



Universiteit
Leiden
The Netherlands

Analyzing cancer patients' experiences with embedding-based topic modeling and LLMs

Ionescu, T.C.; Han, L.; Suasnabar, J.H.; Stiggelbout, A.; Verberne, S.

Citation

Ionescu, T. C., Han, L., Suasnabar, J. H., Stiggelbout, A., & Verberne, S. (2026). Analyzing cancer patients' experiences with embedding-based topic modeling and LLMs. *Computational Linguistics In The Netherlands Journal*, 15, 121-142. Retrieved from <https://hdl.handle.net/1887/4309309>

Version: Publisher's Version

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/4309309>

Note: To cite this publication please use the final published version (if applicable).

Analyzing Cancer Patients' Experiences with Embedding-based Topic Modeling and LLMs

Teodor-Călin Ionescu*

Lifeng Han*†

Jan Heijdra Suasnabar†

Anne Stiggelbout†

Suzan Verberne*

L.HAN@LUMC.NL

J.M.HEIJDRASUASNABAR@LUMC.NL

A.M.STIGGELBOUT@LUMC.NL

S.VERBERNE@LIACS.LEIDENUNIV.NL

*Leiden Institute of Advanced Computer Science (LIACS), Leiden, The Netherlands

†Leiden University Medical Center (LUMC), Leiden, The Netherlands

Corresponding authors: LH and SV

Abstract

This study investigates the use of neural topic modeling and LLMs to uncover meaningful themes from patient storytelling data, with the goal of offering insights that could contribute to more patient-oriented healthcare practices. We analyze a collection of transcribed interviews with cancer patients (132,722 words in 13 interviews). We first evaluate BERTopic and Top2Vec for the purpose of individual interview summarization, by using similar preprocessing, chunking, and clustering configurations to ensure a fair comparison on Keyword Extraction. LLMs (GPT4) are then used for next step topic labeling. Their outputs for a single interview (I0) are 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. We then analyzed the full interview collection with the best model setting. 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 “Coordination and Communication in Cancer Care Management” and “Patient Decision-Making in Cancer Treatment Journey”. 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. Affiliated resources created in this work will be shared publicly at <https://github.com/4dpicture/TM4health> including codes on preprocessing and stopword list prepared.

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 (Ren et al. 2025). 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. These have been reflected by recent work on shared decision making

(SDM) and patient-centered carepath design (Griffioen et al. 2021, Kidanemariam et al. 2024, Bak et al. 2025).

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 paper, we use neural topic modelling techniques to extract relevant topics from cancer patient storytelling data and analyze them systematically. We aim 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. We compare two topic modeling algorithms, BERTopic (Grootendorst 2022) and Top2Vec (Angelov 2020, Angelov and Inkpen 2024), to determine which one performs better at extracting relevant topics from patient storytelling data. By optimizing these models, we 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. For instance, for a collection of patient consultation transcripts, this methods can help to identify cross-patient population level concerns effectively.

We address two research questions in this paper:

1. **What key themes can current neural topic modeling models extract from patient storytelling data, which model is the best choice for this task, and if the integration of clinical domain embeddings help?**
2. **Based on the LLM-labeled topics using key word lists from the first step, 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. The dataset used in this study was provided by Erasmus Medical Center (Erasmus MC) in Rotterdam, the Netherlands (Erasmus MC 2025) through an EU healthcare research project. It originates from the Metro Mapping Project, a design-driven initiative to support cancer patients in navigating their care journey and preparing for significant medical decisions (Griffioen et al. 2017, TU Delft 2021). By analyzing the cancer patient storytelling data, we aim to contribute to that mission, proving insights that reflect the patients’ lived experiences and can be used to further improve patient-centered healthcare practices.

2. Background and Related Work

2.1 Model Evolvment of Topic modeling

Topic modeling is an unsupervised technique for identifying themes in large text collections and forms the foundation of this research. A widely used traditional method is Latent Dirichlet Allocation (LDA) (Blei et al. 2003), a Bayesian algorithm for extracting topics from text. LDA has been applied in various contexts, such as analyzing topic evolution during the COVID-19 pandemic in Swedish newspapers (Griciūtė et al. 2023). However, LDA suffers from limitations, including difficulty capturing semantic relationships, sensitivity to common words, and issues with short-text data (Albalawi et al. 2020). These shortcomings include non-deterministic outputs, the need for a predefined number of topics, data sparsity, and inability to model topic relations.

More recent research introduced embedding-based models like **BERTopic** (Grootendorst 2022) and **Top2Vec** (Angelov 2020) including its contextual-embedding variation from Angelov and Inkpen (2024), which overcome many LDA limitations by using dense vector representations, dimensionality reduction, and clustering. These models require minimal preprocessing, automatically determine topic numbers, and excel with short or context-rich texts. Studies show they outperform classical methods in coherence and interpretability: Egger and Yu (2022) found BERTopic

produced the most distinct topics on COVID-19 Twitter data, while Top2Vec also surpassed LDA. Similarly, Wahbeh et al. (2025) reported BERTopic achieving the highest coherence scores across datasets, with Top2Vec ranking slightly lower but still ahead of LDA. Overall, embedding-based models consistently deliver more coherent and interpretable topics than traditional approaches.

