Understanding cancer symptom management using Amazon reviews: a human-annotated corpus

Liwei Wang, Qiuhao Lu, Rui Li, Taylor Harrison, Heling Jia, Ming Huang, Rui Zhang, Jungwei Fan, Hongfang Liu

Submitted to: Journal of Medical Internet Research
on: January 11, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript 4
Supplementary Files 28
Figures 29
Figure 1 30
Figure 2 31
Figure 3 32
Figure 4 33
Figure 5 34
Figure 6 35
Figure 7 36
Figure 8 37
Figure 9 38
Figure 10 39
Figure 11 40
Figure 12 41
Multimedia Appendixes 42
Multimedia Appendix 1 43
Multimedia Appendix 2 43

Understanding cancer symptom management using Amazon reviews: a human-annotated corpus

Liwei Wang[1]; Qiuhao Lu[1]; Rui Li[1]; Taylor Harrison[2, 3]; Heling Jia[2, 3]; Ming Huang[1]; Rui Zhang[4]; Jungwei Fan[2]; Hongfang Liu[1]

[1] McWilliams School of Biomedical Informatics The University of Texas Health Science Center at Houston Houston US
[2]Department of Artificial Intelligence and Informatics Mayo Clinic Rochester US
[3]Bioinformatics and Computational Biology University of Minnesota Twin Cities US
[4]Department of Surgery University of Minnesota Twin Cities US

Abstract

Background: Complementary therapies are being increasingly used by cancer patients. As a channel for customers to share their feelings, outcomes, ideas, and perceived knowledge about the products purchased from e-commerce platforms, Amazon online reviews are a valuable real-world data source for health care studies.

Objective: In this study, we aim to highlight the potential of using Amazon consumer reviews in mining the outcomes of cancer symptom management, provide a freely accessible corpus, and develop natural language processing (NLP) baseline models to demonstrate the usability of the annotated dataset.

Methods: We preprocessed the Amazon review dataset and conducted content analysis. We then designed an annotation guideline, annotated 159 reviews, and developed baseline models based on deep learning and large language model (LLM) for name entity recognition and text classification tasks.

Results: The annotation labels were designed to capture cancer types, indicated symptoms, and symptom management outcomes. The resulting annotation corpus contains 2,067 labels from 159 Amazon reviews. It's publicly accessible, together with the annotation guideline through the Open Health Natural Language Processing (OHNLP) Github. Our baseline model, bert-base-cased, achieved highest weighted average F1, i.e., 66.92%, for NER, and LLM gpt4-1106-preview-chat achieved the highest F1 for text classification tasks, i.e., 66.67% for "Harmful outcome", 88.46% for "Favorable outcome" and 73.33% for "Ambiguous outcome".

Conclusions: Results showed the potential of using Amazon reviews in mining the outcomes of cancer symptom management. The annotation corpus and baseline models provide a foundation for future enhanced methodology development to facilitate cancer symptom management in cancer patients using Amazon consumer reviews.

(JMIR Preprints 11/01/2025:71102)
DOI: https://doi.org/10.2196/preprints.71102

Preprint Settings

Original Manuscript

Understanding cancer symptom management using Amazon reviews: a human-annotated corpus

Authors:
Liwei Wang, M.D., Ph.D.[1], Qiuhao Lu, Ph.D.[1], Rui Li, Ph.D.[1], Taylor B. Harrison, M.B.A.[2,3], Heling Jia, M.D.[2,3], Ming Huang, Ph.D. [1], Rui Zhang, Ph.D. [4], Jungwei W. Fan, Ph.D.[2], Hongfang Liu, Ph.D.[1]

[1] McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, USA
[2] Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
[3] Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN, USA
[4] Department of Surgery, University of Minnesota, Minneapolis, MN, USA

Corresponding author: Hongfang Liu, PhD
Postal address: 7000 Fannin Street #Suite 600, Houston, TX 77030
E-mail: hongfang.liu@uth.tmc.edu
Telephone: 713-500-3900

Word count: 3609

ABSTRACT

Background: Complementary therapies are being increasingly used by cancer patients. As a channel for customers to share their feelings, outcomes, ideas, and perceived knowledge about the products purchased from e-commerce platforms, Amazon online reviews are a valuable real-world data source for health care studies.

Objectives: In this study, we aim to highlight the potential of using Amazon consumer reviews in mining the outcomes of cancer symptom management, provide a freely accessible corpus, and develop natural language processing (NLP) baseline models to demonstrate the usability of the annotated dataset.

Materials and Methods: We preprocessed the Amazon review dataset and conducted content analysis. We then designed an annotation guideline, annotated 159 reviews, and developed baseline models based on deep learning and large language model (LLM) for name entity recognition and text classification tasks.

Results: The annotation labels were designed to capture cancer types, indicated symptoms, and symptom management outcomes. The resulting annotation corpus contains 2,067 labels from 159 Amazon reviews. It's publicly accessible, together with the annotation guideline through the Open Health Natural Language Processing (OHNLP) Github. Our baseline model, bert-base-cased, achieved highest weighted average F1, i.e., 66.92%, for NER, and LLM gpt4-1106-preview-chat achieved the highest F1 for text classification tasks, i.e., 66.67% for "Harmful outcome", 88.46% for "Favorable outcome" and 73.33% for "Ambiguous outcome".

Conclusion: Results showed the potential of using Amazon reviews in mining the outcomes of cancer symptom management. The annotation corpus and baseline models provide a foundation for future enhanced methodology development to facilitate cancer symptom management in cancer patients using Amazon consumer reviews.

Keywords: Real-world data; cancer research; natural language processing; annotation; baseline models; deep learning; large language model

Introduction

Managing distressing cancer symptoms, such as pain, fatigue, weakness, anorexia, constipation, anxiety, dyspnea, nausea, and vomiting, is critical for improving the quality of life in cancer patients. Complementary approaches like acupuncture, mind-body practices, massage, and dietary supplements offer the potential to alleviate such symptoms when conventional treatments do not work. It is worth noting that cancer patients consume more dietary supplements than healthy populations.[1,2] Vitamin/mineral supplements and herbal therapies are among the most common complementary products used by individuals with cancer.[3] In one study, 69.3% of patients reported using dietary supplements after their cancer diagnosis.[4]

Online reviews are featured with reduced anonymity and personal opinions.[5] As a channel for customers to share their feelings, outcomes, ideas, and perceived knowledge about the purchased products based on the e-commerce platform, Amazon online reviews are a valuable real-world data source for health care studies. Existing studies on using health care products as complementary therapies based on Amazon reviews mainly focus on specific domains other than cancer, such as erectile dysfunction and testosterone imposters[6], eye health[7], and chronic pain[8]. This highlights a significant gap in cancer research, as there are limited explorations of how consumers experience complementary therapies, including effectiveness and adverse events. Amazon product reviews may provide implicit patterns and knowledge for cancer symptom management identified through natural language processing (NLP) techniques, where annotated data is an important asset for algorithm development and evaluation. However, the existing annotation corpora of Amazon reviews are limited to the sentiment analysis task. [9-11] There is a lack of manually annotated cancer-focused datasets that could facilitate investigating cancer symptom management from the new dimension of consumers.

In this study, we aim to highlight the potential of using Amazon consumer reviews in mining the outcomes of cancer symptom management, and achieve the two milestones for analyzing cancer outcomes using Amazon reviews, 1) provision of a freely accessible annotation corpus using the real-world data from Amazon reviews; 2) development of baseline models based on deep learning, large language model (LLM) and the annotated data.

Methods

Data source

We used the preprocessed dataset of Health & Personal Care category containing reviews and metadata from Amazon between May 1996 - July 2014.[12] This dataset has been de-duplicated, consisting of 2,982,326 reviews and 263,032 metadata. Review data includes reviewer ID, the Amazon Standard Identification Number (ASIN) which Amazon uses to identify products, reviewer name, helpfulness of rating, review text, overall rating (1–5 stars), summary of review, and review time. Metadata of the reviews include ASIN, title, price, image url, what items the customer also bought, what items the customer also viewed, what items the customer bought together, sales rank, brand, and categories. ASIN is the primary key to link review text and metadata.

Study design

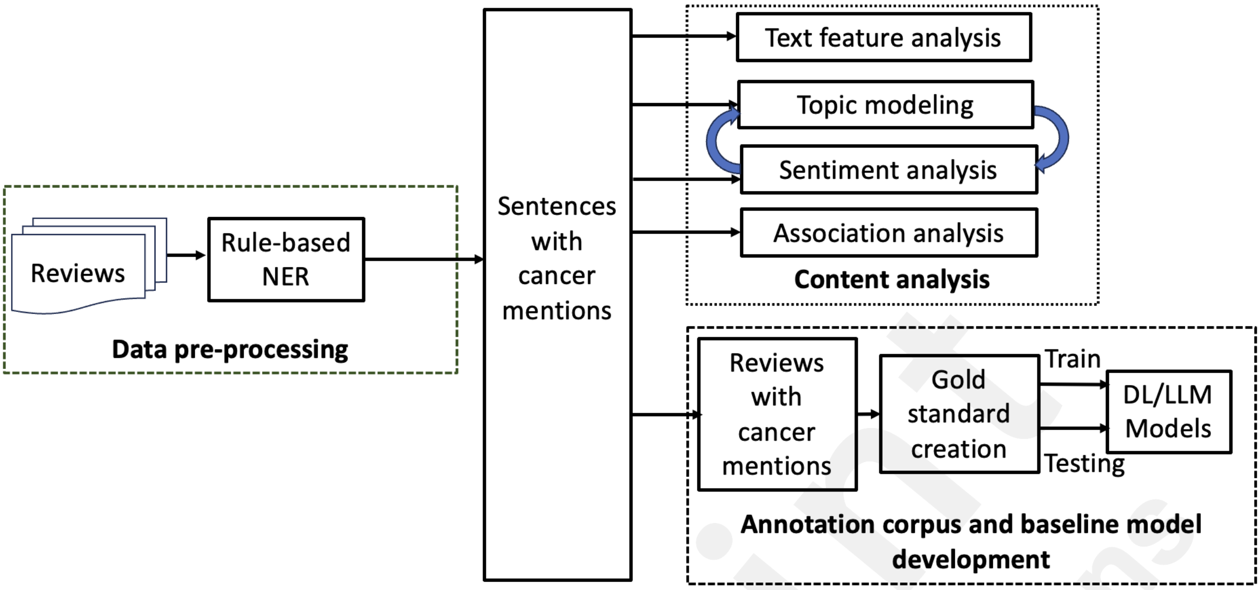


Figure 1. Study design. NER: named entity recognition; DL: deep learning; LLM: large language models

Figure 1 shows the study design. Multiple methodologies have been developed to identify named entities in texts, i.e., machine learning, deep learning, hybrid, and rule-based methods.[13] In the first step, we used a rule-based method to identify a set of review texts with cancer mentions for a high-level content analysis. We then created an annotated corpus from the set of review texts and developed baseline NLP models, including deep learning and LLM, for named entity recognition (NER) and text classification.

Data pre-processing with rule-based NLP method

To identify the reviews with cancer mentions, we prepared a cancer dictionary based on the cancer branch of the Disease Ontology. It includes cell type cancer and organ system cancer integrated from different terminologies and vocabularies including the Catalogue of Somatic Mutations in Cancer, The Cancer Genome Atlas, International Cancer Genome Consortium, Therapeutically Applicable Research to Generate Effective Treatments, Integrative Oncogenomics and the Early Detection Research Network.[14] In total, there are 4,343 cancer term variants corresponding to 1,535 cancer concepts. The cancer terms were prepared into the symbolic lexicon format compatible with the Open Health Natural Language Processing (OHNLP) Toolkit's NLP engine MedTagger.[15] The open-source clinical NLP pipeline analyzed review texts and identified cancer-related medical concepts along with the assertion status of the cancer concept including certainty (i.e., positive, negative, hypothetical and possible). We kept only positive cancer concept mentions for further analysis.

Content analysis

We summarized the features of texts containing sentences with cancer mentions, conducted sentiment analysis, topic modeling, and visualization of cancer types and symptoms association for the review sentences with cancer mentions to gain insights into the prevailing themes and mood surrounding discussions related to cancer within the dataset.

Text feature analysis

To understand the text features of review data, we performed text complexity analysis to summarize review texts containing the sentences with cancer mention, including number of review texts, number of sentences, and number of words. For comparison purposes, the above metrics were also calculated for the entire collection of reviews from the Health & Personal Care category.

Sentiment analysis

Bert-base-multilingual-uncased-sentiment[16] is a fine-tuned model from a bertbase-multilingual-uncased model for sentiment analysis on product reviews in six languages including English. Based on 5,000 held-out product reviews for English, the accuracy (exact), i.e., exact match for the number of stars is 67%. Accuracy (off-by-1), i.e., the percentage of reviews where the number of stars the model predicts differs by a maximum of 1 from the number given by the human reviewer is 95%. The fine-tuned model was used for sentiment analysis of review sentences with cancer term mentions. This model predicts the sentiment of input text as a number of stars (between 1 and 5). The higher the sentiment score, the more overall positive. The lack of context has been one major challenge in sentiment analysis that can affect the interpretation of sentiment.[17] We consider that identifying customer attitudes based on the sentence containing cancer mentions instead of the whole review text can be better constructive in understanding consumers' efficacy and safety perceptions. The sentiment of the review sentences with cancer mentions detected by Medtagger was further analyzed to identify positive or negative attitudes toward the product.[18,19] We analyzed the distribution of sentiment scores across the review sentences with cancer mentions, and the trend of average sentiment score between 1996 and 2014.

Topic modeling

We employed a sentence embedding model (i.e., bge-small-en)[20] to transform the textual content of reviews into numerical embeddings. These embeddings capture the semantic essence of each document in a high-dimensional space. We then applied UMAP (Uniform Manifold Approximation and Projection)[21] to the embeddings for dimensionality reduction. This step is crucial for visualization, as it converts high-dimensional data into a 2-dimensional format suitable for plotting. The core of the analysis is performed by BERTopic,[22] a model that identifies distinct topics within the text data. BERTopic relies on sub-models for embeddings (provided by SentenceTransformer of bge-small-en), dimensionality reduction (UMAP), and hierarchical clustering (HDBSCAN).[23] Additionally, a quantized Large Language Model (LLM) (i.e., openhermes-2.5-mistral-7b)[24] is incorporated for topic label generation. After fitting the data to the BERTopic model, topics are extracted along with their probabilities. Each topic is then assigned a label generated by the LLM based on a predefined prompt. These labels are designed to be concise, with a maximum of five words, and describe the essence of the documents within each topic.

The chosen sentences were preprocessed by removing stop-words, special characters, and numbers and removed sentences with pets (dog, cat, etc.). We detected topics based on all sentences with cancer mentions, as well as the sentences from 5 sentiment score groups. We then visualized the results respectively.

