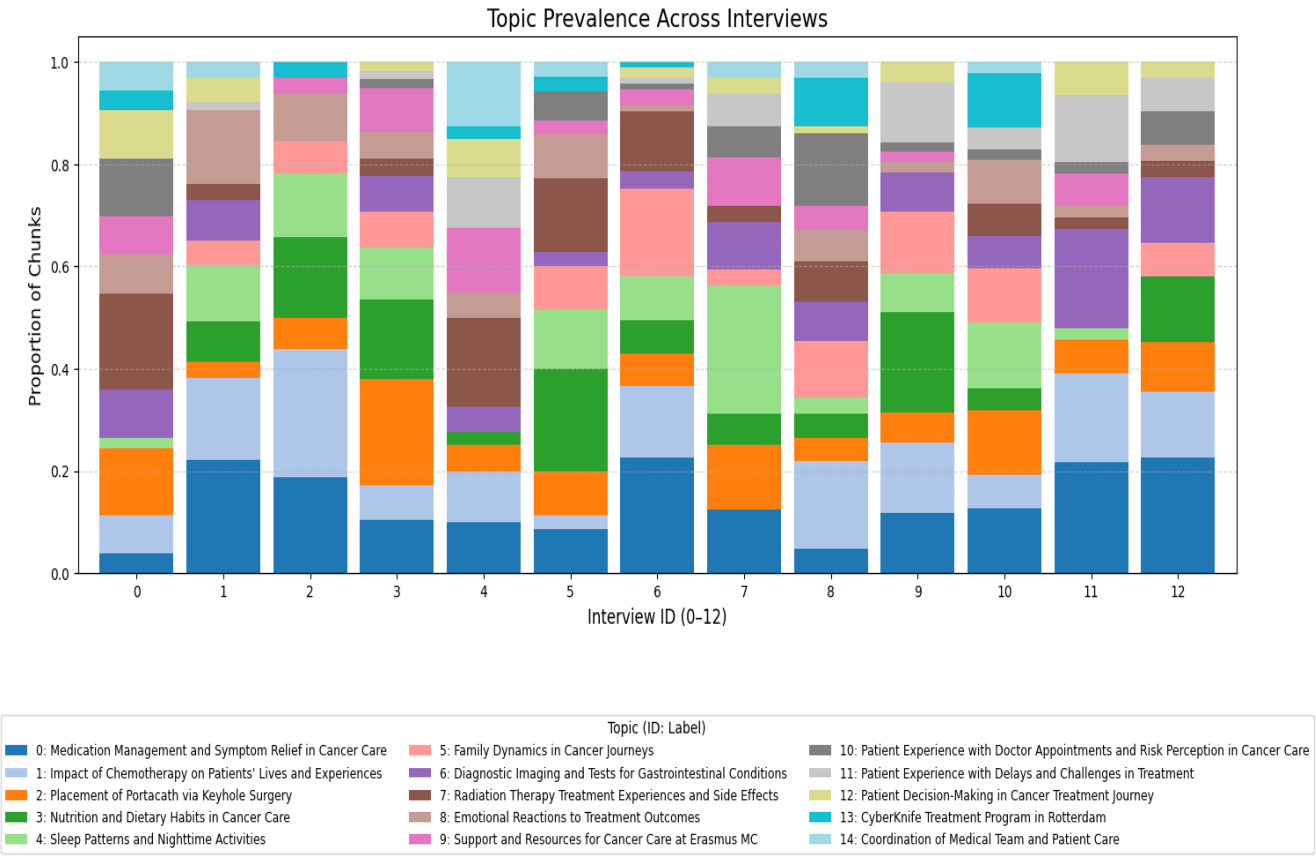


Figure 2: Topic prevalence across all 13 interviews, where each color represents a topic, and each bar represents one of the 13 interviews.

with the existing interview identifiers to create a structured dataframe, with each row representing a single chunk and containing three key columns: the interview it belonged to, the chunk text, and the topic assigned to it. The next step is to count how many chunks in each interview had been assigned to each topic. Because interview length varies, the raw counts were normalized by dividing them by the total number of chunks in that interview. This normalization ensures that topic proportions could be meaningfully compared across interviews regardless of their size. The result was a per-interview topic prevalence profile, where each topic's proportion represents how much of the interview's content is dedicated to that theme. Finally, these normalized proportions are organized into a matrix where each row corresponds to an interview and each column to a labeled topic. This matrix is visualized using a stacked bar chart, with consistent topic colors and labels to allow for easy visual comparison of themes across interviews, which is presented in Figure 2.

Approximate distribution offers a probabilistic perspective on the presence of topics within each interview. Rather than assigning a single topic to each chunk, this method uses the soft output of the BERTopic model. This functionality is provided directly by BERTopic through its `approximate_distribution()` method, which estimates the topic probabilities for each chunk without requiring re-fitting the model. This allows for the possibility that a chunk touches on multiple themes to varying degrees. By averaging these probability distributions across all chunks

