Extracting cancer knowledge from crowd wisdom with NCDs Listener, a social listening tool for non-communicable diseases

Wuttichai Khamna, Ratchanont Thippimanporn, Kannika Wiratchawa, Thanapong Intharah*
Visual Intelligence Laboratory, Department of Statistics, Faculty of Science,
Khon Kaen University, Khon Kaen, 40002, THAILAND
wuttichai.kha@kkumail.com, ratchanont.t@kkumail.com, kannikaw@kkumail.com, thanin@kku.ac.th
(corresponding author)*

Abstract

Non-communicable diseases (NCDs) are chronic conditions that cause over 41 million deaths annually, with most occurring in low- and middle-income countries. Social media has become a critical platform for individuals to share experiences and access information about NCDs. Social (media) listening offers valuable insights by analyzing user discussions, but existing tools are closed-source and commercial. This study seeks to simplify the extraction of NCD-related knowledge from social media, making it easier for the public to understand and access information. It also explores how the NCD community shares its lived experiences online.

We proposed an open-source social (media) listening tool called NCDs Listener to collect, analyze, summarize, and visualize data. The comments about NCDs were collected from public posts on Facebook and Reddit. The K-means method was used to separate the conversation topics. The preliminary data was analyzed using descriptive statistics. Additionally, a Generative AI model summarizes the extracted knowledge in human-readable sentences.

In our case study, The results from 1,402 Facebook comments found that most commenters were female and often shared the experiences of caregivers with cancers. 29 diseases and 24 symptoms were mentioned. The treatment method was mostly surgery. The conversations focused on talking about cancer, symptoms before and after cancer, and expressed concerns about cancer. Reflecting the online community for cancer patients as there is an ongoing exchange of information and encouragement.

Our NCDs Listener tool is open source and can extract knowledge related to NCDs using the wisdom of crowds approach. This knowledge can be used as a guideline for treatment or the development of effective care to meet patients' needs. Collected social media data can also be used in future statistical analysis and information data science approaches.

Keywords: Generative AI, Data Visualization, Crowd Wisdom, Social Listening, Non-Communicable Diseases

1. Introduction

Non-communicable diseases (NCDs) are chronic conditions that cannot be spread from person to person and are not caused by infectious agents like bacteria or viruses. Instead, NCDs often result from unhealthy lifestyle choices and typically develop over long periods. Globally, NCDs are responsible for 41 million deaths annually, accounting for 74% of all deaths. Of these, 17 million individuals die before the age of 70, with 86% of these premature deaths occurring in low- and middle-income countries. In fact, 77% of all NCD deaths are concentrated in these regions. Cardiovascular diseases claim the highest toll, with 17.9 million deaths each year, followed by cancers (9.3 million), chronic respiratory diseases (4.1 million), and diabetes (2.0 million, including kidney disease associated with diabetes). Collectively, these four disease groups account for over 80% of all premature NCD-related deaths [1]. Decisions regarding the treatment or care of non-communicable diseases (NCDs) often raise numerous concerns, such as identifying the behaviors that contribute to these diseases, recognizing symptoms before and after diagnosis, and determining the best course of treatment. Despite the abundance of available information on NCDs, relatively few platforms present this knowledge in a friendly, accessible way for patients and caregivers. Those with firsthand experience—either living with NCDs or caring for someone who does—are often best suited to address these concerns. Social media has become essential for connecting these individuals, transforming how patients learn about their conditions, forming peer support networks, and sharing personal experiences in recent years [2].

Between January 2023 and January 2024, the global number of social media users increased from 4.72 billion to 5.04 billion, an 8% growth, representing an additional 320 million users. This growth is expected to continue. As social media platforms evolve to meet an ever-wider range of consumer needs [3], social (media) listening has emerged as a valuable technique for tapping into users' insights. Social media data can be leveraged to make informed decisions driven by the Wisdom of Crowds (WoC)—a concept where diverse knowledge and viewpoints converge to produce collective intelligence. In many cases, the aggregated judgment of a group surpasses that of any individual member [4]. Studies have shown that diversity in social media communications enhances this crowd wisdom, making it a reliable predictor for various tasks [4], [5]. This approach has become increasingly popular in marketing, offering valuable insights for decision-making [6].

Previous social (media) listening studies have demonstrated how platforms like Facebook can be used to explore public perceptions of virtual reality (VR) in healthcare, highlighting its potential for patient education, therapy, and rehabilitation [7]. In these studies, tools like ATLAS.ti were employed instead of NLP and qualitative analysis. While some research has focused on developing social listening systems capable of analyzing sentiment and tracking trending topics [8], much of this work has been centered on industries like automobiles rather than healthcare. Additionally, many social listening tools, such as Brand24 and YouScan, require expensive subscriptions, making them inaccessible to the general public or independent researchers. Moreover, these tools are typically designed to serve large organizations, limiting their usefulness for individuals wishing to analyze social media data personally [9].

To extract knowledge from public posts on social media into an easy-to-understand format, which will help reduce the time spent searching for information for those who want to study non-communicable diseases. In addition, it explores how the NCD community describes the lived experience on social media platforms to provide perspectives and insights into living with or caring for someone with NCDs. It also helps the general public to access and understand information about non-communicable diseases. Our contributions include:

- Develop NCDs Listener web application, a social (media) listening tool for extracting and analyzing public comments on NCDs-related posts, displaying the information in a customizable dashboard.
- Applying natural language processing (NLP) techniques to derive meaningful insights from social media data.
- Integrating large language models (LLMs) to enhance the social (media) listening process.
- Leveraging the "wisdom of crowds" (WoC) approach to extract knowledge from social media discussions, specifically focusing on cancer-related insights.

2. Related Work

We will discuss research on social listening for non-communicable diseases and the applications of social media listening.

2.1 Social listening for Non-communicable diseases

Many researchers are working on expanding knowledge on social media listening for non-communicable diseases in literature, and some key contributions are providing support for finding user behaviors and situations in different cases worldwide. Some of the essential papers are included in this section.

For disease-specific social listening studies and various social media listening tools and data collection and analysis techniques, Cook et al. [10] used social media listening tools (Social Studio) to analyze patient experiences with dry eye disease across Twitter, blogs, and forums. They explored symptoms, treatments, and quality of life. Rodrigues et al. [11] focused on European social media conversations to understand the experiences of lung cancer patients, caregivers, and healthcare professionals. Using Talkwalker and SocialStudio. Chauhan et al. [12] analyzed the experiences of melanoma patients across 14 European countries using SocialStudio and Talkwalker. They identified significant impacts on patients' daily lives and emotions. Manuelita Mazza et al. [13] used social listening on Twitter, patient forums, and blogs to explore metastatic breast cancer patient experiences. Zinaida Perić et al. [14] investigated GVHD patient needs and lifestyles across Europe using Talkwalker to collect data from Twitter, Facebook, Instagram, and YouTube. Wolffsohn et al. [15] examined social media discussions about presbyopia across seven countries using Social Studio. They focused on symptoms and quality of life impacts. Erica Spies et al. [16] studied lupus patients' experiences using SocialGist to collect data from blogs and forums. The research combined quantitative and qualitative methods to analyze quality of life, treatment efficacy, and unmet needs.