2.2 Patient-Narrative Clinical NLP

Clinical Natural Language Processing (NLP) applies NLP techniques, such as topic modeling, to extract and analyze medical text from sources like unstructured health data, discharge summaries, surveys, and patient files. Modern research highlights its integration into healthcare systems, particularly for *clinical decision support* and *clinician-patient communications* (Demner-Fushman et al. 2009, Wang et al. 2018). For instance, in the statistical era, the review by Demner-Fushman et al. (2009) documents the use of the Linguistic Inquiry and Word Count (LIWC) tool for analyzing patient personality through linguistic style, enabling applications such as predicting cancer adjustment, mental health outcomes after bereavement, and differentiating suicidal from non-suicidal patients. Complementing these clinician-facing applications, Van Buchem et al. (2022) introduced the AI-PREM, an open-ended patient experience questionnaire paired with an NLP pipeline that combines topic modeling and sentiment analysis to thematically cluster short patient narratives into actionable summaries for clinicians. These examples underscore the role of linguistic feature analysis in diagnosis, management, and prognosis. However, Clinical NLP faces challenges in *clustering* tasks, as noted by Sheikhalishahi et al. (2019) and Ghosh (2024), including difficulties in processing clinical notes, ambiguities, context-dependent, and domain specific knowledge.

Recent advancements emphasize *patient-narrative* data, aligning with this study’s focus. Ohno et al. (2025) developed a BERT-based model for monitoring symptoms from patient interviews in a Japanese hospital, achieving superior performance despite limitations like data scarcity and single-source training. Furthermore, Alon et al. (2025) analyzed Hebrew clinician speech using word frequency and Large Language Models (LLMs) to uncover cognitive paradigms, finding reliance on heuristics and intuitive reasoning, though limited by small datasets and lack of multilingual frameworks. Similarly, Somani et al. (2023) used an AI pipeline combining BERTopic, dimensionality reduction, and spectral clustering to organize 10,233 patients’ statin-related Reddit posts into six overarching themes, and then applied a pretrained BERT sentiment model, which found the discussions were predominantly neutral to negative. Overall, literature shows Clinical NLP’s potential to transform healthcare by extracting insights from both patient and clinician narratives, yet challenges persist due to privacy concerns, linguistic variability, and limited annotated datasets (Raj et al. 2023, Han et al. 2024, Bak et al. 2025). This study addresses these issues by exploring neural topic modeling for coherent, patient-centered topics, promoting reproducibility through shared and anonymized datasets and models, and investigating multilingual approaches using specialized embeddings and translation tools to advance linguistically inclusive Clinical NLP.

3. The Dataset - Patient Storytelling

The data we use consists of 13 anonymized transcribed .docx files (I0 to I12), each corresponding to a different cancer patient. As mentioned in the introduction section, the dataset is provided by the 4D PICTURE consortium, an EU research project across multiple institutes from several countries¹, which aims to improve the cancer patient journey and ensure personal preferences are respected. The data originates from the Metro Mapping Project (TU Delft 2021) published in the work of Griffioen et al. (2021). 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,

1. <https://4dpicture.eu> “Design-based Data-Driven Decision-support Tools: Producing Improved Cancer Outcomes Through User-Centred Research”

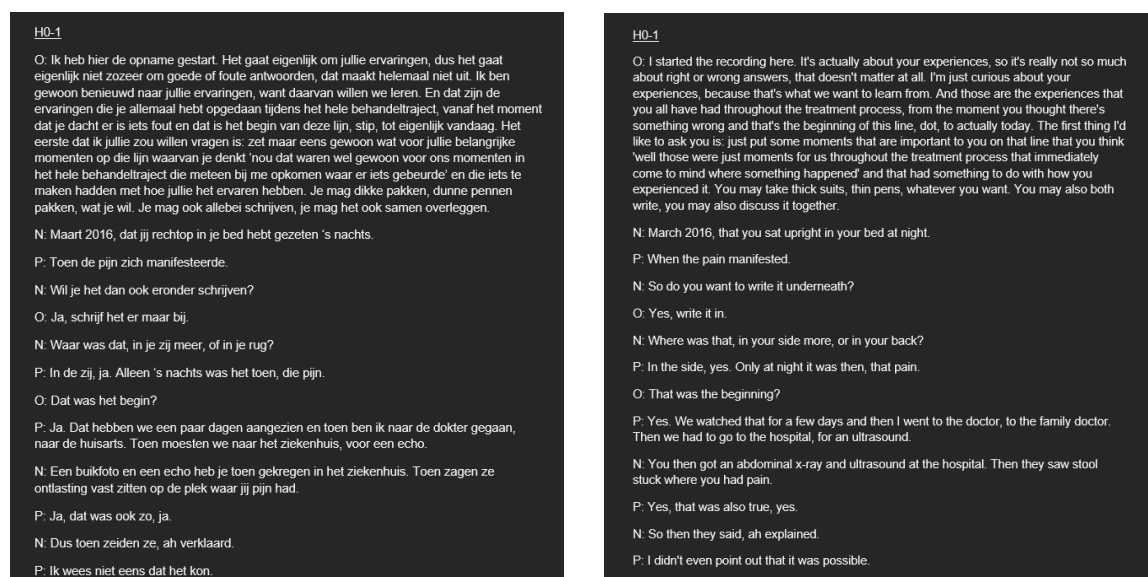


Figure 1: Snippets of the I0 interview in Dutch (left image) and English translation (right image)

with the shortest document containing 5,596 words (approximately 34,000 characters), and the longest document containing 12,875 words (approximately 58,000 characters). The total collection size is 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 and emotional support. 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. An example is shown in Figure 1.

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 many pre-trained domain-specific embedding models tend to perform best on English-language data or are even exclusively English (especially in the biomedical domain), which may impact the quality and consistency of the results when working with Dutch transcripts.

