

# Framework for NLP-Driven Insights into Patient-Centered Communication in Oncology

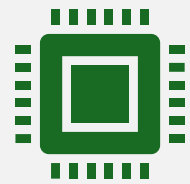
Grace Donovan

MSDE/MSDS Practicum 1  
Regis University, Denver CO  
June 26, 2025

# Overview



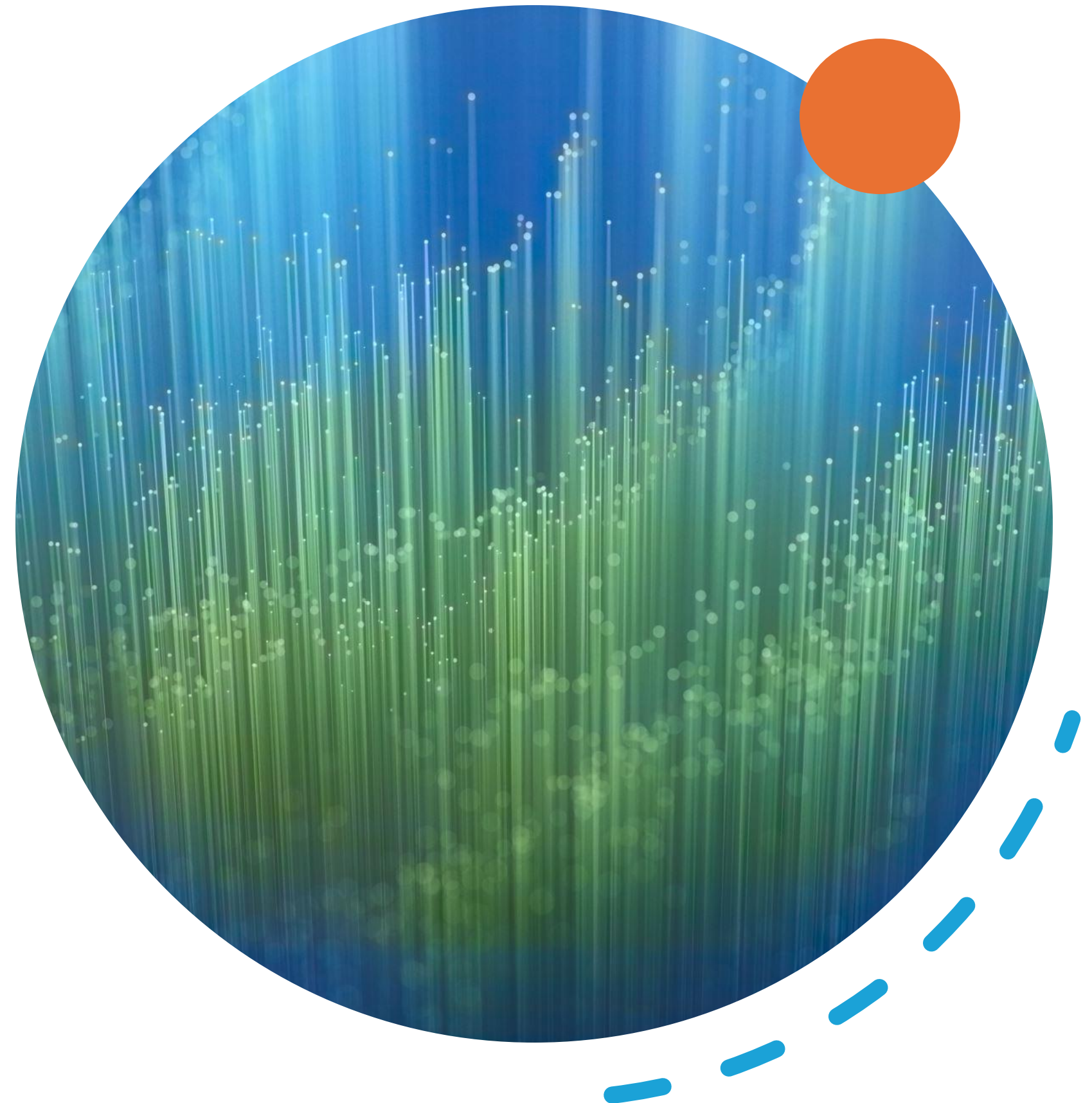
Introduction: Background, Objective, and Key Questions



Framework overview: Data Retrieval, Data Processing, and Data Analysis



Conclusion: Insights, Future Improvements



# Background

1

## Patient-Centered Care

Patient-focused communication is vital - driving treatment adherence, boosting satisfaction, and improving health outcomes (Sharkiya (2023), Krist et al. (2017), Becker, C., et al. (2021)).

2

## Research Challenges

Synthesizing evolving communication practices and associated attitudes from the rapidly expanding medical literature can be difficult, complicating efforts to uphold high-quality patient care (Links, M., et al. (2020)).

3

## Automated Insights

Automation of literature harvesting and analysis, can provide clinicians, researchers, and policymakers with real-time, data-driven decision support (van Dinter, R., et al. (2021)).