2.2 Applications of Social Media Listening

Different studies have employed various tools and techniques to analyze social media data, focusing on understanding public sentiment and opinions. For analysis techniques employing NLP, Kamaran H. Manguri et al. [17] analyzed Twitter data from the COVID-19 pandemic using the hashtags #coronavirus and #COVID-19. They applied NLP and Sentiment Analysis. Julie A. Mahoney et al. [18] applied NLP using the NETbase tool to assess sentiment in social media discussions about agricultural events. Burzyńska J, Bartosiewicz A, and Rękas M. [19] used the SentiOne tool for NLP to analyze social media data on COVID-19 in Poland. Their analysis revealed increased public discussions and information sharing as the pandemic progressed . For analysis techniques utilizing Latent Dirichlet Allocation (LDA) and Topic Modeling, Shoults C. C. et al. [20] employed LDA and t-SNE to analyze social media discussions about telemedicine on Reddit and Twitter. Golos AM, Guntuku SC, and Buttenheim AM [21] used LDA to examine parental concerns and misinformation about COVID-19 vaccines from data on regulations.gov. They identified three key themes: personal beliefs against vaccination, distrust in pharmaceutical companies and the government, and concerns about vaccine ingredients and risks. Furthermore, sentiment analysis and clustering techniques were used to analyze social media data; Sanders C. A. et al. [22] conducted sentiment analysis and clustering of Twitter data to study public opinions on face masks during COVID-19. Additionally, quantitative and qualitative analysis, along with bot detection, were employed for social media data analysis, Spitale G., Andorno B. N., and Germani F. [23] analyzed Telegram conversations about the Green Pass in Italy using both quantitative and qualitative methods. Scannell D. et al. [24] used Talkwalker for data collection and a Botometer for bot detection to study opinions on COVID-19 vaccines. Their analysis aimed to understand sentiment and persuasion strategies among different opinion groups

From reviewed literature, Most research on social (media) listening primarily used ready-made platforms like Talkwalker and SocialStudio for data collection and analysis. Few studies applied NLP techniques, such as tokenization with spaCy and nltk.tokenize, stopword removal with NLTK.corpus, and sentiment analysis with packages like feel-it and VADER. Some also employed topic modeling methods, including K-means and Latent Dirichlet Allocation (LDA).

In contrast, our newly developed NCDs Listener tool employs NLP and keyword-matching principles to extract key insights. Techniques used include tokenization, stopword removal, lemmatization, and normalization. For topic modeling, K-means clustering is applied. Extracted data is presented through descriptive statistics, summarized with large language models (LLMs), and visualized in a dashboard format.

3. Research Methodology

Our NCDs Listener system was developed to collect and analyze data related to non-communicable diseases (NCDs) from public posts. Users can input the URL of the post they wish to examine into the NCDs Listener tool. The system then extracts the data and performs preliminary processing, consolidating redundant comments and filtering excessively brief texts. The data is subsequently analyzed using NLP techniques to prepare and extract relevant information. Users can specify particular diseases of interest via our adjusted data function, and the NCDs Listener tool will present the analysis results through an accessible dashboard, which includes detailed summaries generated by the advanced Generative AI model. Extracted data can be leveraged to use in other statistical and data science tools. This process is illustrated in Figure 3.

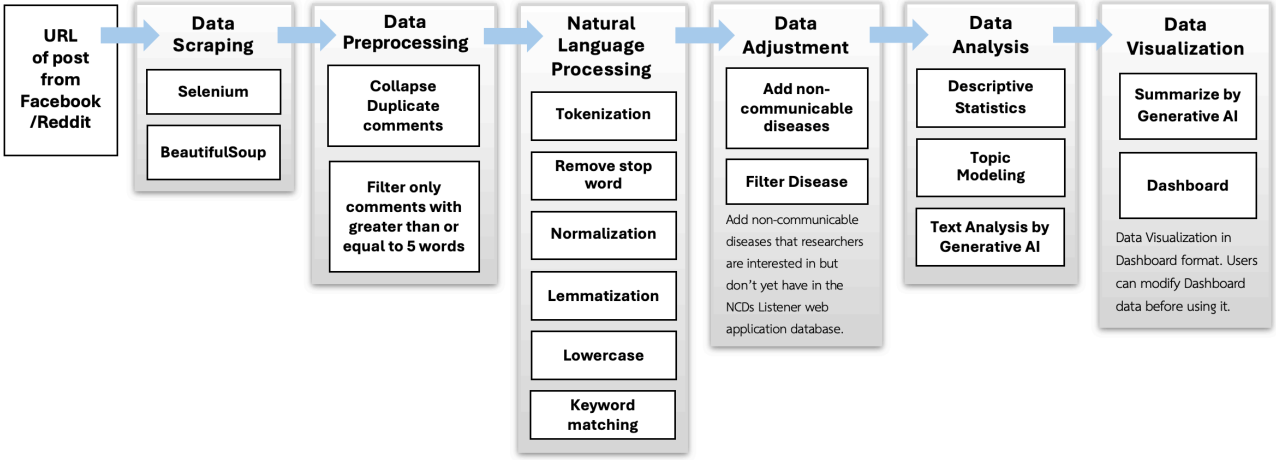

URL of post from Facebook /Reddit
Data Scraping
Selenium
BeautifulSoup
Data Preprocessing
Collapse Duplicate comments
Filter only comments with greater than or equal to 5 words
Natural Language Processing
Tokenization
Remove stop word
Normalization
Lemmatization
Lowercase
Keyword matching
Data Adjustment
Add non-communicable diseases
Filter Disease
Add non-communicable diseases that researchers are interested in but don't yet have in the NCDs Listener web application database.
Data Analysis
Descriptive Statistics
Topic Modeling
Text Analysis by Generative AI
Data Visualization
Summarize by Generative AI
Dashboard
Data Visualization in Dashboard format. Users can modify Dashboard data before using it.

Figure 1. Flow diagram process of the NCDs Listener tool.