Cancer type and symptom association

We constructed bipartite graphs to visualize the relationships between cancer type and symptoms. The state-of-the-art LLM for NER, i.e., UniversalNER-7b-all model was used to identify symptoms in the relevant reviews via 0-shot strategy. We then calculated the numbers of cancer type and symptom pairs. The bipartite graphs were built using the pairs of cancer types and symptoms, where each type of cancer and symptom was represented as a node, with edges indicating their association frequency. Nodes were positioned to ensure even distribution and alignment in their respective groups. Edge widths were normalized and scaled to reflect the frequency of the cancer type-symptom pairs, providing a visual indication of the strength of each association.

Development of gold standards and baseline models

Gold standard creation

We developed an annotation guideline (Appendix File 1) for labeling the target data elements and associated class/type from customer reviews. The guideline was designed to be as brief as possible, aiming to minimize annotators' cognitive load while maximizing the potential use of the annotated dataset for future information extraction tasks. Supplemental Table 1 shows the schema of the annotated labels. The target concepts included Cancer type, Indicated symptoms, Favorable outcome, Harmful outcome, and Product, while each concept is assigned a class/type. The cancer type concept has either the class of human or pet. Indicated symptoms, favorable outcome and harmful outcome are assigned with classes of either cancer related or other, and the product concept has a type of either itself or other. Cancer type, indicated symptoms, favorable outcome and harmful outcome were also labeled with one of the four certainties, e.g., positive, negative, hypothetical, and possible. For example, in the sentence: "I've had salivary gland cancer", "salivary gland cancer" was highlighted as the cancer type concept, associated with a human class and positive certainty. In the sentence: "Some people say it might prevent cancer", "might prevent cancer" was labeled as the favorable outcome concept, associated with the cancer related class and hypothetical certainty.

MedTator,[25] a free and open-source annotation tool, was used for this annotation task. Two trained abstractors with medical and informatics backgrounds were first trained to annotate following the annotation guideline. When they had the same understanding of the annotation task indicated by an inter-annotator agreement (IAA) of 0.9, they started to label the reviews independently. All disagreements were discussed in the adjudication process, and a final " consensus" gold standard was derived. IAA was calculated based on F1 value.

We then randomly selected 200 review texts with cancer mentions detected from the first step. In the annotation process, we focused on customer perspectives and excluded reviews of literature contents. As a result, 159 reviews were chosen for annotation.

Development of baseline models

In our study, the annotated data is used for two distinctive NLP tasks, i.e., NER and text classification. The NER task aimed to identify and classify entities from customer comments on products, focusing on cancer types, indicated symptoms, and product mentions. Note that only human cancers (not pet cancers) and cancer-related symptoms in the review are used in the development of NER models. The Product entity refers to the anaphors of products. Specifically, cancer types entity includes specific cancers such as "breast cancer", "leukemia," "lymphoma," and "melanoma". In developing baseline models, the entity type is limited to human-related cancer types where the certainty of the mention is positive. Indicated symptoms refer to the indications that the product was used for, e.g., "affected her eye" in "She had cancer that affected her eye". Product entity means direct mentions of the product names or indirect references such as "this" are captured. For text classification, the task was to categorize user review comments into three categories, including favorable outcome, harmful outcome, and ambiguous outcome. Specifically, Favorable outcomes are comments where the product is noted to positively affect a cancer-related condition, with varying degrees of certainty. Harmful outcomes are comments indicating a negative impact on cancer-related conditions. Ambiguous outcomes include comments with possible and hypothetical impacts, reflecting the speculative nature of the feedback. We used the data on cancer-related outcomes.

We developed two types of baselines for the NER and text classification tasks. The first baseline is based on the supervised fine-tuning (SFT) of BERT-like models on the training data, with two different classification heads on top, i.e., token classification and sequence classification, respectively. We tested the performance of two popular BERT-like models, i.e., bert-base-cased and Bio_ClinicalBERT. The second type of baseline is based on LLM where we instruct gpt4-1106-preview-chat to generate the desired labels in zero-shot, few-shot (using 5 examples), many-shot (using all training data) [26] in-context learning settings. Appendix File 2 shows the prompts used for NER and text classification. For NER and text classification, 80% of the datasets were used for training, and 20% were designated for testing.

RESULTS

A total of 4,703 sentences were detected with positive cancer mentions, corresponding to 3,349 reviews and 2,589 products. These reviews contained 26,078 sentences and 500,087 words, with an average of 149.3 words per review. In contrast, there are 10,469,336 sentences and 199,501,964 words in the 2,982,326 reviews from the Health & Personal Care category, with an average of 66.9 words per review. Figure 2 shows the distribution of product categories with review sentences of cancer mentions.

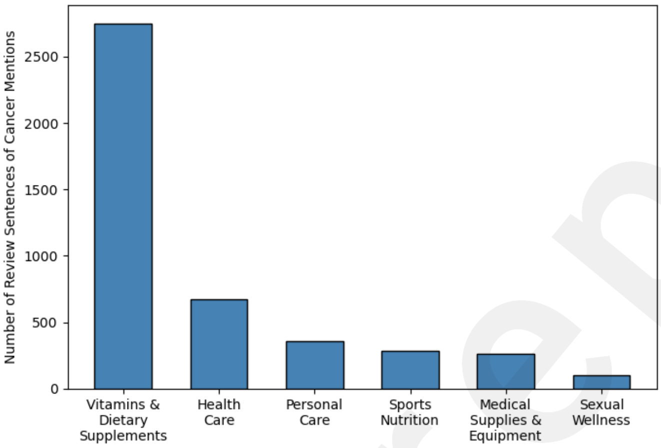


Figure 2. Distribution of product categories with review sentences of cancer mentions

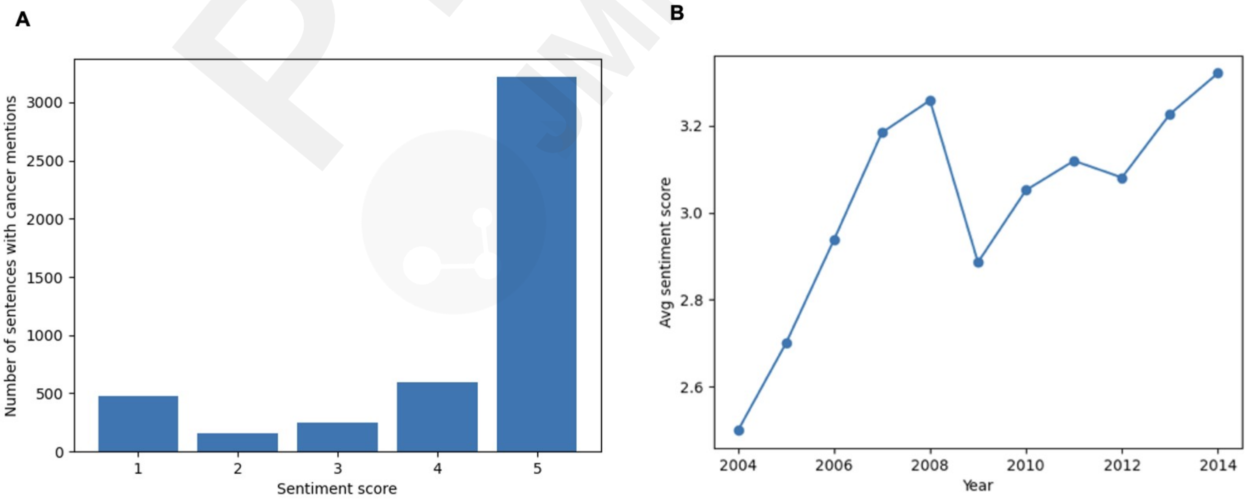


Figure 3. Analysis of sentiment scores in review sentences with cancer mentions. A. Distribution of sentiment scores across the review sentences. B. Trend of average sentiment score during May 1996 - July 2014

Figure 3A shows the distribution of sentiment scores across the review sentences with cancer mentions, where score 5 prevailed as the most common sentiment. Figure 3B shows the trend of average sentiment score between 1996 and 2014. In general, increased trends can be observed before and after a dip around 2008.

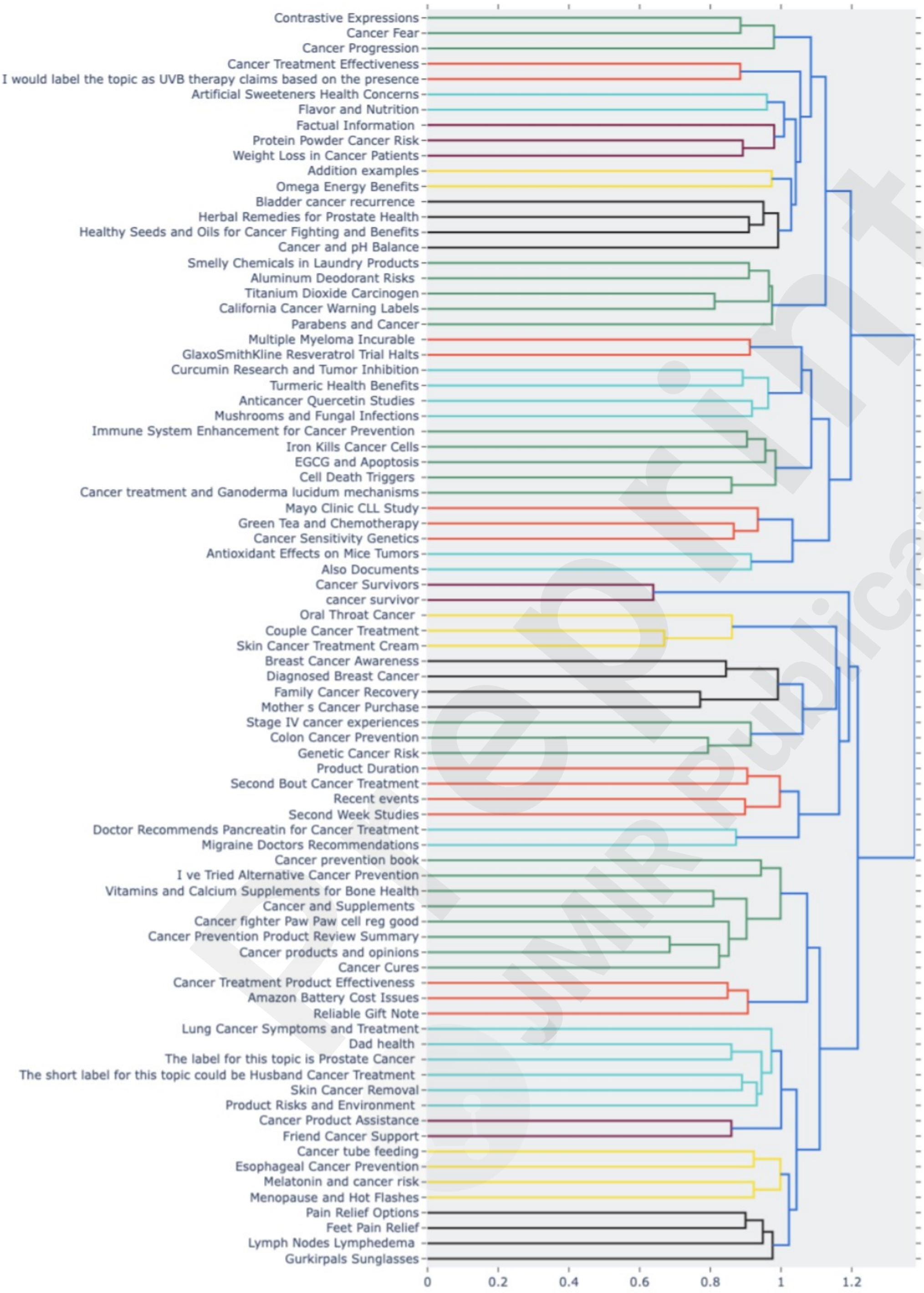


Figure 4. Hierarchical clustering of topics based on all sentences with cancer mentions.

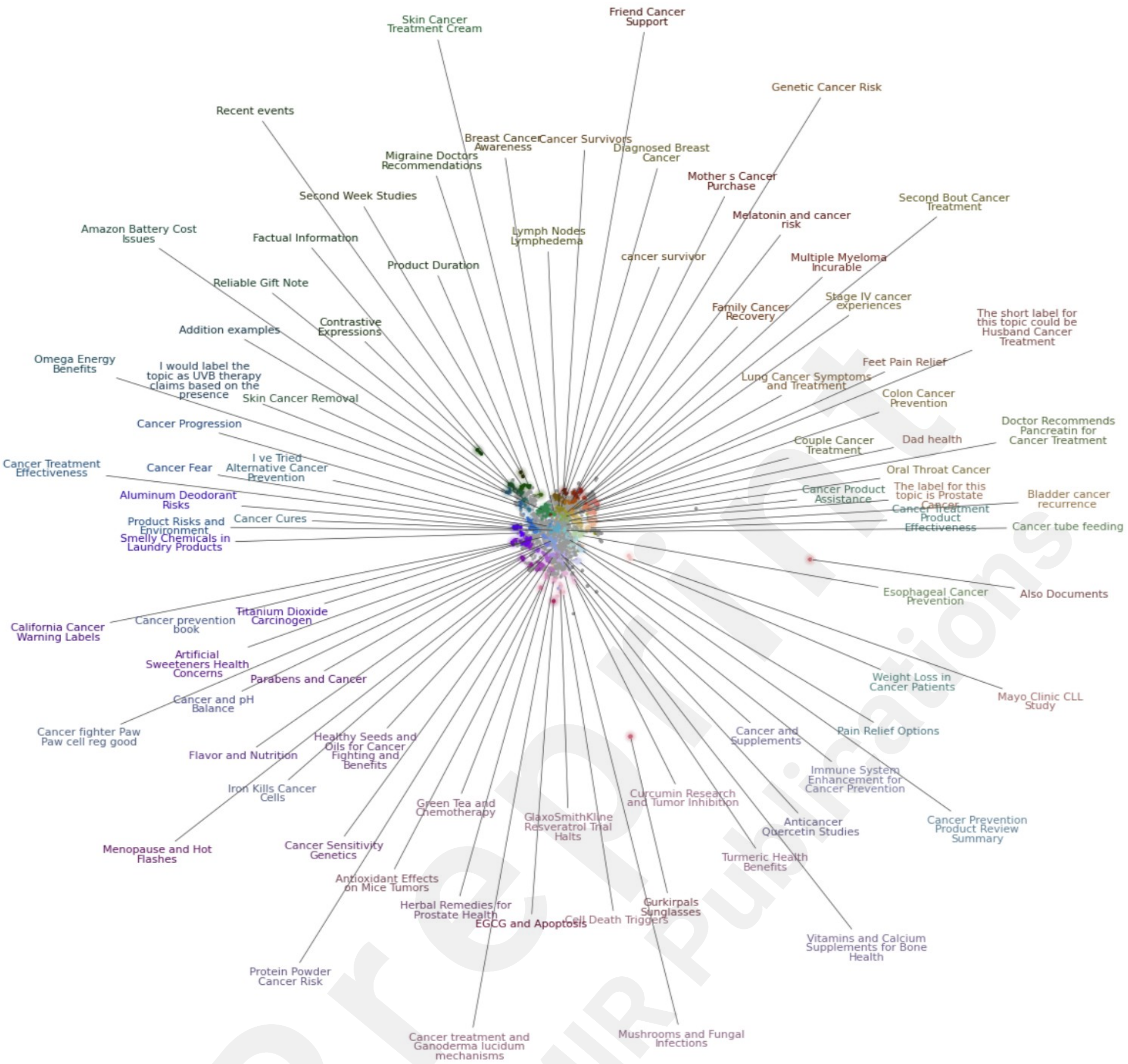


Figure 5. Topic modeling with openhermes-2.5-mistral-7b based on all sentences with cancer mentions

The hierarchical clustering of topics identifies (Figure 4) the subgroups of topics. For example, protein powder and cancer risk is clustered with weight loss from cancer; breast cancer survivor is clustered with estrogen positive cancer; vitamins and calcium for health is clustered with cancer prevention. Figure 5 shows the results from topic modeling based on all sentences with cancer mentions extracted from the dictionary method. Meaningful insights are revealed, e.g., green tea and chemotherapy, cancer prevention treatment, breast cancer product recommendations, post cancer oral issues, antioxidant effects on tumor blood vessels. Figure 6-10 (at the end of the munuscript) show the hierarchical clustering and topic modeling based on the sentences from 5 sentiment score groups. Crucial revelations could be found respectively in various sentiment score groups, e.g., more topics on cancer risks appeared in the group of sentiment score 1 than other groups of sentiment scores, including carcinogen ingredients, California warning label products, artificial sweetener risks, beware cheap Amazon products, ingredients and toxicity opinion, etc. More topics on benefits for cancer appeared in the group of sentiment score 5 than other groups of sentiment scores, such as iodine and thyroid health benefits, calcium vitamin supplements and bone health, flaxseed health benefits, cancer survivorship and thriving, sleep aid for cancer treatment, anticancer supplements for cancer patients, etc.