“Effective communication is a cornerstone of quality healthcare. Communication helps providers bond with patients, forming therapeutic relationships that benefit patient-centered outcomes.”  
- Samer H. Sharkiya (2023)





## Objective

Develop framework using open-source tools to automate the synthesis and analysis of literature in an evolving research landscape



## Topical Focus

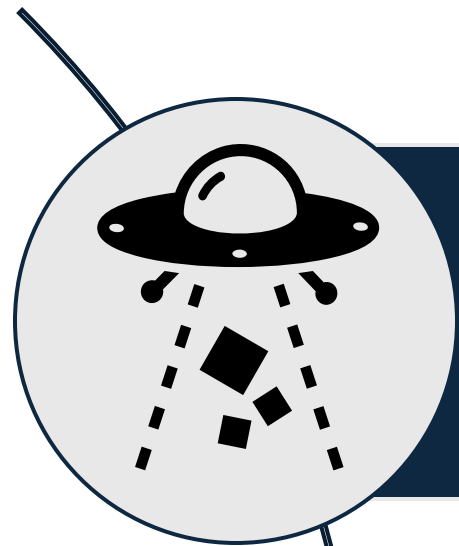
Exploring patient-centered communication themes in peer-reviewed oncology literature



## Key Questions

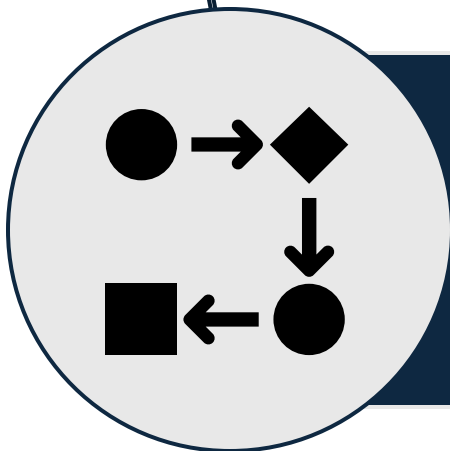
1. Which patient-centered communication topics emerge most prevalent in oncology research?
2. How does author sentiment vary across these topics?

# Framework Overview



## 1. Data Retrieval

- Related publication identification
- Publication full-text data scraping



## 2. Data Processing

- Cleaning, standardizing, sub-setting, and transforming scraped HTML data to JSON dictionary
- Preprocessing cleaned text data for analysis (stop word removal, POS tagging, lemmatization, etc.)



## 3. Data Analysis

- Exploratory data analysis using word frequency distributions and n-grams
- Aggregating topic modeling for theme identification and sentiment analysis for sentiment classification of identified themes

# Related Publication Identification and Initial Retrieval using xDD API

## Component 1 – Data Retrieval

- **What is xDD?**

- A cyberinfrastructure that compiles data on published literature and provides users with the ability to perform full text searches of published literature through the xDD API (Peters et al. 2025).

- **Workflow:**

- **Keyword Query:** publications containing **both** “patient-centered communication” and “cancer” keywords
- **Filter:** articles with  $\geq 2$  total keyword hits, published between 2020 and 2025.
- **Extracted Fields:** Publication name, publisher name, title, date, DOI (Digital Object Identifier), author names, text snippets (“highlights”) where keywords appeared, and count of keyword occurrences per document
- **Output:** Consolidated 130 publication metadata entries into a single JSON file for downstream analysis or visualization.

```
1_data_retrieval > data > {} xdd_pubs.json > {} 128 > # hits
1  [
2      {
3          "pubname": "Supportive Care in Cancer",
4          "publisher": "Springer",
5          "title": "Oncology patients\u2019 communication experiences during COVID-19",
6          "publication_date": "2022 06",
7          "doi": "10.1007/s00520-022-06897-8",
8          "authors": "Street, Richard L; Treiman, Katherine; Kranzler, Elissa",
9          "highlight": [
10             "The <em class=\"hl\">Patient</em>-<em class=\"hl\">Centered</em> communication",
11             "<em class=\"hl\">Patient</em>\u2019-centered communication",
12             "<em class=\"hl\">Patient</em>-centered communication",
13             "<em class=\"hl\">Patient</em>-centered communication",
14             "Psychometric evaluation and design of <em class=\"hl\">patient-centered",
15         ],
16         "hits": 5
17     },
18     {
19         "pubname": "Supportive Care in Cancer",
20         "publisher": "Springer",
21         "title": "Hodgkin lymphoma survivor perspectives on their engagement in research",
22         "publication_date": "2022 02",
23         "doi": "10.1007/s00520-021-06538-6",
24         "authors": "Murphy-Banks, Rachel; Kumar, Anita J.; Lin, Mingqian; Saviola, David",
25         "highlight": [
26             "Implications for <em class=\"hl\">cancer</em> survivors\u2019",
27             "Keywords\u2019 Late effects \u2019 Hodgkin lymphoma \u2019",
28             "<em class=\"hl\">Patient</em>-<em class=\"hl\">centered</em> communication",
29             "This contributes to our position that more attention on <em class=\"hl\">patient-centered",
30             "Epstein RM, Street RL, Jr (2007) <em class=\"hl\">Patient</em>-centered",
31         ],
32         "hits": 5
33     },
34     {
35         "pubname": "Journal of Perinatology",
36         "publisher": "Springer",
37         "title": "Communication between neonatologists and parents when prognosis is uncertain",
38         "publication_date": "2020 09",
39         "doi": "10.1038/s41372-020-0673-6",
40         "authors": "Drach, Laura L.; Hansen, Debra A.; King, Tracy M.; Sibinga, Elizabeth",
41         "highlight": [
42             "Study design Guided by the National <em class=\"hl\">Cancer</em> Institute",
43             "<em class=\"hl\">centered</em> communication",
44             "The National <em class=\"hl\">Cancer</em> Institute (NCI) <em class=\"hl\">centered",
45             "The <em class=\"hl\">Patient</em> <em class=\"hl\">Centered</em> communication",
46             "<em class=\"hl\">Patient</em>-<em class=\"hl\">centered</em> communication",
47         ],
48         "hits": 5
49     },
50 ]
```

# Full-text Data Scrapping of Related Publication Webpages

## Component 1 – Data Retrieval

Constructed publication DOI URLs from DOI identifiers collected from xDD API

Used Selenium WebDriver and BeautifulSoup for scraping full-text from publication webpages

Specific sections targeted for scraping (abstract, methods, results, conclusion, etc.)

