Doctoral Dissertation

Medical Needs Extraction of Breast Cancer Patients and Social Needs Extraction in COVID-19 Pandemic

Masaru Kamba

March 27, 2023

Graduate School of Science and Technology
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Masaru Kamba

Thesis Committee:

Professor Eiji Aramaki	(Supervisor)
Professor Yoshinobu Sato	(Co-supervisor)
Associate Professor Shoko Wakamiya	(Co-supervisor)
Assistant Professor Shuntaro Yada	(Co-supervisor)

Medical Needs Extraction of Breast Cancer Patients and Social Needs Extraction in COVID-19 Pandemic*

Masaru Kamba

Abstract

It is essential to identify the medical and social issues affecting patients as soon as possible. For example, Japan faces a wide range of medical issues, including a shortage of physicians, overwork of physicians, medical fees, pressure on social security costs, and assurance of the quality of medical care. In addition, the achievement of the Sustainable Development Goals is one of the current social issues. While these are issues that have already been identified, individual medical issues faced by the general patient population may change from time to time and era to era, and social issues that arise when unexpected events occur, such as the COVID-19 pandemic, may not yet have been fully identified. To solve such issues, they must first be identified.

In this thesis, we collected data from social media and used natural language processing techniques to collect and extract human needs. Because the voices and thoughts of the general public are accumulated in social media and question and answer (QA) sites, people's potential and trending needs can be mined there. For example, if the name of a specific cancer type is used as a search query on a QA site, the text of information needed by patients with that type of cancer can be obtained, and medical needs can be extracted from that information. If the question is for information on the side effects of a particular drug, the inability to access that information may be an issue. Therefore organizing the database

*Doctoral Dissertation, Graduate School of Science and Technology,
Nara Institute of Science and Technology, March 27, 2023.

in such a way that cancer patients can easily access required information would fulfill their needs.

In the early days of the COVID-19 pandemic, the term “Corona no-sei” (meaning, e.g., “because of the new coronavirus” and “because of COVID-19” in Japanese) was often posted on social media to indicate that the patient was restricted. Using this term as a search query and aggregating co-occurring words will highlight numerous restrictions. That is, it is possible to visualize cancelled plans and activities that could not be done. If we can extract needs and problems from such information, and find and organize social issues, it will be useful information in this era of the new coronovirus.

Keywords:

Medical Needs, Social Needs, Breast Cancer, COVID-19, Natural Language Processing (NLP), Medical Informatics

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

There are many unsolved medical and social issues across countries worldwide. In particular, Japan needs to solve a wide range of medical issues, such as inadequate numbers and overwork of physicians, medical fees, pressure on social security costs, and ensuring the quality of medical care [1]. The social issues include the achievement of the Sustainable Development Goals (SDGs) [2]. In addition to these already pointed out identified issues, we are tasked with managing novel issues. New medical and social issues may arise with the passage of time or when unexpected events occur, such as the COVID-19 pandemic. Therefore, we must first identify the issues. Because the new issues arise from people's needs, it is essential to first identify the needs that exist in the real world. Indeed, the importance of the real world has been recognized in recent years, and its utilization has been promoted, especially in the field of medicine [3, 4]. Research institutions[1] and companies[2] are additionally utilizing real world data for the development of next-generation medicine. Moreover, the government is considering the use of real world data to improve the efficiency of clinical trials [5]. In this thesis, we attempt to extract medical and social needs. Figure 1.1 shows the flow from needs to issues identification and resolution. Using figure as basis, we assert that extracting people's needs is the first step in solving medical and social issues. For example, if a cancer patient requires information on the side effects of a particular medication, the issue might be that it is difficult for the patient to access that information. A database of side-effects information that is accessible to all and easy to understand would be a necessary response to this issue, and

[1] https://www.rwd.kuhp.kyoto-u.ac.jp/
[2] https://prime-r.inc/

would provide cancer patients with access to this information, which would be useful for them to make treatment choices and fulfill their needs. In the early days of the COVID-19 pandemic, activities such as remote work and schools, and online medical care were identified as needs because around the world people were required to practice social distancing. If we can identify and organize the resolution of social issues emerging from such needs and problems, it will be useful information during the COVID-19 pandemic.