3.1 Data Scraping

We collected social media comments on public posts provided by users related to NCDs and that feature interaction or shared experiences. In our case study, data was collected from Facebook. We employed web scraping techniques using Python Selenium and BeautifulSoup to extract variables such as the commenter's name, comments, likes, and replies, understanding that a single commenter may leave multiple comments. After extraction, each comment was saved as a text file and converted into a comma-separated values (CSV) format. Our tool can analyze textual data in both Thai and English.

3.2 Data Preprocessing

We used Python 3.11.9 to access, clean, and analyze the data. To streamline the dataset, we consolidated comments from the same commenter into a single entry and filtered out comments with fewer than five words, considering which were considered. We converted the number of likes and replies from their alphabetical representations (e.g., “1.1k”) to numeric values (e.g., “1,100”).

3.3 Natural language processing

3.3.1 text processing

For data cleaning, we applied different Natural Language Processing (NLP) techniques based on languages. For English, we utilized the NLTK tool [25]. We performed word tokenization, converted text to lowercase, removed stop words, eliminated spaces and special characters from tokens, and performed lemmatization.

For Thai, we used PyThaiNLP [26] for data cleaning. We performed tokenization, removed stop words, cleaned up spaces and special characters, and addressed misaligned or misspelled text with normalization techniques.

3.3.2 Keyword Matching

We identify various data characteristics to extract knowledge or insights from the comments. The variables to be extracted are experience, gender, disease, symptoms, cell type, and treatment methods. The keywords are extracted by matching the token with the NCDs Listener's disease database. After extracting the data, we use descriptive statistics to analyze the data.

The variable "Experience" is data used to classify whether the patient or the caregiver commented. It will be used to show the proportion of disease experience-sharing charts, which have three analysis sequences as follows:

1) Identification of Third-Person Mentions:
- For English: Terms like "son", "grandfather."
- For Thai: Terms like "พ่อ" (father) and "แม่" (mother). If such terms are present, the comment is classified as describing the "experience of others."

2) Identification of First-Person Mentions:
- For English: Terms like "my", "myself."
- For Thai: Terms like "ฉัน" (I) and "ผม" (I - male). If these terms are found, the comment is categorized as describing "one's own experience."

3) Indeterminate Cases: If neither condition is met, the comment is classified as "unable to identify."

The variable "*Gender*" is the data used to separate the gender of cancer patients, which will be divided into two genders: male and female. This variable set will be used to display a chart of the number of comments and the gender of the patients, which have three analysis sequences as follows:

1) Third-Person Gender Identification:
- Gender is inferred from terms such as "mother" (female), "son" (male), etc.

2) Self-Referencing Gender Identification:
- For English: Gender cannot be specified if the comment refers to oneself.
- For Thai: Terms like "ค่ะ" (female) and "ครับ" (male) are used to infer gender.

3) Indeterminate Cases: If the gender cannot be specified, the comment is categorized as "Cannot be specified."

The variable "*Disease*" is the data used to separate the diseases mentioned in the comments since this case study uses cancer.

1) Disease Identification for Thai:
- Words like "มะเร็ง" (cancer) are captured and combined with adjacent words to identify specific types, e.g., "มะเร็งปอด" (lung cancer).

2) Disease Identification for English:
- Similar to Thai, but with terms like "my cancer" and "lung cancer." The comment is labeled accordingly if no match is found or if multiple diseases are mentioned.

The variable "*Symptoms*" is the data used to identify the mentioned symptoms. Both Thai and English tokens will capture words that mention symptoms such as "ไข้" (fever) and "chronic fever", etc. When these words are found, they will be checked against the NCDs Listener disease database to find out what symptoms these words are and then identify them according to the symptoms. For example, if the word "fever" is found, it will be specified as chronic fever. This data set will be used to display a chart showing the number of comments and symptoms that are mentioned.

The variable "*Treatment Method*" is the data that specifies the treatment that is mentioned. Both Thai and English tokens will capture words that mention treatment, such as "ผ่าตัด", "Surgery", "operation", "incision", "radiation" "ฉายแสง", "คีโม", and "chemotherapy". When these words are captured, they will be checked against the NCDs Listener disease database to determine what symptoms these words represent and then be specified according to them. For example, if the word "Surgery" is found, it will be specified as "Surgery". This data set will be used to display a chart of the number of comments and treatments that are mentioned.

The variable "*Cell type*" The grouping of cancers in this study followed the Cancer Research UK guidelines. The groups include carcinoma, sarcoma, leukemia, lymphoma, myeloma, and brain and spinal cord cancers. Comments mentioning specific cancers are grouped into these categories; for example, cervical and lung cancers are categorized as carcinoma [27].

3.4 Data Adjustment

This component is utilized to add or filter diseases after the execution of keyword matching, enabling the identification of diseases that align with the researcher's specific requirements. The implementation of this step relies on the pandas package, a data manipulation and analysis tool for Python.

3.5 Data Analysis

3.5.1 Descriptive statistics

We used the Python package pandas to perform descriptive data analysis in the form of frequencies and percentages and simultaneously display the results to examine the characteristics of the data in each variable.

3.5.2 Topic Modeling

To perform topic modeling, we used word embeddings to represent the comments. The Python packages sentence_transformers and the model 'paraphrase-multilingual-mpnet-base-v2' were employed for this purpose. Subsequently, we applied the K-means clustering algorithm from the sklearn.cluster module. For this research, we set the number of clusters to three (K=3) to simplify the topic assignment.

Each cluster was assigned a topic based on the frequency of the top five tokens identified within the cluster. This process ensured that the most representative topics were associated with each group of comments. Token frequency analysis was conducted using nltk.probability, facilitating the efficient identification of key terms within each cluster.

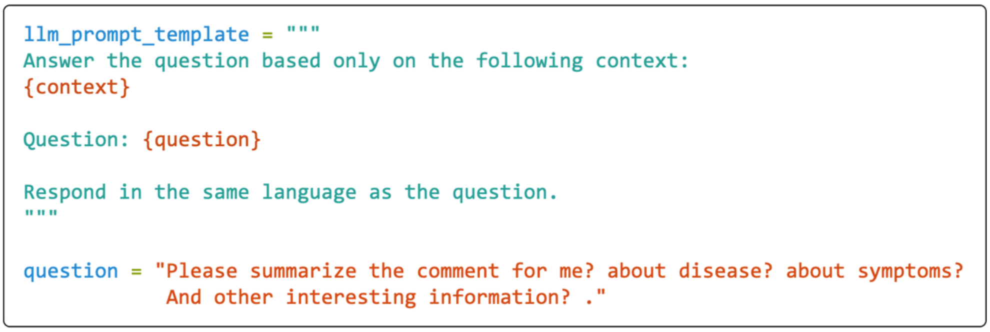
3.5.3 Text Analysis by Generative AI

We utilize the Google Gemini 1.5 pro through Langchain [28] to incorporate a Generative AI model. This model employs Retrieval-Augmented Generation (RAG) to define the knowledge scope based on the collected comments. NCDs Listener uses a set of pre-defined questions to assist Gemini in summarizing disease and symptom information derived from the analyzed comments, and for analysis, number each comment sequentially. For instance, comment 1, which states "Was diagnosed at 22 with Hodgkin Lymphoma and went through 16 rounds of chemo…", is formatted as "1. Was diagnosed at 22 with Hodgkin Lymphoma and went through 16 rounds of chemo…".