Filters included for handling inconsistencies and scraping issues

89 publications successfully scraped and saved as HTML files





## Component 2 – Data Processing

# 89 HTML Files (Raw Text Data)

```

Lama_revised.docx pages 7 to 1000_050202-02-0708-03mm-1 G#2
    <-Abstract>
        >>Standard reliability reports (SRR) are designed to communicate information between doctors. With many patients having instantaneous access to SRRs on patient portals, interpretation without guidance from doctors can cause anxiety and potential harm. In this pilot study, we designed a patient-centred prostate MRI template report (PACSR) to address some of these challenges and tested whether PACSRs improve patient knowledge and experience.
        >>Patients booked for clinical prostate MRI were randomly assigned to SRR or SRR + PACSR. Questionnaires included multiple-choice (that targeted a domain understanding, usefulness, next steps, emotional experience) hypothesized to improve with patient-centred reports and short answer questions, testing knowledge regarding MRI results. Clinical encounters were observed and recorded to explore whether adding practical PACSR improved communication. Likert scaled-responses and short-answer questions were compared using Mann-Whitney U test and Kruskal-Wallis test.
        >>Of the 48 participants, the majority were MRI naïve (76%). Patients receiving a PACSR had higher scores in the categories of patient understanding (mean: 4.37 vs. 3.39, p<0.001), confidence in their doctor's advice (mean: 4.74 vs. 3.47, p<0.001) and identifying next steps (mean: 3.48 vs. 2.87, p<0.001) but not emotional experience (mean: 4.04 vs. 3.79, p=0.22). PACSR participants found the layout and design more patient friendly (mean: 4.74 vs. 2.61, p<0.001) and easier to understand (mean: 4.37 vs. 2.36, p<0.001). In the knowledge section, overall, the PACSR was answered correctly 87% vs. 50%, p<0.001>.
        >>Conclusion:
            >>With the addition of practical MRI PACSRs, participants had better understanding of their results and felt more prepared to involve themselves in discussions with their doctor.
    <-Materials and methods>
        >>Patients booked for clinical prostate MRI were randomly assigned to SRR or SRR + PACSR. Questionnaires included multiple-choice (that targeted a domain understanding, usefulness, next steps, emotional experience) hypothesized to improve with patient-centred reports and short answer questions, testing knowledge regarding MRI results. Clinical encounters were observed and recorded to explore whether adding practical PACSR improved communication. Likert scaled-responses and short-answer questions were compared using Mann-Whitney U test and Kruskal-Wallis test.
        >>Results:
            >>Of the 48 participants, the majority were MRI naïve (76%). Patients receiving a PACSR had higher scores in the categories of patient understanding (mean: 4.37 vs. 3.39, p<0.001), confidence in their doctor's advice (mean: 4.74 vs. 3.47, p<0.001) but not emotional experience (mean: 4.04 vs. 3.79, p=0.22). PACSR participants found the layout and design more patient friendly (mean: 4.74 vs. 2.61, p<0.001) and easier to understand (mean: 4.37 vs. 2.36, p<0.001). In the knowledge section, overall, the PACSR was answered correctly 87% vs. 50%, p<0.001>.
            >>Conclusion:
                >>With the addition of practical MRI PACSRs, participants had better understanding of their results and felt more prepared to involve themselves in discussions with their doctor.

```

[illegible][illegible]

```

6      implementation into clinical practice, therefore reaching more patients. >?
7      >?Methods<?/2>
8      >?Results<?/2>
9      >?Nine clinical trials, 29 observational studies, and 1 case study were identified. The articles were categorized into one area within Epstein and Street's areas of communication. Many of the articles examined the patient's and provider's acceptability and feasibility of technology-based methods of communication, while other articles examined their efficacy. >?
10     >?Conclusions<?/2>
11     >?While research studies were identified in each of the areas of communication, the majority of technology-based communication strategies were focused on the exchange of information between patients and their providers. Further research and the development of technology-based communication interventions assessed through clinical trials are needed in the areas of healing relationships and making decisions in cancer care. Additionally, the communication strategies found effective at improving health outcomes in advanced cancer should begin implementation into clinical practice, therefore reaching more patients. >?