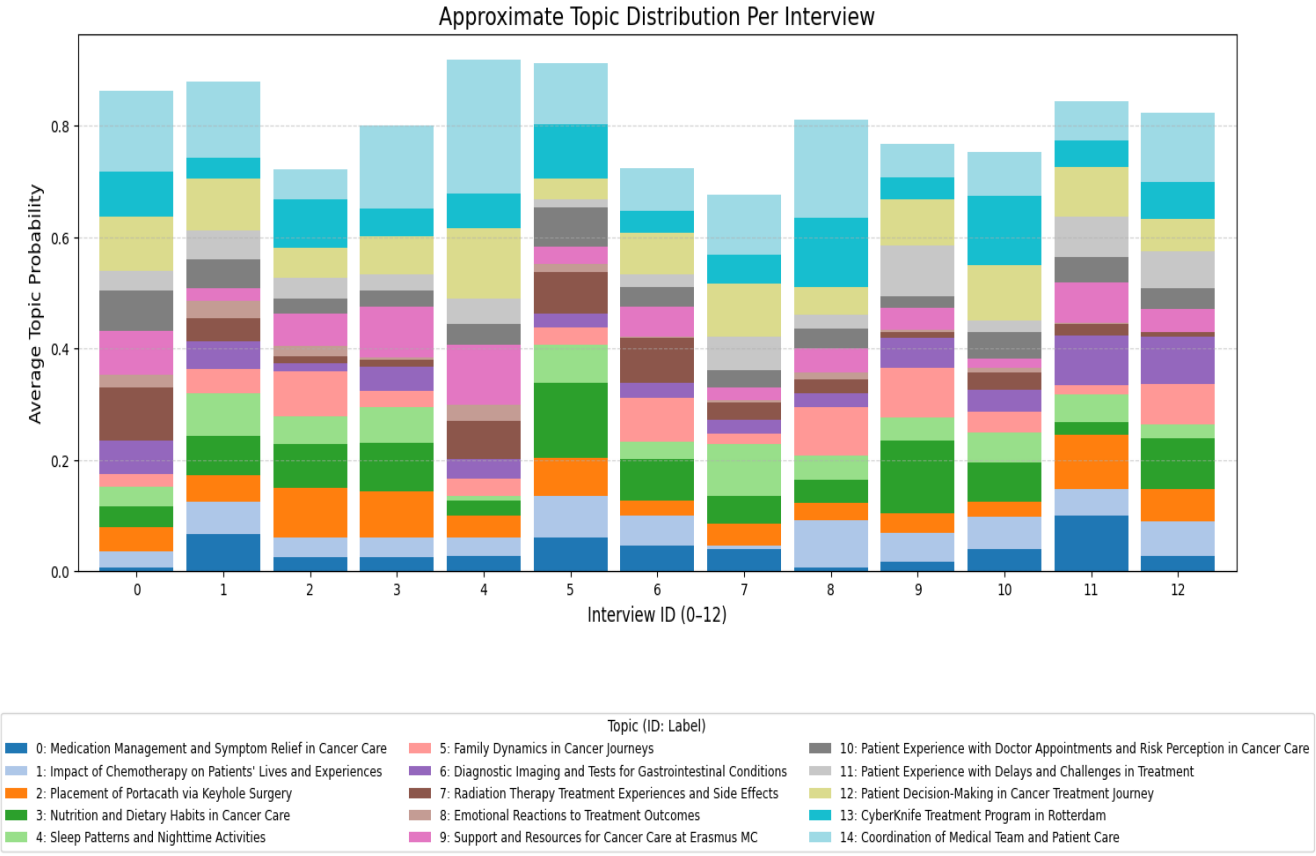


Figure 3: Approximate distribution across all 13 interviews, where each color represents a topic, and each bar represents one of the 13 interviews.

in a given interview, an interpretable vector is produced that captures how strongly each topic is present throughout the interview as a whole. This approach is useful for capturing thematic overlap and uncertainty that may be missed with hard topic assignments. To compute this, the same chunks used for the topic prevalence analysis are passed through the model's approximate distribution function to obtain their topic probabilities. Each resulting vector contains values corresponding to the model's estimated confidence for each topic per chunk. Because each chunk is already tagged with its source interview during the earlier segmentation stage, these vectors could be grouped by interview and averaged to produce a single topic distribution per patient. The final results can be visualized within a similar plot as the topic prevalence one, and is presented in Figure 3.

While topic prevalence and approximate distribution are derived from the same chunks, they differ due to how they handle ambiguity. Topic prevalence assigns each chunk to a single topic, highlighting only the dominant themes per interview. On the other hand, approximate distribution considers the full probability spectrum of each topic for every chunk, capturing subtler or overlapping themes that may not appear as top assignments. As a result, a topic may appear highly prevalent in one metric but less so in the other. For instance, a theme mentioned consistently but never as the main focus might score low in the prevalence, but rank higher in the approximate distribution. This difference is useful, as prevalence emphasizes key focal points, while approximate

distribution uncovers background or secondary topics. Interpreting both side by side allows for a fuller understanding of the overarching themes in each interview.