3.6 Data Visualization

The NCDs Listener presents the analyzed data through a visually interactive dashboard that includes the following types of graphs:

1) Pie graph: This graph displays variables with no more than three possible answers, including the number of comments, the patient's gender, and the proportion of people sharing their disease experiences.
2) Bar graph: This graph displays variables with more than three possible answers, such as the number of comments on the disease mentioned, the number of comments on the symptom mentioned, and the number of comments on the treatment mentioned.
3) Basic statistics table: This graph displays basic statistical data on comments, including the maximum, minimum, and average values of likes, replies, and words, as well as the number of commenters compared to the total number of comments, to analyze the frequency of repeated comments.
4) Prompt template and predefined questions: The prompt template serves as a structured framework to guide the responses of generative artificial intelligence. In this context, the generative AI is expected to address inquiries based on the aggregated comments and provide responses in English. For pre-defined inquiries, the generative AI is instructed to furnish answers in Thai and English. The answer will address aspects of the disease in question, its associated symptoms, potential treatment modalities, and any noteworthy or intriguing information pertinent to the subject matter. Shown in Figure 2.



```
llm_prompt_template = """
Answer the question based only on the following context:
{context}

Question: {question}

Respond in the same language as the question.
"""

question = "Please summarize the comment for me? about disease? about symptoms?
            And other interesting information? ."
```

Figure 2. Prompt template on Facebook for summarizing data about disease and symptoms of cancer.

3.7 Development of the NCDs Listener Service

The NCDs Listener Service is designed to collect, process, and analyze social media comments related to non-communicable diseases (NCDs). The service extracts relevant information, applies natural language processing (NLP) techniques, and generates insights using a Generative AI model. Our NCDs Listener tool can only extract knowledge from Facebook and Reddit platforms.

3.7.1 Overview of system implementation

Our NCDs Listener is a social (media) listening tool developed using HTML in the development structure, decorated with CSS for web pages and Python 3.11.9 using Flask API to create a backend system and PolyDash to create a dashboard on the website. NCDs Listener has a database of chronic diseases and symptoms (as a reference database and specify the characteristics of opinion data) collected from Google Gemini AI generator, collected opinion and book. The researcher checked the data again before saving it to the database. NCDs Listener uses Google Gemini 1.5 Pro via Langchain using the Retrieval-Augmented Generation (RAG) principle to define the scope of knowledge for Generative AI to answer questions that match the opinion collected in a Text file.

3.7.2 System functionality

The NCDs Listener performs the following functions:

3.7.2.1 Social Media Data Extraction Function

Users can extract the URL of social media posts. Currently, only Facebook and Reddit can be extracted with posts that have Q&A in the comments. Moreover, there must be posts that are mainly about NCDs. Two tools are used to extract this function: Selenium and BeautifulSoup. The researcher used Selenium to open the website via URL and open the hidden comment data such as "more", "additional comments", etc. When all the hidden data are opened, BeautifulSoup will be used to extract comment data such as "name", "comment", "likes", and "number of replies".

3.7.2.2 Disease and Symptom Database Update Function

Users can reduce the number of diseases and symptoms they do not want by having NCDs Listener filter the data only to include those that mention the disease and symptoms they want. Users can also add diseases and symptoms that the original database could not capture. Once added, the NCDs Listener will detect the new diseases and symptoms that are identified and then filter the data again to include only those that mention the disease and symptoms they want.

3.7.2.3 Generative AI Data Summarization Function

Users will get all the desired feedback summaries from the Generative AI model Google Gemini 1.5 pro latest, with the summary focusing mainly on diseases and symptoms.

3.7.2.4 Knowledge Extraction Result Presentation Function

Users will be presented with a summary of the knowledge extraction results in the form of a dashboard. The dashboard will include:

- Customization Section: Users can select the data displayed using checkboxes and tag name boxes. For instance, by filtering out data where gender or disease is unspecified, the dashboard will only display comments where gender and disease are specified.
- Display Section with Charts: The section will present data in various formats.
- Summary Section: The final section of the Dashboard is a concise summary of the chart results so that users can understand the meaning of the data correctly and efficiently.

3.7.2.5 Data export function

Users can export feedback data in two formats:

1) A CSV file that shows data in a table format, both before extraction and after extraction, for further use.
2) A report will be presented in PDF format after the data is presented in Dashboard format.

3.7.3 Data sources

The NCDs Listener tool collects data from Facebook and Reddit, including commenter names, comment content, likes, and replies. It retrieves all available comments from the posts but cannot handle overlapping comments across both platforms. The time required for data retrieval varies based on the platform and comment volume, with more comments or overlaps resulting in longer retrieval times. For instance, fetching data from a Facebook post with 1,600 comments may take up to 32 minutes, whereas a post with 60 comments may only take 1-2 minutes.

4. Results of the case study

The performance of the NCDs Listener tool was evaluated based on its ability to collect, process, and analyze data from Facebook posts related to non-communicable diseases (NCDs). The following summarizes the key findings and effectiveness of the tool. For this study, we present the results of knowledge extraction on general cancer and cancer-specific to the Facebook platform.

4.1 Dataset: Facebook posts on cancer

In this comprehensive case study, data was meticulously collected from three Facebook posts [29-31], resulting in a substantial total of 2,238 comments. After a rigorous filtering process, 1,402 comments were selected for further analysis and visualization, ensuring the highest quality of data for our study.

4.2 The Analysis

The data analysis and the demographic characteristics of the commenters on all three Facebook posts. It was found that 564 comments, or 40.2 percent of the total comments (n=1,402), could identify the narrators of cancer stories shown in Figure 3A. Most of the comments were about the experiences of others (52 percent of the comments were with an identified narrator). Regarding gender, shown in Figure 3B, 917 comments, or 65.4 percent of the total comments, could be identified. The majority were female (78% of the gender-specific comments).