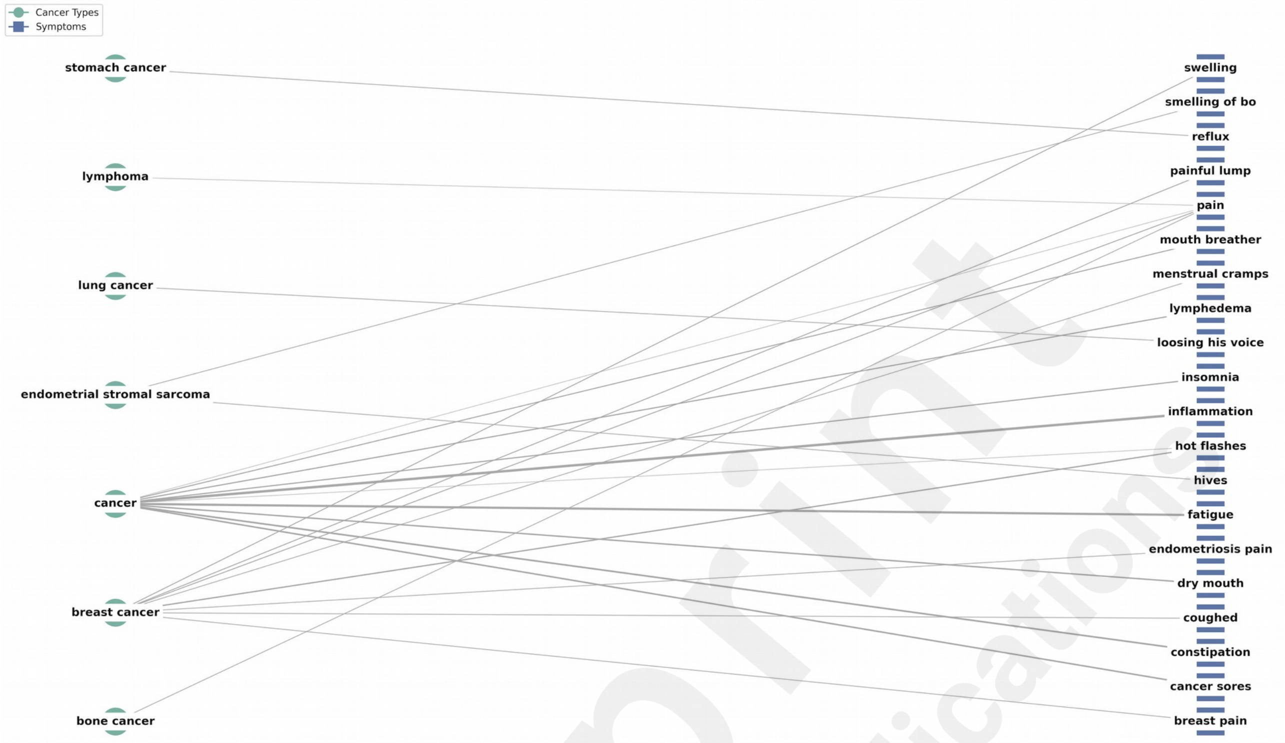


Figure 11. Bipartite graph of cancer types and symptoms extracted by LLM.

Figure 11 shows the bipartite graphs of cancer types with symptoms. The bipartite graphs is used to show the association between cancer types and symptoms instead of causal relations. Zero-shot LLM extracted detailed symptoms, such as pain, inflammation, fatigue, constipation, etc. Results showed associations between stomach cancer and reflux, breast cancer and menstrual cramps, bone cancer and pain, etc.

Figure 12 shows the top 15 symptoms in reviews, with pain being the most frequent symptom, followed by inflammation, fatigue, hot flashes, dry mouth, constipation, cancer sores, nausea, insomnia, etc.

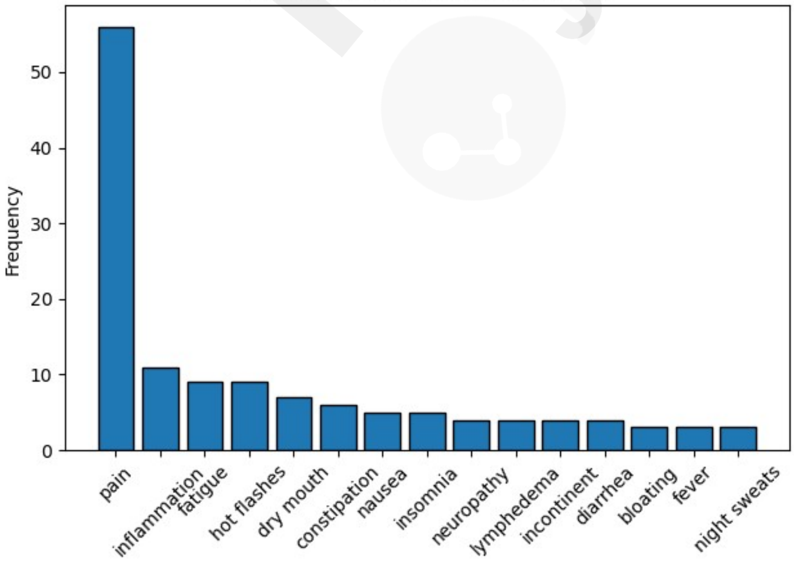


Figure 12. Top 15 symptoms in reviews

Table 1 shows the statistics for the resulting annotated corpus for each concept and associated classes (type) and certainties. In total, 2,067 labels were generated from 159 reviews. Supplemental Table 2 shows the inter-annotator agreements for each concept annotation, with the overall inter-annotator agreement being 0.86. IAA for cancer type is the highest (0.97) and harmful outcome is the lowest (0.63). The annotated corpus is publicly accessible through the OHNLP Github.[27]

Table 1. Statistics of the resulting annotated corpus.

Concepts	**Class (Type)/Certainty**							
Cancer_type	Human				Pet			
	Positive	Negative	Hypothetical	Possible	Positive	Negative	Hypothetical	Possible
	131	9	100	2	18	0	3	0
	Cancer_related				Other			
	Positive	Negative	Hypothetical	Possible	Positive	Negative	Hypothetical	Possible
Indicated_symptom	105	1	1	0	80	0	2	0
Harmful_outcome	16	0	5	0	23	0	1	0
Favorable_outcome	145	0	51	1	242	3	15	0
Product	Itself				Other			
	1015				98			

In our study, the annotated data is used for two distinctive NLP tasks, i.e., named entity recognition and text classification. The dataset for the NER task included 1054 annotated samples, with 80% (843 samples) used for training the model and 20% (211 samples) designated for testing its accuracy. For text classification, the dataset consists of 218 annotated samples, with 80% (174 samples) allocated for training the model and 20% (44 samples) reserved for testing its accuracy. Table 2 shows the statistics of annotation entity labels for model development.

Table 2. Statistics of annotation entity labels for model development.

Task	**Target**	**Criteria**	**No. Label**
NER	Cancer_type	human, positive	131
	Indicated_symptom	cancer_related, positive	105
	Product	itself	1015
Text classification	Favorable_outcome	cancer_related, positive	145
	Harmful_outcome	cancer_related, positive	16
	ambiguous_outcome	cancer_related, hyperthetical	57

		and possible	

Table 3 shows the performance of bert-base-cased, Bio_ClinicalBERT and gpt4-1106-preview-chat in NER. In general, bert-like models outperformed LLM, with 0.6692 weighted average F1 for bert-base-cased, 0.6558 for Bio_ClinicalBERT, and the best performance of gpt4-1106-preview-chat was 0.5077 weighted average F1 through many-shot strategy. Among the three entities, “indicated symptom” showed consistent lower performance across all baseline models compared with the other two entities, i.e., “cancer type” and “product”, implying the difficulty of extracting this entity.

Table 4 show the performance of baseline models in text classification. The performance of LLM gpt4-1106-preview-chat using many-shot strategy exceeded bert-like models. Specifically, the performance of bert-base-cased and Bio_ClinicalBERT in classifying “Harmful outcome” was zero. This could be explained by the limited number, i.e., 16, of “Harmful outcome” labels in the gold standard. In addition, the IAA of harmful outcome is the lowest during annotation, implying that “Harmful outcome” classification is the most difficult classification task among all. In contrast, LLM excelled in the scenario of the limited labels, achieving the highest F1 for the three classes, i.e., 0.6667 for “Harmful outcome”, 0.8846 for “Favorable outcome” and 0.7333 for “Ambiguous outcome”.

Table 3. Performance of baseline models in NER.

Model	Learning strategy	entity	precision	recall	f1-score
bert-base-cased	SFT	Cancer_type	0.5366	0.6286	0.5789
		Indicated_symptom	0.1667	0.1429	0.1538
		Product	0.6773	0.7161	0.6962
		Micro avg	0.6514	0.6905	0.6704
		Macro avg	0.4602	0.4959	0.4763
		Weighted avg	0.6495	0.6905	0.6692
Bio_ClinicalBERT	SFT	Cancer_type	0.5349	0.697	0.6053
		Indicated_symptom	0.3000	0.2143	0.2500
		Product	0.695	0.6583	0.6762
		Micro avg	0.6675	0.6462	0.6567
		Macro avg	0.5100	0.5232	0.5105
		Weighted avg	0.6684	0.6462	0.6558
gpt4-1106-preview-chat	Zero-shot	Cancer_type	0.2885	0.6818	0.4054
		Indicated_symptom	0.0759	0.4615	0.1304
		Product	0.3529	0.3243	0.338
		Micro avg	0.2776	0.3619	0.3142
		Macro avg	0.2391	0.4892	0.2913
		Weighted avg	0.3334	0.3619	0.3333

	Few-shot	Cancer_type	0.3148	0.7727	0.4474
		Indicated_symptom	0.0536	0.2308	0.087
		Product	0.4743	0.5405	0.5053
		Micro avg	0.3857	0.5447	0.4516
		Macro avg	0.2809	0.5147	0.3465
		Weighted avg	0.4394	0.5447	0.4791
	Many-shot	Cancer_type	0.4000	0.6364	0.4912
		Indicated_symptom	0	0	0
		Product	0.5672	0.5135	0.539
		Micro avg	0.5079	0.4981	0.5029
		Macro avg	0.3224	0.3833	0.3434
		Weighted avg	0.5242	0.4981	0.5077

Table 4. Performance of baseline models in text classification

Model	Learning Strategy	Sentiment	precision	recall	f1-score
bert-base-cased	SFT	Harmful_outcome	0	0	0
		Favorable_outcome	0.6470	0.8800	0.7457
		Ambiguous_outcome	0.7000	0.4375	0.5384
Bio_ClinicalBERT	SFT	Harmful_outcome	0	0	0
		Favorable_outcome	0.6486	0.96	0.7741
		Ambiguous_outcome	0.8571	0.375	0.5217
gpt4-1106-preview-chat	Zero-shot	Harmful_outcome	0.6667	0.6667	0.6667
		Favorable_outcome	0.7368	0.56	0.6364
		Ambiguous_outcome	0.4545	0.625	0.5263
	Few-shot	Harmful_outcome	0.5	0.6667	0.5714
		Favorable_outcome	0.6429	0.72	0.6792
		Ambiguous_out	0.3333	0.25	0.2857

		come			
	Many-shot	Harmful_outcome	0.6667	0.6667	0.6667
		Favorable_outcome	0.8519	0.92	0.8846
		Ambiguous_outcome	0.7857	0.6875	0.7333

DISCUSSION

Complementary therapies are being increasingly used by cancer patients. Especially in breast cancer patients, the ratio of dietary supplement use was ranging from 67% to 87%.[1,28] However, complementary therapies are not usually included in the health care systems, and clinical research conducted on human subjects is still limited, as patients did not intend to report their complementary use to their provider.[29,30]

In this study, we first highlighted the potential of using Amazon consumer reviews in mining the outcomes of cancer symptom management through a content analysis focused on the following aspects. First, topic clustering identified meaningful subgroups, such as 1) protein powder, cancer risk, and weight loss from cancer, 2) breast cancer survivor and estrogen positive cancer, 3) vitamins, calcium for health, and cancer prevention. Second, more topics on cancer risks appeared in the group with a lower sentiment score, and more topics on the benefits for cancer symptom management appeared in the group with a higher sentiment score (Figure 6-10). Third, associations between cancer types (identified through the rule-based dictionary method), and detailed symptoms (identified through zero-shot LLM) could be explicitly revealed. Fourth, the top 15 symptoms reflected common cancer symptoms to be managed, with pain being the most frequent. In this content analysis, no evaluation was conducted for the zero-shot LLM, as it was used only to extract potential symptoms.