7.2.4 Global Analysis Results and Interpretation

The results of topic distribution across interviews, based on topic prevalence and approximate distribution, are visualized, as mentioned previously, in Figure 2 and Figure 3, respectively. These two perspectives offer complementary insights into how themes manifest across the dataset. From the topic prevalence plot (2), several interviews are strongly dominated by a small number of topics. For instance, Interview 7 is overwhelmingly characterized by **Topic 4** (Sleep Patterns and Nighttime Activities), while Interview 2 is dominated by **Topics 0** (Medication Management and Symptom Relief in Cancer Care) **and 1** (Impact of Chemotherapy on Patient Experience and Expectations). This suggests highly focused conversations within those interviews, centered on specific experiences regarding medication management, chemotherapy, and sleep patterns respectively. In contrast, the approximate distribution plot (3) provides a more nuanced picture. **Topic 14** (Coordination and Communication in Cancer Care Management) appears predominantly across several interviews (e.g. Interviews 0, 1, 3, 4, 8), even in those where it was not one of the top three in prevalence view (most notably, Interview 3). This suggests that while it may not have been the main focus in any one interview, it is a persistent theme that underlies many conversations, with the coordination of the medical team likely being a common underlying conversational topic among all patients. Another shift of this sort occurs with **Topics 12** (Navigating Treatment Decisions with Specialist Nurses) **and 13** (CyberKnife Treatment Program in Rotterdam), which are rarely top topics by prevalence, but frequently show up in the approximate distribution. This implies these themes may appear in shorter or more subtle forms, mentioned briefly, but across a wide range of patients. Meanwhile, the sharply focused topics (such as **Topics 0 or 2**) remain central only to select individuals.

By comparing these two perspectives, we can conclude that topic prevalence excels at highlighting what individual interviews are mostly about, capturing dominant and specific themes. Approximate distribution excels at revealing underlying, or recurring themes that may not dominate, but still reoccur throughout the corpus. Topics that rank highly in both metrics (such as **Topic 3**) are likely both recurrent and dominant, making them important targets for clinical or narrative analysis. These observations suggest that while each interview offers a distinct narrative, there are common themes that tie the experiences together, supporting the idea that both individualized and systemic elements are important in understanding patient journeys. Table 12 lists the five most frequently occurring topics across the entire dataset based on the topic prevalence metric, while Table 18 (in Appendix C) shows the top three topics for each interview using the same metric. In contrast, Table 13 and Table 19 (in Appendix C) present the corresponding results using the approximate distribution metric.

These results suggest that **Topic 14** (Coordination and Communication in Cancer Care Management) is the most recurring topic throughout all interview conversations in a more subtle form, while **Topic 0** (Medication Management and Symptom Relief in Cancer Care) is the most dominant topic throughout all interviews in terms of amount of chunks assigned with the topic as being the most dominant, meaning that the coordination of medical team and patient care (**Topic 14**) is the most recurring theme throughout the interviews, without being explicitly talked about,

Table 12: Most Occurring Topics Overall (Topic Prevalence)

Topic ID	Count
0	92
1	77
2	56
3	54
4	51

Table 13: Most Occurring Topics Overall (Approx. Distribution)

Topic ID	Mean Avg. Probability
14	0.118
12	0.079
3	0.071
13	0.070
2	0.053

while medication management and symptom relief (**Topic 0**) is the most dominant topic explicitly debated throughout the interviews.

8 Discussion

The experiments conducted in this study offer insight into the capabilities of topic modeling tools, specifically BERTopic and Top2Vec, for extracting meaningful and clinically relevant themes from cancer patient interview transcripts. From a broader perspective, the findings suggest that BERTopic, particularly when configured with clinically oriented embedding models and sentence-based chunking, has a stronger capacity to capture nuanced and useful topics. These results prove that topic modeling has the potential to be used as a backbone for a clinical feedback tool, which could help medical staff navigate lengthy and unstructured patient documents without the need to read the entire file themselves, therefore saving valuable time and shifting the focus to a more patient-oriented approach to healthcare.

The comparison between BERTopic and Top2Vec revealed substantial differences in the quality and interoperability of the generated topics. While both models are able to generate fairly coherent topics, BERTopic consistently produces topics that are perceived as more precise, with less overlap, which is essential in a medical context where clarity is critical. Moreover, the ability to utilize domain-specific embedding models, such as BioClinicalBERT, enabled BERTopic to adapt more efficiently to the clinical context, suggesting that the selection of embedding models plays a crucial role in performance. This suggests that domain adaptation, even without additional fine-tuning, can enhance topic coherence and relevance in specialized tasks.

However, the process highlights certain limitations and challenges that also need to be addressed for the sake of transparency and efficiency, in order to enable future progress in this field. One recurring

challenge encountered is the sensitivity of topic modeling to the chunking strategy. While smaller chunks lead to more granular and diverse topics, they occasionally fragment longer narratives, consequently losing broader contextual coherence. This trade-off between granularity and context preservation proves to be a recurring theme in the experimental setup for both BERTopic and Top2Vec. Another limitation of this study stems from my limited expertise in natural language processing and topic modeling. While substantial effort was made to understand and implement state-of-the-art techniques, it is possible that alternative approaches or configurations, particularly in areas such as preprocessing, embedding selection, or clustering strategies, could have yielded improved or different results. As such, some methodological choices may reflect practical constraints or a learning curve, rather than optimal design decisions.

An important component of this study involves a global-level analysis of topic relevance across interviews, using both the topic prevalence and approximate distribution metrics. These complementary measures provide different perspectives on how dominant certain themes were throughout the entire dataset. While topic prevalence offers a clearer view of the most frequently occurring topics in a discrete manner, the approximate distribution metric gives a more nuanced, probabilistic understanding of thematic presence. The global analysis reveals which themes consistently appeared across interviews, and which ones were more unique to certain patient narratives. This dual approach proves to be useful for identifying both dominant patterns and subtle topic associations, forming a richer interpretation of the dataset.