In consideration of the above, it is beneficial to clarify issues by focusing on their underlying needs. In the current age of computer networking, large amounts of real voices and thoughts of the general public are accumulated in social media and question and answer (QA) sites, and we project that there are potential needs for this information. Therefore, this thesis attempts to collect data on social media and extract information regarding various needs using natural language processing techniques.

This thesis challenged to integrate informatics and sociology by approaching sociological issues using informatics technology to create new services using social media data such as Twitter and QA data. Conversely, there are several academic disciplines that use data to study the science of society, such as computational social science and social informatics. Computational social science is the study of acquiring, processing, analyzing, and modeling large-scale social data using information technology to quantitatively and theoretically understand human behavior and social phenomena. Social informatics is the study of information and communication tools in cultural or institutional contexts. Our study differs from these studies in that our goal is to reach for developing real-world applications or services by utilizing research results and discoveries in social computing.

1.2 Target Needs

In this thesis, we attempt to extract medical and social needs.

First, for medical needs, we select diseases from the TOBYO database[3], which collects a large number of disease journals. Table 1.1 shows the top 20 diseases as of December 2022. Diseases for which treatment tend to be prolonged occupy

[3] `https://www.tobyo.jp/`

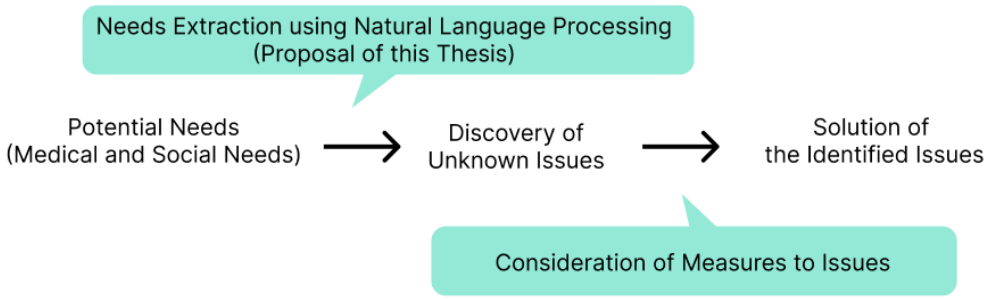

Figure 1.1: Issue Solving Process

the top positions, with breast cancer being the disease that is featured in the most number of disease journals. Hence, we chose to identify the medical needs of breast cancer patients.

Next, we consider social needs. Since the index case of the COVID-19 virus infection was observed in Japan in January 2020, the people's way of life has been changed drastically. Therefore, because we think that many issues remain to be resolved in the COVID-19 pandemic, we attempt to identify the social needs arising from the COVID-19 pandemic.

1.2.1 Medical Needs of Breast Cancer Patients

In this section, we set a background for extracting the medical needs of breast cancer patients.

As a related study, the Shizuoka Cancer Center conducted three surveys with the aim of improving the quality of life of cancer patients [6–8]. Based on the results of these surveys, a database called Cancer Problem Classification (CPC) [4] was manually created and is maintained at the Shizuoka Cancer Center. The CPC was created by collecting complaints from cancer patients throughout Japan and categorizing them. The CPC systematizes the worries and burdens of cancer patients surveyed through telephone consultations and other means. A total of 7,855 people participated in 2003 and 4,054 in 2013.

[4] https://www.scchr.jp/cancerqa/start_shizuoka.html

Table 1.1: Diseases that there are a number of disease journals (top 20 categories).

