Sentiments in Oncology: A cancer treatment sentiment dataset

Souaad Hamza-Cherif[1,*], Aissa Benfettoum Souda[1], Nesma Settouti[2]

[1] Biomedical Engineering Laboratory, Faculty of Technology,
University of Tlemcen, Algeria
[2] LabISEN - Yncréa Ouest, Caen, France
* souad.hamzacherif@univ-tlemcen.dz

1 Domain Background

Sentiment analysis (SA) is the process of collecting, classifying, and interpreting opinions, emotions, and sentiments expressed in text. As a critical subfield of natural language processing (NLP), SA has gained significant traction, driven by the exponential growth of online information-sharing platforms and social media [PL08]. These platforms have transformed modern communication, enabling people to openly express their feelings, thoughts, and opinions on a wide array of topics, ranging from products and services to more personal and sensitive subjects such as health.

In healthcare, understanding patients' sentiments can be particularly valuable. Textual data derived from patient reviews, social media posts, or specialized health forums provide a rich resource for analyzing how patients perceive their treatment experiences [WHO24]. By leveraging SA techniques, researchers and healthcare providers can gain deeper insights into patients' attitudes towards treatments, identify common concerns, and evaluate the overall effectiveness and impact of medical interventions. This is crucial in designing personalized healthcare strategies and improving patient outcomes.

In this context, our study focuses on understanding the sentiments of cancer patients by analyzing their opinions and experiences with three specific drugs used in cancer treatment: Afinitor, Folfox, and Aromasin. These drugs are commonly prescribed for different cancer types and treatment stages. By compiling a diverse dataset of patient narratives, we aim to shed light on the emotional and experiential dimensions of cancer treatment. Our goal is to provide a resource for the NLP community and healthcare professionals that facilitates the development of advanced models capable of accurately capturing and interpreting the nuanced sentiments expressed by cancer patients [DFCM09]. This, in turn, could inform drug development, patient support services, and personalized treatment approaches.

2 Description of the dataset

Natural language processing (NLP), particularly in the domain of sentiment analysis (SA), presents significant challenges due to the complexity and variability of human language. In the context of healthcare, and more specifically cancer treatment, these challenges are amplified by the sensitivity and complexity of the subject matter [WW22]. The task in this study can be outlined in two main aspects:

Technical Aspect

The technical challenge revolves around identifying and implementing appropriate lexicon-based and machine learning models capable of accurately classifying patient comments according to their sentiment orientation (e.g., positive, negative, neutral). This involves selecting suitable features, fine-tuning models, and ensuring robust performance, even when faced with the nuanced and sometimes ambiguous language that patients may use when describing their experiences. The technical task also includes handling issues such as class imbalance, domain-specific terminology, and the subtleties of medical language.

Contextual Aspect

The healthcare domain, particularly when dealing with cancer, requires careful consideration due to its delicate nature. Cancer is a life-altering disease, and patients' narratives often include a mix of hope, frustration, fear, and relief. This study focuses on understanding the feedback of patients receiving specific treatments (Afinitor, Folfox, and Aromasin) for invasive and recurrent cancers. This is crucial, as patient perceptions can significantly influence their treatment adherence and overall outlook on recovery. In addition, some patients may hold preconceived notions or biases about certain cancer treatments, particularly when dealing with advanced or recurrent stages. Understanding these sentiments can therefore help in demystifying these treatments, offering more clarity, and possibly improving patient support mechanisms.

The value of this dataset extends beyond sentiment classification. It also serves as an essential resource for understanding the broader impact of cancer treatments from a patient-centric perspective. Analyzing and sharing this information can be beneficial for multiple stakeholders:

- **Patients:** They can find support in shared experiences, gaining insights and emotional validation from others facing similar challenges.
- **Healthcare providers:** Doctors and oncologists can gain a better understanding of how their patients perceive the treatments they administer,

allowing them to tailor communication and support strategies more effectively.

- **Pharmaceutical companies:** Insights from patient feedback can inform drug development, helping these companies understand real-world patient reactions and potentially adjust treatment formulations or communication strategies accordingly.

This dataset therefore aims to bridge the gap between patients' emotional experiences and the technical aspects of sentiment analysis, providing a valuable tool for enhancing the understanding and impact of cancer treatments.