```

## 1 JSON File (Standardized, Cleaned, and Formatted Text Data)

```
2_data_processing > data > {} pub_content_filtered.json > {} > doi
1 {
2 {
3 "doi": "10.1016/j.jpsychores.2021.110440",
4 "conclusion": "Physiological responses of doctors predicted patients' responses differently depending on relationship length. Importantly by influencing patients' physiological responses on a moment-to-moment basis, doctors may have even more influence over patients' physiology and health than previously known."
5 },
6 {
7 "doi": "10.1007/s00520-020-05615-6",
8 "conclusion": "Uninsured cancer patients with low educational attainment have higher SC-needs and receive lower quality of PCC than their counterparts. Health services should face these challenges to reduce inequalities in Mexico."
9 },
10 {
11 "doi": "10.1016/j.jvir.2021.02.026",
12 "conclusion": "Although the total number shows an upward trend, women interventional radiologists are still underrepresented. Education level, geographic areas, and other socioeconomic factors may simultaneously influence the population size of women interventional radiologists in China."
13 },
14 {
15 "doi": "10.1111/eje.12477",
16 "conclusion": "Either workshops or training programmes with a combination of teaching techniques were effective in terms of enhancing their OHL or PCC. The more frequent follow-up might increase the long-term effectiveness of the learning programme."
17 },
18 {
19 "doi": "10.1016/j.jneb.2021.07.013",
20 "conclusion": "Nutrition interventions tailored toward nutrition literacy deficits may play an important role in improving patient diet behaviors."
21 },
22 {
23 "doi": "10.1016/j.jss.2021.09.006",
24 "conclusion": "More frequent interpreting services per day during peri-operative admission are associated with shorter length of stay in adjusted analysis. The findings merit further study in an intervention to increase use of interpreting services for surgical patients with limited English proficiency to study the impact of increased frequency of culturally competent care."
25 },
26 {
27 "doi": "10.1016/j.clinimag.2021.01.042",
28 "conclusion": "In this cohort, there was a high rate of non-significant incidental findings and normal further investigations. However, these risks are likely to be outweighed by the high number of cancer diagnoses and significant non-cancer findings."
29 },
30 {
31 "doi": "10.1007/s00520-021-06766-w", Generate code: \sEnter
32 "conclusion": "Oncology nurses believed that the four SCP components were helpful to the long-term management of CRC survivors, supported SCP provision, and expressed their perceived responsibilities for preparing and delivering SCPs. The findings suggested opportunities for oncology nurses to play a significant role in developing and implementing SCPs. However, additional efforts are needed to expand nurses' roles in survivorship care and establish practice guidelines that will facilitate integration of SCPs into nursing practice."
33 },
34 {
35 "doi": "10.1038/s41372-020-0673-6",
36 "conclusion": "Families and neonatologists value principles of patient centered communication but report challenges implementing this practice. Incorporating a multidisciplinary approach in settings of prognostic uncertainty to foster patient centered communication, may enhance communication surrounding NICU care."
37 },
38 }
```



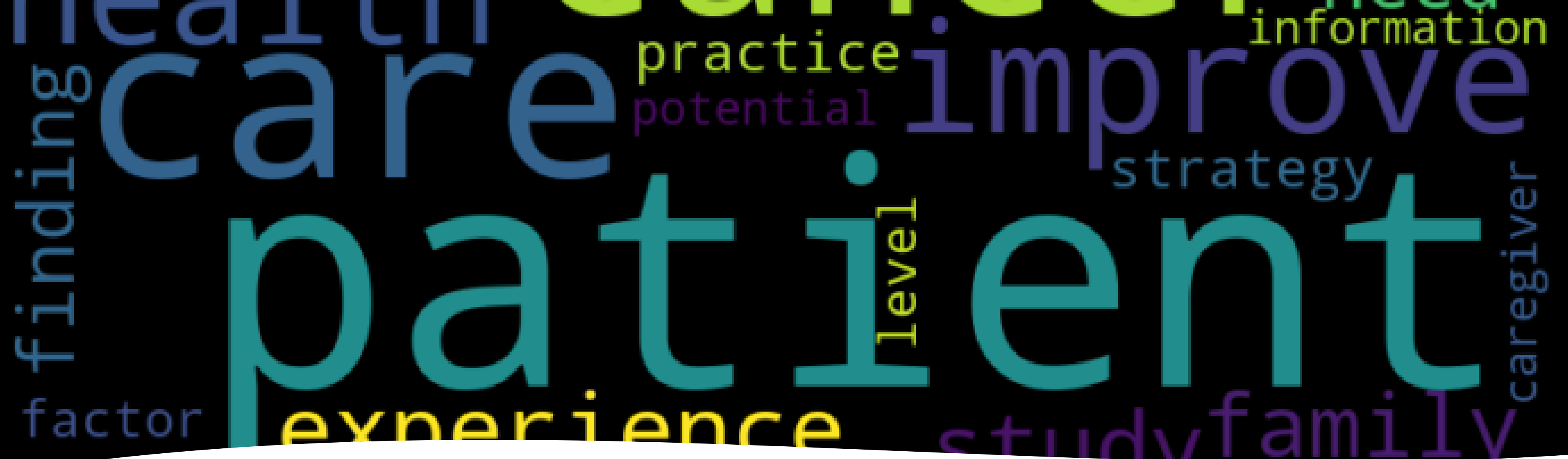
# Preprocessing Conclusion Data for Analysis

Component 2 – Data Processing

## Process:

- **Tokenization:** Sentence → word level
- **Lowercasing:** Normalize to lowercase
- **Stop word removal:** stop words + domain-specific terms
- **Length filtering:** Remove words <3 characters
- **POS filtering:** Remove adverbs (less contextually relevant)
- **Lemmatization:** Reduce words to root forms