Diseases	Number of Disease Journals
Breast cancer	7193
Depression	3094
Infertility	2198
Cervical cancer	1320
Ovarian cancer	1166
Rheumatoid arthritis	1139
Panic disorder	1024
Type 1 diabetes	1013
Uterine fibroids	963
Schizophrenia	940
Ulcerative colitis	757
Malignant lymphoma	711
Gastric cancer	709
Systemic lupus erythematosus	698
Pancreatic cancer	677
Lung cancer	673
Uterine cancer	636
Bipolar disorder	633
Colorectal cancer	564
Acute myelogenous leukemia	546

Because the CPC was created manually by specialists, its update and maintenance cost and time required to create a new one were cited as problems. However, with the rapid increase in internet penetration in recent years, a large number of illness-related concerns have been accumulated in Japan via blog posts which are very popular. By contrast, TOBYO has collected many diaries and blogs specialized in struggles against diseases (about 63,000 in number), dealing with about 1,500 diseases. Among these, blogs on breast cancer, for which treatment tends to be prolonged, account for more than 10% (approximately 6,900 entries), and are particularly numerous. Yahoo Q&A service (YJQA) [5] is one of the leading QA services in Japan, and a search using the keyword breast cancer yields approximately 60,000 questions. In this way, a vast archive of patients' complaints has been created on the Internet.

In this context, many recent studies have utilized accumulated information [9–13]. Thus, using patients' narratives on the web can provide a qualitative and timely understanding of needs from the patient's perspective and be considered a type of patient-reported outcome, which may help transform health care in terms of patient-centered care [14, 15].

1.2.2 Social Needs during COVID-19 Pandemic

In this section, we set the background for extracting social needs during the COVID-19 pandemic.

The COVID-19 disease has become a worldwide epidemic, causing a major impact, especially in social and economic sectors all over the world [16]. In the early stages of the pandemic, health authorities recommended social distancing to control the spread of the virus, reduce cases and avoid overwhelming the healthcare facilities [17–19]. Each country has its own strategy for dealing with COVID-19. A survey conducted in six countries illustrates the public's perception regarding the measures taken in each country as a response to COVID-19 [20]. Different results were observed as a result of social distancing policies in several aspects of life, including economic activities [21], and consumer behavior, such as reductions in mobility [22]. There was an association between the implementation of some

[5] `https://chiebukuro.yahoo.co.jp/`

mitigation policies in response to COVID-19 and the outcomes regarding public mobility [23], one of which was also seen in Japan.

Immediately after the Japanese government confirmed the index COVID-19 case on January 16, 2020, the number of infected cases quickly escalated within three months, which compelled the government to declare a state of emergency to prevent further spread. This measure had a significant impact on the daily routines and social life of residents in Japan, as they were forced to refrain from going out, close schools, work from home, and were restricted from visiting crowded locations, such as departmental stores and movie theaters. The first state of emergency effectively reduced the number of COVID-19 cases [24], however, at a high cost of public mental well-being, education quality, and economy. The case number quickly rebounded after the state of emergency was lifted, suggesting that the government was confronted by the dilemma of mitigating the social and economic impact of lockdown and stopping the spread of COVID-19 [25]. Owing to the fluctuations in COVID-19 cases, the government declared other states of emergency, knowing that the impact of a COVID-19 lockdown could be profound at societal and economic levels.

There have been various investigations into the states of emergency. One of them was a report on the prediction of COVID-19 infection using state-space models [26]. There was also a study showing the effect of a state of emergency on mental health [27]. In the aspect of mobility, studies showed that state of emergency suppressed social activities between the masses [28]. Furthermore, it also changes people's behavior, such as complying with the request to stay at home, which was also supported by cell phone location data [29–31].

A disruption caused by restrictions imposed during a pandemic can be interpreted as a failure to do what was planned. However, detecting this disruption was quite challenging. For example, it was difficult to obtain behavioral data on trips that individuals could not take or events they could not attend due to the restriction. As a result, social media, which people use to share their activities and thoughts, can be a great source of information. Twitter has proven helpful in summarizing peoples' responses about the pandemic and its measures, showing challenges experienced throughout the pandemic [32]. Previously, Twitter had been used to determine public opinion about COVID-19 in Korea and Japan,

showing trending words during the pandemic [33].

As people are actively sharing their day-to-day on Twitter, Twitter has the potential as a data source to investigate the impacts of restrictions on the public.

