NCDs Listener: A Social Listening Tool for Non-communicable Diseases

Ratchanont Thippimanporn, Wuttichai Khamna, Kannika Wiratchawa, and Thanapong Intharah*
Visual Intelligence Laboratory, Department of Statistics, Faculty of Science,
Khon Kaen University, Khon Kaen, 40002, THAILAND
*Corresponding author's email: thanin@kku.ac.th

Abstract:

Over 41 million people worldwide die from NCDs (non-communicable diseases) annually, the majority of which occur 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, which is available as an open-source project. The NCDs Listener to collect, analyze, summarize, and visualize data. Comments in Thai or English about NCDs can be collected from public posts. This time, we studied the characteristics of comments from Facebook and Reddit posts that mentioned NCDs to demonstrate the NCDs Listener tool. Use keyword matching and the BERT Model to extract knowledge from comments. The preliminary data was analyzed using descriptive statistics. Additionally, a Generative AI model summarizes the extracted knowledge in human-readable sentences. Our NCDs Listener tool is open-source and can extract knowledge related to NCDs. This knowledge can be used as a guideline for treatment or the development of effective care to meet patients' needs. Our findings demonstrate that aggregated social media data not only provides immediate insights but also serves as a springboard for advanced statistical analyses and cutting-edge data science approaches, opening new avenues for understanding complex social phenomena and predicting emerging trends.

Keywords: Social listening; Non-communicable diseases; Text classification; Data visualization; Generative AI

Introduction

Non-communicable diseases (NCDs) are chronic conditions caused by unhealthy lifestyles, with cardiovascular diseases being the highest, followed by cancers, chronic respiratory diseases, and diabetes (1). Treatment decisions for NCDs often involve a variety of concerns, such as understanding contributing behaviors, recognizing symptoms, and determining effective interventions. While information on NCDs is abundant, accessible and layman-friendly platforms remain limited. Social media has recently played a key role in educating patients, building peer support networks, and sharing experiences (2).

Social (media) listening studies have primarily focused on developing systems that analyze sentiment and track trending topics (3). Plenty of research has been conducted through social listening tools. However, the tools, such as Brand24 and YouScan (4), require expensive monthly subscriptions, which makes them inaccessible to the general public or independent researchers. Furthermore, these tools are often designed to cater to large organizations, limiting their applicability to individuals who wish to analyze social media data on a personal level (4). We developed an open-source tool to analyze public social media posts about NCDs and present the extracted insight as a dashboard. This helps researchers study how the NCD community shares their real-life experiences while reducing the time needed to find relevant information and improving public access to understanding these conditions. The primary distinction between the NCDs Listener and other

existing social media listening tools are as follows; NCDs Listener supports Thai and English languages and is designed especially for extracting knowledge about non-communicable diseases.

This work highlights:

- Developing a Web application, NCDs Listener, a social (media) listening tool, to extract and analyze public sentiment from NCD-related posts, displaying the data in a customizable dashboard.
- Utilizing Natural Language Processing (NLP) techniques and Bidirectional Encoder Representations from Transformers (BERT) to derive meaningful insights from social media data.
- Integrating large language models (LLMs) to enhance the social media listening process

Related Work

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, Rodrigues et al. (5) 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. (6) 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. (7) used social listening on Twitter, patient forums, and blogs to explore metastatic breast cancer patient experiences. Zinaida Perić et al. (8) investigated GVHD patient needs and lifestyles across Europe using Talkwalker to collect data from Twitter, Facebook, Instagram, and YouTube. The research combined quantitative and qualitative methods to analyze quality of life, treatment efficacy, and unmet needs. Our work proposed the utilization of a recent large language model technique to summarize all knowledge in a human-readable paragraph.

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. (9) analyzed Twitter data from the COVID-19 pandemic using the hashtags #coronavirus and #COVID-19. They applied NLP and Sentiment Analysis. Burzyńska J, Bartosiewicz A, and Rękas M. (10) 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 CC. et al. (11) employed LDA and t-SNE to analyze social media discussions about telemedicine on Reddit and Twitter, Sanders C. A. et al. (12) 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. (13) analyzed Telegram conversations about the Green Pass in Italy using both quantitative and qualitative methods.