We examine 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. Design of Methods

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

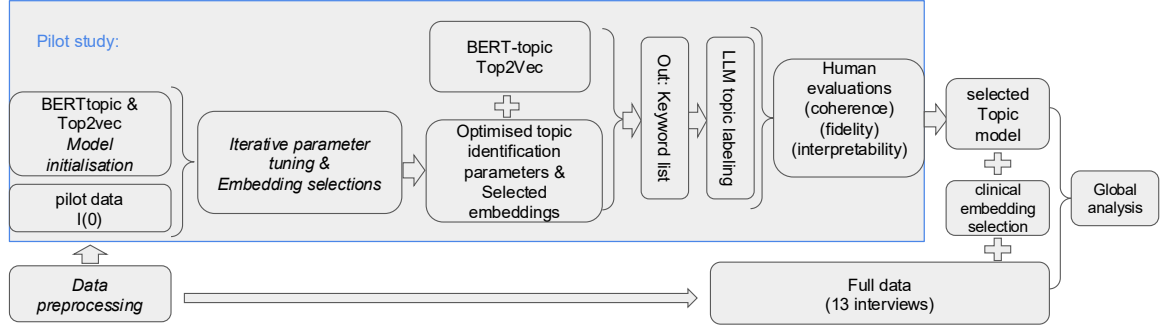


Figure 2: Framework Design of Overall Experimental Investigation

medical staff. To achieve this, BERTopic and Top2Vec are applied on a single anonymized cancer patient interview at a time, with the pilot-study interview being **Interview I0**, and their outputs are evaluated in order to determine their coherence and relevancy in a medical context, as shown in the system development framework in Figure 2.

Out of available topic modeling methods, we first select **BERTopic** 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.² 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 HDBSCAN as clustering algorithm, which determines the number of clusters based on the distribution and density of the embedded documents. For comparison purpose, we select **Top2vec** as an alternative neural topic model.³ 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 (Cer et al. 2018).

As shown in Figure 2, from bottom-left to right, the dataset is firstly 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 in terms of context length. In the Pilot-study, 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. Moreover, we tested several different embedding models as backbones at the “iterative parameter tuning and embedding selections” phase, but due to the differences between BERTopic and Top2Vec, the focus is placed on a common embedding model, namely **all-mpnet-base-v2** (Reimers and Gurevych 2020), in order to fairly evaluate their performance on as much common ground as possible. During this phase, 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.

Subsequently, the optimized topic identification parameters and selected embeddings (available within the two neural models) will be used for both neural topic models to produce keyword lists. We use an LLM to summarize topics from the keyword list into a descriptive label. Then human evaluations will be carried out on topic model selection by looking into features of coherence, fidelity, and interpretability.

2. available at <https://github.com/MaartenGr/BERTopic>

3. available at <https://github.com/ddangelov/Top2Vec>

Table 1: Preprocessing Pipeline

Preprocessing Step	Purpose
Conversion to plain text	Converting the document to a clean <code>.txt</code> file by using the <code>python-docx</code> package (Foundation 2024)
Document Translation	Translation from Dutch to English with DeepL (DeepL GmbH 2024) for the application of domain-specific backbone models and interpretation by non-Dutch speakers.
Speaker Label Removal	Remove speaker tags (e.g., <code>P:</code> , <code>N:</code>) that can confuse embedding models.
Section Header Removal	Remove structural markers (e.g., <code>I0-1</code>) that do not carry any semantic meaning.
Contraction Expansion	Prevent broken tokens (e.g., <i>wasn</i> , <i>t</i>) by expanding contractions (e.g., <i>wasn't</i> to <i>was not</i>).
Custom Stop Words List	Remove uninformative and meaningless words (e.g., “uh”, “yeah”, “says”).

For analyzing the entire interview data, we will further explore clinical domain embedding models, beyond the default embeddings from BERT-topic and Top2Vec themselves, together with the selected topic model framework from the pilot study, for a global analysis, in Section 7.

5. Pilot-study and Optimizing Topic Models

5.1 Data Preprocessing

As stated in Section 3, the source data consists of 13 anonymized cancer patient interviews stored as `docx` files, 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. Our preprocessing pipeline is listed in Table 1 with examples, including conversation to plain text conversion, document translation, speaker label removal, section header removal, contraction expansion, and removing custom stop words list.

For model development and pilot-testing, we first apply each model to a single 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, detailed in next sections separately for two neural models.

5.2 Model Development and Optimization: BERTopic

To manage the length of storytelling transcripts and ensure compatibility with embedding model input limits, we apply a sentence-based chunking strategy. After processing, the text is split into sentences using a regex-based approach, chosen for its simplicity and sufficient accuracy given the well-punctuated nature of the transcripts. These sentences are then grouped into fixed-size chunks, and different chunk sizes are tested to assess their impact on topic generation and coherence. Table 2 shows that the number of chunks directly influences the number of topics produced.

We start with BERTopic’s default settings and the all-mpnet-base-v2 embedding model, briefly testing alternatives like MiniLM-L6-v2, which we discard due to token limits and inconsistent outputs. We adjust UMAP parameters (`n_neighbors`, `min_dist`, `n_components`) to balance topic separation and cohesion, and tune HDBSCAN settings such as `min_cluster_size` and `min_samples` to control granularity and noise. The vectorizer is optimized to include bigrams, improving inter-

Table 2: 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

pretability by capturing multi-word expressions common in clinical narratives. Finally, we experiment with `min_topic_size`, finding the default value (10) best preserves both broad and specific themes without excessive fragmentation.

5.3 Model Development and Optimization: Top2Vec

To compare BERTopic and Top2Vec fairly, we consider both the original Top2Vec and its contextualized variant (C-Top2Vec) from Angelov and Inkpen (2024), which incorporates transformer-based embeddings such as `all-mpnet-base-v2` for richer, context-aware topics. However, we exclude C-Top2Vec from the final comparison because it lacks support for viewing representative documents (document-to-topic mapping), which is crucial for interpretability, manual verification, and the envisioned clinical support. We apply the same sentence-based chunking strategy to ensure both models receive identical input.