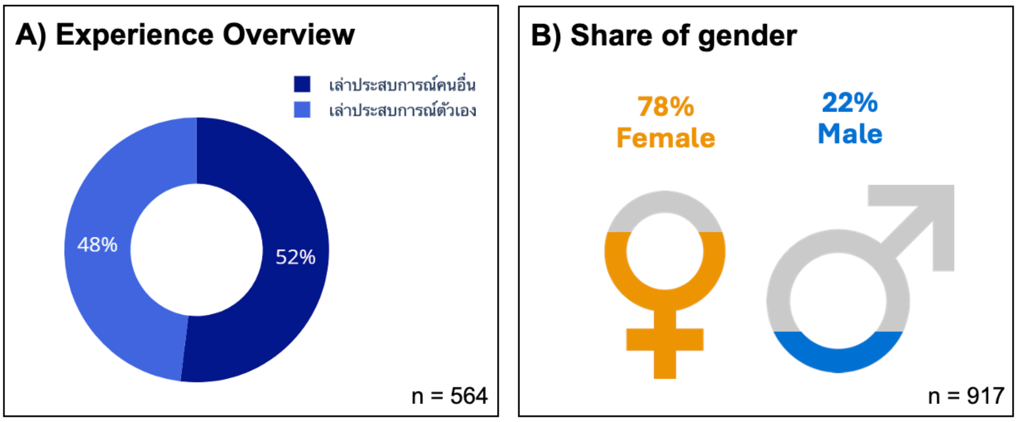


Figure 3. Data visualization showing cancer-relevant charts on Facebook posts; (A) types of experiences shared by commenters; (B) gender proportion of commenters.

In-depth analysis using the NCDs listener revealed that 29 diseases were mentioned, with breast cancer being the most mentioned (80 comments), followed by lung cancer (67 comments), lymphoma (60 comments), liver cancer (53 comments), and colon cancer (43 comments), respectively. This is shown in Figure 4.

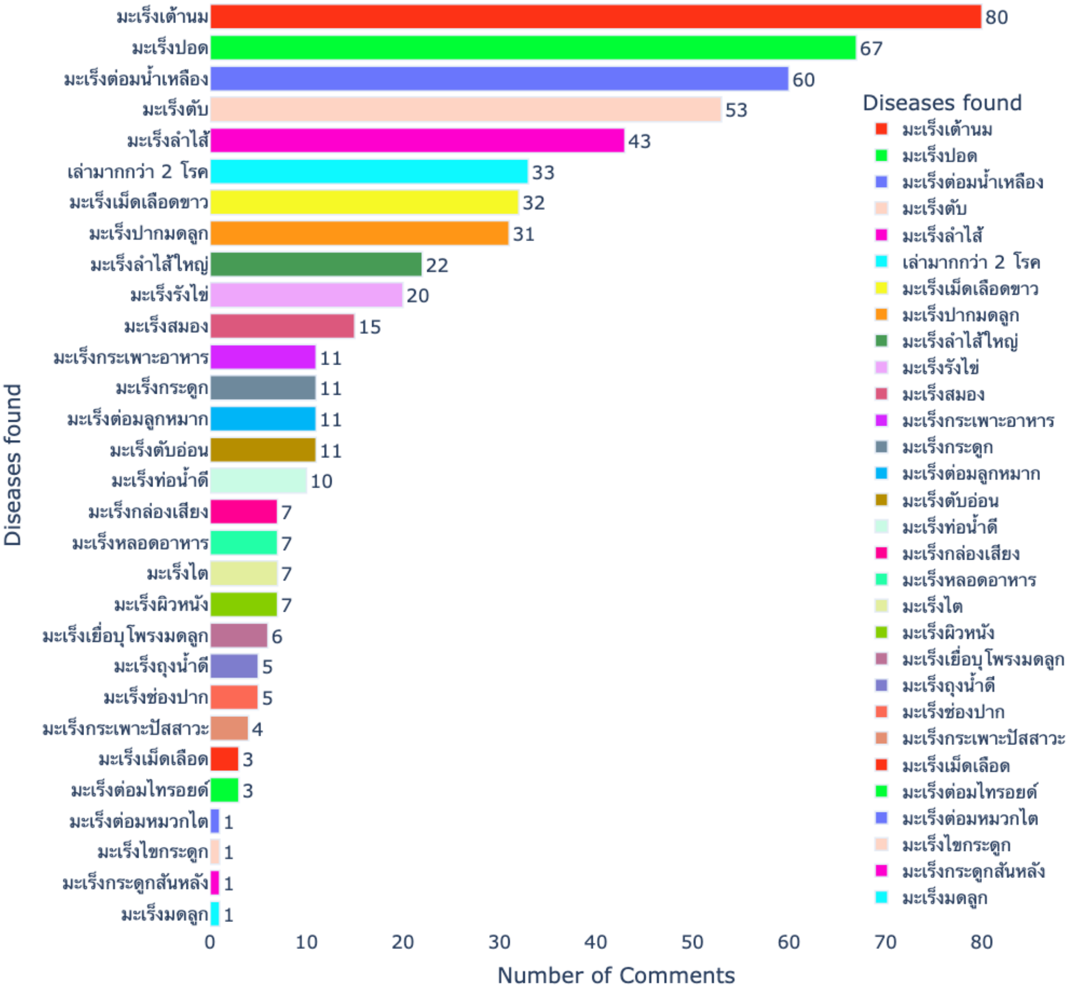


Figure 4. Name of cancers mentioned by commenters.

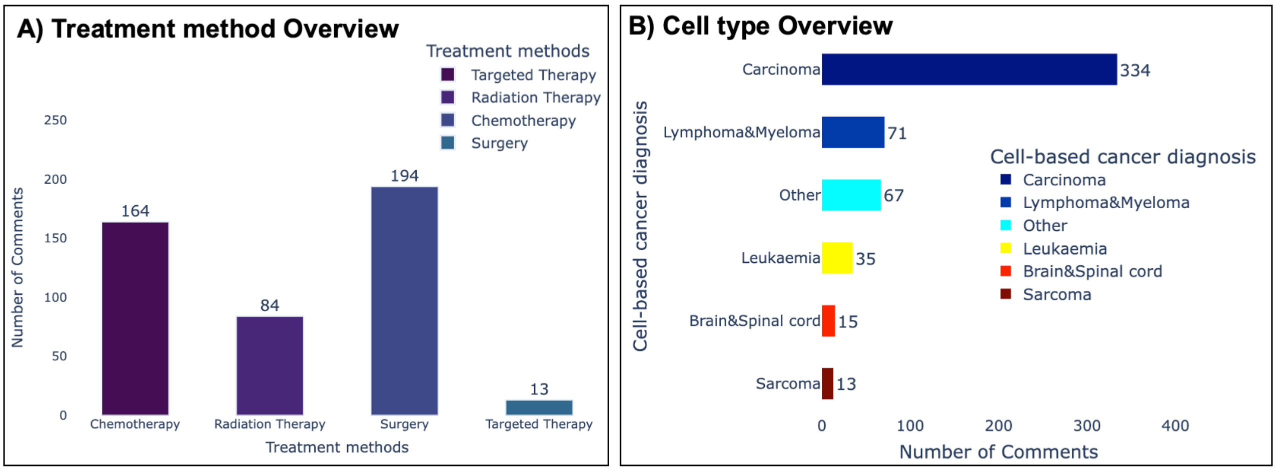


Figure 5. Data visualization showing cancer relevant chart on Facebook posts; (A) treatment methods of cancer mentioned by commenters; (B) name of cell type (group cancers by cell type) mentioned by commenters.