3 Treatment overview

The dataset focuses on patient feedback for three specific cancer treatments. The goal is to collect and analyze patient comments to understand their experiences, perceptions, and emotions related to these treatments. By examining this feedback, we aim to uncover insights that can help improve patient care, inform healthcare providers, and guide pharmaceutical companies in developing better support strategies. Here's a simple explanation of each treatment covered in the dataset::

Aromasin (Exemestane): This medication is used for treating certain types of breast cancer in postmenopausal women. It works by blocking the production of estrogen, a hormone that can help some breast cancers grow [Aro24]. Aromasin is often prescribed:

- As a follow-up treatment for early-stage breast cancer after patients have already been treated with another drug called tamoxifen for 2 to 3 years.
- For advanced breast cancer when other hormone treatments have not been effective.

Afinitor (Everolimus): Afinitor is a cancer drug that targets and blocks a specific protein (mTOR) which helps cancer cells grow and spread. By stopping this protein, Afinitor slows down the growth of tumors [Vid24]. It is used for treating:

- Certain types of breast cancer in postmenopausal women. Some kidney cancers.
- Tumors in the pancreas, gastrointestinal tract, or lungs.

Folfox: Folfox is a combination chemotherapy treatment used for various types of cancer. It includes three drugs: leucovorin (folinic acid), fluorouracil, and oxaliplatin. It is commonly used for treating cancers, especially those that affect the digestive system [Fol24]. Folfox is sometimes referred to as "Oxaliplatin de Gramont" or "OxMdG," which stands for a modified version of this regimen.

The aim of this dataset is to offer a comprehensive resource that reflects patient sentiments, helping to build models that accurately classify and interpret the emotions and reactions associated with these treatments. This, in turn, can support healthcare professionals, researchers, and pharmaceutical companies in enhancing patient care and communication.

4 Description of dataset

4.1 Data collection

The dataset was collected by scraping various multilingual websites (in English, French, and German), including Reddit, Drugs, AskPatient, and WebMD (for English), Carenity and Espoir de Vie Cancer du Sein (for French), as well as Medzin Forum Medecine (for German). The scraping focused on people's comments regarding the three cancer treatments: Afinitor, Aromasin, and Folfox. The extracted data were compiled into a local CSV database (Table 1), which organizes comments related to various cancer treatments (Afinitor, Aromasin, and Folfox) and includes information on the type of cancer being treated when available. Additionally, the most frequent words used in these comments are summarized in Table 2. The scraping process gathered comments from relevant websites, covering a period from 2005 to 2024.

Table 1: Overview of Medication Comments Dataset

ID	Disease	Comment	Medicament
1	Breast	I have had several discussions with my doctor ...	Aromasine
2	Breast	I was actually a participant in the clinical t...	Aromasine
3	Breast	Had a lumpectomy for Stage I Advanced Breast C...	Aromasine
4	Breast	I was on Arimidex for 6 months and started hav...	Aromasine
5	Breast	After surgery I was pronounced cancer free but...	Aromasine

In order to standardize the database, the data in French and German were translated into English using the automatic translation model: the Google Translate module, also called googletrans, which is a Python library that provides a user-friendly interface to use the Google Translation API.

Rank	Common Words	Count
0	get	5004
1	go	3824
2	take	3654
3	like	3482
4	would	3374
5	well	2999
6	one	2903
7	good	2844
8	make	2832
9	year	2791
10	time	2785
11	day	2740
12	people	2674
13	treatment	2660
14	cancer	2652
15	know	2564
16	say	2537
17	also	2382
18	think	2259
19	work	2150

Table 2: Top 20 most common words and their Counts

4.2 Data pre-processing

Data cleaning is an important process in an NLP task, especially when the processed information comes from the web and is noisy with meaningless and unstructured comments. It is then necessary to improve the quality of the data by detecting and removing noise in order to refine the sentiment analysis. The main steps followed in the preprocessing are:

- Lowercase conversion: This step is necessary to avoid duplicates, i.e. even if the meaning of words such as “Health” and “health” is the same, they are treated as separate lexical units if they are found in different cases.

- Duplicate line removal: Duplicate line removal is a crucial step in data preprocessing for NLP tasks, as it can lead to inaccuracies in the models and increase the processing time.

- Punctuation removal: This involves removing all punctuation from the text that does not provide any useful information is a process that helps improve the data classification performance.

- Removal of stop words: these are very common words in the language studied but that do not provide any informative value to understand the "meaning" of a document or corpus, so they are removed.

• Removal of rare and common words: to avoid the negative impact that rare and common words can have on the classification, we remove them by counting their frequency of appearance in the text, this helps to reduce the noise they generate in the text.

