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

Grace Donovan

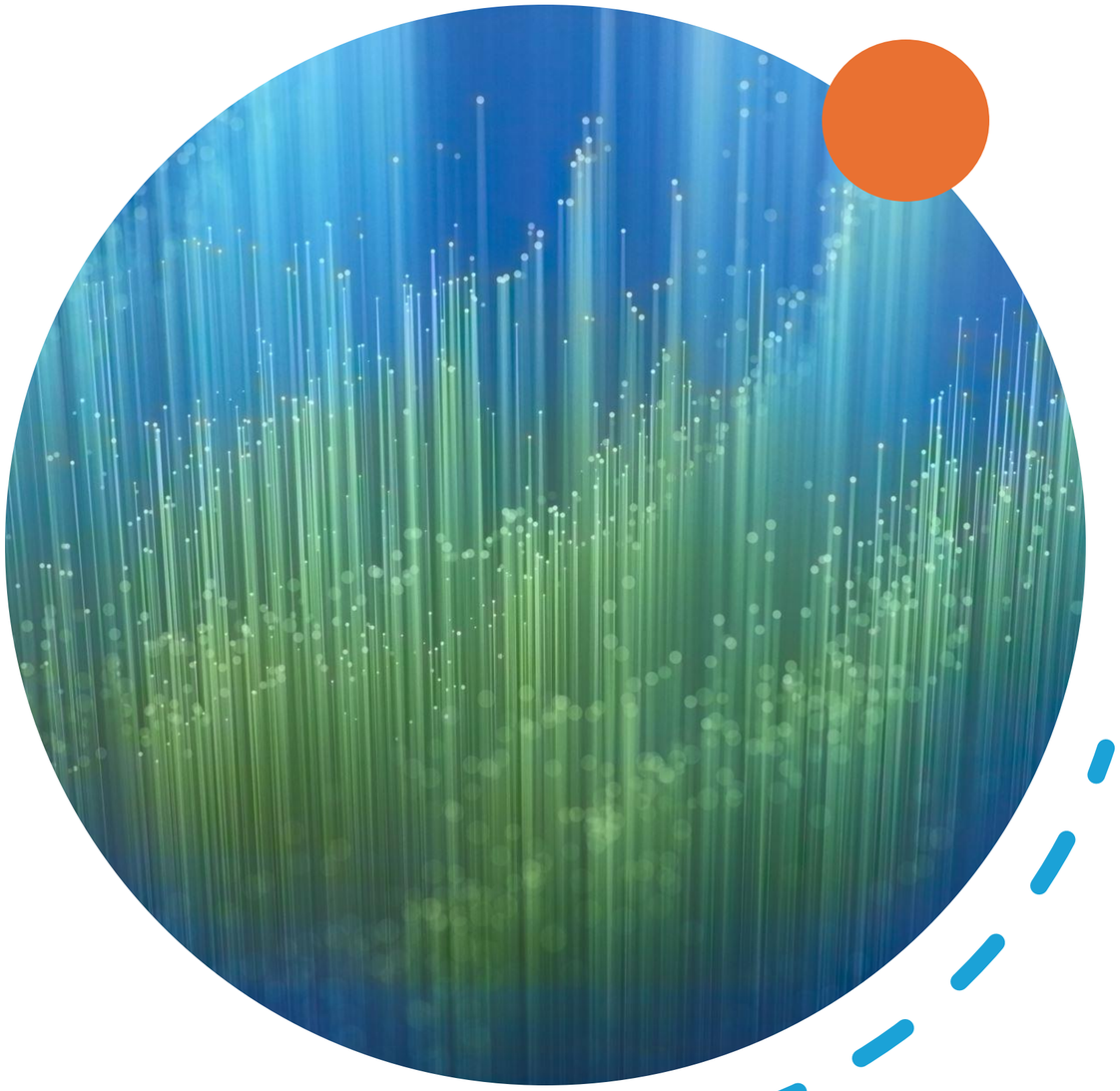
MSDE/MSDS Practicum 1
Regis University, Denver CO
June 26, 2025

Overview

Introduction: Background, Objective, and Key Questions

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

Conclusion: Insights, Future Improvements



Background

> “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)

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)).