One interesting personal observation that emerged during both the evaluation phase and the experimentation phase was the seemingly reduced significance of keyword lists in understanding the extracted topics. Unlike traditional topic modeling methods, where keywords often serve as the primary cues for understanding topic content, in this study, the keywords alone were frequently insufficient to convey the full meaning or context of the associated topic. Instead, the representative documents played a more central role in making sense of each topic, especially within the nuanced and emotionally complex narratives found in patient interviews. It is important to acknowledge, however, that my own limited experience with topic modeling may influence this observation. A more seasoned researcher in this field might be better equipped to interpret keyword lists more effectively or to refine the modeling process in a way that improves their clarity and usefulness. Additionally, the evaluation process itself introduced certain constraints. Although the human-centered survey offered valuable insights into the perceived quality of the extracted topics, its small scale reduces the generalizability of the findings. Moreover, none of the evaluators had a clinical background, meaning that the judgements were based on general interpretability rather than professional applicability in a medical context.

Another important consideration lies in the nature of the dataset. The original interviews were conducted in Dutch and then translated into English before being processed by the models. While this translation is necessary due to the tooling, model availability, and my personal limitations in the Dutch language, it may have introduced inaccuracies or subtle changes in meaning that could impact the quality and authenticity of the output. In theory, topic modeling on the original Dutch texts, using a multilingual or Dutch-specific medical embedding model, might yield more faithful representations of the patients' narratives. However, no suitable multilingual embedding model with proven clinical expertise was found during the experimentation phase, and I was unable

to secure help from a fluent Dutch speaker with medical knowledge to verify the outputs using a general multilingual embedding model. This highlights a broader challenge in the field: the lack of accessible, high-quality, and domain-specific resources for languages other than English.

Ultimately, the findings of this study suggest a promising real-world application of topic modeling in clinical contexts. By enabling the automatic extraction of relevant and interpretable themes from lengthy patient interviews, models like BERTopic, when configured appropriately, could serve as the foundation for clinical support tools aimed at improving workflow efficiency and enhancing the patient-doctor relationship, shifting the focus to a more patient-oriented approach to healthcare. Such tools could enable medical professionals to quickly navigate large amounts of narrative data, identify key concerns, and prioritize patient needs without having to review every document manually. While the current system remains a proof of concept, it lays the groundwork for future implementations that could meaningfully support healthcare delivery by putting patient voices at the center of healthcare processes.

9 Conclusion

The goal of this study was to explore how neural topic modeling techniques could be applied to cancer patient storytelling data, with a focus on two main questions: first, what kinds of topics can current models extract from these types of interviews, and second, how the information gained from these topics might help improve existing healthcare frameworks or procedures by making patient perspectives more accessible.

In response to the first question, the results show that current neural topic modeling techniques, namely Top2Vec and BERTopic, can extract a variety of relevant themes from patient interviews. These include emotional experiences, treatment details, personal struggles, and reflections on the treatment processes. While both techniques produce fairly coherent and easily interpretable topics, BERTopic, especially when paired with an embedding model pretrained on large amounts of clinical data, such as BioMedicalBERT, and a sentence-based chunking strategy, delivers more refined and focused results, which better represent the patients' experiences and concerns expressed during the interviews. This makes BERTopic more promising for practical use in clinical environments, as opposed to Top2Vec, given the specific experimental setup and dataset.
As for the second question, the extracted topics suggest several ways in which topic modeling could support and improve healthcare processes. Most importantly, they could help clinicians identify and understand key moments in a patient's narrative without having to read every transcript manually. This could save time, reduce the workload of clinical staff, and give more visibility to the patient's voice, especially in cases where emotional or psychological concerns might otherwise be overlooked. Although not empirically tested in this study, this type of automated workflow for analyzing patient documents could also help reduce the risk of overlooking important details in a patient's history. By eliminating the need for clinicians to read through lengthy transcripts manually, the system may lessen the chance of missing key information due to time constraints or fatigue. Although this study did not involve a clinical trial or professional assistance, the structure of the output, combined with feedback from the conducted small-scale human evaluation, shows clear potential for integrating topic modeling into tools that support patient-centered care. Moreover,

the global analysis of all 13 interviews reveals recurring themes discussed by all patients, which could potentially assist in identifying patterns in cancer patient treatment journeys in order to mitigate common issues.

For future work, it will be important to test the system with clinical experts to better understand how the generated topics can be applied in a real-world healthcare setting. Beyond that, moving past translated text is an important next step. This includes not only exploring Dutch-language embedding models suitable for processing the original interviews, but also gathering new datasets in both English and other native languages. Doing so would serve the purpose of supporting the development and training of more robust multilingual, domain-specific embedding models, but it would also allow this approach to be tested on native English data to see how it performs without the distortions that come with translation. Additionally, future experiments could explore more adaptive or dynamic chunking strategies, which might better balance granularity and contextual coherence, especially across interviews of different lengths and structures. Lastly, the global dataset analysis could be expanded to track how the identified global themes evolve across different patient populations over time, and offer solutions to help clinical staff solve or minimize common complaints from cancer patients.

References

[AI24] Dimo Angelov and Diana Inkpen. Topic modeling: Contextual token embeddings are all you need. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[AMB+19] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Yuhao Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[Ang20] Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[Ang24] Dimo Angelov. Top2vec: Distributed representations of topics, 2024. Accessed: 2025-03-04.

[ANL+25] Yaniv Alon, Etti Naimi, Chedva Levin, Hila Videl, and Mor Saban. Leveraging natural language processing to elucidate real-world clinical decision-making paradigms: A proof of concept study. *Journal of Biomedical Informatics*, 136:104829, 2025.