# Exploratory Data Analysis

Component 3 –  
Data Analysis

## Word Frequency Analysis

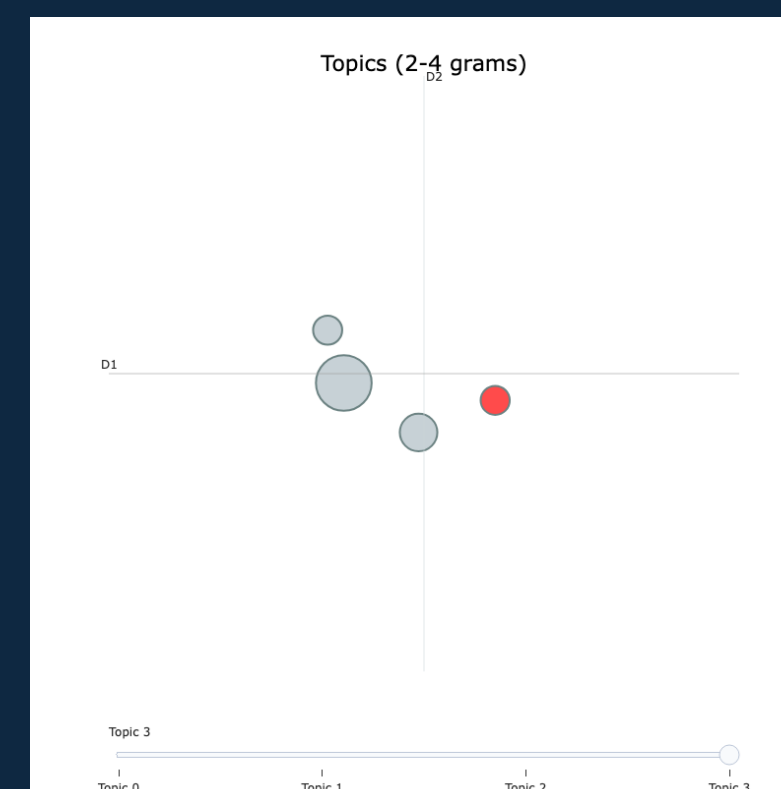
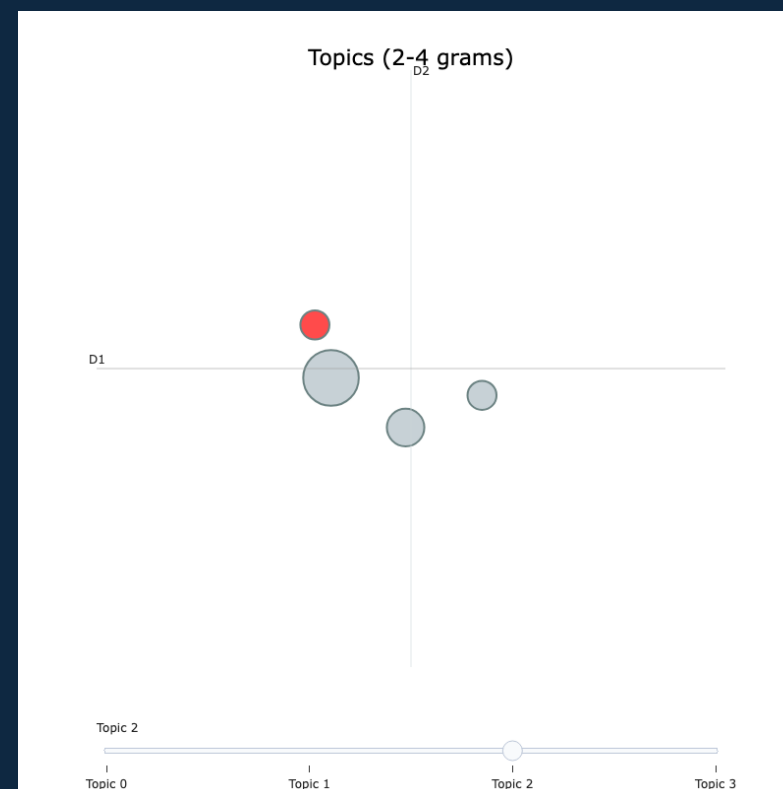
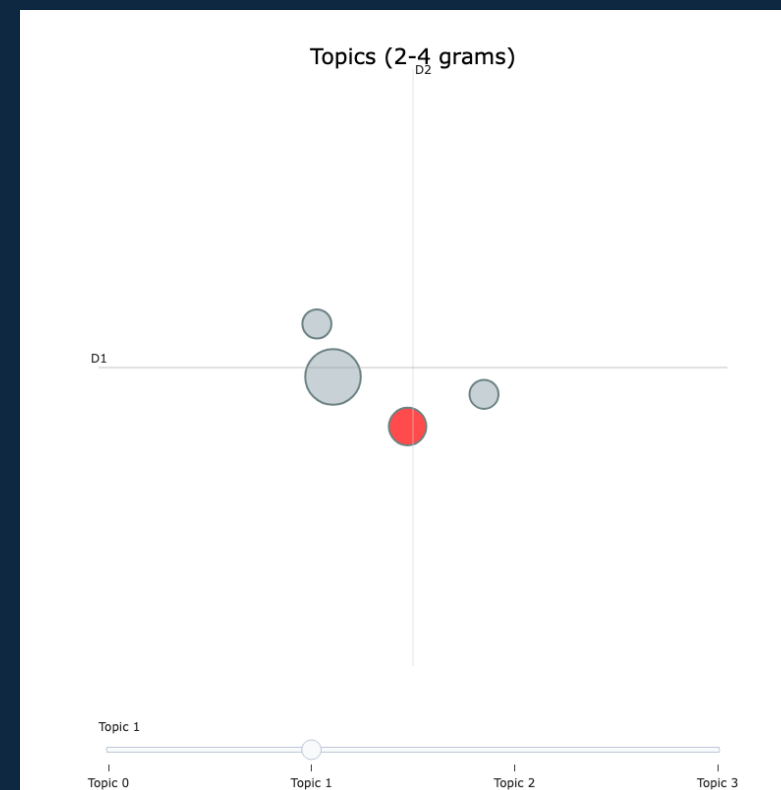
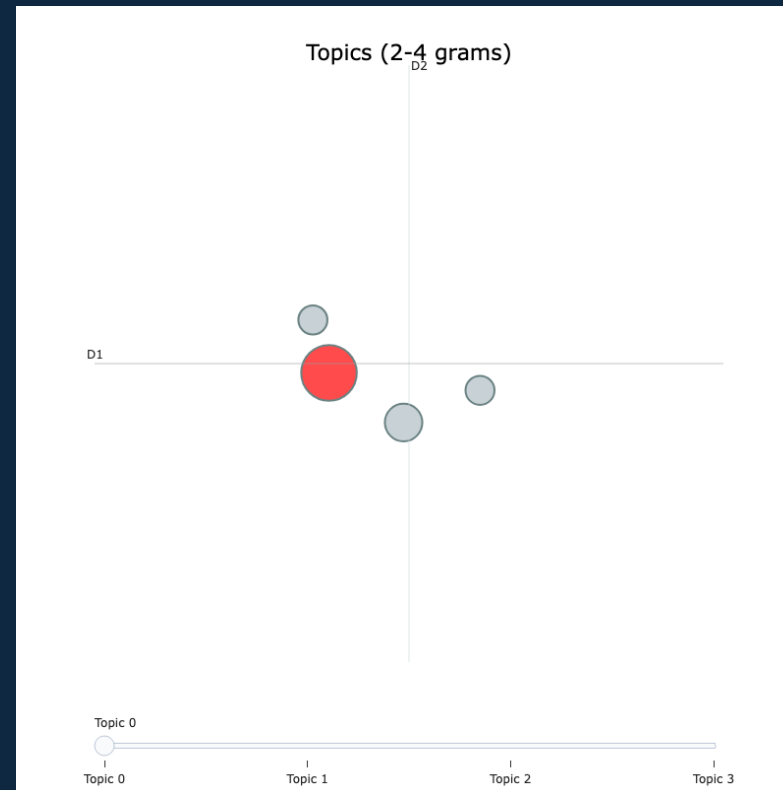
- Top terms: patient (54), care (25), cancer (24), communication (22)
- "Patient" occurrence confirmed research focus alignment