Our key contributions lie in developing a manually annotated dataset with 159 reviews, and baseline models for NER and text classification to demonstrate the usability of the annotation dataset. It is worth noting that we leveraged the rule-based dictionary method to secure relevant reviews with cancer mentions for the following annotation task and used zero-shot LLM to extract potential symptoms for preliminary content analysis. The annotation labels were designed to capture more variants of specific cancer types, such as “cancer in his bone”, as well as the nuances in customer feedback related to health impacts. Therefore, the models built on top of the annotated dataset could potentially enable NER at a finer granularity and more detailed analysis of consumers’ perceptions of outcomes of cancer symptom management. Our baseline model, bert-base-cased, achieved highest weighted average F1 for NER, LLM gpt4-1106-preview-chat achieved the highest F1 for text classification tasks.

The performance of LLMs in this study aligns with prior findings, indicating that LLMs are not particularly strong at NER tasks.[31,32] However, LLMs excel at text classification tasks, as reflected in Table 4. For example, many-shot learning achieved the best performance across the table, with an F1-score of 0.8846 for “Favorable outcome” and 0.7333 for “Ambiguous outcome.” Moreover, LLMs demonstrate a significant advantage in data-scarce scenarios. For instance, in the case of "Harmful outcome," where only 16 samples are available, the gpt4-1106-preview-chat achieved an F1-score of 0.6667 in the zero-shot setting, while fine-tuning BERT/Bio_ClinicalBERT resulted in an F1 of 0. This emphasizes the LLM's ability to effectively generalize and provide meaningful predictions in low-resource situations, a crucial capability in tasks with limited annotated data.

Our study has one limitation in the sentiment analysis, which is usually domain-dependent.[33] Though the majority of available pre-trained language models have the ability to classify text by sentiment, few can be found targeting medical or health domains. To perform a high-level analysis, we employed the existing sentiment analysis model, with the accuracy (exact), i.e., exact match for the number of stars for product reviews being 67%, on the sentences with cancer mentions. Thus, a misalignment of the sentiment score occurred in some sentences. For example, the sentence "My husband took this for early stage CLL and after 9 months is in remission." was assigned a sentiment score of 1, while it indicated a very positive sentiment.

In the future, we will continue to annotate data to enrich the resource, weighing in the balance with regard to the number of positive and negative sentences. In this study, we used the Amazon Reviews dataset between May 1996 and July 2014. Recently a new version of review texts was released ranging from May 1996 to Sep 2023, 245.2% larger than the version we used. [34] As the new data became available, these annotation resources could be leveraged to develop new models, mine the bigger body of review texts, and provide insights into cancer symptom management using complementary therapies.

CONCLUSION

Our results showed the potential of Amazon consumer reviews in mining the outcomes of cancer symptom management. We presented the design and a first study of the annotation corpus from Amazon consumer reviews, publicly accessible through the OHNLP GitHub, focusing on cancer type, indicated symptoms, and symptom management outcomes. The annotation corpus and the developed baseline models laid the foundation for future enhanced methodology development to facilitate cancer symptom management in cancer patients using Amazon consumer reviews. In addition, we revealed the potential of using Amazon consumer reviews in mining the outcomes of cancer symptom management, thereby defining a promising starting point for any future argumentation analysis.

Acknowledgments

This project is supported by the Cancer Prevention Research Institute of Texas (CPRIT) RR230020, and NIH/NCCIH 2R01AT009457. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author Contribution

L.W.: conceptualized and designed the study, designed annotation guideline, developed NLP models, analyzed the data, and drafted the manuscript; Q.L.: designed the study, developed NLP models, analyzed the data and drafted the manuscript; R.L.: analyzed the data and revised the manuscript; T.B.H.: designed annotation guideline, conducted annotation and revised the manuscript; H.J.: designed annotation guideline and conducted annotation; M.H.: designed the study, and revised the manuscript; R.Z.: revised the manuscript; J.W.F.: advised on the study design and revised the manuscript; H.L.: conceptualized and designed the study, and revised the manuscript.

Conflicts of Interest

The authors do not have conflicts of interest related to this study.

Abbreviations

ASIN: Amazon Standard Identification Number
LLM: large language model
NER: named entity recognition
NLP: natural language processing
OHNLP: open health natural language processing

SFT: supervised fine-tuning
UMAP: uniform manifold approximation and projection
IAA: inter-annotator agreement

Data Availability

The annotated dataset is available at the OHNLP GitHub.[27]

References

1. Velicer CM, Ulrich CM. Vitamin and mineral supplement use among US adults after cancer diagnosis: a systematic review. *Journal of clinical oncology*. 2008;26(4):665-673.

2. Giovannucci E, Chan AT. Role of vitamin and mineral supplementation and aspirin use in cancer survivors. *Journal of clinical oncology*. 2010;28(26):4081.

3. Anderson JG, Taylor AG. Use of complementary therapies for cancer symptom management: results of the 2007 National Health Interview Survey. *The journal of alternative and complementary medicine*. 2012;18(3):235-241.

4. Ferrucci LM, McCorkle R, Smith T, Stein KD, Cartmel B. Factors related to the use of dietary supplements by cancer survivors. *The Journal of Alternative and Complementary Medicine*. 2009;15(6):673-680.

5. Erkan I, Evans C. The influence of eWOM in social media on consumers' purchase intentions: An extended approach to information adoption. *Computers in human behavior*. 2016;61:47-55.

6. Balasubramanian A, Thirumavalavan N, Srivatsav A, Yu J, Hotaling JM, Lipshultz LI, Pastuszak AW. An analysis of popular online erectile dysfunction supplements. *The journal of sexual medicine*. 2019;16(6):843-852.

7. Alsoudi AF, Loya A, Abouodah H, Koo E, Rahimy E. An Evaluation of Popular Online Eye Health Products on Amazon Marketplace. *Ophthalmic Surgery, Lasers and Imaging Retina*. 2023;54(3):147-152.

8. Fan JW, Wang W, Huang M, Liu H, Hooten WM. Retrospective content analysis of consumer product reviews related to chronic pain. *Frontiers in Digital Health*. 2023;5:958338.

9. Hu M, Liu B. Mining and summarizing customer reviews. 2004:168-177.

10. Ding X, Liu B, Yu PS. A holistic lexicon-based approach to opinion mining. 2008:231-240.

11. Boland K, Wira-Alam A, Messerschmidt R. Creating an annotated corpus for sentiment analysis of german product reviews. 2013;

12. He R, McAuley J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. 2016:507-517.

13. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*. 2018;77:34-49.

14. Wu T-J, Schriml LM, Chen Q-R, et al. Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. *Database*. 2015;2015:bav032.

15. Liu H, Bielinski SJ, Sohn S, et al. An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings*. 2013;2013:149.

16. Huggingface. Accessed March 12th, 2024. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

17. RADHA P, BHUVANESWARI NS. OPTIMIZING SENTIMENT ANALYSIS OF AMAZON PRODUCT REVIEWS USING A SOPHISTICATED FISH SWARM OPTIMIZATION-GUIDED RADIAL BASIS FUNCTION

NEURAL NETWORK (SFSO-RBFNN). *Journal of Theoretical and Applied Information Technology*. 2023;101(11)

18. Adams DZ, Gruss R, Abrahams AS. Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International journal of medical informatics*. 2017;100:108-120.

19. Babu NV, Kanaga EGM. Sentiment analysis in social media data for depression detection using artificial intelligence: a review. *SN computer science*. 2022;3(1):74.

20. Xiao S, Liu Z, Zhang P, Muennighof N. C-pack: packaged resources to advance general Chinese embedding. 2023. *arXiv preprint arXiv:230907597*. 2023;

21. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*. 2018;

22. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:220305794*. 2022;

23. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. *J Open Source Softw*. 2017;2(11):205.

24. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. *arXiv preprint arXiv:231006825*. 2023;

25. He H, Fu S, Wang L, Liu S, Wen A, Liu H. MedTator: a serverless annotation tool for corpus development. *Bioinformatics*. 2022;38(6):1776-1778.

26. Agarwal R, Singh A, Zhang LM, et al. Many-shot in-context learning. *arXiv preprint arXiv:240411018*. 2024;

27. OHNLP. Amazon review annotation. December 20, 2024. https://github.com/OHNLP/Amazon-review-annotation

28. Kwan ML, Weltzien E, Kushi LH, Castillo A, Slattery ML, Caan BJ. Dietary patterns and breast cancer recurrence and survival among women with early-stage breast cancer. *Journal of Clinical Oncology*. 2009;27(6):919.

29. Eisenberg DM, Davis RB, Ettner SL, Appel S, Wilkey S, Van Rompay M, Kessler RC. Trends in alternative medicine use in the United States, 1990-1997: results of a follow-up national survey. *Jama*. 1998;280(18):1569-1575.

30. Frenkel M, Ben-Arye E, Baldwin CD, Sierpina V. Approach to communicating with patients about the use of nutritional supplements in cancer care. *South Med J*. 2005;98(3):289-94.

31. Lu Q, Li R, Wen A, Wang J, Wang L, Liu H. Large language models struggle in token-level clinical named entity recognition. *arXiv preprint arXiv:240700731*. 2024;

32. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*. 2024:ocad259.

33. Pang B, Lee L. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*. 2008;2(1–2):1-135.

34. Hou Y, Li J, He Z, Yan A, Chen X, McAuley J. Bridging language and items for retrieval and recommendation (2024). *CoRR, abs/240303952*.

Supplemental Tables and Figures

Supplemental Table 1. Schema of the annotated labels.

Concepts	**Class/Type**		**Certainty**			
Cancer_type	Human	Pet	Positive	Negative	Hypothetical	Possible
Indicated_symptom	Cancer_related	Other	Positive	Negativ	Hypothetica	Possible

				e	l	
Harmful_outcome	Cancer_related	Other	Positive	Negative	Hypothetical	Possible
Favorable_outcome	Cancer_related	Other	Positive	Negative	Hypothetical	Possible
Product	Itself	Other	NA	NA	NA	NA

Supplemental Table 2. Inter-annotator agreements

Concept	F1
Overall	0.86
Cancer_type	0.97
Indicated_symptom	0.81
Harmful_outcome	0.63
Favorable_outcome	0.70
Product	0.91

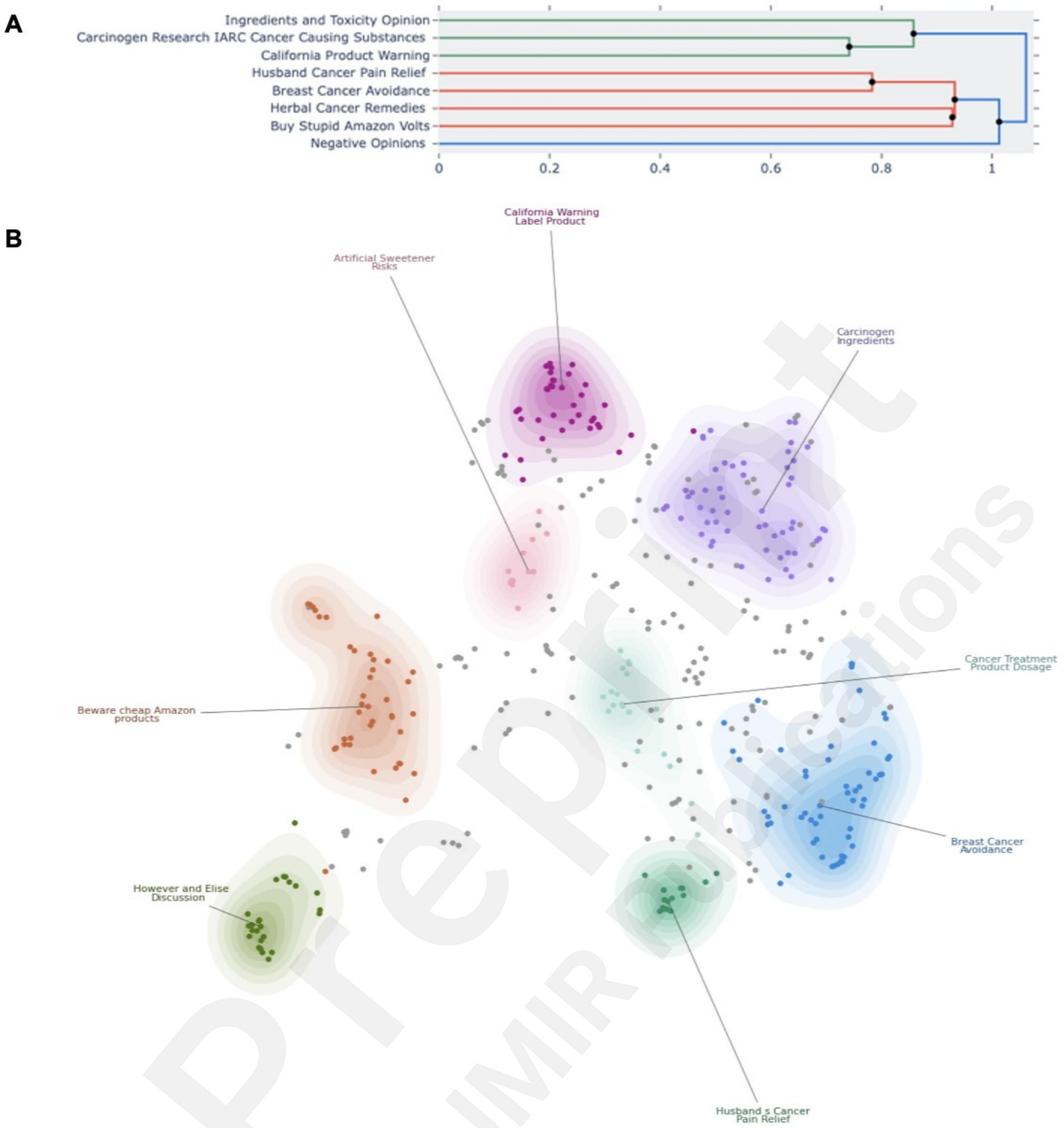


Figure 6. Hierarchical clustering (A) and topic modeling (B) based on sentences with semantic score 1