[AYB20] Rania 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.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

[CYK+18] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dee24] DeepL GmbH. DeepL Translator. https://www.deepl.com/translator, 2024. Accessed: 2025-05-13.

[DFCM09] Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009. Biomedical Natural Language Processing.

[Era25] Erasmus MC. Erasmus university medical center. https://www.erasmusmc.nl/en/, 2025. Accessed: 2025-05-12.

[EY22] Roman Egger and Jie Yu. Interpretable topic modeling for social media analysis: Covid-19 and travel on twitter. *Frontiers in Sociology*, 7:850586, 2022.

[Fou24] Python Software Foundation. python-docx 0.8.11 documentation, 2024. Accessed: 2025-05-13.

[GHN23] Bernadeta Griciūtė, Lifeng Han, and Goran Nenadic. Topic modelling of swedish newspaper articles about coronavirus: a case study using latent dirichlet allocation method. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 627–636, 2023.

[GMSS17] Ingeborg Griffioen, Marijke Melles, Anne Stiggelbout, and Dirk Snelders. The potential of service design for improving the implementation of shared decision-making. *Design for Health*, 1(2):194–209, 2017.

[Gro22] Maarten Grootendorst. Bertopic: Neural topic modeling with class-based tf-idf. *arXiv preprint arXiv:2203.05794*, 2022.

[Gro24] Maarten Grootendorst. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics, 2024. Accessed: 2025-03-04.

[GTC+20] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

[HM17] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. In *To appear*, 2017.

[JM25] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 3 edition, 2025.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[Med] MedicalAI. Clinicalbert - huggingface. https://huggingface.co/medicalai/ClinicalBERT. Accessed: 2025-05-22.

[MHM25] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. https://umap-learn.readthedocs.io/en/latest/parameters.html, 2025. Accessed: 2025-05-12.

[Nat23] National Cancer Institute. Folfirinox. https://www.cancer.gov/about-cancer/treatment/drugs/folfirinox, 2023. Accessed: 2025-05-17.

[OAN+25] Yukiko Ohno, Tohru Aomori, Tomohiro Nishiyama, Riri Kato, Reina Fujiki, Haruki Ishikawa, Keisuke Kiyomiya, Minae Isawa, Mayumi Mochizuki, Eiji Aramaki, and Hisakazu Ohtani. Performance improvement of a natural language processing tool for extracting patient narratives related to medical states from japanese pharmaceutical care records by increasing the amount of training data: Nlp analysis and validation study. *JMIR Medical Informatics*, 2025.

[Ope24] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. Accessed: 2025-05-12.

[PVG+24a] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. scikit-learn: Feature extraction — stop words, 2024. Accessed: 2025-05-11.

[PVG+24b] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. scikit-learn: Tfidfvectorizer documentation, 2024. Accessed: 2025-05-11.

[RG20a] Nils Reimers and Iryna Gurevych. Sentence-bert: all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2020. Accessed: 2025-05-17.

[RG20b] Nils Reimers and Iryna Gurevych. Sentence-bert: all-mpnet-base-v2. https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 2020. Accessed: 2025-05-17.

[sci23] scikit-learn developers. HDBSCAN — scikit-learn 1.6.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.HDBSCAN.html, 2023. Accessed: 2025-05-12.

[SMD+19] Sara Sheikhalishahi, Riccardo Miotto, Joel Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, 7(2):e12239, 2019.

[TU 21] TU Delft. Better prepared for big decisions in healthcare. https://www.tudelft.nl/en/ide/delft-design-stories/better-prepared-for-big-decisions-in-healthcare, 2021. Accessed: 2025-05-12.

[TUR50] A. M. TURING. I.—computing machinery and intelligence. *Mind*, LIX(236):433–460, 10 1950.

[WAREg+25] Abdullah Wahbeh, Mohammad Al-Ramahi, Omar El-gayar, Ahmed Elnoshokaty, and Tareq Nasralah. Evaluating topic models with openai embeddings. Technical report, University of Hawai'i at Mānoa, 2025.

[Wei66] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, January 1966.

A Volunteer Survey Questionnaire

This section contains the volunteer survey questionnaire, analyzed in Section 6.3, which I used in order to evaluate the outputs of the two chosen topic modeling techniques. The volunteers were presented with interview I0, along with the following survey paper to complete:

Context: This survey is part of a research project investigating the use of automated topic modeling techniques on cancer patient interview transcripts. The broader goal of this thesis is to explore how topic modeling can help clinical staff quickly extract relevant insights, such as emotional responses, symptoms, experiences, etc., from lengthy patient narratives without having to read entire files. You will be presented with one anonymized patient interview and the resulting topic outputs from each model. Your feedback will help evaluate the clarity, relevance, and usefulness of the topics, contributing to an assessment of how well these models could support future clinical decision-making tools.

On the next page, you will find the extracted topics and keywords generated by each model. Please read the interview first before proceeding with the questions.

1. **Please rate each topic on a scale from 1 to 5, where:**

   - 1 = Not coherent / Not useful
   - 5 = Very coherent / Very useful

   For each rating, please add a brief explanation if needed.
   Example: T1: 4 – Useful topic, but title could be more precise.
   T2: 1 - The topic makes no sense, this was never talked about during the interview.