1.3 Objective

As stated in Section 1.2, in this thesis, we aim to identify the medical needs of breast cancer patients and social needs during the COVID-19 pandemic.

In the extraction of medical needs of breast cancer patients, we collected data of questions from breast cancer patients posted on the Yahoo Q&A service (YJQA) because patients' claims on the web can provide a qualitative and timely understanding of needs from the patient's perspective, again as stated in Section 1.2.1.

However, there are some limitations to using the accumulated information, the biggest problem being the difficulty in examining large amounts of data. Therefore, this thesis proposes a method for automatically classifying the needs of breast cancer patients using natural language processing technology. This study aimed to extract the needs of patients with breast cancer from the YJQA data and categorize them into cancer problem classification (CPC) categories. Here the CPC categories are hierarchically structured from the first-level to the fourth-level. Table 1.2 shows examples of the CPC's first-level categories through fourth-level categories for the part of category number 1. In the CPC's first-level through third-level categories, the problem granularity is coarse, and it is difficult to understand the specific issues. Therefore, this study attempted a fourth-level categorization to more concretely grasp patients' problems.

In the extraction of social needs during the COVID-19 pandemic, we utilize data from Twitter because it can be a great source of information to summarize people's activities and thoughts as mentioned in Section 1.2.2. When the Japanese government announced a state of emergency, the phrase "Corona no-sei" was frequently tweeted on Twitter, followed by a sentence describing restricted actions. The words co-occurring with the phrase "Corona no-sei" are considered to represent disrupted events and actions. Therefore, we identify restrictions and extract social needs by tabulating co-occurring words from Twitter posts containing the

word “Corona no-sei”. Thus, the purpose of this study is to explore and visualize the cancellation of events due to COVID-19 measures in Japan.

1.4 Outline

The remainder of this thesis is structured as follows. Chapter 2 describes the method used to identify the medical needs of breast cancer patients and presents the results. Using two corpora (the cancer problem classification (CPC) corpus and the Yahoo Q&A service (YJQA) corpus), we present an algorithm to classify text from the YJQA service into the CPC categories. The identified categories are used to show how to identify patients' needs. Chapter 3 describes the method and results of social needs extraction, focusing on COVID-19. We identify the social needs by extracting words that co-occur with “Corona no-sei” on Twitter. This allows for the identification of disrupted plans. In addition, we evaluate the relationship between the number of Twitter posts and the number of positive cases of COVID-19, utilizing the daily COVID-19 case numbers from the special site for COVID-19 by Nippon Hoso Kyokai (Japan Broadcasting Corporation, abbreviated as NHK)[6]. Chapter 4 presents proposals for the identified medical and social needs. Finally, Chapter 5 summarizes the thesis and discusses issues for further research.

[6] https://www3.nhk.or.jp/news/special/coronavirus/data/

Table 1.2: The CPC's first-level categories through fourth-level categories for the part of category number 1

1st-level		2nd-level		3rd-level		4th-level	
Code	Name	Code	Name	Code	Name	Code	Name
1.	Outpatient	1.1.	Choice of hospital and doctor	1.1.1.	Difficulties and hesitations in choosing a hospital and doctor	1.1.1.1.	Difficulty in obtaining information for selecting hospitals and doctors
						1.1.1.2.	Difficult to choose a hospital
						1.1.1.3.	conditions for hospital selection (access, facilities, etc.)
						1.1.1.4.	Hospital selection for future cancer screening
						1.1.1.5.	I can't find a hospital that can handle the aftereffects (lymphedema, etc.)
						1.1.1.6.	I can't find a hospital that can receive special treatment (bone marrow transplantation, etc.) and advanced medical care.
		1.2	Outpatient consultation	1.2.1.	Outpatient treatment	1.2.1.1.	Long waiting time
						1.2.1.2.	Difficult outpatient treatment for anticancer drugs

2 Medical Needs of Breast Cancer Patients

2.1 Background and Related Work

In this chapter, we aimed to extract the medical needs of patients with breast cancer from the Yahoo Q&A service (YJQA) data and classify them into the cancer problem classification (CPC) categories. We adopted the fourth-level CPC categories described in the above table for the classification of patients' medical needs. In the CPC's first-level categories, the problem granularity is coarse, and it is difficult to understand the specificity of the issues. For example, while the CPC's first-level category is outpatient, the corresponding fourth-level categories include "1.1.1.1. Difficulty in obtaining information to select a hospital or doctor," and "1.1.1.2. Difficulty in hospital selection." Therefore, we attempted to classify the fourth-level categories to more concretely grasp patients' problems.