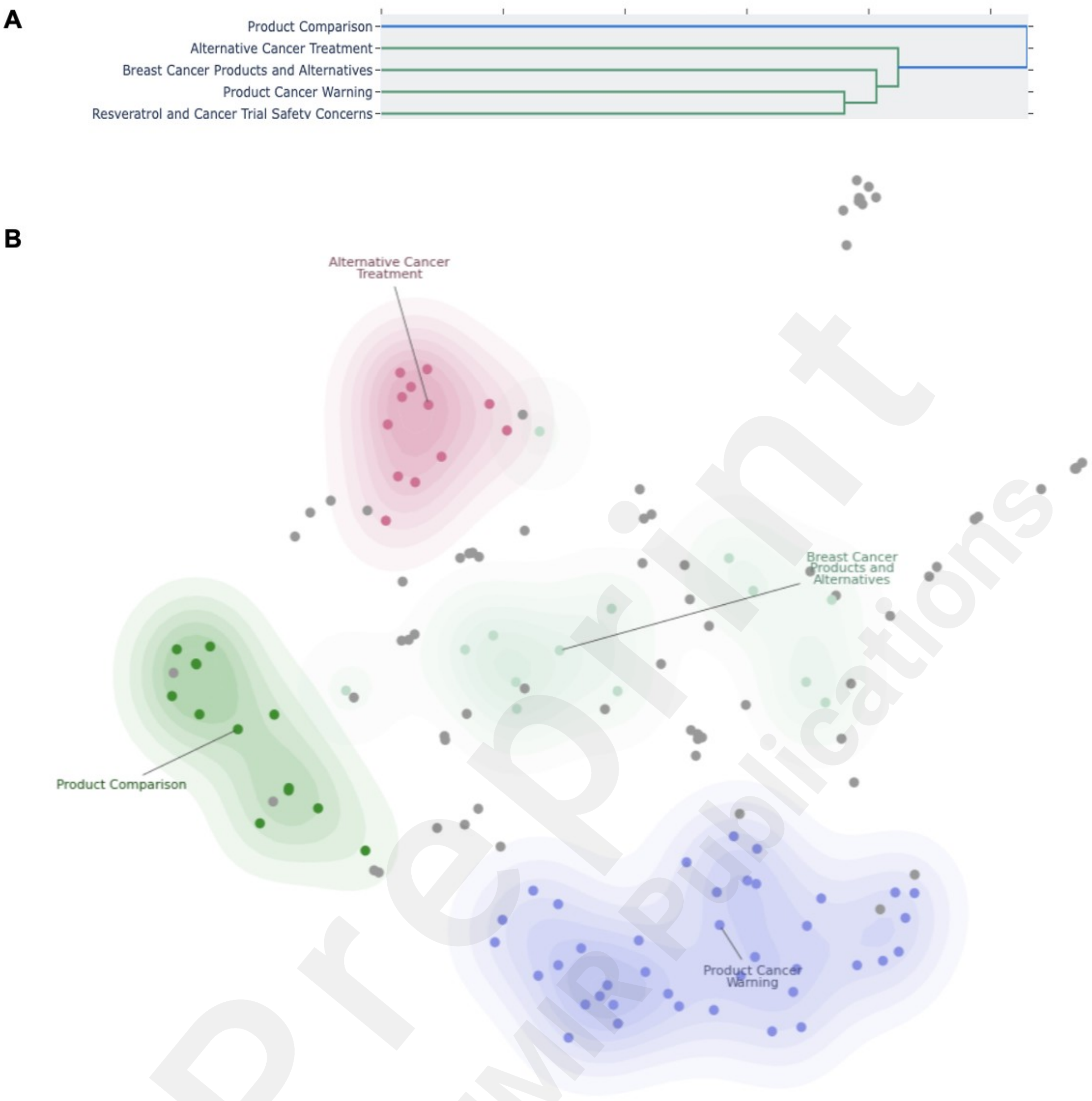


Figure 7. Hierarchical clustering and topic modeling based on sentences with semantic score 2

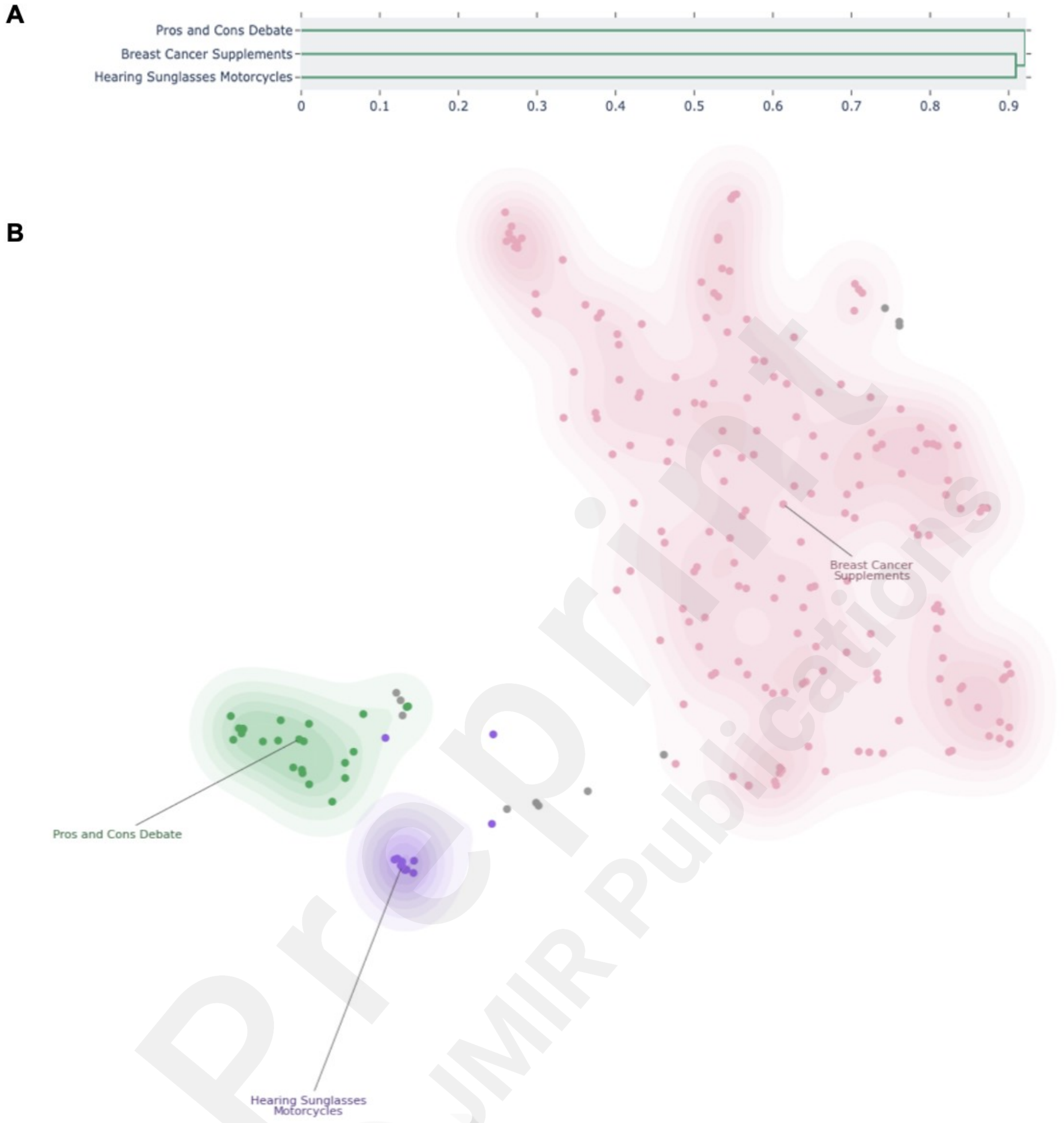


Figure 8. Hierarchical clustering and topic modeling based on sentences with semantic score 3

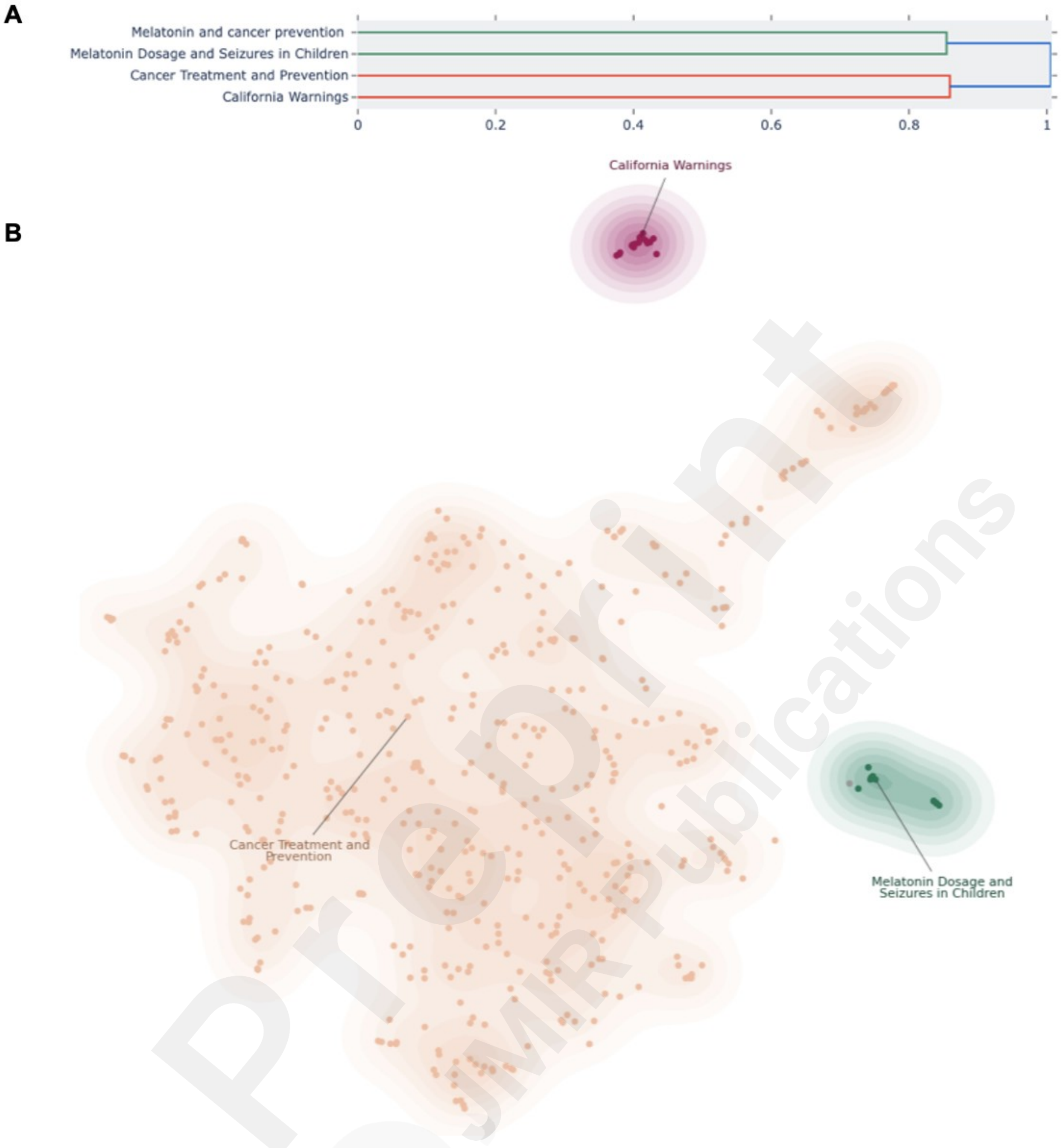


Figure 9. Hierarchical clustering and topic modeling based on sentences with semantic score 4

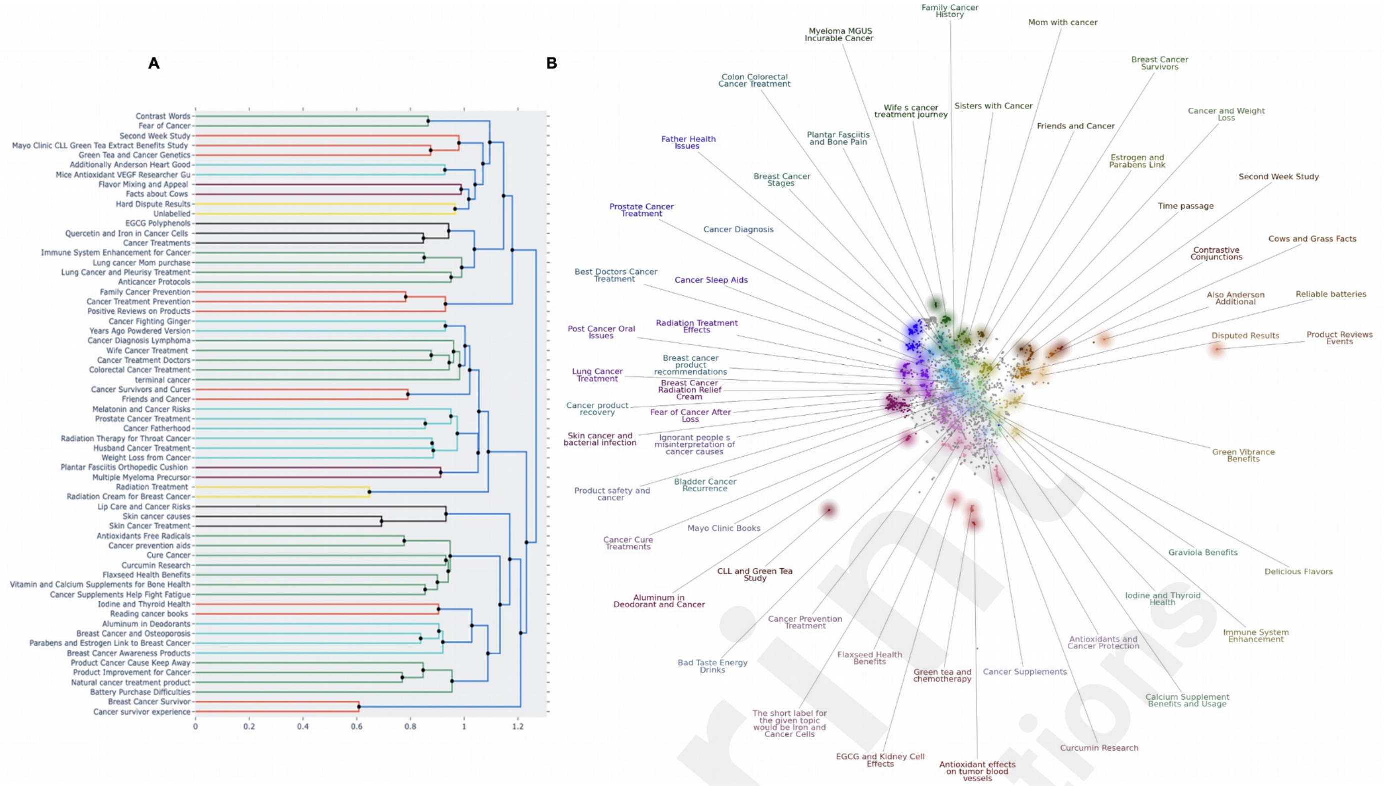
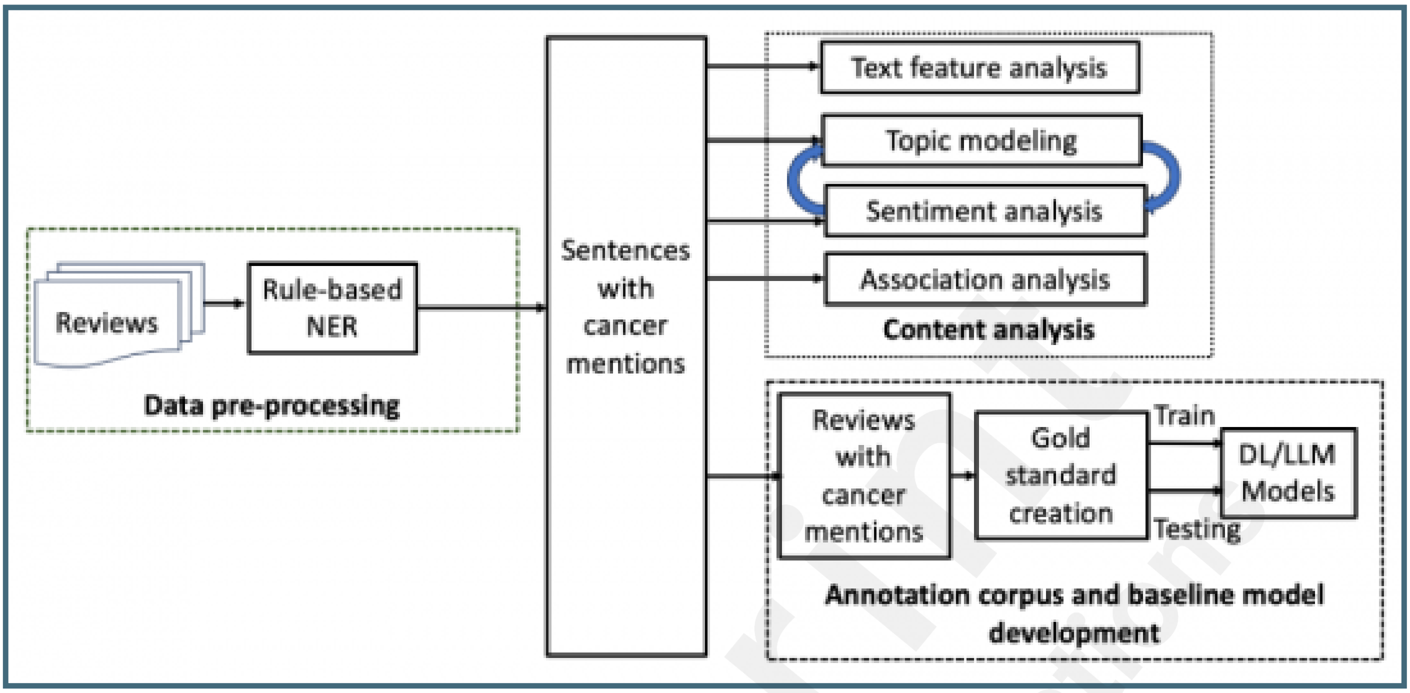