Similar to the tuning process employed for BERTopic, Top2Vec is optimized through iterative refinements of key parameters. This time, however, we start from the parameters of the already-tuned BERTopic model 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. We evaluate the same UMAP parameters explored during BERTopic tuning, namely `n_neighbors`, `min_dist` and `metric`, to observe their influence on topic formation. Similarly, we modify the HDBSCAN parameters such as `min_cluster_size` and `min_samples` in order to balance specificity and generalization within the topic clusters.

In our experiments, the default Top2Vec settings cause clustering errors, but tuning UMAP and HDBSCAN parameters resolves this and produces meaningful topics. Unlike BERTopic, Top2Vec *does not* allow custom n-gram or stop-word settings, limiting keyword flexibility, though contextual embeddings still yield relevant terms. After iterative refinements, the optimized Top2Vec models generate topics suitable for comparative analysis with BERTopic.

5.4 Topic Labeling with an LLM on Keyword Lists

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, we use a large language model (LLM) to label the topics in order to automate the process. The specific LLM model that we 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 (OpenAI 2024). At first, we 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.

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 absence of document context substantially impacts the quality of the generated labels, which is an impediment for individual-interview analysis. We therefore adapt the prompt *including document snippets* as additional *contextual* evidence to help interpret the topics more accurately. The final prompt 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:”.

6. Pilot-study Results and analysis

6.1 BERTopic Results

We first evaluate BERTopic on the preprocessed content of interview I0. The final model uses the **all-mpnet-base-v2** embedding model for semantic encoding, UMAP for dimensionality reduction, and HDBSCAN for clustering. Using the optimized 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. A comprehensive overview of the topics, along with their automatically generated labels and keywords, is presented in Table 3. The **all-mpnet-base-v2** embedding model manages to produce rich semantic representations that help to distinguish nuanced narratives.

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 (National Cancer Institute 2023) 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.

The pilot study also shows limitations of BERTopic. 1) 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

Table 3: Output of the Optimized BERTopic Model (Interview I0): pilot study

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

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. 2) 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.

6.2 Top2Vec Results

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.

Using the optimal 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 4. 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.⁴ While both C-Top2Vec and the original Top2Vec implementations are explored during the experimentation process, we 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.

6.3 Model Evaluation and Comparison with Human Perspectives

To evaluate the coherence, relevance, and clinical interpretability of the topics produced by BERTopic and Top2Vec, we conducted a small-scale user study involving three volunteer participants. Each participant was provided with the anonymized cancer interview I0 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. While the volunteers are not 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). They were also asked to identify any essential themes that the models may have missed. The outputs that they were provided with were Tables 3 and 4. 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.

The results for Q1 (topic coherence and usefulness) are in Figure 3. It shows that the ratings for BERTopic are on average higher and there are fewer topics that score very low compared to Top2Vec. Overall, BERTopic performed strongly, with 12 out of the 17 topics receiving a coherence score of 4.0 or higher. The results for Q3 (keyword descriptiveness) are in Figure 4. While Top2Vec achieved

4. 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.

Table 4: Output of the Optimized Top2Vec Model (Interview I0): pilot study

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, folfinrox, 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

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. 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**. 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 *potential* topics than BERTopic, it might fail to capture nuanced themes, which are crucial for the primary goal of this research.

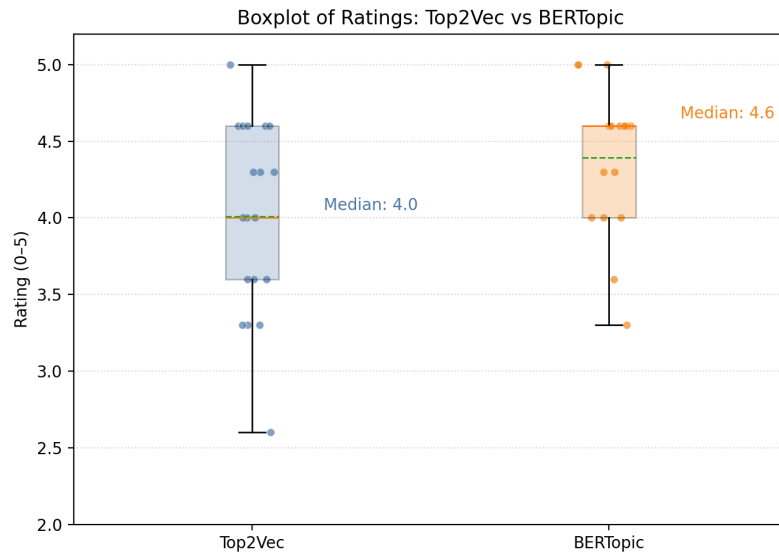


Figure 3: Distribution of ratings for Top2Vec and BERTopic topics (Q1), averaged over 3 raters. Note that Top2Vec has generated 19 topics and BERTopic has generated 17 topics.

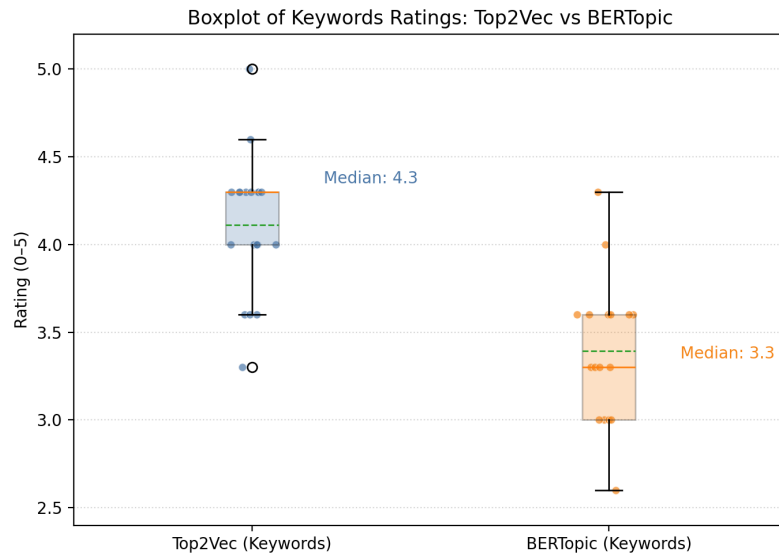


Figure 4: Distribution of ratings for Top2Vec and BERTopic keywords (Q3), averaged over 3 raters. Note that Top2Vec has generated 19 topics and BERTopic has generated 17 topics.