Objective

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

Topical Focus

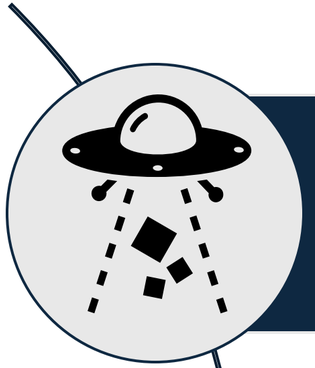
Exploring patient-centered communication themes in peer-reviewed oncology literature

Key Questions

1. Which patient-centered communication topics emerge most prevalent in oncology research?

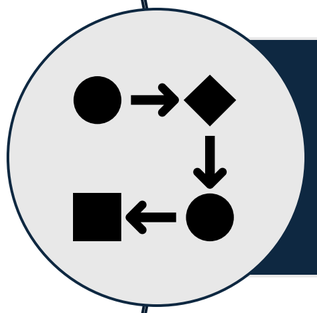
2. How does author sentiment vary across these topics?

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

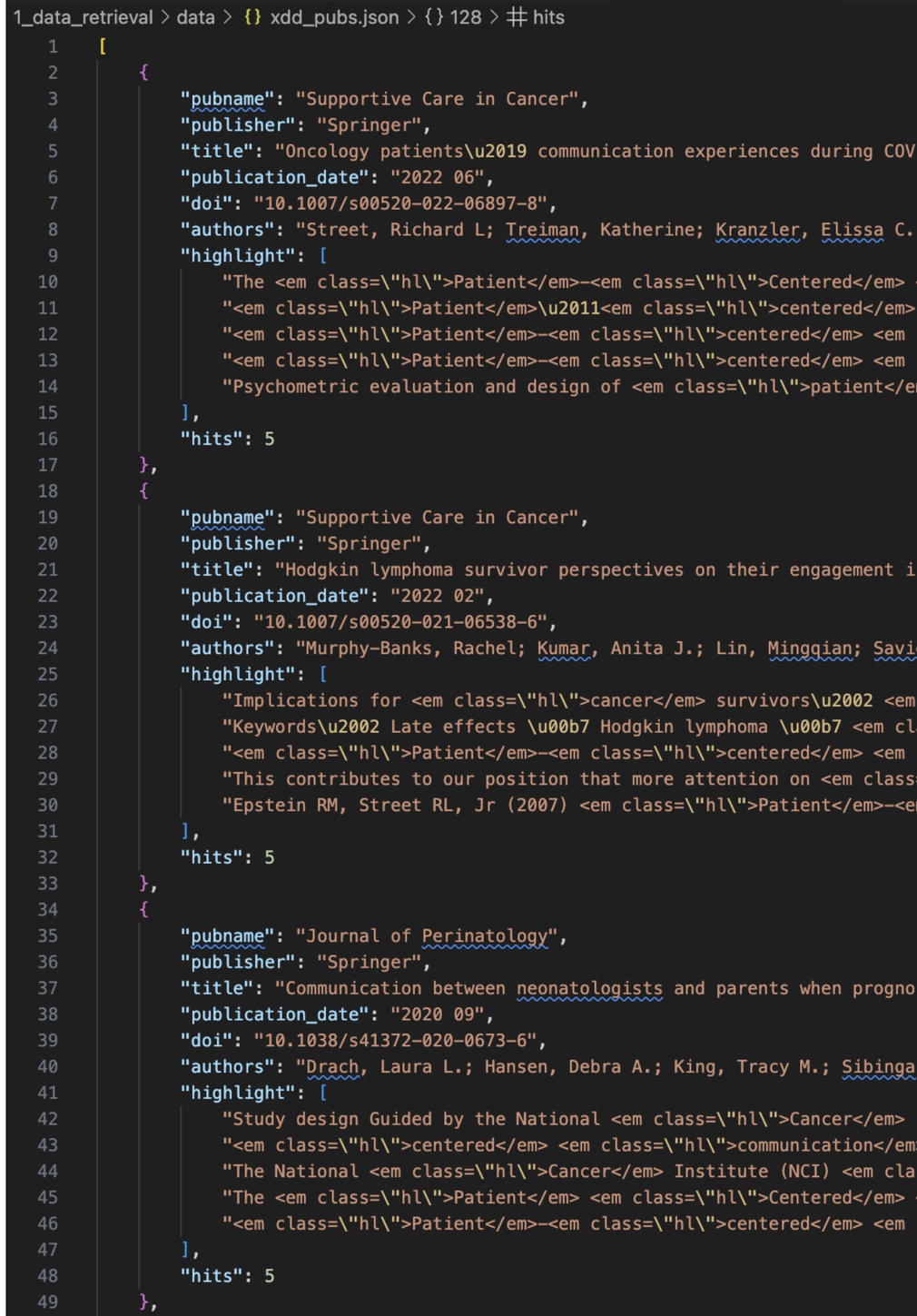
- Exploratory data analysis using word frequency distributions and n-grams
- Aggregating topic modeling for theme identification and sentiment analysis for sentiment classification of identified themes