2. **Overall, how well do the extracted topics represent the content of the interview you read?**
   (1 = Not at all, 5 = Very accurately)

3. **How helpful/accurate were the keywords under each topic for understanding what the topic was about?**
   (1 = Not helpful, 5 = Very helpful)
   Example: T1: 4 - Useful keywords, but it has one irrelevant word in it: "the")

4. **Were there any important themes, ideas, or aspects of the interview that were missing from the extracted topics?**

   - Yes

- No

If yes, please specify.

5. **Do you have any additional feedback or suggestions about the topics or the overall experience?**

Each participant received one survey document for each model, which had the same questions, with the only difference being the topic output.

B Clinically-Oriented Embedding Model Outputs

The analysis of the three clinically-oriented embedding models, namely **BioClinicalBERT**, **ClinicalBERT**, and **MSR BiomedBERT**, previously known as **PubMedBert**, can be found in Section 7.1. The settings for the final topic outputs are the following:

- **ClinicalBERT**: Baseline settings + 7-sentence chunking + `min_df = 3` + `n_components = 8`
- **BioClinicalBERT**: Baseline settings + 7-sentence chunking + `min_df = 3`
- **MSR BiomedBERT**: Baseline settings + 6-sentence chunking + `min_df = 3`

The output for each model, which includes every topic label and top 15 keywords for each topic, can be found below for each of the embedding models. As with the rest of the outputs showcased throughout this research, the following outputs are also obtained using interview I0.

Table 14: Final Output with the ClinicalBERT-Tuned BERTopic Model (Interview I0)

Topic ID	**Topic Label**	**Top 15 Keywords**
0	Challenges in Patient-Doctor Conversations Within Limited Treatment Spaces and Options	conversation, room, started, probably, little bit, little, wait, cures, eye, let, june, 19, long time, things, called
1	Patient Engagement in Cancer Care: Importance of Asking Questions and Seeking Information During Appointments	questions, person, especially, felt, question, head, making, ask, asked, appointments, guys, mind, pancreatic, talk, patient
2	Challenges with FOLFIRINOX Treatment, Size Discrepancies, and Information Gaps in Patient Experience	size, folfirinox, example, eventually, 19, new, information, let, day, prick, people, happens, happy, rotterdam, happen
3	Emotional Impact of Cancer Diagnosis and Treatment Experiences at Daniel den Hoed Hospital	head, lying, daniel den, hoed, den hoed, daniel, den, grumpy, understand, write, going happen, couple times, rest, throat, point
4	Delays in Chemotherapy Start Due to Treatment Coordination and Malignant Diagnosis Concerns	poked, june, malignant, radiotherapist, long time, 21, fact, able, chemo, follow, ultrasound, appointment, certain, radiation treatments, asked
5	Delay in Tumor Marker Evaluation and Persistent Pain Leading to Further Medical Intervention	pain, ultrasound, tumor marker, marker, days, week, tumor, possible, weeks later, wait, contact, eye, rest, right, malignant
6	Challenges in Scheduling Appointments and Blood Puncture Procedures During Cancer Treatment	appointments, december, puncture room, puncture, talk, appointment, end, anymore, need, blood, september, results, hand, patient, day
7	Patient Experience and Choices in Hospital Transitions and Care Interactions	notice, people, real, sweet, choice, quite true, does matter, secretary, showed, nice, results, allowed, patient, matter, exactly
8	Challenges and Experiences with Port-a-Cath Insertion and Blood Draw Procedures	looks, markers, port cath, cath, port, markers placed, prick, size, poked, blood, times, placed, hand, explained, happen
9	Conversations with Doctors and Hospital Visits in April Regarding Patient Care	april, doctor doctor, conversation, touch, place hospital, hospital, erasmus, doctor, surgeon, idea, read, space, new, went, away
10	Rising Tumor Markers and Scan Results Impacting Treatment Decisions in December 2017	december, year, anymore, tumor marker, marker, blood, cure, tumor, scan, possible, takes, clear, sat, gee, heard
11	Challenges in Communication with Medical Staff and Coping with Cancer Journey	secretary, difficult, life, learned, doing, prepared, rotterdam, knows, mean, possible, nurses, life coach, coach, hold, pick
12	Navigating Hope and Fear in Pancreatic Cancer Diagnosis and Treatment Experiences	hope, pancreatic cancer, pancreatic, beginning, information, pain, sit, lot, mean, cancer, look, possible, ask, bad, definitely
13	Waiting Room Experiences and Communication About Test Results in Cancer Care	getting, sitting waiting, guys, showed, television, taken, waiting room, metastases, waiting, hour, true, scan, half, sitting, moment
14	Importance of Direct Consultation with Expert Surgeon in Rotterdam for Cancer Treatment	friday, rotterdam, does matter, matter, called, sitting, appointment, surgeon, sure, make sure, saying, brought, examination, definitely, follow

Table 15: Final Output with the BioClinicalBERT-Tuned BERTopic Model (Interview I0)