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

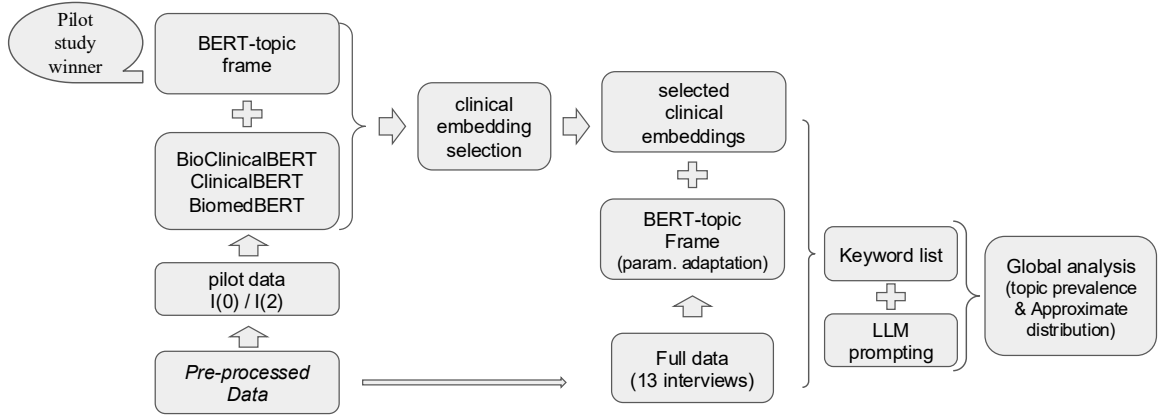


Figure 5: BERT-topic Deployment on all Interviews with Clinical LM Embeddings

sive topics from the I0 interview, the BERTopic model’s output is way more *precise* and contained *less overlap*.

Taking the evaluation process, as well as the experimentation and results, into consideration, we decide to focus the analysis on BERTopic further. Not only did BERTopic demonstrate that it can consistently produce topics of high overall quality, but it also offers greater flexibility and extensibility because it is compatible with a wide range of publicly available embedding models. This 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.

7. Investigating the Generalizability with Global Analysis

7.1 Domain-specific Clinical/Medical Embedding Model Selection

To explore whether clinically oriented embeddings improve BERTopic’s performance on patient storytelling data, we experimented with three widely used models on I(0) interview: **BioClinicalBERT** (Alsentzer et al. 2019), **ClinicalBERT** (Liu et al. 2025), and **MSR BiomedBERT** (formerly PubMedBERT) (Gu et al. 2020) for domain-specific embedding model selection, as shown in Figure 5. All experiments used the same BERTopic configuration as in Section 5.2 to ensure consistency. ClinicalBERT produced 17 topics initially, but coherence issues persisted even after tuning parameters such as chunk size and dimensionality reduction. 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 document chunks 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 is probably less aligned with narrative-style patient data.

In contrast, BioClinicalBERT consistently generated more interpretable and clinically meaningful topics, likely due to its pretraining on real clinical notes. Its outputs captured both technical medical experiences and patient perspectives with greater nuance, outperforming the other models in terms of coherence and relevance. These findings suggest that embedding models trained on clinical narratives provide *stronger semantic representations* for this type of data.

Further tests on shorter interviews (such as I(2), which is the shortest in the dataset at approximately 5596 words) revealed that chunk size also plays a critical role: Reducing the chunk size to 6 sentences increases the number of topics to I(2), 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. This indicates that a dynamic chunking strategy, combined with clinically oriented embeddings such as BioClinicalBERT, may be essential for optimizing topic modeling across datasets with varying lengths and complexities.

7.2 Global Analysis on All Interview Data

7.2.1 MODEL SETUP

As shown in Figure 5 (middle-horizontal), to gain deeper insights into the full dataset of 13 interviews, we applied topic modeling to the entire corpus rather than individual interviews, aiming to uncover overarching themes and recurring patterns. For this global analysis, we used **BioClinicalBERT**, identified in Section 7.1 as the most effective domain embedding model for producing coherent and clinically meaningful topics. Its domain-specific training on biomedical and clinical text makes it well-suited for capturing general themes. Unlike the per-interview approach, which prioritized fine-grained topics, this phase focused on **broader themes**.

For model configuration, we increased the chunk size from 6 to 7 sentences to provide more *context* per document, as larger chunks tend to yield fewer but more generalized topics. The BERTopic pipeline was tuned/adapted accordingly: UMAP parameters were adjusted to promote broader clustering (`n_neighbors = 16`, `min_dist = 0.2`, `n_components = 4`), and HDBSCAN settings were modified to favor stable clusters (`min_cluster_size = 11`, `cluster_selection = eom`). These changes bias the model toward extracting higher-level themes that recur across interviews. We also revised the LLM-prompt labeling process to suit global themes. Instead of using representative documents, the new LLM prompt focuses *exclusively on keywords* for each topic:

“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.”