• Removal of short sentences: filtering out small texts or texts below a certain length is also an important step in data preprocessing. Such texts (short sentences), usually do not contain enough information and can therefore be removed to improve the quality of the dataset. The process consists of determining the minimum length of texts to keep in the dataset, identifying texts below the threshold and removing them.

• Tokenization: this is the act of analyzing a text into tokens, in other words, the text is segmented into linguistic units such as words, punctuation marks, numbers, alphanumeric data. Each element corresponds to a token that will be useful for the analysis.

• Lemmatization: this consists of replacing each word with its canonical form, for example, the word "known" is replaced by its canonical form "know". This step is useful for the thematic classification of texts because it treats different variants resulting from the same form or root as a single word.

4.3 Dataset features

The dataset is designed to capture a broad range of patient sentiments regarding three major cancer treatments: Afinitor, Aromasin, and Folfox. These comments, sourced from multiple multilingual platforms, offer insights into patient experiences, perceptions, and expectations before and after treatment. By processing this data, we gain a unique opportunity to analyze how patient sentiments evolve over time, especially in response to specific treatments. Figure 1 illustrates the distribution of comments across each treatment, distinguishing between comments recorded before and after treatment. This distribution highlights both the volume and diversity of patient feedback. Table 3 further summarizes the dataset by showing the number of comments initially collected and the subset retained after preprocessing for each treatment, spanning over a period ranging from 2005 to 2024.

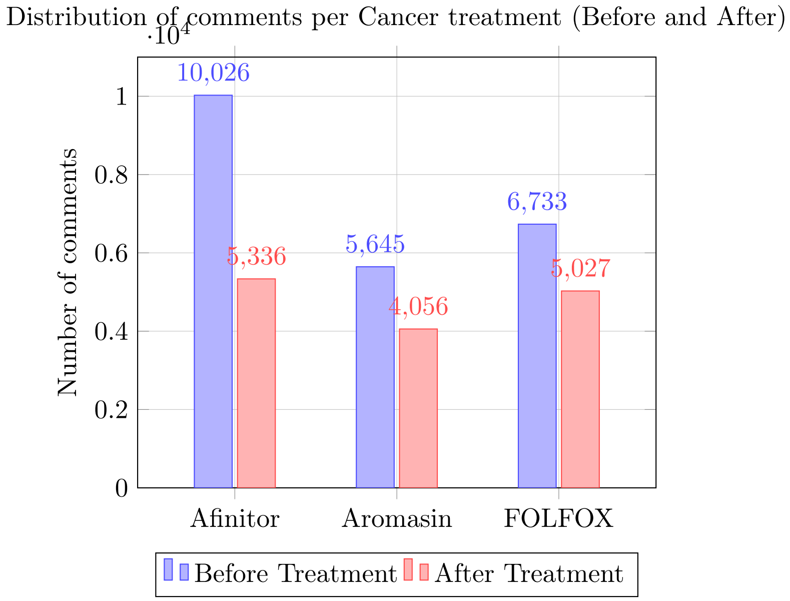

Figure 1: Number of comments distribution per Cancer treatment before and after processing

Treatment	**# Comments Before**	**# Comments After**	**Period**
Afinitor	10026	5336	2012-2024
Aromasin	5645	4056	2005-2024
Folfox	6733	5027	2019-2024
Total	22404	14419	

Table 3: Number of comments before and after preprocessing for each treatment

The potential applications of this dataset are extensive. For example, it can aid in building machine learning models capable of classifying patient sentiments and identifying trends or common themes in treatment feedback. This resource is also invaluable for healthcare professionals who wish to understand patient concerns and preferences better, as well as for pharmaceutical companies seeking real-world feedback to improve drug development and patient care strategies. The emergent potential of this dataset lies in its ability to serve as a foundation for advanced sentiment analysis models that could contribute to more personalized and responsive cancer care. By capturing a diverse array of patient experiences, this dataset enables a deeper exploration of the emotional and psychological dimensions of cancer treatment, providing actionable insights to improve patient support systems and communication in the healthcare industry.

References

[Aro24] Base de données publique des médicaments - fiche information aromasine, 2024. Last modification on September 2024.

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

[Fol24] Cancer research uk - folfox information, 2024. Last modification on September 2024.

[PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[Vid24] Vidal. Gamme de médicaments afinitor, 2024. Last modification on August 2024.

[WHO24] World health organization - cancer, 2024. Accessed on October 2024.

[WW22] Hooman Wu and Annie Wen. *Natural Language Processing for Healthcare: Advances and Applications*. Academic Press, 1st edition, 2022.