Most research on social media listening relies on platforms like Talkwalker and SocialStudio, with limited use of NLP techniques such as tokenization, stop words removal, sentiment analysis, and topic modeling. In contrast, our open-source NCDs Listener tool uses advanced NLP methods, including tokenization, lemmatization, normalization, and BERT for topic classification. The tool provides insights through descriptive statistics, LLM-based summaries, and dashboard visualizations.

The NCDs Listener Tool

Systems Design

The NCDs Listener system was developed using Python 3.11.9 to collect, analyze, and summarize social media posts related to NCDs. It helps turn these posts into meaningful insights. The system implements a five-step workflow:

1. Data Scraping step: When a user enters a social media post URL into the NCDs Listener tool, the system leverages web scraping tools like Selenium and Beautiful Soup to extract relevant data from the post.

2. Data Preprocessing step: The system cleanses the collected data by removing duplicate comments and filtering out extremely short text entries. Preprocessed data can be exported in CSV format for future analysis.

3. Knowledge Extraction step: The system employs natural language processing (NLP) techniques, including tokenization, stop word removal, normalization, lemmatization, keyword matching, and application of the Bidirectional Encoder Representations from Transformers (BERT) model (14) to identify the type of comments.

4. Data Adjustment step: Users can add diseases and symptoms they wish to include or exclude those they do not want. The system then filters the comments based on the specified diseases and symptoms, tailoring the data to the user's requirements

5. Data Visualization step: The enhanced data is summarized and displayed on an interactive dashboard, featuring a comprehensive summary generated by a Generative AI model. Users can also explore the data from a summary statistics table about the post or export the results to a PDF. The workflow is illustrated in **Figure 1**.

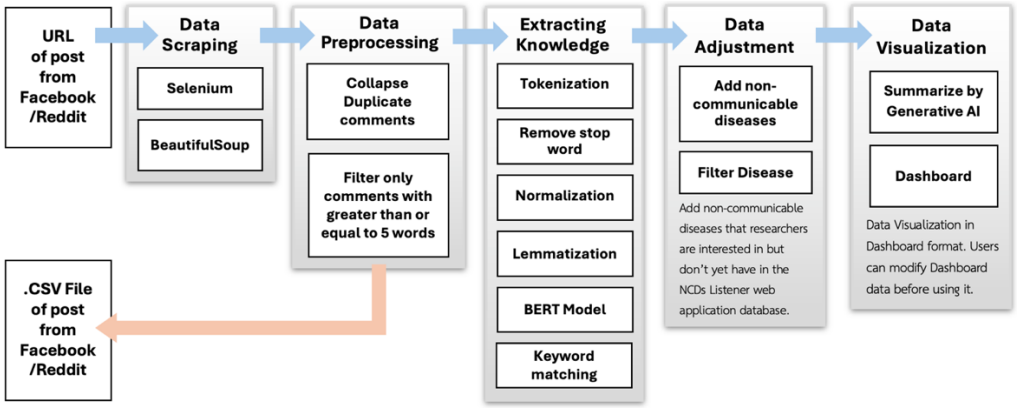

Figure 1. Flow of the NCDs Listener tool.

Artificial Intelligence Models for Insight Extraction

In this work, we employed two AI models in the insight's extraction. First, BERT was used for topic classification. Second, LLMs was used for content summarization

BERT for Topic Classification

The system employs machine learning models to classify the types of comments, enabling users to identify which comments are likely to provide useful information. The BERT model was trained on over 3,000 NCD-related comments in Thai and English. This model can categorize comments into three categories:

1. Experience Stories: Comments that share personal experiences with diseases, including symptoms (before or after diagnosis), treatments, and how the disease has affected quality of life. Such comments often cover all or at least 2 to 3 of these aspects.
2. Questions: Comments that inquire about various aspects of NCDs, such as symptoms, treatments, or disease impacts.

3. Irrelevant or Non-Useful Comments: Comments that are not related to NCDs or lack meaningful information, such as simply stating the name of a disease or unclear descriptions of symptoms.