Treatment methods were also mentioned, as seen in Figure 5A, with surgery being the most mentioned (194 comments), followed by chemotherapy (164 comments), radiation therapy (84 comments), and targeted therapy (13 comments) respectively. When classifying cancer by cell type of origin, Carcinoma was the most mentioned (334 comments), while Sarcoma was the least mentioned (13 comments), as shown in Figure 5B.

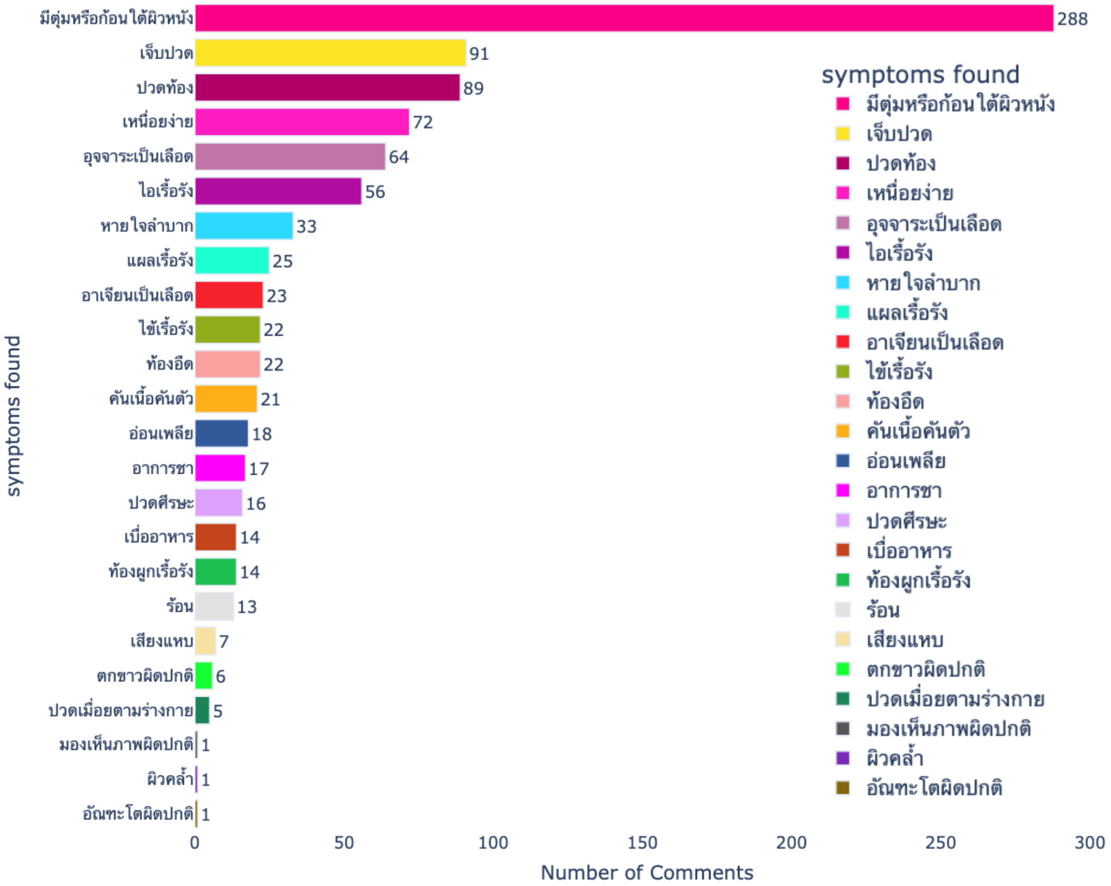


Figure 6. Symptoms of cancer mentioned by commenters.

Figure 6. shows that 24 symptoms were mentioned, with the most mentioned being malignant tumor or the presence of a lump under the skin (288 comments), followed by pain (91 comments), abdominal pain (89 comments), fatigue (72 comments), and bloody stool (64 comments).

When we fed the data to the Google Gemini Generative AI model to summarize all the comments, we got the following conclusions: *"The comments describe various cancer experiences, including breast, liver, lung, ovarian, and colorectal cancers. Symptoms vary depending on the type and stage but include lumps, pain, fatigue, weight loss, bowel changes, and unexpected bleeding. Treatments include surgery, chemotherapy, radiation, and targeted therapies. Many emphasize the importance of early detection, regular checkups, and healthy lifestyle choices. Some share emotional challenges and coping mechanisms, highlighting the impact of cancer on mental well-being. The comments collectively provide a glimpse into the diverse realities of cancer patients."*

This shows that data set describes different types of cancer, along with descriptions of symptoms and treatment methods. Interestingly, some commenters who had cancer reported that they did not experience any symptoms before being diagnosed. In addition, many commenters emphasized the importance of taking care of their health holistically, both physically and mentally, by recommending exercise, getting enough rest, stress management, and encouraging regular health check-ups.

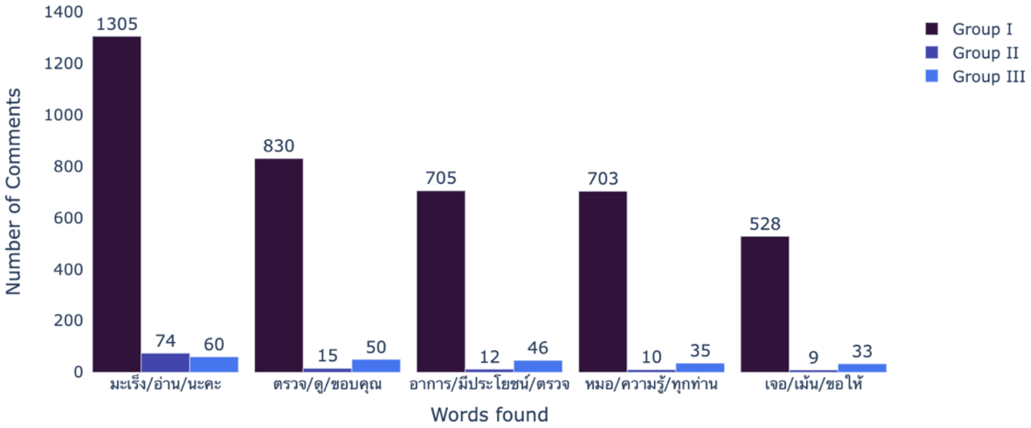


Figure 7. Frequency of Top Keywords by Group and Rank chart on Facebook posts.