Figure 10. Hierarchical clustering and topic modeling based on sentences with semantic score 5

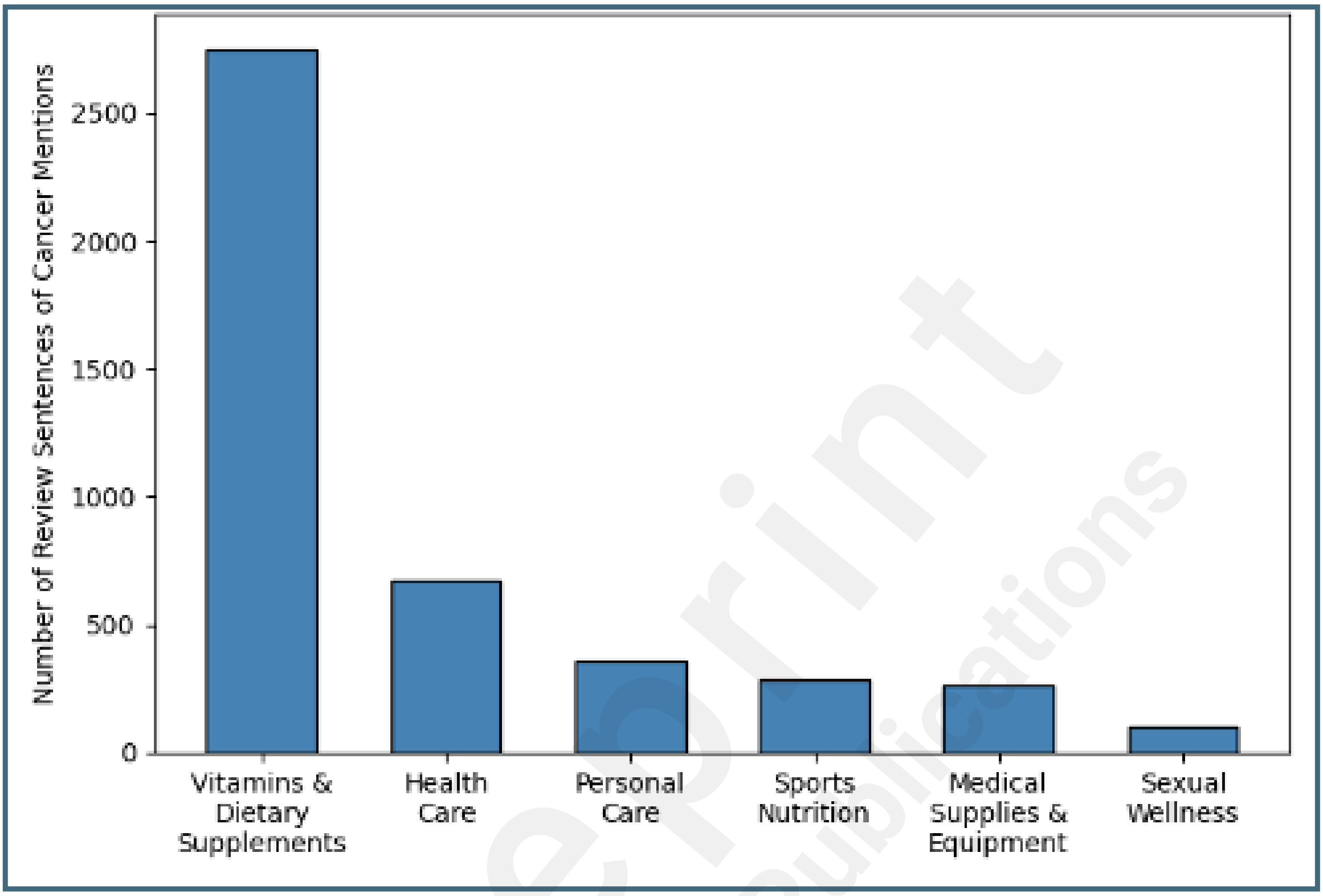
Supplementary Files

Figures

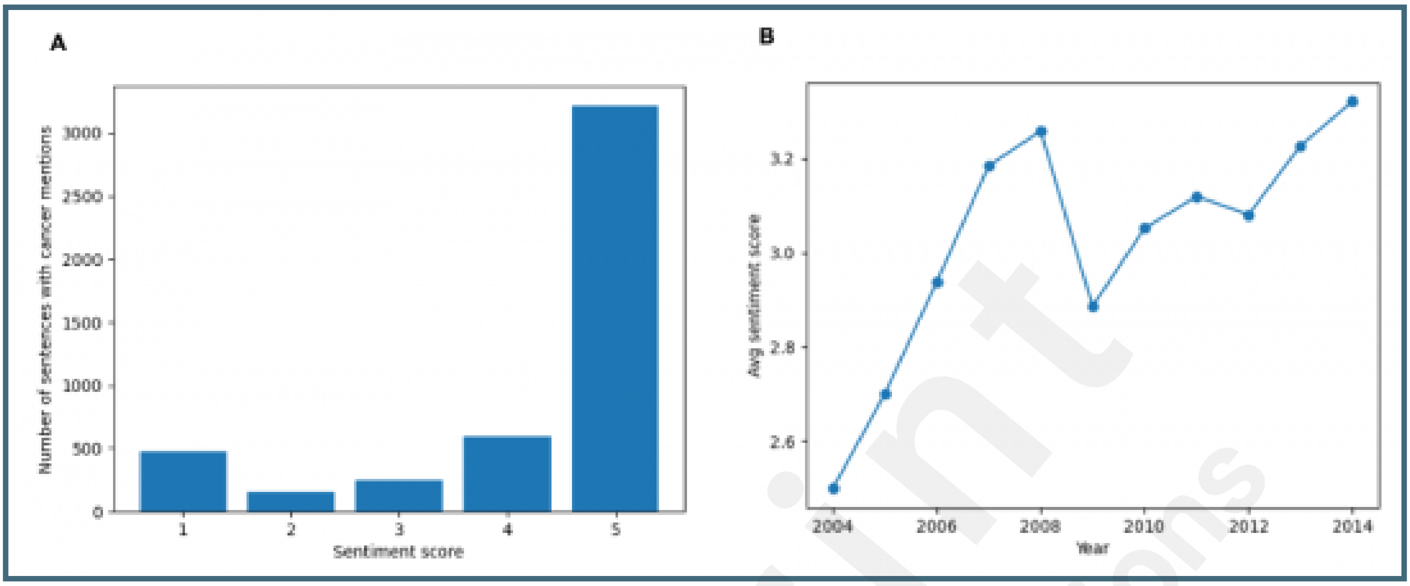
Study design. NER: named entity recognition; DL: deep learning; LLM: large language models.



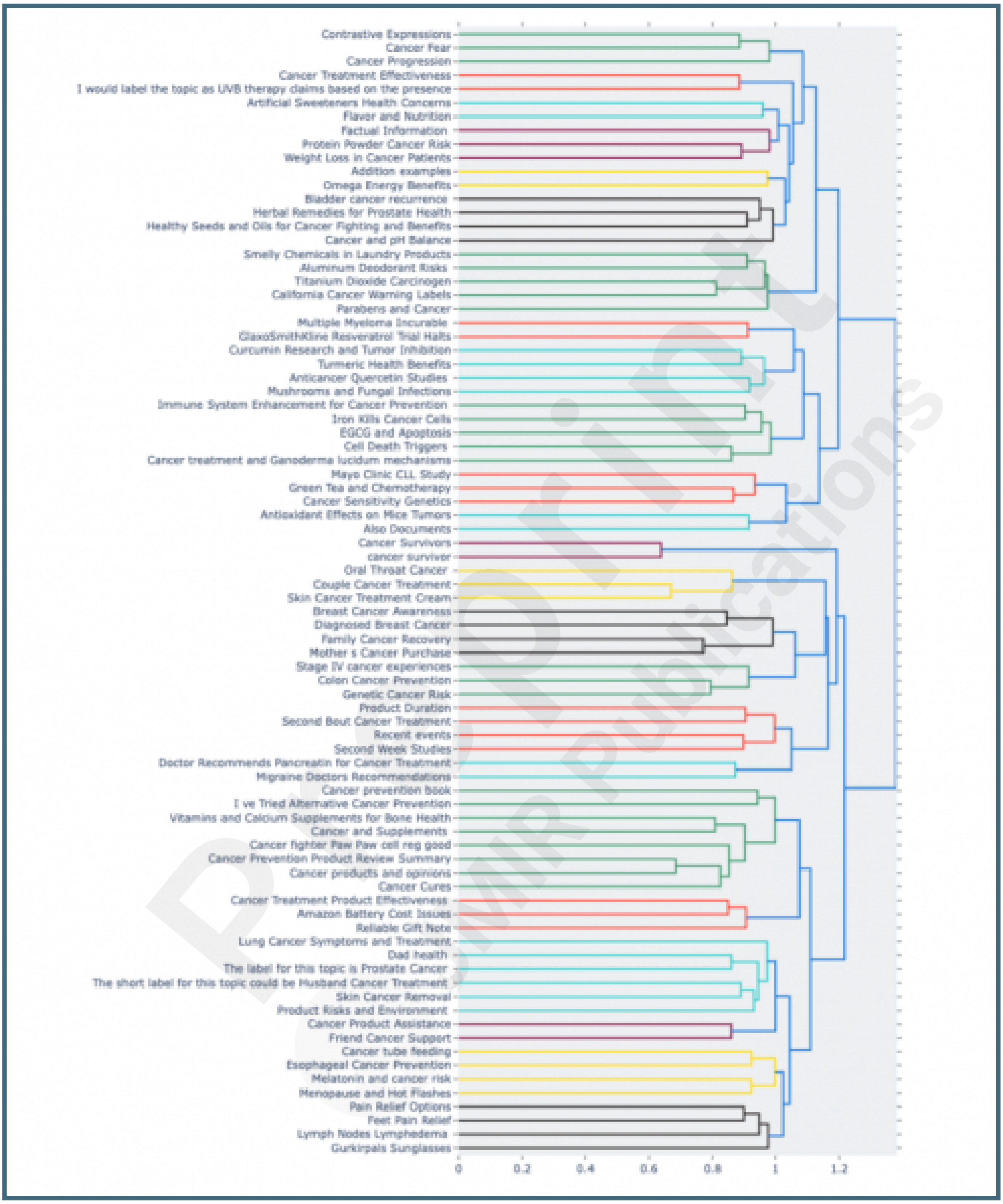
Distribution of product categories with review sentences of cancer mentions.