As mentioned Section 1.2.1, the Shizuoka Cancer Center conducted three surveys and created the CPC database with the aim of improving the quality of life of cancer patients [6–8].

Rosenblum and Yom-Tov [9] investigated how people search for information related to attention-deficit/hyperactivity disorder using the Microsoft Bing search engine[1] and Yahoo! Answers[2], a web QA site, Park et al. [10] investigated the use of medical concepts regarding diabetes from the textual data of blogs and QA sites, whereas Yom-Tov and Gabrilovich [11] investigated the side effects of medications from web search queries. Tsuya et al. [12] demonstrated that cancer patients share information about their diseases, including diagnosis, symptoms,

[1] `https://www.bing.com/`

[2] Yahoo Answers has shut down as of May 4, 2021.

and treatments via Twitter[3], and Hong [13] explored whether patients could accurately and adequately express their information needs on Chinese health QA websites.

For breast cancer patients, Hongru et al. [34] explored information needs from database sites, including Web of Science[4] and PubMed[5]. Cristina and Anne [35] investigated information and emotional needs for long-term survivors of breast cancer. Dean et al. [36] investigated unmet needs in breast cancer survivors. However, these studies are based on data from clinical studies and generally do not reflect the thoughts of patients. As mentioned in Section 1.2.1, we think that patient needs can be extracted from social media data, more so in the modern era.

Hence, we investigate the medical needs for breast cancer patients utilizing social media data.

2.2 Materials and Methods

2.2.1 Materials

We built a dataset of 7,993 questions submitted to the YJQA service between January 1, 2018, and July 31, 2020. The CPC has been systematized to extract the problems and burdens of cancer patients, consisting of 16 first-level categories and 631 fourth-level categories. We utilized two corpora for training, the CPC and YJQA corpora.

The CPC corpus is a large collection of pairs of cancer survivors' worries and their labels. The label consists of the CPC category code and the CPC category name (hereafter, both are collectively referred to as CPC categories), obtained from the CPC database. Unless otherwise noted, the CPC categories represent fourth-level categories. An example from the CPC corpus is presented in Table 2.1.

The YJQA corpus is a labeled corpus of 1000 randomly selected questions on breast cancer posted to the YJQA service from January 1, 2018, to June 9,

[3] https://twitter.com/

[4] https://clarivate.com/solutions/web-of-science/

[5] https://pubmed.ncbi.nlm.nih.gov/

Table 2.1: Cancer survivors' worries, CPC category code and name.

Cancer survivors' worries	CPC code	CPC name
I was worried because we had to make decisions based on my limited knowledge and emotions, without any information or indicators to judge whether the hospital's policies and techniques were accurate, especially whether my doctor was trustworthy.	1.1.1.1	Difficulty in obtaining information for selecting hospitals and doctors.

2020. Because multiple different worries are possible, each question is manually assigned to up to 3 different CPC categories. An example from the YJQA corpus is presented in Table 2.2.

We assigned CPC categories to 456 of the 1000 cases, while the remaining 546 cases had no corresponding CPC categories. Thus, the total number of cumulatively classified questions was 661, which were assigned to 133 CPC categories. Table 2.3 summarizes the most frequent categories, up to the 20th (top 20), regarding the number of questions classified. For example, the most frequent category was "worrying about cancer with subjective symptoms," with 24.2% of the labeled data falling into this category. Moreover, the category "difficulty in expressing questions and concerns to doctors" was included in the top 10 categories, suggesting that people submitted questions to the YJQA because they had difficulty expressing their concerns to their doctors.