Large Language Model (LLM) for Summarization

We integrate Generative AI for comment summarization, employing the Google Gemini 1.5 flash model through LangChain (15). This approach applies to a method called Retrieval-Augmented Generation (RAG) to ensure accurate domain-specific and minimize hallucinations in Generative AI. Within the NCDs Listener system, the Generative AI's question-answering capability relies on a predefined set of questions to guide Gemini in generating summaries. Each comment summary reflects the characteristics of the overall community alongside suggestions and noteworthy information related to NCDs.

Prompt template and predefined questions: The prompt template serves as a structured framework to guide the responses of Generative AI. In this context, Generative AI addresses inquiries based on the aggregated comments and responds in English. For predefined questions, the generative AI delivers responses in both Thai and English to meet specific contextual requirements.

A Case Study: Lung Cancer Knowledge from the Crowd

To show the functionalities of the NCDs Listener tool, we focused on lung cancer as a case study. For this analysis, three cancer-related Facebook posts (16-18) were chosen, including a news post discussing cancer mortality rates, a post from a cancer community group, and a personal experience-sharing post. These posts collectively contained 1,797 comments, subsequently filtered to identify 30 comments specifically mentioning lung cancer. The system's data visualization and discovery capabilities were demonstrated through results such as knowledge extraction via keyword matching, classification using the BERT model, and Generative AI-based summarization using Google Gemini.

Results of Knowledge Extraction

From the data analysis and demographic characteristics of the commenters on the three Facebook posts (16,17,18) mentioning lung cancer, **Figure 3A** showed that 21 comments, or 67.74% of all comments (n=31), were classified as real experience sharing. However, a large portion of the comments consisted of name tagging and encouragements, which were classified as not informative. Regarding gender distribution of the patients, as shown in **Figure 3B.** 24 comments, or 77.40% of all comments, included identifiable gender information, with the majority being males (48.40% of comments specific to gender). It is expected that the reason for identifying a large number of male patients may be because the commenters were females who wanted to share their experiences about close individuals such as grandfathers, fathers, husbands, and brothers, who are groups of people with the highest risk behaviors for lung cancer, such as smoking, alcohol consumption, working in close proximity to pollutants such as construction sites. (19, 20)

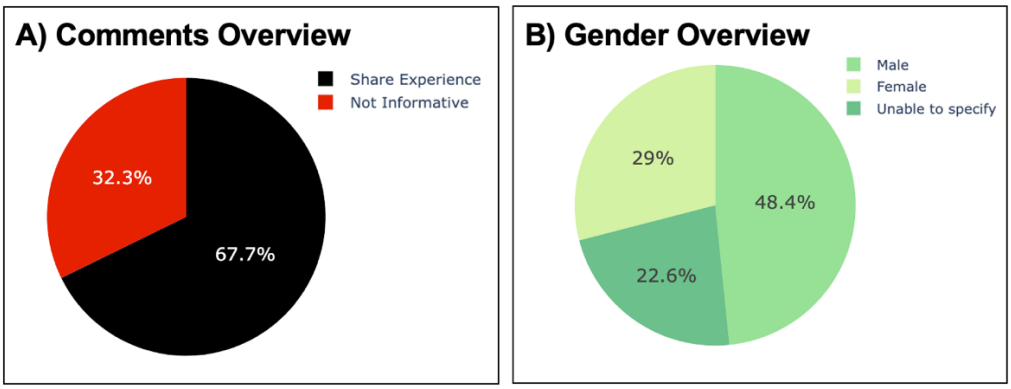

Figure 3. Data visualization shows the demographic characteristics of the commenters; (A) types of comments shared by commenters; (B) gender proportion of commenters.

An in-depth analysis using the NCDs listeners revealed that the most mentioned pre-lung cancer lifestyle behavior was non-smoking (10 comments). However, some comments included other behaviors such as smoking, not drinking, and exercising, as shown in **Figure 4A**. The results reflect that the actual causes of lung cancer may not correspond to the generally accepted risk factors, such as smoking, for lung cancer. (5, 21) Furthermore, **Figure 4B** indicates that three treatment options were mentioned: Chemotherapy (5 comments), Surgery (2 comments), and Radiation Therapy (1 comment). This reflects the perspective that lung cancer is commonly treated with chemotherapy and surgery, aligning with the public perception that these are the primary methods of cancer treatment. (21, 22)