This approach ensures more objective and consistent labels for overarching themes. Additionally, we expanded the stop word list to reduce noise, though translation errors and linked words across interviews made perfect filtering impractical without manual review.

7.2.2 MODEL OUTPUT: TOPIC ANALYSIS

Using the above configuration and parameters, we fitted the topic model on the entire corpus of 13 interviews. The resulting output (Table 5) reveals overarching themes that span all interviews, offering a broad view of the cancer treatment experience. Unlike granular per-interview models, which capture individual nuances, the global model identifies recurrent concerns and *shared points across patients*, enabling detection of systemic issues and common emotional or physical pain points. For example, **Topic 0** includes keywords such as “oxycontin”, “medications”, and “nausea”, highlighting widespread struggles with medication side effects. **Topic 2** reflects how patients recall surgical interventions, specifically keyhole surgery, while **Topic 7** clusters references to timelines and appointments, suggesting opportunities for visual treatment timelines. Similarly, **Topic 4** (Sleep Patterns and Nighttime Activities) emphasizes recurring sleep-related discussions, indicating that sleep is a significant aspect of the cancer journey and warrants deeper analysis to uncover common issues affecting rest and well-being.

Table 5: BERT-topic Global Analysis using BioClinicalBERT embedding on 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, live, 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

Other topics capture *institution*-specific and emotional insights. **Topic 9** focuses on “Support and Resources for Cancer Care at Erasmus MC”, referencing amenities like food, lighting, and buildings. These are elements that, while non-clinical, influence patient experience and could inform facility improvements. *Emotional* responses also emerge prominently: **Topic 8**, labeled “Coping with Treatment Setbacks and Emotional Reactions”, groups keywords related to coping mechanisms, underscoring the need for mental health integration in oncology care. Finally, **Topic 12** (Navigating Treatment Decisions with Specialist *Nurses*) highlights the critical role of nurses in guiding patients through complex choices, providing both clinical clarity and emotional reassurance. Overall, the globally tuned model captures high-level patterns across medical procedures, emotional resilience, logistics, and support systems, complementing per-interview analyses. While minor issues like repetitive keywords (e.g., “day day” in **Topic 0**) persist, they are less pronounced, confirming the effectiveness of this broader modeling approach.

7.2.3 TOPIC DISTRIBUTION IN THE WHOLE CORPUS

Figure 6 shows the distribution of topics over interviews. Rather than assigning a single topic to each chunk, the distribution is based on the soft output of the BERTopic model.⁵ This allows for the possibility that a chunk touches on multiple themes to varying degrees. By averaging these probability distributions across all chunks 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

5. 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.

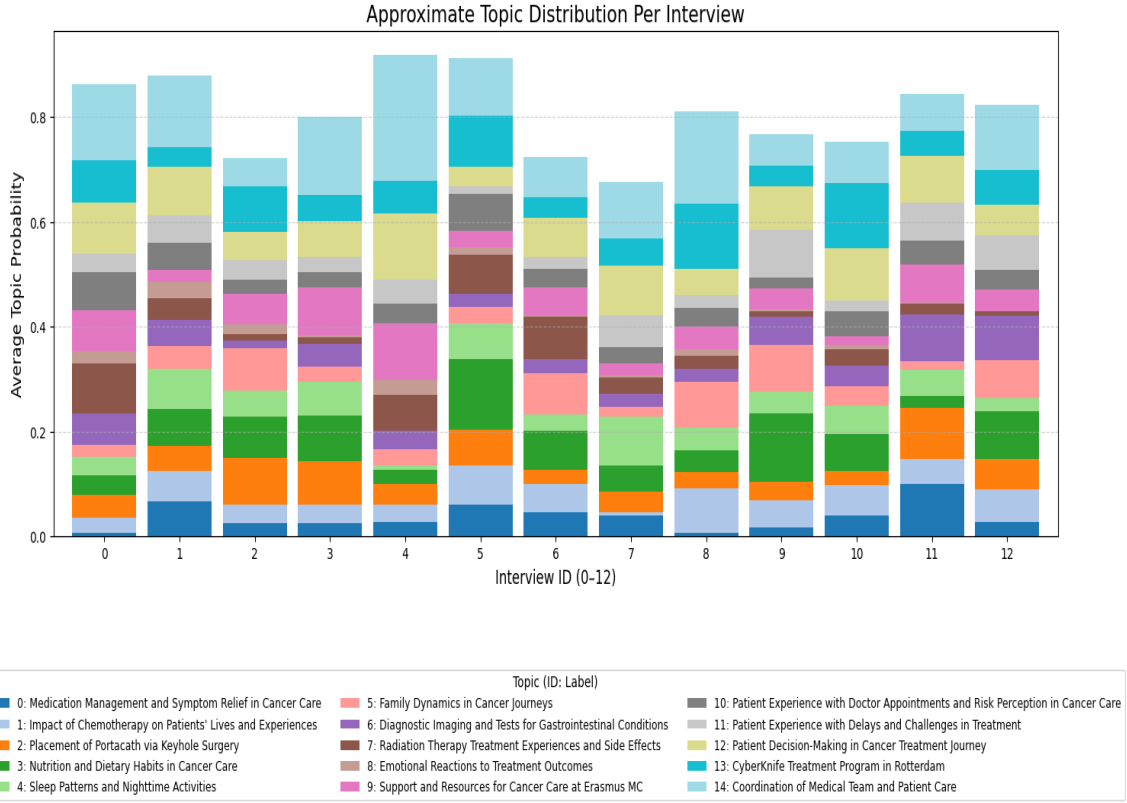


Figure 6: Approximate distribution across all 13 interviews, where each color represents a topic, and each bar represents one of the 13 interviews.

is useful for capturing **thematic overlap** and uncertainty that may be missed with hard topic assignments. Table 6 lists the 5 highest probability topics across all 13 interviews.