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 ≥ 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.



Full-text Data Scraping of Related Publication Webpages

Component 1 – Data Retrieval

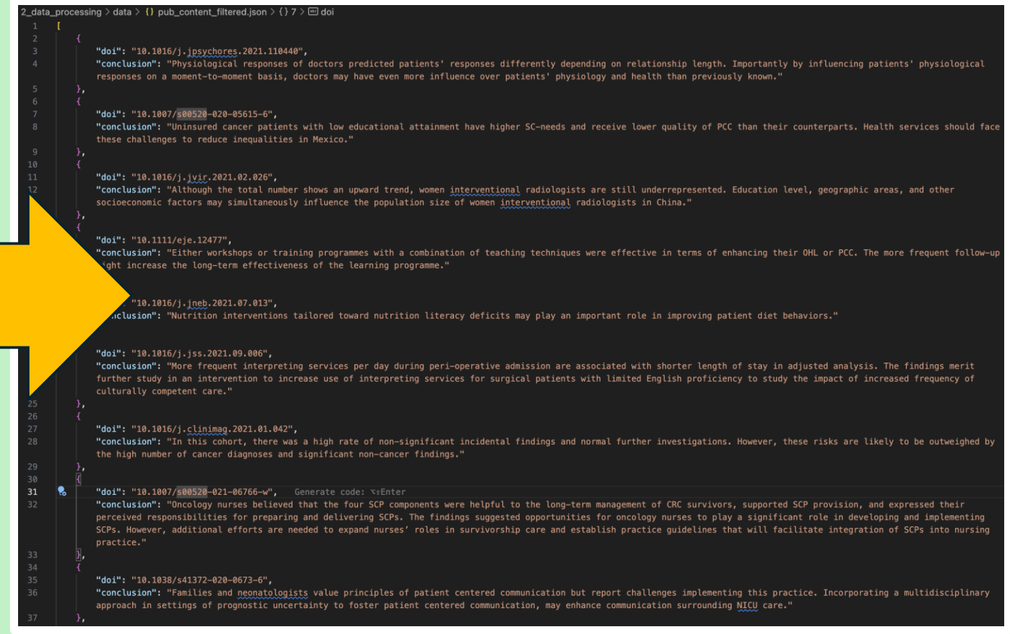
- Constructed publication DOI URLs from DOI identifiers collected from xDD API
- Used Selenium WebDriver and BeautifulSoup for scraping full-text from publication webpages
- Specific sections targeted for scraping (abstract, methods, results, conclusion, etc.)
- Filters included for handling inconsistencies and scraping issues
- 89 publications successfully scraped and saved as HTML files