Figure 7. Show Facebook data with 1,402 comments was divided into three groups using K-means. The Feature Vector is derived by converting Text Embeddings from transforming Comment Texts into Feature Vectors. This extraction process is accomplished using the 'paraphrase-multilingual-mpnet-base-v2' Sentence Transformers model [32]. Subsequently, the resultant Text Embeddings undergo a clustering procedure. Then, the topics were determined by the number of the five most common words and summarized using Google Gemini.

1. Group I. The words found are: มะเร็ง / ตรวจ / อาการ / หมอ / เจอ

Summarized by Google Gemini: *"These comments share personal experiences with cancer, detailing diagnoses, symptoms, treatments, and outcomes for individuals or their loved ones. They highlight diverse cancer types, stages, and treatments, often emphasizing the importance of early detection, regular checkups, and healthy lifestyle choices."*

This cluster can interpret the words found as a topic as "*Cancer Topics.*"

2. Group II. The words found are: อ่าน / ดู / มีประโยชน์ / ความรู้ / เม้น

Summarized by Google Gemini: *"He comments are generally short and express a desire to read the post later, share it with others, or tag friends and family to do the same. Many simply state their intention to read, while others use phrases like "good information," "useful," or "let's read." The comments highlight a focus on personal health and well-being, with some urging others to get checkups and prioritize their health."*

This cluster can interpret the found words as a topic as "*Interest in posts.*"

3. Group III. The words found are: นะคะ / ขอบคุณ / ตรวจ / ทุกท่าน / ขอให้

Summarized by Google Gemini: *"The comments are a collection of personal anecdotes about cancer, offering advice, support, and sharing experiences. Many describe specific symptoms experienced by themselves or loved ones, while others express gratitude for the information shared and offer well wishes. Some comments highlight the importance of regular check-ups, early detection, and lifestyle choices. Overall, the comments create a sense of community and shared concern about cancer, emphasizing the importance of awareness and support."*

This cluster can interpret the found words as a topic as "*Cancer Concerns.*"

We can explain that all clusters discuss cancer-related content, provide knowledge, and exchange experiences. This includes identifying types of cancer, symptoms, and lifestyle recommendations. Additionally, the cluster highlights short comments expressing a desire to read posts later, share them with others, or tag friends and family, demonstrating the widespread dissemination of information and increased awareness. Furthermore, the cluster emphasizes comments reflecting personal experiences with cancer, offering advice and emotional support. Gratitude is expressed for the shared information, demonstrating community and solidarity among cancer patients. This highlights the importance of awareness and mutual support within the online community.

5. Conclusions

Social media sources are extremely important and are constantly changing based on user needs. The idea that a large group of people is smarter than individual experts reflects the potential of social media to be large and rich in valuable information. The nature of the information obtained from this idea includes both primary and in-depth data, which can be used to create knowledge for the public and are essential for developing effective approaches or tools that meet needs. However, collecting and analyzing useful opinions from a large group of people is very resource-intensive. The tools have been developed to help collect, analyze, and summarize the wisdom of the crowd, called the social (media) listening tool.

In this study, the NCDs Listener tool was developed to crowd social media websites and summarize opinions from the website to extract knowledge and share experiences related to NCDs. The results of using this tool found that the majority of Facebook users were women and shared their experiences of breast cancer. They tend to present stories that often emphasize sharing pre- and post-cancer symptoms, inviting others to learn about the experiences and concerns of cancer patients, and providing support and advice. Interestingly, many emphasize early cancer screening and quality of life in terms of physical and emotional health. Therefore, we can use the knowledge we have gained to develop and expand in the future. This tool can potentially support research and innovation in public health and healthcare.

The NCDs Listener tool is designed with the user in mind. It can extract data from social media using only the URL of public posts related to NCDs. Users can easily visualize the information and store the extracted data for later use. They can also filter or add information about diseases and symptoms as needed. Once the desired data is obtained, a summary presentation of the knowledge extraction results can be created in the form of a dashboard consisting of intuitive charts with captions. The data can also be exported as a CSV file or PDF report, making it a user-friendly tool for all.

The NCDs Listener tool is not just a tool, it's a comprehensive solution for NCD research. It includes the creation of crowd wisdom from comments on social media using the generative AI model Google Gemini to summarize comments related to the diseases and symptoms mentioned. It also presents interesting information in an easy-to-understand format. In summary, NCDs Listener is a tool that provides both basic and in-depth knowledge about NCDs, as well as creating cloud wisdom with generative AI to provide users with maximum benefit.

6. References

[1] World Health Organization. Noncommunicable diseases. [document on the Internet]. World Health Organization; 2023 [cited 2024 Aug 29]. Available from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases..

[2] Kapoor KK, Tamilmani K, Rana NP, Patil P, Dwivedi YK, Nerur S. Advances in social media research: Past, present and future. Information Systems Frontiers. 2018;20:531-58

[3] Chaffey D. Global social media statistics research summary 2024. [document on the Internet]. Smart Insights; 2024 [cited 2024 Aug 29]. Available from:

https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

[4] Bhatt S, Minnery B, Nadella S, Bullemer B, Shalin V, Sheth A. Enhancing crowd wisdom using measures of diversity computed from social media data. In: Proceedings of the International Conference on Web Intelligence (WI '17); 2017; New York, NY, USA. p. 907–913. Association for Computing Machinery.

[5] Hong H, Du Q, Wang G, Fan W, Xu D. Crowd wisdom: The impact of opinion diversity and participant independence on crowd performance. In: Proceedings of the Twenty-second Americas Conference on Information Systems (AMCIS 2016); 2016; San Diego.

[6] Berger J, Packard G, Boghrati R, Hsu M, Humphreys A, Luangrath A, Moore S, Nave G, Olivola C, Rocklage M. Marketing insights from text analysis. Marketing Letters . 2022;33:365-377.