## N-gram Pattern Discovery

- Bigrams: "cancer patient" (5), "cancer care" (4), "family caregiver" (4), "lung cancer" (4), "health information" (3)
- Trigrams: "health information platform" (2), "cancer communication guideline" (2), "patient support person" (2), "positive health outcome" (2), "experience racial discrimination" (2)
- "Patient" occurrence confirmed research focus alignment

# Topic Modeling Implementation

## Component 3- Data Analysis



### BERTopic Clustering:

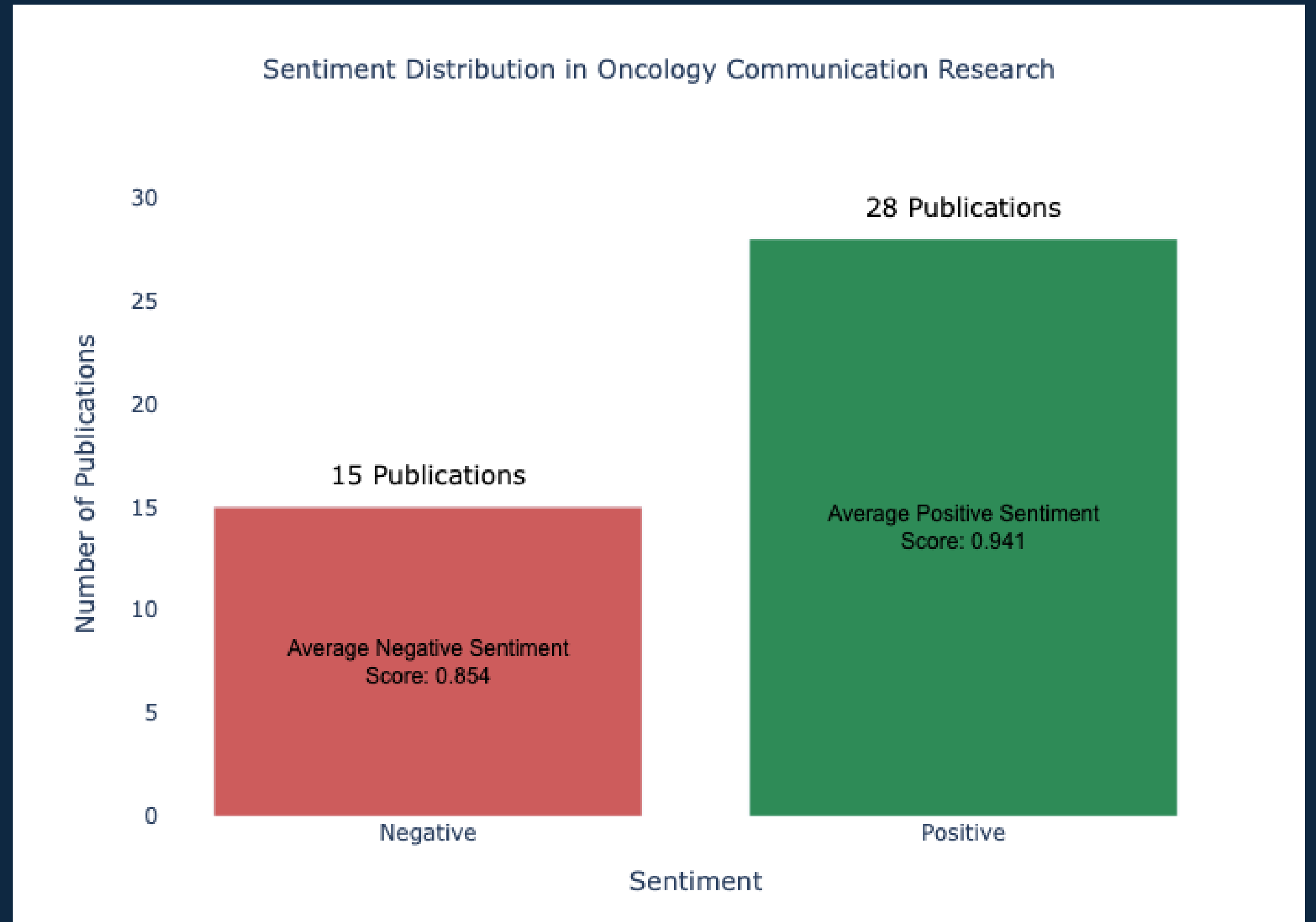
- Multi-word theme extraction using n-grams (2-4 word phrases)
- UMAP dimensional reduction for semantic clustering
- HDBSCAN density-based topic formation
- Minimum cluster size: 3 documents (coherent topics)