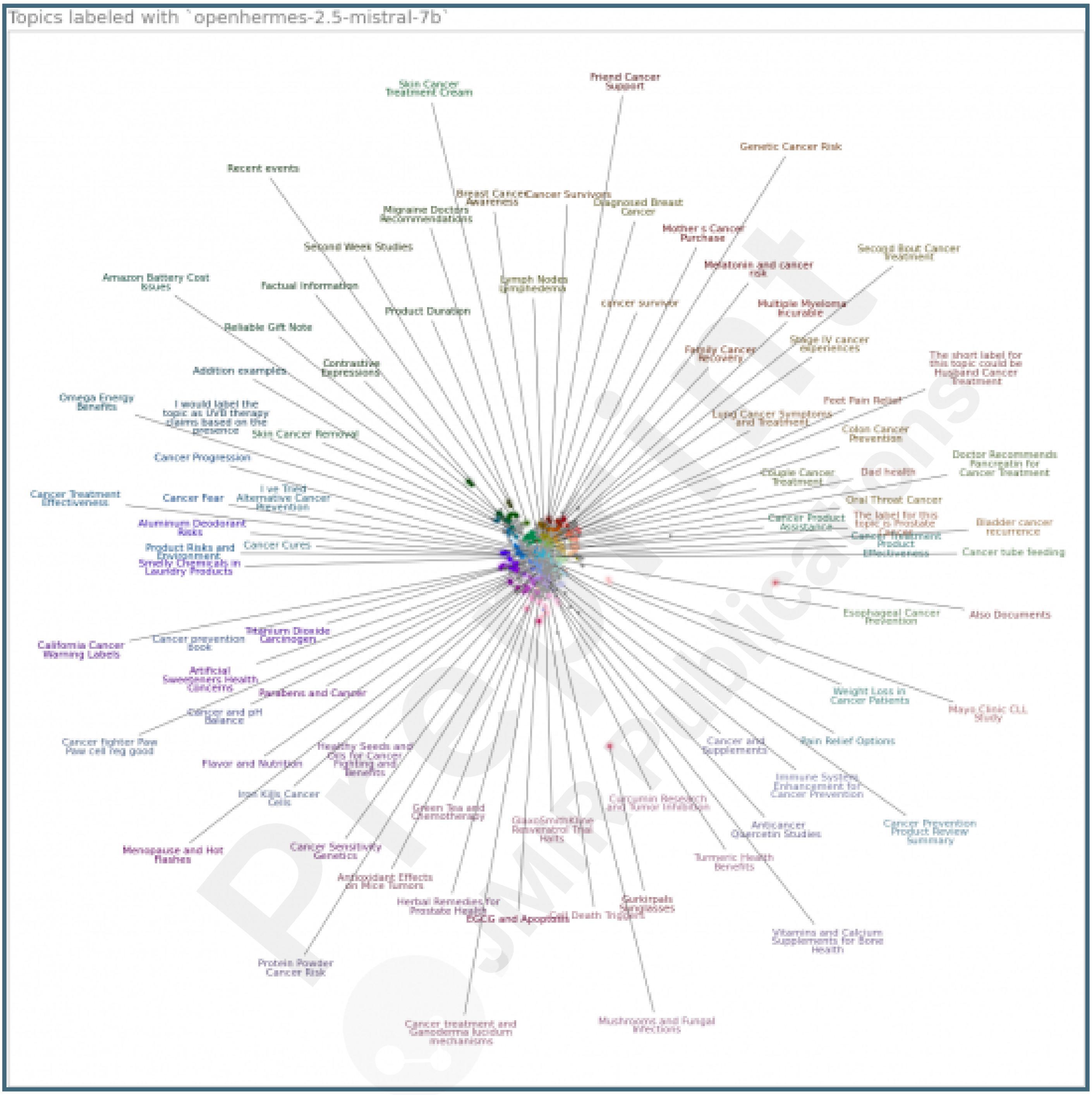
Analysis of sentiment scores in review sentences with cancer mentions. A. Distribution of sentiment scores across the review sentences. B. Trend of average sentiment score during May 1996 - July 2014.



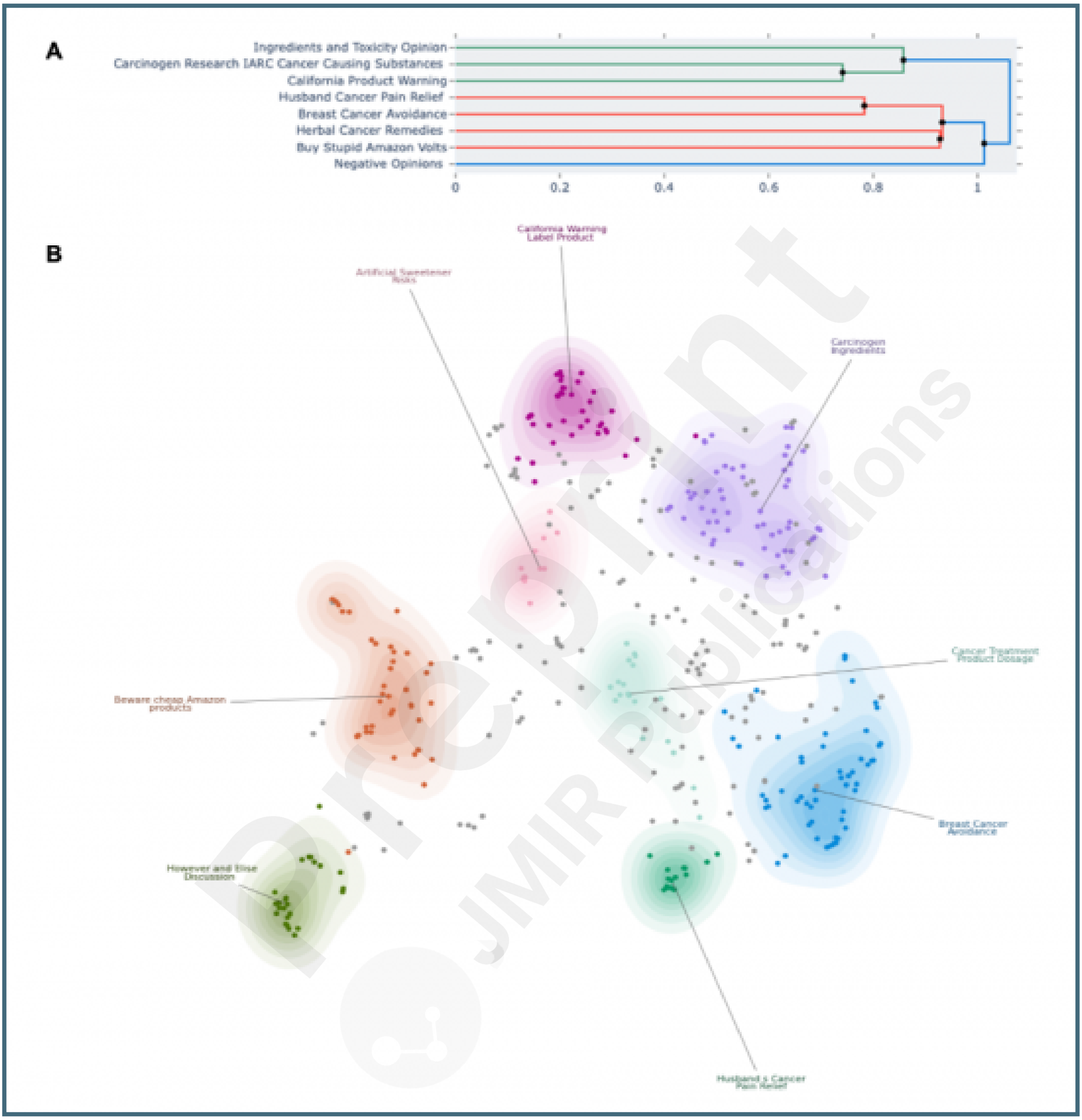
Hierarchical clustering of topics based on all sentences with cancer mentions.



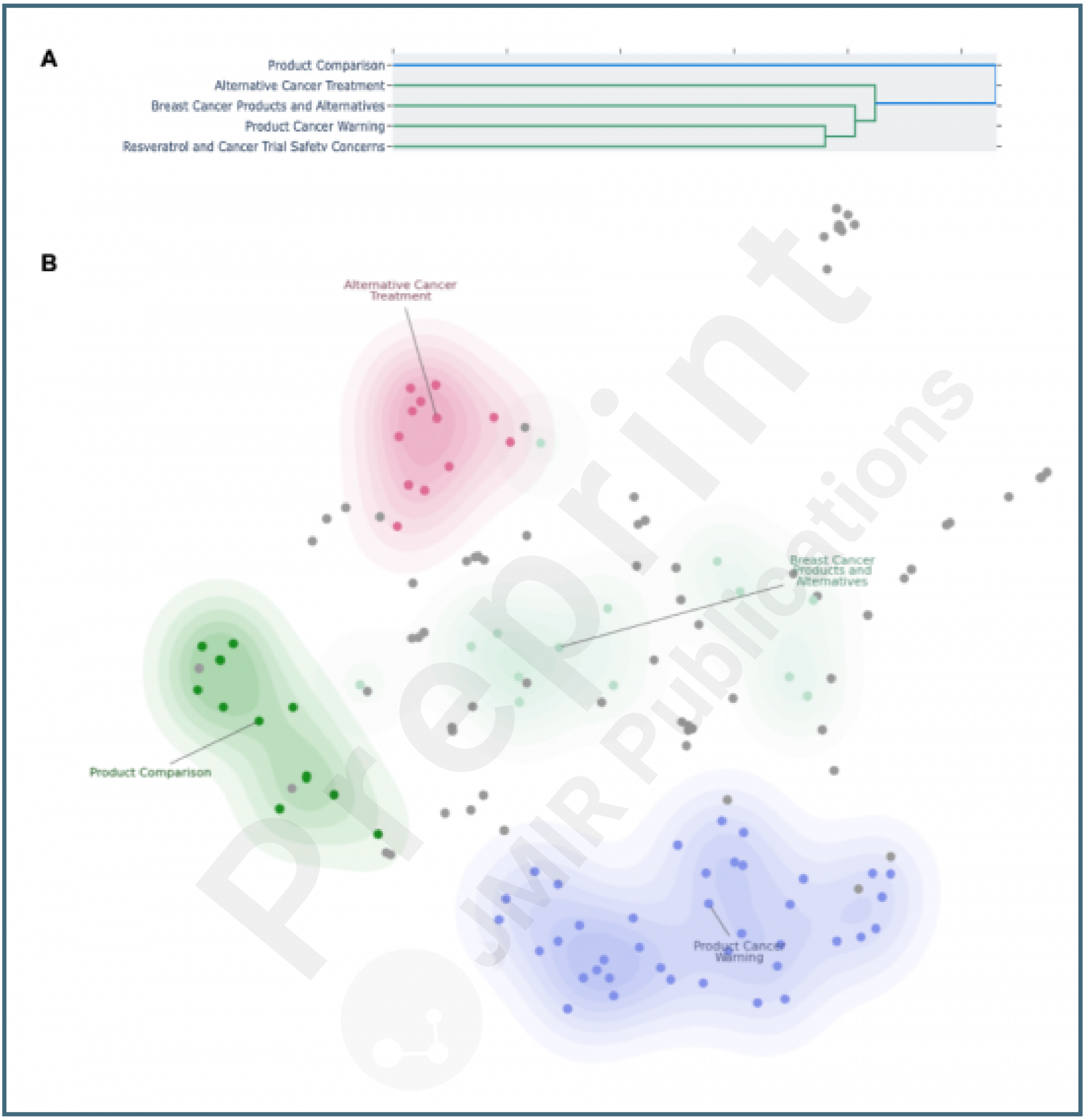
Topic modeling with openhermes-2.5-mistral-7b based on all sentences with cancer mentions.



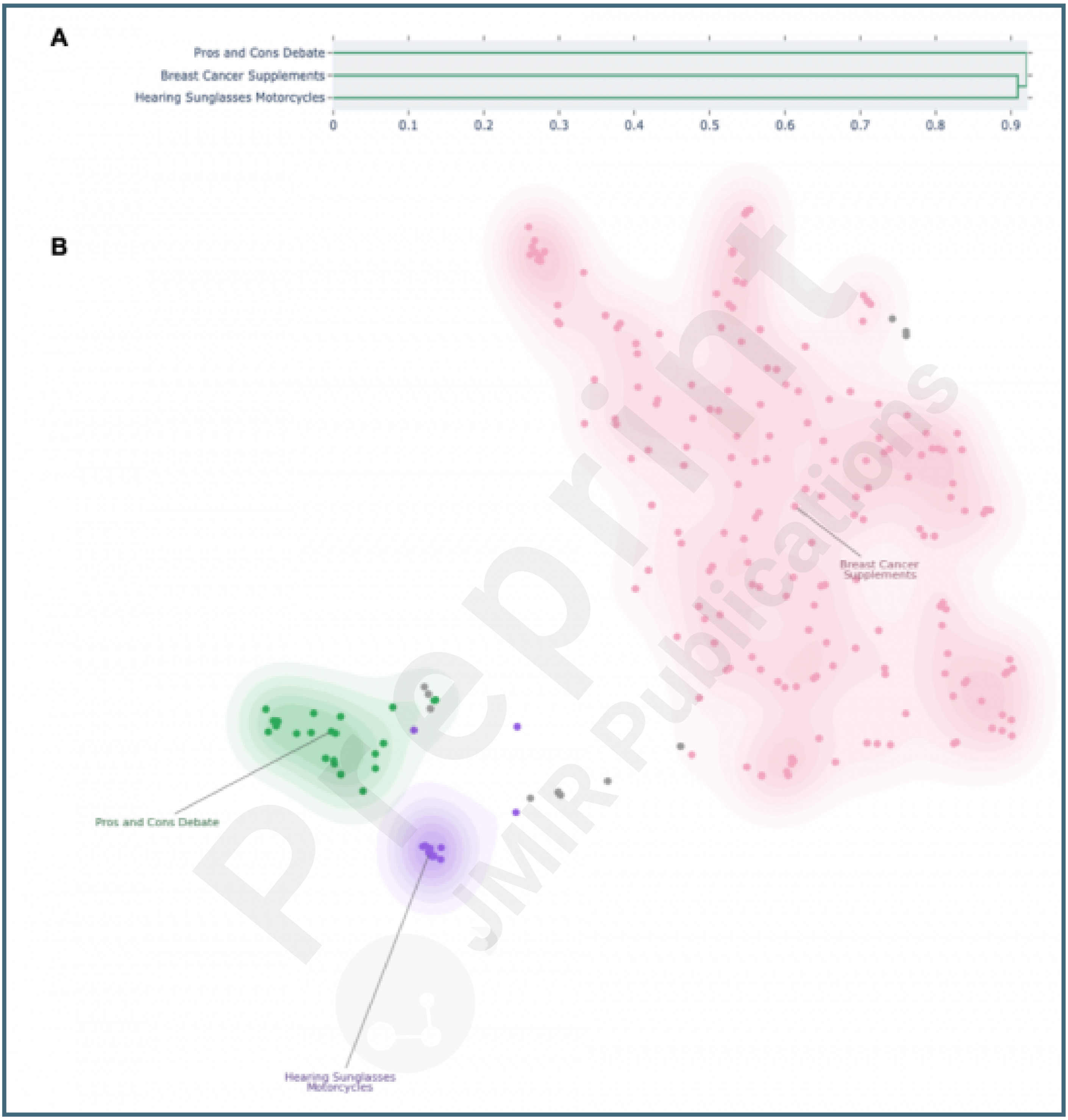
Hierarchical clustering (A) and topic modeling (B) based on sentences with semantic score 1.