[7] Keller MS, Park HJ, Cunningham ME, Fouladian JE, Chen M, Spiegel BMR. Public perceptions regarding use of virtual reality in health care: a social media content analysis using Facebook. Journal of Medical Internet Research. 2017;19(12):e419

[8] Sakamoto D, Matsushita N, Noda M, Tsuda K. Social listening system using sentiment classification for discovery support of hot topics. Procedia Computer Science. 2018;126:1526-1533.

[9] Geyser W. Top 25 social media listening tools for 2024 [document on the Internet]. Influencer Marketing Hub; 2024 [cited 2024 Sep 3]. Available from: https://influencermarketinghub.com/social-media-listening-tools/.

[10] Cook N, Mullins A, Gautam R, Medi S, Prince C, Tyagi N, Kommineni J. Evaluating Patient Experiences in Dry Eye Disease Through Social Media Listening Research. Ophthalmology and Therapy. 2019;8(4):407-420.

[11] Rodrigues A, Chauhan J, Sagkriotis A, Hughes R, Kenny T. Understanding the lived experience of lung cancer: a European social media listening study. BMC Cancer. 2022;22:(475).

[12] Chauhan J, Aasaithambi S, Márquez-Rodas I, Formisano L, Papa S, Meyer N, Forschner A, Faust G, Lau M, Sagkriotis A. Understanding the lived experiences of patients with melanoma: real-world evidence generated through a European social media listening analysis. JMIR Cancer. 2022;8(2).

[13] Mazza M, Piperis M, Aasaithambi S, Chauhan J, Sagkriotis A, Vieira C. Social media listening to understand the lived experience of individuals in Europe with metastatic breast cancer: a systematic search and content analysis study. Frontiers in Oncology. 2022;12:863641.

[14] Perić Z, Basak G, Koenecke C, Moiseev I, Chauhan J, Asaithambi S, Sagkriotis A, Gunes S, Penack O. Understanding the needs and lived experiences of patients with graft-versus-host disease: real-world European public social media listening study. JMIR Cancer. 2023;9.

[15] Wolffsohn JS, Leteneux-Pantais C, Chiva-Razavi S, Bentley S, Johnson C, Findley A, Tolley C, Arbuckle R, Kommineni J, Tyagi N. Social media listening to understand the lived experience of presbyopia: systematic search and content analysis study. Journal of Medical Internet Research. 2020;22(9).

[16] Spies E, Andreu T, Hartung M, Park J, Kamudoni P. Exploring the perspectives of patients living with lupus: retrospective social listening study. JMIR Formative Research. 2024;8.

[17] Manguri KH, Ramadhan RN, Amin PR. Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research. 2020;5:54–65.

[18] Mahoney JA, Widmar NJO, Bir CL. #GoingtotheFair: a social media listening analysis of agricultural fairs. Translational Animal Science. 2020;4(3):1-13.

[19] Burzyńska J, Bartosiewicz A, Rękas M. The social life of COVID-19: early insights from social media monitoring data collected in Poland. Health Informatics Journal. 2020;26(4):3056-3065.

[20] Shoults CC, Dawson L, Hayes C, Eswaran H. Comparing the discussion of telehealth in two social media platforms: social listening analysis. Telemedicine Reports. 2023;4(1):236–248.

[21] Golos AM, Guntuku SC, Buttenheim AM. “Do not inject our babies”: a social listening analysis of public opinion about authorizing pediatric COVID-19 vaccines. Health Affairs Scholar. 2024;2(7)

[22] Sanders AC, White RC, Severson LS, Ma R, McQueen R, Alcântara Paulo HC, Zhang Y, Erickson JS, Bennett KP. Unmasking the conversation on masks: natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. AMIA Summits on Translational Science Proceedings. 2021;2021:555-564.

[23] Spitale G, Biller-Andorno N, Germani F. Concerns around opposition to the Green Pass in Italy: social listening analysis by using a mixed methods approach. Journal of Medical Internet Research. 2022;24(2).

[24] Scannell D, Desens L, Guadagno M, Tra Y, Acker E, Sheridan K, Rosner M, Mathieu J, Fulk M. COVID-19 vaccine discourse on Twitter: a content analysis of persuasion techniques, sentiment and mis/disinformation. Journal of Health Communication. 2021;26(7):443–459.

[25] Bird S, Loper E. NLTK: The Natural Language Toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions; 2004; Barcelona, Spain. p. 214–217. Association for Computational Linguistics.

[26] Phatthiyaphaibun W, Chaovavanich K, Polpanumas C, Suriyawongkul A, Lowphansirikul L, Chormai P, Limkonchotiwat P, Suntorntip T, Udomcharoenchaikit C. PyThaiNLP: Thai Natural Language Processing in Python. In: Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023); 2023; Singapore. p. 25–36. Association for Computational Linguistics.

[27] Cancer Research UK. Types of cancer [document on the Internet]. London: Cancer Research UK; 2023 [cited 2024 Sep 5]. Available from: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts/types-of-cancer.

[28] Chase H. LangChain: Build context-aware reasoning applications [document on the Internet]. GitHub; 2024 [updated 2024 Sep 14; cited 2024 Sep 14]. Available from: https://github.com/langchain-ai/langchain

[29] Chawalit A. โพสต์นี้ดีมาก มีประโยชน์ [document on the Internet]. Facebook; 2023 December 6 [cited 2024 Jul 19]. Available from: https://www.facebook.com/chawalit.atchulacancer/posts/pfbid023gqcSi5bD9soLZb3vNvGwqMAzvWi8w1uHpVnYFtPVi8bVVpzVGHbvt3tuo5yvgDFl?_rdc=2&_rdr

[30] Chuenchom W. จะดีมากๆเลยค่ะ ถ้าผู้ป่วยมะเร็งมาแชร์อาการก่อนตรวจพบว่าเป็นมะเร็ง [document on the Internet]. Facebook; 2022 October 20 [cited 2024 Jul 19]. Available from: https://www.facebook.com/chawalit.atchulacancer/posts/pfbid023gqcSi5bD9soLZb3vNvGwqMAzvWi8w1uHpVnYFtPVi8bVVpzVGHbvt3tuo5yvgDFl?_rdc=2&_rdr

[31] โรคร้ายผ่านได้ถ้าไม่ยอมแพ้. อาการเริ่มต้นแบบไหนถึงไปตรวจมะเร็ง มะเร็ง [document on the Internet]. Facebook; 2024 April 5 [cited 2024 Jul 19]. Available from: https://www.facebook.com/story.php?story_fbid=738492865098283&id=100068127288896&rdid=KSfcPd478i4jBurb

[32] Hugging Face. Sentence Transformers: paraphrase-multilingual-mpnet-base-v2 [Internet]. Hugging Face; 2024 March 2024 [cited 2024 Sep 9]. Available from: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2