Of the 7993 questions submitted to the YJQA service, 6993 were used as the YJQA corpus data classified using CPC categories, excluding the 1000 labeled questions (training data).

Table 2.2: Questions in YJQA, CPC category code and name.

Questions in YJQA	CPC code	CPC name
Choosing a hospital for breast cancer treatment. I'm wondering if I'm making a mistake in choosing the first hospital. Is there any problem in choosing the university hospital that is closest to my house?	1.1.1.1.	Difficulty in obtaining information for selecting hospitals and doctors.
I had a breast cancer screening and had to be retested for a suspected breast mass. My mother had breast cancer. I will have a mammogram next month. Is the chance of getting breast cancer high? I am very scared and worried.	3.2.2.1./16.3.2.1.	I'm worried about finding out the test results/Concerns regarding suspicion of cancer (other)

2.2.2 Classification Algorithm

The classification algorithm consists of the following steps:

1. Preprocessing: Convert the two corpora (CPC corpus and YJQA corpus) into term frequency (TF)-inverse document frequency (IDF)-weighted word vectors (corpus word vector) as document-term representation.

2. STEP1: Given an unknown problem, convert the problem into TF-IDF-weighted word vectors (problem word vector).

3. STEP2: Classify the target problem into the most relevant CPC category based on cosine similarity between the problem word vector from STEP1 and corpus word vectors from the two corpora.

Table 2.3: Results of manual classification of YJQA questions (top 20 categories).

CPC code	CPC name	n (%)
16.3.1.1.	Worrying about cancer with subjective symptoms	160 (24.2)
16.2.1.1.	Matters related to cancer screening	85 (24.2)
12.2.4.1.	Anxiety due to lack of knowledge about cancer	42 (12.9)
16.3.2.1.	Concerns regarding suspicion of cancer (other)	39 (6.4)
9.1.2.2.	Difficulty in asking questions or expressing concerns to the doctor	17 (2.6)
3.2.2.2.	Worrying about the results and their trends	17 (2.6)
12.1.1.1.	Anxiety about the possibility of recurrence or metastasis	14 (2.1)
3.2.1.6.	Concerns about undergoing tests (other)	11 (1.7)
3.1.1.1.	Uncertainty about treatment options	10 (1.5)
3.2.2.3.	Issues related to receiving tests (other)	9 (1.4)
14.1.2.14.	Thing about medical expenses (others)	8 (1.2)
3.1.3.5.	The received treatment (choice), whether it is correct	8 (1.2)
12.3.2.2.	Linking illness to cancer	7 (1.1)
3.1.5.1.	Use of folk remedies and health foods	6 (0.9)
5.2.1.3.	I'm not convinced by the doctor's explanation	6 (0.9)
3.1.2.1.	Before treatment: Concerns about future treatment	5 (0.8)
12.3.2.1.	Anxiety increases while waiting for test results	5 (0.8)
3.1.3.11.	I am worried about the effect of anticancer drugs	5 (0.8)
12.3.2.3.	I can't stop thinking about cancer	5 (0.8)
1.1.1.4.	Hospital selection for future cancer screening	5 (0.8)

Using a morphological dictionary mecab–ipadic–NEologd[6], we extract nouns, verbs, and adjectives while excluding symbols and numbers. For the TF–IDF calculation, we utilized the TfidfVectorizer using the default parameters in the sklearn.feature_extraction.text module[7].

Thereafter, we constructed three classification methods using the CPC corpus, the YJQA corpus, and their combined corpus, referred to as the description-based (D-based) method, example-based (E-based) method, and description and example combination-based (D+E-based) methods, respectively.