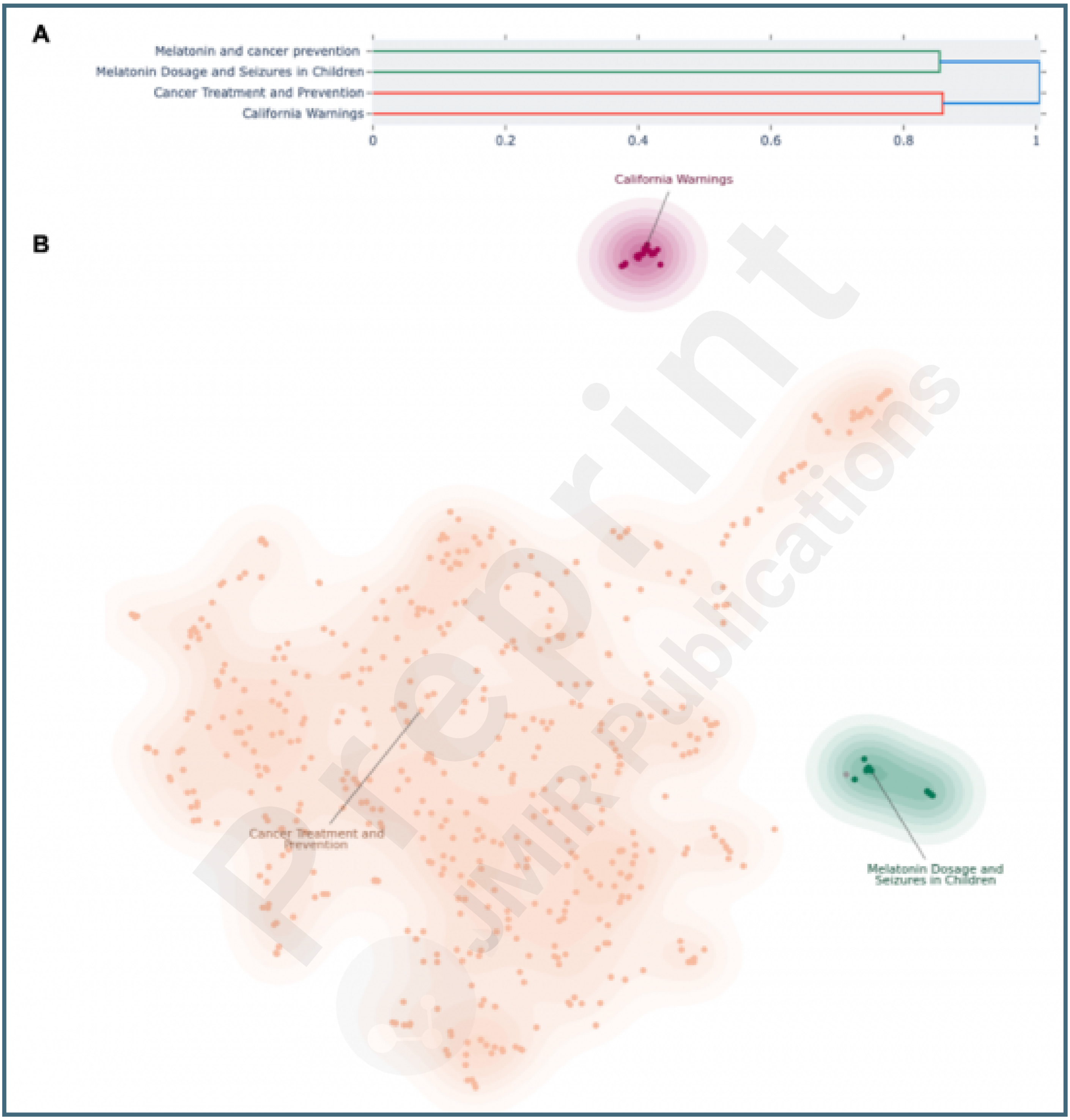
Hierarchical clustering and topic modeling based on sentences with semantic score 2.



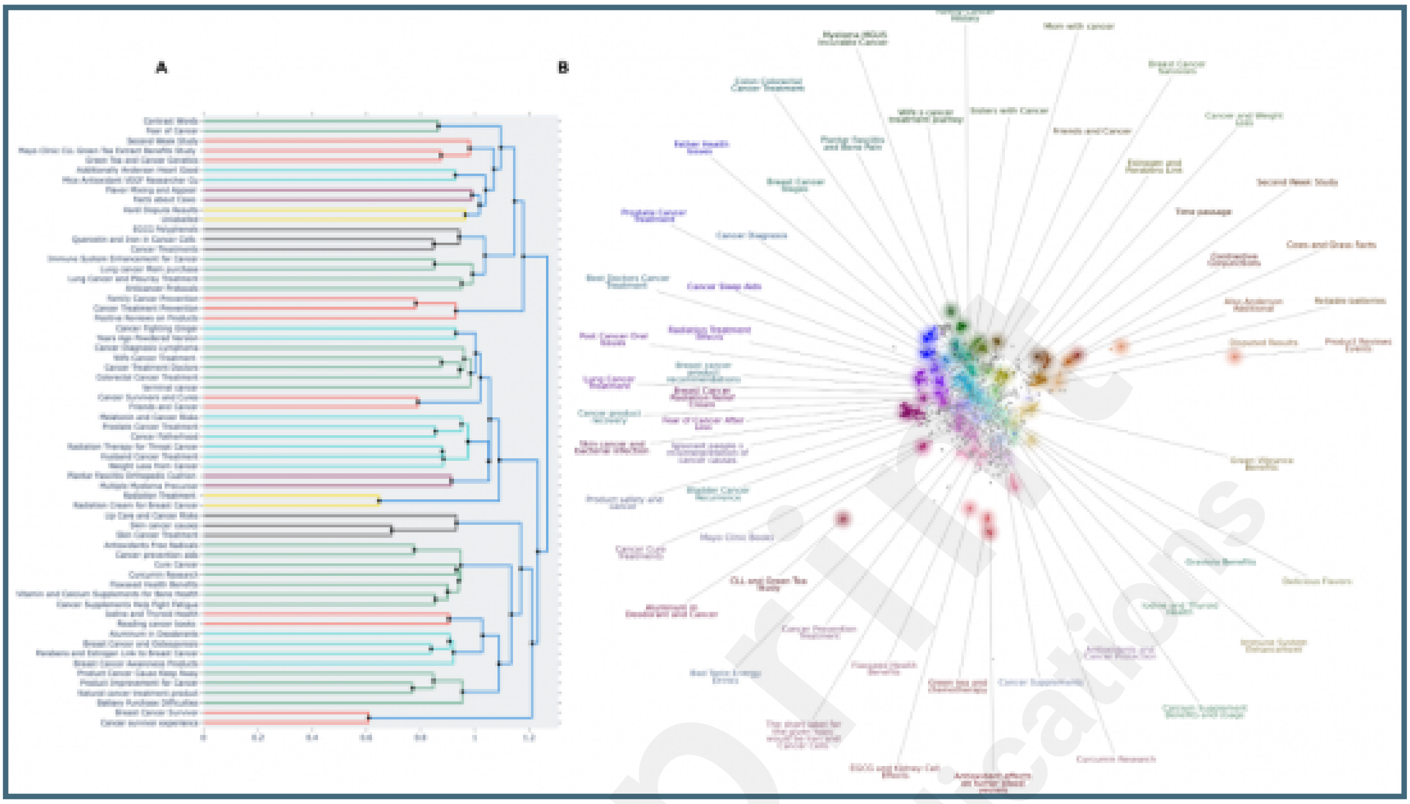
Hierarchical clustering and topic modeling based on sentences with semantic score 3.



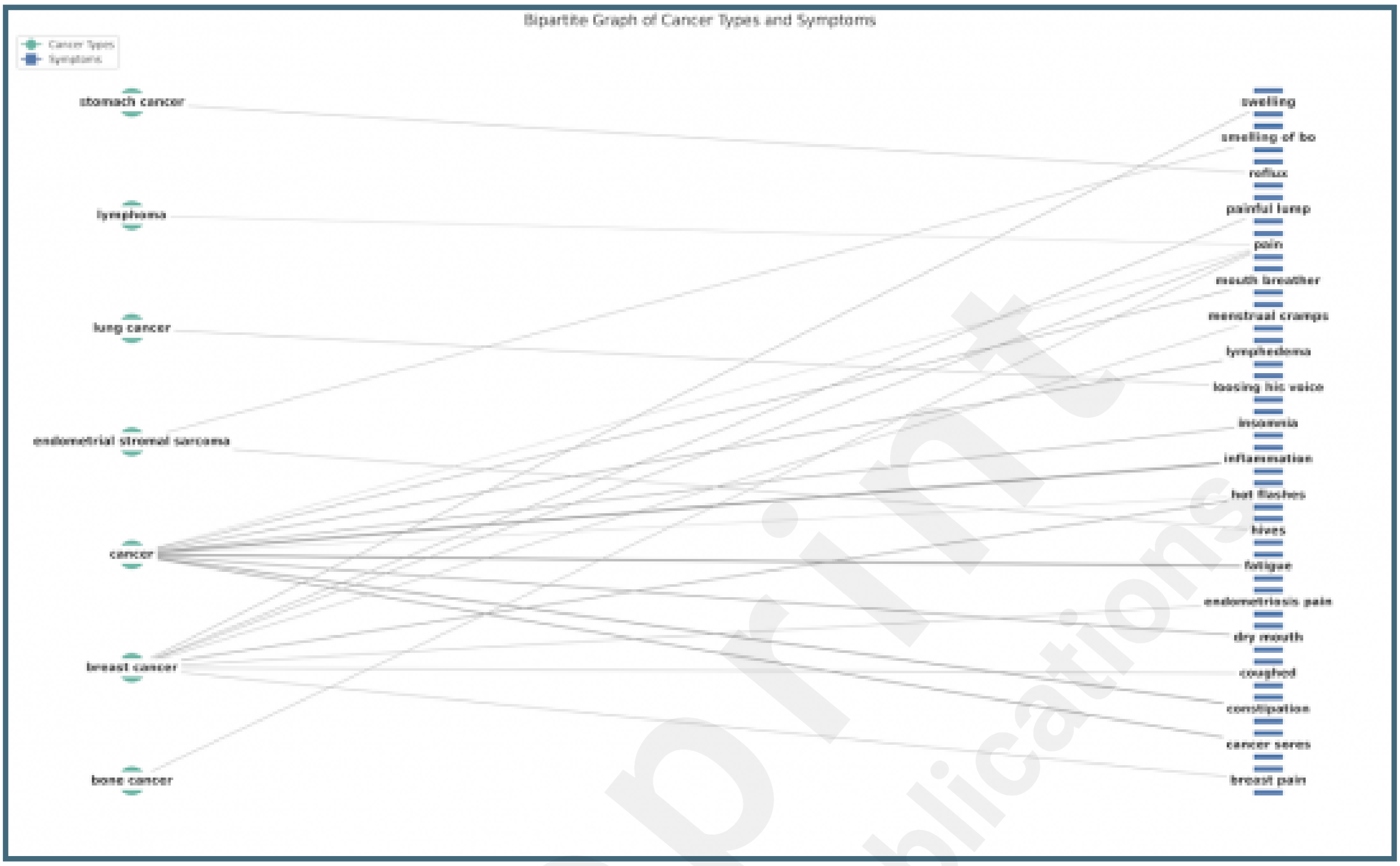
Hierarchical clustering and topic modeling based on sentences with semantic score 4.



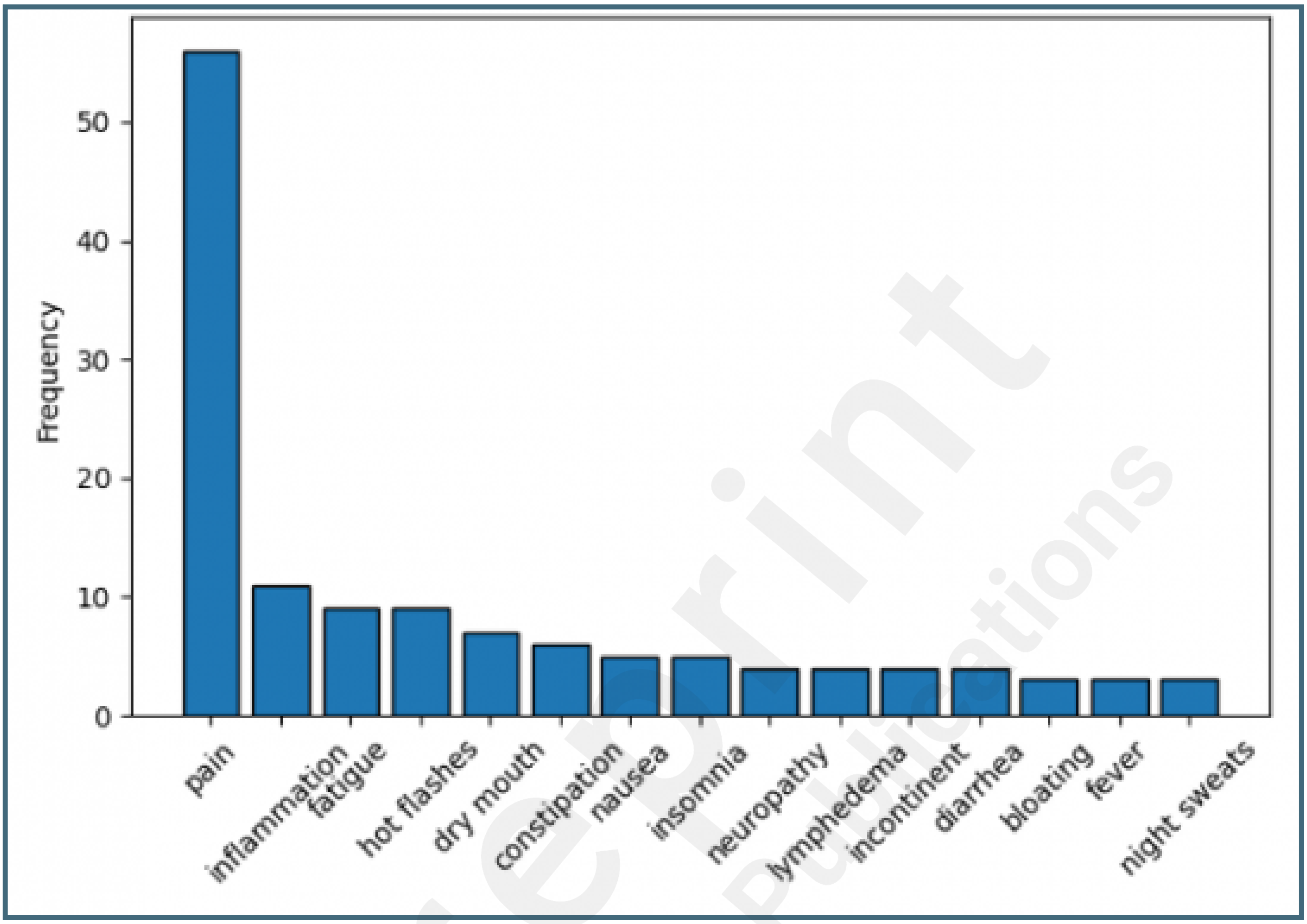
Hierarchical clustering and topic modeling based on sentences with semantic score 5.



Bipartite graph of cancer types and symptoms extracted by LLM.



Top 15 symptoms in reviews.



Multimedia Appendixes

Amazon Review - Annotation Guidelines.
URL: http://asset.jmir.pub/assets/e8653488f008601142672cb70267f1b0.docx

LLM prompting instruction.
URL: http://asset.jmir.pub/assets/d26872131872b888e4df69df43f1388d.docx