# Sentiment Analysis

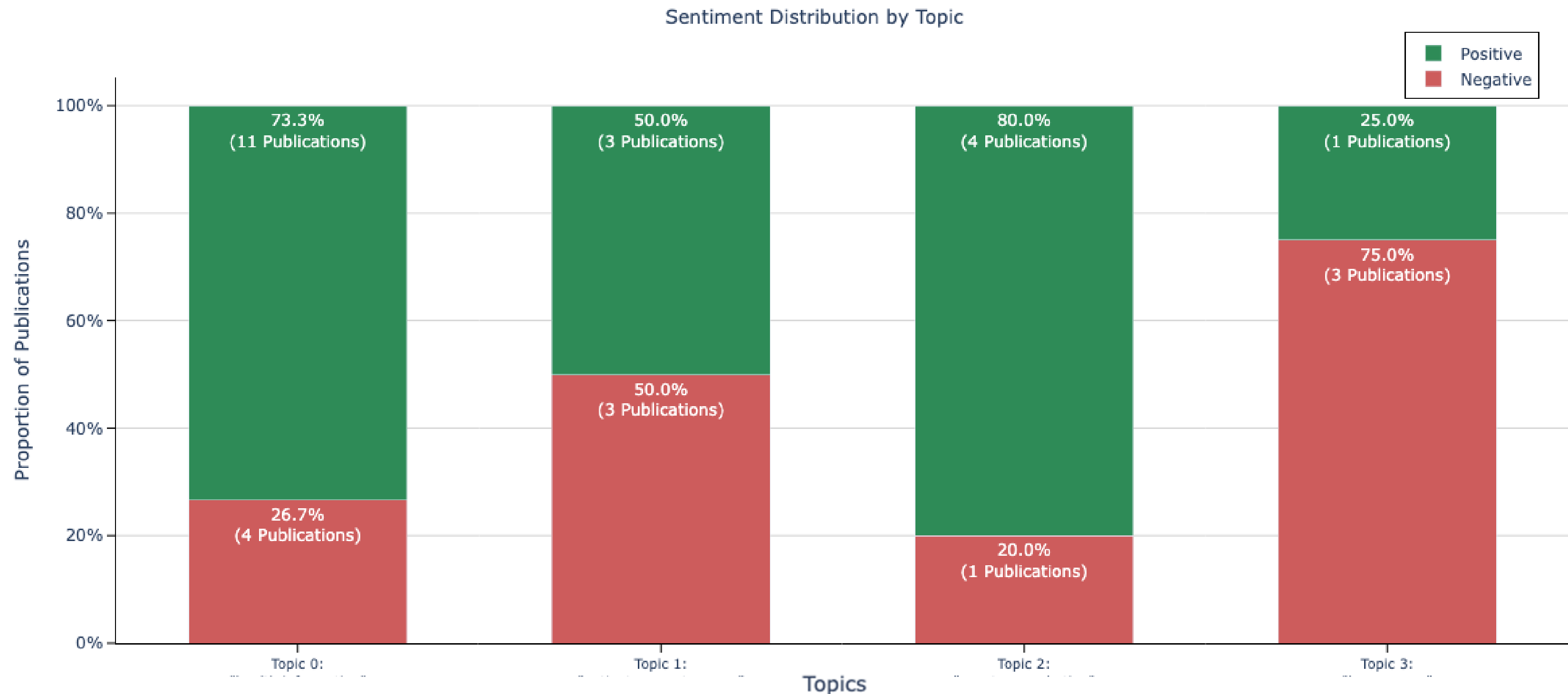
## Component 3 - Data Analysis

- DistilBERT model (distilbert-base-uncased-finetuned-sst-2-english) applied to publication conclusions
- Predominantly positive sentiment identified, revealing research community attitudes toward interventions



# Aggregation of Topic-level Sentiment Distribution Analysis

## Component 3 - Data Analysis



# Aggregation of Topic-level Sentiment Distribution Analysis

## Component 3 - Data Analysis

### Topic 0 Top Keywords:

- “health information”
- “secure message”
- ”improve patient”
- “oncology nurse”