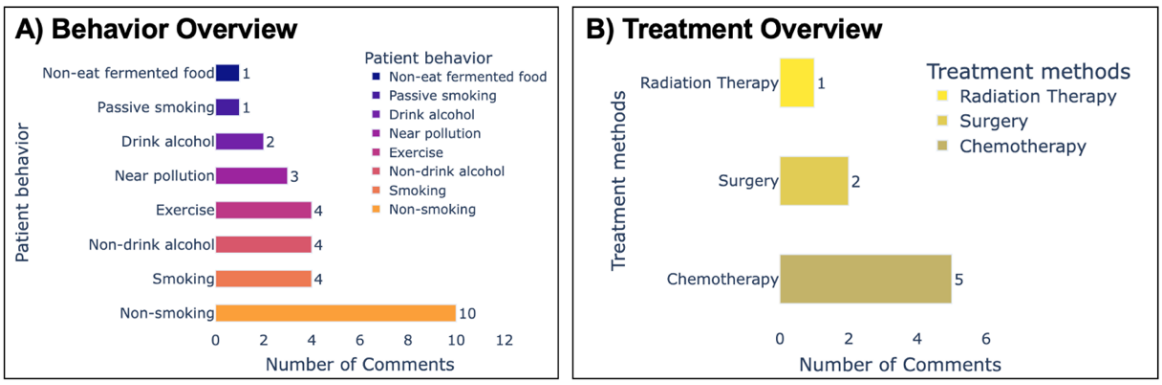

Figure 4. Data visualization shows the relevant cancer chart; (A) patient behavior shared by commenters; (B) treatment mentioned by commenters.

Summarization by the Generative AI Model

While visual data summarization can provide valuable insights, it may not deliver comprehensive conclusions for users. We propose utilizing Generative AI for data summarization based on RAG principles to enhance efficiency. This approach prioritizes accuracy management and minimizes hallucinations by leveraging comments related to lung cancer. The use of comments as meticulously crafted recommendation templates direct data to ensure maximum relevance and precision:

"Answer the question based only on the following context: {context} Question: {question} Respond in the same language."

The result of the comment summary will be generated based on the question as follows:

"Would you summarize the key characteristics of the comments?"

Finally, the results are summarized in a text format that is linked to the user's desired comment (the comment that mentions lung cancer) as follows:

> **"***The comments express a range of opinions and experiences related to lung cancer, predominantly focusing on the seemingly unpredictable nature of the disease. Many commenters express surprise and concern that lung cancer can affect non-smokers, healthy individuals, and even those who exercise regularly. There's a significant amount of shared personal experience, including stories of loved ones battling or succumbing to the disease, and a search for understanding regarding its causes. Genetic predisposition, environmental factors like PM2.5 and asbestos, and even seemingly unrelated factors like diet are discussed as potential contributing elements. Overall, the online conversation reflects a mixture of fear, empathy, and a quest for answers regarding lung cancer's etiology***"**

Based on both the visualization results and the summary of comments generated by Generative AI, most comments predominantly reflected the experiences of others, most commonly men, whose pre-cancer behavior was primarily as non-smokers. Discussions on treatment mainly

focused on chemotherapy. However, the majority of comments expressed concern about the possibility of developing cancer, even among healthy individuals or non-smokers.

Discussion & Conclusion

The NCDs Listener web application is designed to monitor social media data related to NCDs. The results incorporate social data obtained through keyword matching, BERT-based comment extraction, and Generative AI-based comment summarization for comprehensive insights, enabling users to understand societal contexts better. This application is a valuable tool for researchers studying the social characteristics of diseases of interest (5-7). Unlike existing tools primarily focused on marketing (3), this tool adopts a knowledge extraction approach that combines keyword matching with BERT to identify significant data characteristics. It sets it apart from most current research, emphasizing pre-trained models for analyzing emotional responses in comments (5,9,10). Furthermore, the NCDs Listener employs data visualization to present its findings, aligning with methodologies used in comparable studies (5,7,9,10).

Social media serves as a vast and continually evolving source of information, making social media information sources highly significant and valuable in addressing the needs of the masses. The information derived from social media encompasses both primary and in-depth data, which can be utilized to generate knowledge for the public and is crucial for developing effective methods or tools that cater to these needs. However, gathering and analyzing meaningful insights from many individuals is resource-intensive. Consequently, tools have been developed to assist in collecting, analyzing, and summarizing, known as social listening (media) tools.