Topic ID	Topic Label	Top 15 Keywords
0	Patient Experience and Information Gaps in Pancreatic Cancer Treatment Choices and Communication	questions, especially, choice, guys, quite, information, pancreatic cancer, saying, quite true, pancreatic, kind, eventually, real, hold, story
1	Challenges and Experiences with Port-a-Cath Usage During Cancer Treatment	prick, maybe, story, pain, port, port cath, cath, mean, prepared, sit, cure, beginning, look, possible, blood
2	Experience of Radiation Treatment: Challenges, Waiting Times, and Patient Comfort During Procedures	idea, lie, weird, hour, stop, sit, radiation, exciting, people, helped, fine, conversation, end, gee, certain point
3	Experiencing Anxiety and Uncertainty During Throat Examination and Treatment Processes	throat, quickly, going happen, make, understand, rest, couple times, lying, lie, happens, information, tubes, happen, times, doing
4	Frustrations with Treatment Plans and Follow-Up in Cancer Care Conversations	plan, little bit, bit, probably, takes, wait, poked, cures, look, june, television, radiation treatments, guys, stop, place
5	Monitoring Tumor Markers and Blood Sampling Experiences Over Time in Cancer Care	prick, blood, year, cures, tumor marker, couple, waiting room, december, times, tumor, weeks, waiting, marker, scan, option
6	Frustration with Appointment Changes and Desire for Consistency in Doctor Consultations	grumpy, appointment, called, does matter, different, doctor doctor, matter, make sure, sure, monday, surgeon, nurse, afternoon, researcher, end
7	Early Detection and Monitoring of Tumors Through Ultrasound and Tumor Markers	pain, heard, ultrasound, tumor marker, taken, saw, tumor, marker, beginning, true, hospital place, sitting waiting, gets, scan, real
8	Patient Reflections on Communication and Information During Cancer Diagnosis and Treatment Journey	sorry, especially, remember, information, hear, bed, pretty, throat, lying, minutes, later, question, tubes, examination, room
9	Malignant Biopsy Experiences and Challenges in the Puncture Room at Daniel den Hoed	malignant, puncture room, puncture, den hoed, daniel den, hoed, den, daniel, person, ultrasound, blood, results, saw, hand, came
10	Discussion on Pancreatic Cancer Diagnosis and Treatment Experiences in Rotterdam, April Timeline, and Communication	hear, getting, rotterdam, guys, size, long time, obviously, april, pancreatic cancer, true, surgeon, metastases, pancreatic, remember, important
11	Timeline of Cancer Treatment: Appointments, Chemo Start Dates, and Patient Experience	21, september, chemo, june, walking, follow, end, took, town, appointment, radiation, asked, december, start, long time
12	Understanding Port-a-Cath Placement and Tumor Markers in Cancer Treatment Context	looks, markers placed, placed, rotterdam, markers, cath, port, port cath, explained, exciting, possible, nurse, ask, tumor, happy
13	Navigating Life After Cancer: Learning, Recovery, and Uncertainty in Personal Experiences	life, learned, weird, knows, doing, little bit, bit, pick, took long, hold, half, town, moment, goes, room
14	Impact of Personal Connections and Delays on Patient Experience in Healthcare Settings	notice, people, sweet, showed, secretary, does matter, results, date, nice, matter, patient, does, stuff, feel, important
15	Significant Conversations and Key Dates Related to Doctor Visits at the Hospital	april, date, conversation, place hospital, doctor, space, hospital, read, doctor doctor, called, erasmus, surgeon, good, away, new
16	Key Moments and Experiences During Cancer Treatment Journey: Insights from Patient Perspectives	moments, wrong, times, let, happened, experienced, researcher, gee, happy, certainly, everybody, eventually, appointments, rest, couple times

Table 16: Final Output with the MSR BiomedBERT-Tuned BERTopic Model (Interview I0)

Topic ID	Topic Label	Top 15 Keywords
0	Delay in Chemotherapy Start Date and Emotional Impact on Patient Care	head, tubes, stuff, 21, june, date, saw, gone, chemo, examination, placed, puncture room, puncture, fact, lie
1	Patient Experience and Information Seeking During Cancer Treatment Journey in Rotterdam	comes, stop, mind, prepared, rotterdam, mention, feel, showed, hold, later, pancreatic cancer, pancreatic, googled, bed, kind
2	Sudden Rise in Tumor Marker Levels and Impact on Patient Care Decisions	september, year, tumor marker, suddenly, end, days, blood, tumor, cure, marker, weeks, scan, radiation, went, weeks later
3	Patient Experience with Hospital Procedures and Communication Challenges in Rotterdam	mean, couple times, understand, lie, 19, does, read, rotterdam, allowed, speak, brought, nurses, times, sitting waiting, experienced
4	Importance of Patient Questions and Understanding in Cancer Treatment Process and Port-a-Cath Use	markers placed, questions, important, placed, cath, port, port cath, process, treatment process, moments, markers, ask, kind, moment, looks
5	Coordination of Medical Care: Conversations and Travel to Rotterdam for Pancreatic Cancer Treatment	doctor doctor, friday, rotterdam, conversation, pick, surgeon, googled, definitely, doctor, took, went, plan, april, idea, read
6	Urgent Need for Follow-Up Appointments with Radiotherapist and Surgeon Amidst Treatment Changes	suddenly, radiotherapist, does matter, able, appointment, grumpy, matter, ultrasound, asked, surgeon, talk, room, sitting waiting, right away, waiting room
7	Importance of Patient Experience and Communication in Cancer Care at Hospitals	patient, important, place hospital, sweet, number, researcher, tubes, examination, malignant, doctor, understand, half hour, grumpy, long, den
8	Discussion on Surgical Procedures, Patient Experiences, and Emotional Responses Related to Cancer Treatment	gee, exciting, 2017, puncture room, puncture, saying, surgery, clear, true, quite true, year, poked, 19, working, sorry
9	Monitoring Eye Health and Treatment Progress in Cancer Care Discussions	eye, keeps, june, started, probably, radiotherapist, 19, making, cures, look, knows, maybe, looks, bit, need
10	Experiences with Radiation Treatment at Daniel den Hoed for Pancreatic Cancer	den hoed, hoed, den, daniel, daniel den, idea, radiation, thought, possible, remember, scary, exactly, bed, option, lie
11	Patient's Proactive Approach in Seeking Second Opinions and Treatment Options During Consultations	minutes, ask, look, fine, helped, list, option, definitely, getting, fact, asked, real, treatment process, process, folfirinox
12	Challenges and Experiences with Blood Draws and Port-a-Cath During Cancer Treatments	poked, prick, half hour, blood, results, radiation treatments, waiting, treatments, port, cath, port cath, hour, times, cath port, feel
13	Hope and Uncertainty Surrounding Pancreatic Cancer Diagnosis and Treatment Experiences	hope, cure, days, throat, maybe, bad, pancreatic cancer, pancreatic, does work, number, scary, gone, keeps, explained, couple times
14	Challenges and Decisions in Cancer Treatment: Folfirinox, Metastases, and Surgical Options	folfirinox, metastases, wait, easy, getting, make, surgery, television, probably, way, start, sorry, cath, port, port cath
15	Experiencing Cold, Unwelcoming Spaces and Lack of Choice in Medical Settings	exactly, real, quite true, moments, sitting, anymore, choice, space, felt, read, room, need, allowed, hospital, talk
16	Issues with Tumor Marker Monitoring and Communication of Treatment Plans During Appointments	plan, tumor marker, came, tumor, new, marker, make sure, sure, learned, come, clear, television, appointment, went, certain