The results indicate that **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.

8. Discussion

The experiments conducted in this study provide insight into the capabilities of BERTopic and Top2Vec for extracting clinically relevant themes from cancer patient interviews. Overall, BERTopic – particularly when configured with clinically oriented embeddings and sentence-based chunking – demonstrated a stronger ability to capture nuanced topics. These findings suggest that topic mod-

Table 6: Most Occurring Topics Overall (Approx. Distribution)

Topic ID	Mean Avg. Probability	Topic Label
14	0.118	Coordination of Medical Team and Patient Care
12	0.079	Patient Decision-Making in Cancer Treatment Journey
3	0.071	Nutrition and Dietary Habits in Cancer Care
13	0.070	CyberKnife Treatment Program in Rotterdam
2	0.053	Placement of Portacath via Keyhole Surgery

eling could serve as the basis for clinical feedback tools, enabling medical staff to navigate lengthy, unstructured patient documents efficiently and focus on *patient-centered* care. Comparing BERTopic and Top2Vec revealed clear differences: while both produced coherent topics, BERTopic consistently generated more precise and less overlapping themes, which is critical in medical contexts, while Top2Vec generated more descriptive key terms. The ability to leverage domain-specific embeddings such as BioClinicalBERT further improved performance, confirming that embedding selection plays a key role in topic coherence and relevance.

We would like to point out a few limitations. First, topic modeling proved highly sensitive to chunking strategies: smaller chunks increased granularity but fragmented narratives, while larger chunks preserved context but reduced specificity. This trade-off between granularity and context preservation proves to be a recurring theme in the experimental setup for both BERTopic and Top2Vec. Second, while substantial effort was made to optimize the processing of the data, 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. Third, our evaluation survey was small-scale and lacked clinical experts, meaning that judgments were based on interpretability rather than professional applicability. Dataset characteristics introduced further challenges: interviews were originally in Dutch and translated into English, which may have altered meaning. These issues underscore broader resource limitations for non-English clinical NLP. Despite these challenges, the study demonstrates the potential of topic modeling as a foundation for clinical support tools that streamline workflows and amplify patient voices.

9. Conclusions and Future Work

Overall, this study highlights the value of embedding-based topic modeling for clinical applications on patient care. By automating the extraction of meaningful themes from patient narratives, tools like BERTopic could enhance workflow efficiency, awareness of patient concerns, and strengthen patient-doctor communication.

In response to the first research 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 research 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.

While this work remains a proof of concept, it lays the foundation for future research focused on multilingual capabilities, contextual models, improved evaluation frameworks, and integration into clinical decision-support systems. Such advancements could help transform narrative data into actionable insights, placing patient voices at the center of healthcare delivery.

For future work, it will be important to test the system with clinical experts to better understand how the generated topics can be applied/integrated 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 English and other 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 language 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 (*temporal* dimension), and offer solutions to help clinical staff solve or minimize common complaints from cancer patients.

10. Acknowledgment

We thank Ida Korfage and Sheila Payne for the valuable comments on the Abstract version of this work (presented at CLIN2025). We thank Anne Kuld-Nielsen for valuable feedback and encouragement from reading the manuscript.

This work has been funded by the European Union under Horizon Europe Work Programme 101057332. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The UK team are funded under the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number: 10041120.

References

- Albalawi, Rania, Tet Hin Yeap, and Morad Benyoucef (2020), Using topic modeling methods for short-text data: A comparative analysis, *Frontiers in Artificial Intelligence* **3**, pp. 42. <https://www.frontiersin.org/articles/10.3389/frai.2020.00042/full>.
- Alon, Yaniv, Etti Naimi, Chedva Levin, Hila Videl, and Mor Saban (2025), Leveraging natural language processing to elucidate real-world clinical decision-making paradigms: A proof of concept study, *Journal of Biomedical Informatics* **136**, pp. 104829. <https://doi.org/10.1016/j.jbi.2025.104829>.

- Alsentzer, Emily, John Murphy, William Boag, Wei-Hung Weng, Yuhao Jin, Tristan Naumann, and Matthew McDermott (2019), Publicly available clinical bert embeddings, *arXiv preprint arXiv:1904.03323*.
- Angelov, Dimo (2020), Top2vec: Distributed representations of topics, *arXiv preprint arXiv:2008.09470*.
- Angelov, Dimo and Diana Inkpen (2024), Topic modeling: Contextual token embeddings are all you need, in Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, Miami, Florida, USA, pp. 13528–13539. <https://aclanthology.org/2024.findings-emnlp.790/>.
- Bak, Marieke, Laura Hartman, Charlotte Graafland, Ida J Korfage, Alena Buyx, Maartje Schermer, 4D PICTURE Consortium, et al. (2025), Ethical design of data-driven decision support tools for improving cancer care: embedded ethics review of the 4d picture project, *JMIR cancer* **11** (1), pp. e65566, JMIR Publications Inc., Toronto, Canada.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003), Latent dirichlet allocation, *J. Mach. Learn. Res.* **3** (null), pp. 993–1022, JMLR.org.
- Cer, Daniel, 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 (2018), Universal sentence encoder, *arXiv preprint arXiv:1803.11175*.
- DeepL GmbH (2024), DeepL Translator. Accessed: 2025-05-13.
- Demner-Fushman, Dina, Wendy W. Chapman, and Clement J. McDonald (2009), What can natural language processing do for clinical decision support?, *Journal of Biomedical Informatics* **42** (5), pp. 760–772. Biomedical Natural Language Processing. <https://www.sciencedirect.com/science/article/pii/S1532046409001087>.
- Egger, Roman and Jie Yu (2022), Interpretable topic modeling for social media analysis: Covid-19 and travel on twitter, *Frontiers in Sociology* **7**, pp. 850586.
- Erasmus MC (2025), Erasmus university medical center. Accessed: 2025-05-12.
- Foundation, Python Software (2024), python-docx 0.8.11 documentation. Accessed: 2025-05-13. <https://python-docx.readthedocs.io/en/latest/>.
- Ghosh, Subhajit (2024), Natural language processing: Basics, challenges, and clustering applications, *Federated learning for Internet of Vehicles: IoV Image Processing, Vision and Intelligent Systems*, Bentham Science Publishers, pp. 61–82.
- Griciūtė, Bernadeta, Lifeng Han, and Goran Nenadic (2023), Topic modelling of swedish newspaper articles about coronavirus: a case study using latent dirichlet allocation method, *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pp. 627–636.
- Griffioen, Ingeborg, Marijke Melles, Anne Stiggelbout, and Dirk Snelders (2017), The potential of service design for improving the implementation of shared decision-making, *Design for Health* **1** (2), pp. 194–209, Taylor & Francis.
- Griffioen, Ingeborg PM, Judith AC Rietjens, Marijke Melles, Dirk Snelders, Marjolein YV Homs, Casper H van Eijck, and Anne M Stiggelbout (2021), The bigger picture of shared decision making: a service design perspective using the care path of locally advanced pancreatic cancer as a case, *Cancer Medicine* **10** (17), pp. 5907–5916, Wiley Online Library.