HTML Content Extraction and Cleanup

Component 2 – Data Processing

89 HTML Files (Raw Text Data)

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



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

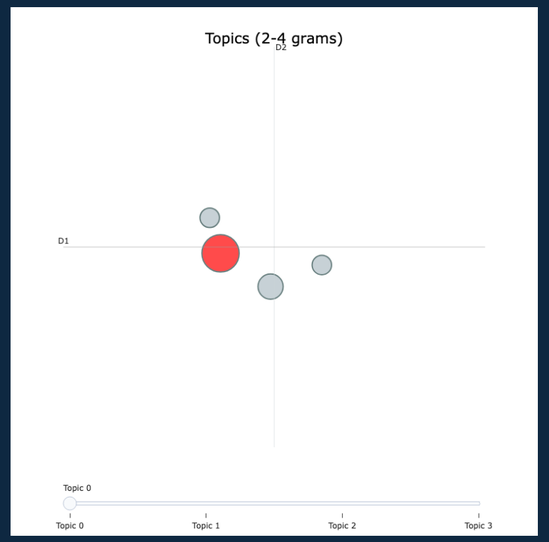
- Top terms: patient (54), care (25), cancer (24), communication (22)
- "Patient" occurrence confirmed research focus alignment

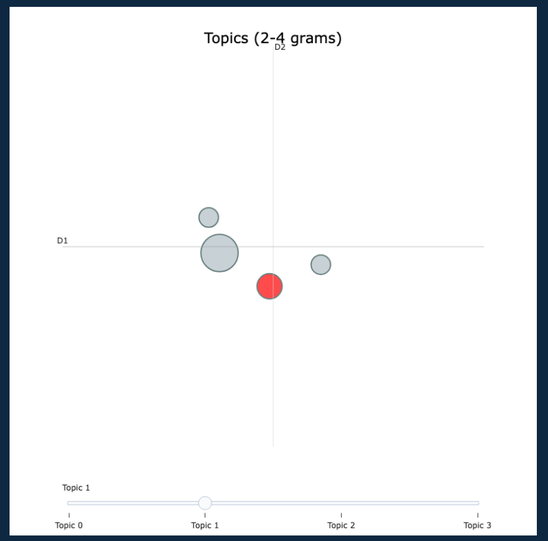
N-gram Pattern Dcanceriscovery

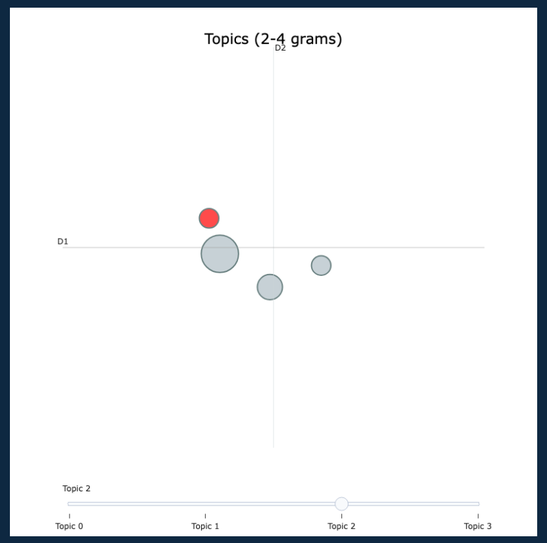
- Bigrams: ”cancer patient" (5), "cancer care" (4), ”family caregiver" (4), “lung cancer” (4), “health information” (3)
- Trigrams: ”health information platform" (2), ”cancer communication guideline" (2), ”patient support person" (2), “positive health outcome” (2), “experience racial discrimination”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