C Global Analysis Data

Table 17: Global Analysis (All 13 Interviews)

Topic ID	Topic Label	Top 15 Keywords
0	Medication Management and Symptom Relief in Cancer Care	knee, day day, pills, oxycodone, medications, diarrhea, medication, symptoms, times day, went doctor, nausea, stomach, prescribed, consultantdoctor, ones
1	Impact of Chemotherapy on Patient Experience and Expectations	chemo, chemotherapy, intense, effects, tomorrow, paper, oncologist, does thats, start chemo, took long, oncologist oncologist, door, meeting, cells, drive
2	Placement of Portacath via Keyhole Surgery	portacath, placed, arm, keyhole surgery, keyhole, anesthesia, surgery, portacath portacath, probe, puncture, puts, run, shower, sedated, poked
3	Nutrition and Dietary Habits in Cancer Care	cook, drinking, eating, sandwich, food drink, taste, eat, eating drinking, food, dietician, weight, eaten, brother, soup, fat
4	Sleep Patterns and Nighttime Activities	sleep, downstairs, bed, couch, awake, lie, bathroom, watch, single, wash, groceries, rest rest, outside, upstairs, cup
5	Family Support and Life Impact in Cancer Journeys	son, sister, joint, mother, project, children, twice, life, lives, times time, live, kind thing, older, child, large
6	Diagnostic Imaging and Tests for Abdominal Conditions	bowel, ultrasound, mri, stomach, examination, appendix, tests, ct, biopsy, admission, ct scan, medium, pain clinic, scan hospital, taken
7	Radiation Therapy Treatment Experiences and Side Effects	radiotherapist, radiation, courses, treatments, poked, december, october, thats possible, abdominal pain, abdominal, operate, markers, september, november, placed
8	Coping with Treatment Setbacks and Emotional Reactions	failed, plan, weeks later, wall, reactions, success, calmly, tremendous, dirty, alive, face, march, nerves, cells, tried
9	Support and Resources for Cancer Care at Erasmus MC Hospital	euros, erasmus, erasmus mc, mc, places, hospitals, hospital erasmus, light, food drink, support, hair, approach, lot people, possibly, building
10	Patient Experience with Doctor Appointments and Risk Assessment	risk, doctor hospital, date, wonder, gosh, appointment doctor, forget, data, rotterdam, touch, tomorrow, space, ended, wife, march
11	Patient Experience with Medical Equipment and Care Delays	pump, ticket, broken, waited, burden, air, walked, nurses, hours, does work, minutes, outpatient, decisions, early, nursing
12	Navigating Treatment Decisions with Specialist Nurses	decisions, treatment process, experiences, trajectory, open, calls, important decision, advise, negative, specialist nurse, shes, cries, real, super, metastatic
13	CyberKnife Treatment Program in Rotterdam	rotterdam, program, cab, stone, cyberknife, quarter past, button, file, quarter, puncture, push, examination, family doctor, liver, record
14	Coordination and Communication in Cancer Care Management	responsible, team, secretary, order, number, personal, doctor come, creon, surgeon, clear, scary annoying, short, conversations, turn, knows

Table 18: Top 3 Topics per Interview (Topic Prevalence)

Interview	Topic I	Topic II	Topic III
1	7	2	10
2	0	1	8
3	1	0	3
4	2	3	0
5	7	9	14
6	3	7	4
7	0	5	1
8	4	0	2
9	1	10	5
10	3	1	0
11	0	2	4
12	0	6	1
13	0	1	3

Table 19: Top 3 Topics per Interview (Approx. Distribution)

Interview	Topic I	Topic II	Topic III
1	14	7	12
2	14	12	4
3	2	13	5
4	14	9	3
5	14	12	9
6	3	14	13
7	7	5	14
8	14	12	4
9	14	13	5
10	3	11	5
11	13	12	14
12	0	2	6
13	14	3	6