- Grootendorst, Maarten (2022), Bertopic: Neural topic modeling with class-based tf-idf, *arXiv preprint arXiv:2203.05794*.
- Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon (2020), Domain-specific language model pretraining for biomedical natural language processing.
- Han, Lifeng, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic (2024), Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning, *Frontiers in Digital Health* **6**, pp. 1211564, Frontiers Media SA.
- Kidanemariam, Martha, Matthijs A Graner, Willem Jan W Bos, Marielle A Schroijen, Eelco JP de Koning, Anne M Stiggelbout, Arwen H Pieterse, and Marleen Kunneman (2024), Patient-clinician collaboration in making care fit: A qualitative analysis of clinical consultations in diabetes care, *Patient Education and Counseling* **125**, pp. 108295, Elsevier.
- Liu, Xiaohong, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. (2025), A generalist medical language model for disease diagnosis assistance, *Nature medicine* **31** (3), pp. 932–942, Nature Publishing Group US New York.
- National Cancer Institute (2023), Folfirinox. Accessed: 2025-05-17. <https://www.cancer.gov/about-cancer/treatment/drugs/folfirinox>.
- Ohno, Yukiko, Tohru Aomori, Tomohiro Nishiyama, Riri Kato, Reina Fujiki, Haruki Ishikawa, Keisuke Kiyomiya, Minae Isawa, Mayumi Mochizuki, Eiji Aramaki, and Hisakazu Ohtani (2025), 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*, JMIR Publications. <https://www.jmir.org/2025/1/e68863>.
- OpenAI (2024), Gpt-4o mini: advancing cost-efficient intelligence. Accessed: 2025-05-12. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- Raj, Minakshi, Kerry Ryan, Philip Sahr Amara, Paige Nong, Karen Calhoun, M Grace Trinidad, Daniel Thiel, Kayte Spector-Bagdady, Raymond De Vries, Sharon Kardia, et al. (2023), Policy preferences regarding health data sharing among patients with cancer: public deliberations, *JMIR cancer* **9** (1), pp. e39631, JMIR Publications Inc., Toronto, Canada.
- Reimers, Nils and Iryna Gurevych (2020), Sentence-bert: all-mpnet-base-v2. Accessed: 2025-05-17.
- Ren, Libo, Yee Man Ng, and Lifeng Han (2025), Malei at multiclinsum: Summarisation of clinical documents using perspective-aware iterative self-prompting with llms, *BioASQ WS at CLEF 2025*.
- Sheikhalishahi, Sara, Riccardo Miotto, Joel Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani (2019), Natural language processing of clinical notes on chronic diseases: Systematic review, *JMIR Medical Informatics* **7** (2), pp. e12239. <https://medinform.jmir.org/2019/2/e12239>.
- Somani, Sulaiman, Marieke Van Buchem, Ashish Sarraju, Tina Hernandez-Boussard, and Fatima Rodriguez (2023), Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure (ai-prem), *JAMA Netw Open* **6** (4), pp. e239747. <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2803988>.

TU Delft (2021), Better prepared for big decisions in healthcare. Accessed: 2025-05-12.

Van Buchem, Marieke, Olaf M. Neve, Ilse M. J. Kant, Ewout W. Steyerberg, Hileen Boosman, and Erik F. Hensen (2022), Artificial intelligence-enabled analysis of statin-related topics and sentiments on social media, *BMC Medical Informatics and Decision Making*. <https://doi.org/10.1186/s12911-022-01923-5>.

Wahbeh, Abdullah, Mohammad Al-Ramahi, Omar El-gayar, Ahmed Elnoshokaty, and Tareq Nasrallah (2025), Evaluating topic models with openai embeddings, *Technical report*, University of Hawai'i at Mānoa. <https://scholarspace.manoa.hawaii.edu/items/4fab27e3-5fad-412c-8f31-0c6f4de3c25d>.

Wang, Yanshan, Liwei Wang, Majid Rastegar-Mojarad, et al. (2018), Clinical information extraction applications: A literature review, *Journal of Biomedical Informatics*.

Appendix A. Volunteer Survey Questionnaire

This section contains the volunteer survey questionnaire, analyzed in Section 6.3, which we 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 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.