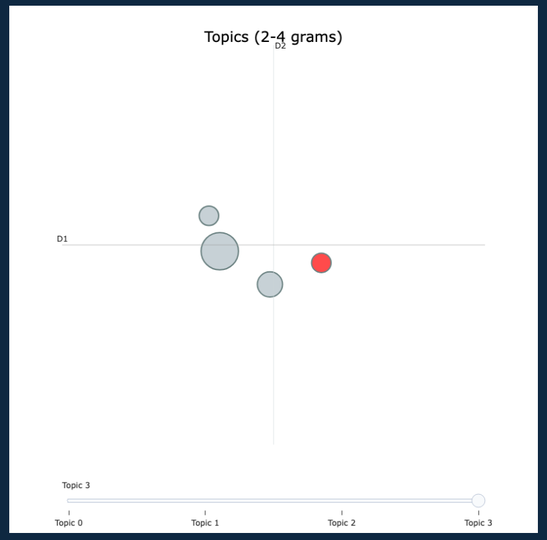
Topic Modeling Implementation

Component 3- Data Analysis









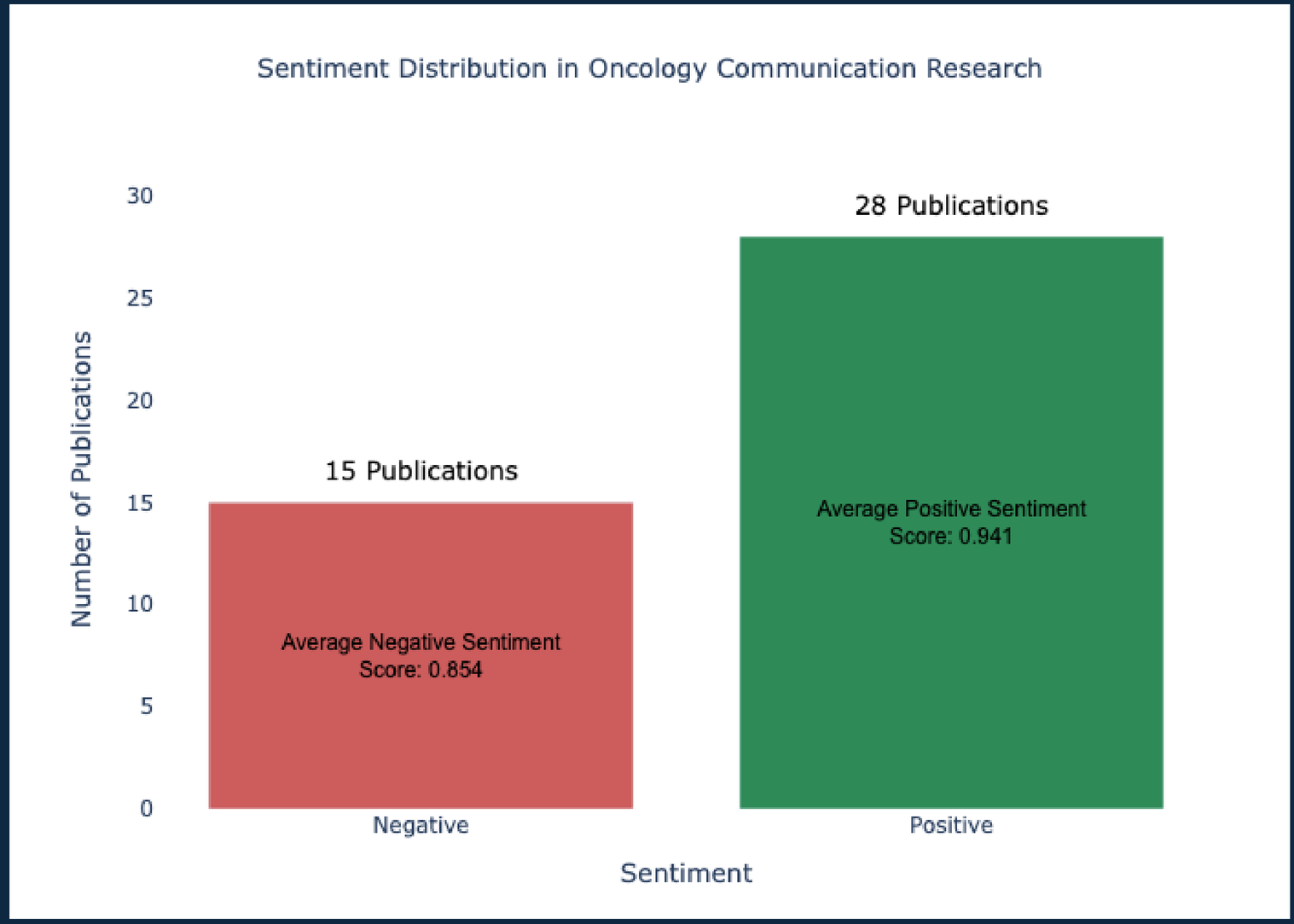
BERTopic Clustering:

- 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

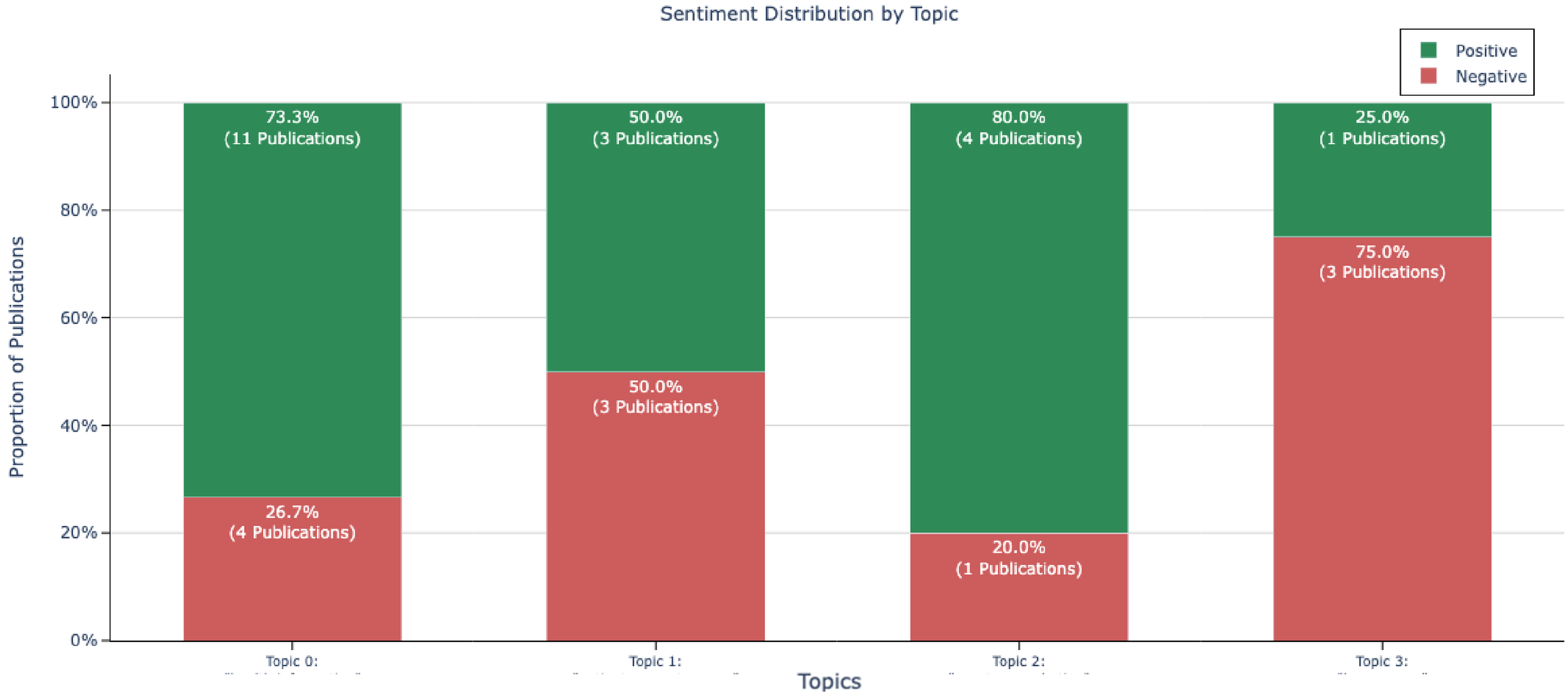
Component 3 - Data Analysis

- 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!