### Topic 1 Top Keywords:

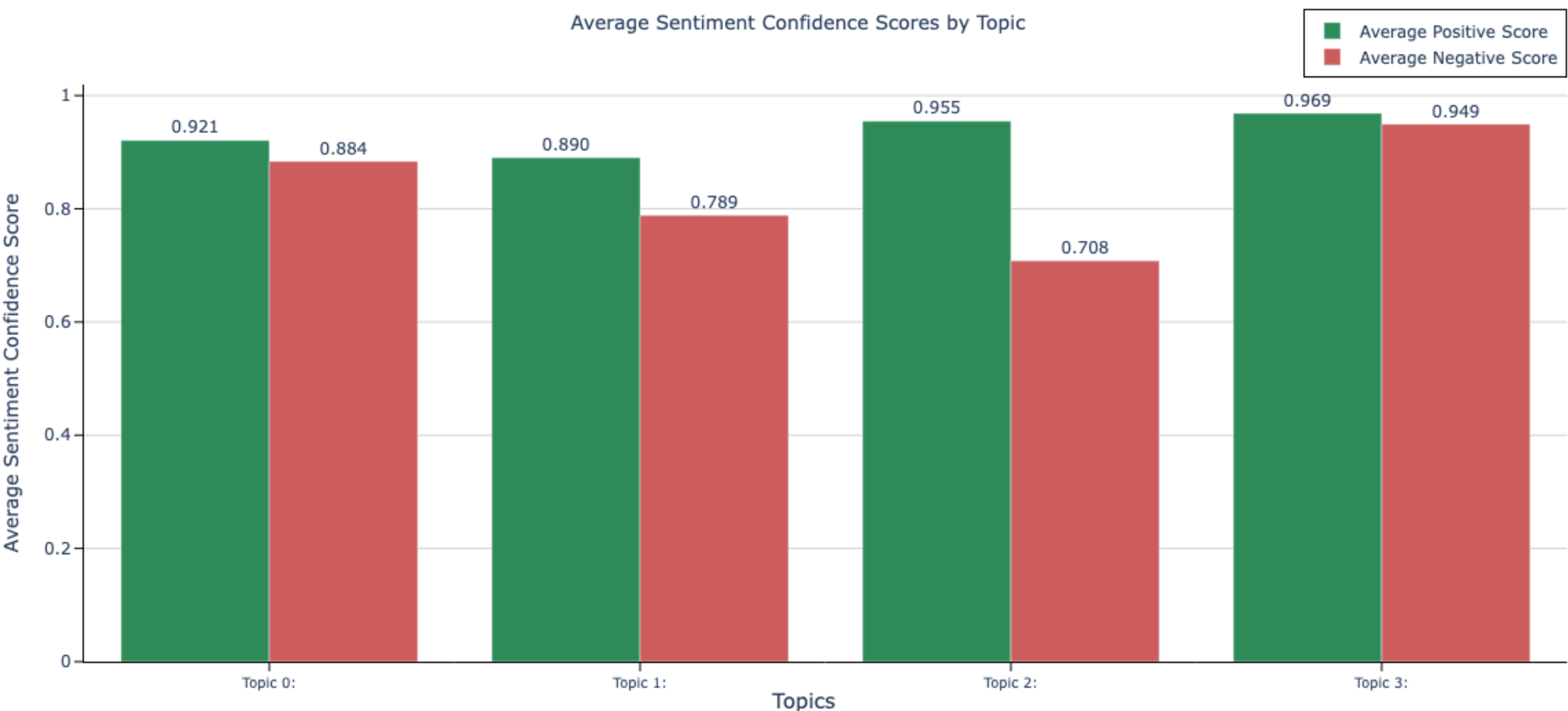
- “patient support person”
- “cancer survivor”
- “support person”
- “experience racial discrimination”

### Topic 2 Top Keywords:

- “onset exacerbation”
- “therapeutic burden”
- “patient crpc”
- “drug product”

### Topic 3 Top Keywords:

- “lung cancer”
- “acute traumatic”
- “radiotherapy advance”
- “mitigate risk adverse”





## Project Insights and Lessons Learned

- **xDD streamlines discovery** of related studies and automates **retrieval of publication metadata**
- Capturing **full-text articles remains challenging**, due to publisher paywalls and inconsistent website structures
- **Multi-word extraction** more effectively captured relevant terms
- A **combined topic modeling and sentiment analysis approach** revealed both optimistic and cautionary tones in key themes

## Future Improvements

- **Performance & Monitoring:** Refined and async scraping
- **Model & Research:** Recalibrate models, add temporal, cancer-type, and outcome analyses, and analyze more publication content (other than conclusions).
- **Deployment:** Migrate to a database backend



# Conclusion



## Approach

- Automated pipeline for literature mining and interpretable outputs
- Combined NLP methods (BERTopic, DistilBERT)



## Key Findings

- Identified dominant themes in patient-centered cancer communication
- Detected mixed sentiment toward communication-focused interventions



## Research Impact

- Offers an evidence-based framework for identifying oncology communication strategies
- Has the potential to reveal best practices that can be consolidated into a unified approach



## Technical Contribution

- Architecture tailored to midsize scientific literature corpora
- Modular components could be reused across domains

# References

- Becker, C., et al. Interventions to Improve Communication at Hospital Discharge and Rates of Readmission: A Systematic Review and Meta-analysis. JAMA network open vol. 4,8 e2119346. (2021), <https://doi.org/10.1001/jamanetworkopen.2021.19346>.
- Krist, A.H., et al. “Engaging Patients in Decision-Making and Behavior Change to Promote Prevention.” Studies in health technology and informatics vol. 240 (2017), <https://pmc.ncbi.nlm.nih.gov/articles/PMC6996004/>.
- Links, M.J., et al. Finding common ground: meta-synthesis of communication frameworks found in patient communication, supervision and simulation literature. BMC Med Educ 20, 45 (2020). <https://doi.org/10.1186/s12909-019-1922-2>.
- Peters, S.E., et al. xDD: About. (2025), <https://geodeepdive.org/about.html>.
- Sharkiya, S.H., Quality communication can improve patient-centred health outcomes among older patients: a rapid review. BMC Health Serv Res 23, 886 (2023). <https://doi.org/10.1186/s12913-023-09869-8>.
- van Dinter, R., et al. A decision support system for automating document retrieval and citation screening, Expert Systems with Applications, (2021), <https://doi.org/10.1016/j.eswa.2021.115261>.
- All images and icons used in the presentation were sourced from PowerPoint.



**Thank you!**