The NCDs Listener tool was developed to collect and summarize basic and in-depth knowledge about NCDs data from social media. This tool utilizes keyword matching and BERT to classify the characteristics of comments and employs Generative AI to generate knowledge by summarizing comments related to social characteristics. Moreover, the RAG principle enhances accuracy and minimizes hallucination in Generative AI outputs. Additionally, the tool presents insights through data visualization, enabling users to comprehend the social characteristics and insights of NCDs effectively.

This tool is designed to assist doctors, researchers, and the general public address various problems. For instance, it enables doctors to conduct surveys by posing questions to gather public opinions comprehensively. Additionally, it aids doctors and researchers in studies related to treatment by collecting data from social media and filtering opinions relevant to specific diseases of interest. Lastly, the tool facilitates information communication to the general public, enabling them to read, interpret, and summarize with clarity. It can serve as a resource for learning or supporting initial decision-making concerning non-communicable chronic diseases.

References

(1) World Health Organization. Noncommunicable diseases. Available at: URL:https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Accessed Aug 29, 2024.

(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) 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.

(4) Geyser W. Top 25 social media listening tools for 2024. Available at: URL :https://influencermarketinghub.com/social-media-listening-tools/. Accessed Sep 3, 2024.

(5) 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).

(6) 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).

(7) 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.
(8) 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.
(9) Manguri KH, Ramadhan RN, Amin PR. Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research. 2020;5:54–65.
(10) Burzyńska J, Bartosiewicz A, Ręka s 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.
(11) 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.
(12) 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.
(13) 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).
(14) Chase H. LangChain: Build context-aware reasoning applications. Available at: URL: https://github.com/langchain-ai/langchain. Accessed Sep 14, 2024.
(15) Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019;4171–86.
(16) Rueng Lao Chao Nee. ตำรวจหนุ่มโพสต์เรื่องราวของตนเอง ผ่านเฟซบุ๊กส่วนตัวเกี่ยวกับโรคมะเร็ง ที่ตนกำลังเผชิญอยู่. Available at: URL: https://www.facebook.com/MorningNewsTV3/posts/pfbid03pjHtzLPnML6x4GU9pvh4Dc8Cmn69Hd3opGcVJ1jKUoaTWs9HoWdys1WuawyWDacl?rdid=1sAmvQWoDYwDJWq4. Accessed Nov 25, 2024
(17) Thyroid Cancer Patient Community of Thailand. แบ่งปันประสบการณ์ได้รับการรักษามะเร็งไทรอยด์ครบสามปี. Available at: URL:https://www.facebook.com/ThaiThyCaCom/posts/pfbid031BwfrFJTy68h3vGpnCpt21Hajv2ZDMpN5cDdaf1nMeHyZ9LKSGoRAbYruVdXeR1Kl?rdid=M5de3aVi1c0y3ckA. Accessed Nov 25, 2024
(18) Pantip. ลูก 8 เดือน จากอาการเหมือนเป็นหวัดสู่โรคทางเดินอาหารสุดท้ายที่มะเร็งสมอง. Available at: URL:https://www.facebook.com/pantipdotcom/posts/pfbid0KAnukjRbzTnNu89338DFSech2H1YLdAoCBT9cQZJd7qtXViPgwZUzGKGj6cXjxcol?rdid=mqnj2q1CrShOz6x7#. Accessed Nov 25, 2024
(19) Social Statistics Division of National Statistical office. The Smoking and Drinking Behaviour Survey 2017. Bangkok: Pimdeekarnpim; 2018.
(20) The Office of Permanent Secretary Ministry of Labour. Labour Statistics Yearbook 2020. Bangkok: Labor Economics Division; 2021
(21) Nattha Phipopchaiyasit. มะเร็งปอด ความเสี่ยงใกล้ตัว. Available at: URL: https://www.nci.go.th/th/New_web2024/service/sv17.html. Accessed Dec 20, 2024
(22) National cancer institute Thailand. การรักษามะเร็งตามหลักสากลที่ปฏิบัติกันอยู่ในประเทศไทย. Available at: URL: https://www.nci.go.th/th/Knowledge/treat.html. Accessed Dec 20, 2024