2.3 Evaluation

2.3.1 Evaluation Methods

We explain how to objectively evaluate the three methods. The accuracy of each method was evaluated by calculating the proportion of correct classifications. The proportion of correct classifications for the D-based method is calculated as follows. First, we find the categories with the highest cosine similarity between the word vectors of the CPC corpus and the manually labeled YJQA corpus (top 1-10). Next, we calculate the proportion of correct categories. A category is counted as correct if at least 1 of the 3 (maximum) categories is included. Based on the highest cosine similarity, the calculated percentage is referred to as the top 1 accuracy (Acc@1). Similarly, using the top 10 cosine similarities, the top 10 accuracies (Acc@10) are calculated. The proportion of correct classifications is calculated using 5-fold cross-validation [37] to evaluate the E-based method. Using the cosine similarity between the training and validation data sets, the proportion of correct classifications is the mean and median of the rate, as in the above calculation. For the evaluation of the D+E-based method, the proportion of correct classifications is calculated by employing the same evaluation method as for the E-based method using both the CPC and YJQA corpora. Note that because the data used for the evaluation is from the YJQA corpus, it is not possible to evaluate all 631 fourth-level CPC categories. That is, these methods are

[6] https://github.com/neologd/mecab-ipadic-neologd

[7] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

evaluated for 133 fourth-level CPC categories from the YJQA corpus. However the remaining categories were not categorized in the extracted data and therefore were not considered more common patient voices currently.

Here, we do not apply the statistical test to compare these methods because the p-value obtained by statistical tests is commonly misused and misinterpreted, and the American Statistical Association[8] announced formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the p-value [38]. In accordance with this statement, it is not appropriate to apply statistical tests in this situation.

Even if the estimates are the same, different precisions of the estimates will result in different p-values. Because the precision of the estimates is affected by sample size, the sample size must be determined in advance and the p-values should be interpreted appropriately. These should be determined before data are obtained; otherwise they can be intentionally manipulated to reach statistical significance, if they are made after the data are available. For example, increasing the number of cross-validations might make them statistically significant.

2.3.2 Classification Accuracy

Table 2.4 shows the proportions of correct classifications calculated using the above evaluation methods. Here, SD means unbiased sample standard deviation. For the D-based method, there are missing values because 5-fold cross-validation is not utilized as described in the evaluation method section. The Acc@1 and Acc@10 of the D-based method were approximately 10% and 30%, respectively. Furthermore, for both the E-based and D+E-based methods, they were approximately 50% and 70%, respectively. The E-based method is an optimized classification method used to classify YJQA questions. However, it does not cover all CPC categories, whereas the D+E-based method does so, and the rate of correct answers is not significantly different from that of the E-based method. Therefore, in this study, we interpret the results of the D+E-based method.

[8] `https://www.amstat.org/`

Table 2.4: Accuracy for each method. The three methods: the description-based (D-based) method, example-based (E-based) method, and description and example combination-based (D+E-based) methods.

Accuracy	Statistics	D-based	E-based	D+E-based
Acc@1	Mean	0.1096	0.4891	0.4781
	SD	–	0.017	0.018
	Median	–	0.4835	0.4725
Acc@10	Mean	0.2946	0.6960	0.7062
	SD	–	0.030	0.201
	Median	–	0.7015	0.7106

2.3.3 Classification Results

We present the classification results of the D+E-based method for the target data to be classified. Table 2.5 lists the top 20 frequent categories, with the top 10 categories accounting for 61.9% of the total. The category with the most frequent questions was "worrying about cancer with subjective symptoms" (1661 questions), which accounted for 23.8% of the total. There were 448 categories classified by the D+E-based method, and the distribution of the top 30 categories is shown in Figure 2.1. The rate of change from the top 1 to the top 2 categories was the largest at 57.7%. Moreover, the rate of change from the top 20 categories was 20% to 40%, after which it was approximately 10%. As a result, the frequency distribution has a long tail.

Table 2.5: Results obtained using the D+E-based method (top 20 categories).

CPC code	CPC name	Frequency (%)
16.3.1.1.	Worrying about cancer with subjective symptoms	1661 (23.8)
16.2.1.1.	Matters related to cancer screening	702 (10.0)
16.3.2.1.	Concerns regarding suspicion of cancer (other)	494 (7.1)
12.2.4.1.	Anxiety due to lack of knowledge about cancer	419 (6.0)
3.1.3.5.	The received treatment (choice), whether it is correct	255 (3.6)
3.2.2.2.	Worrying about the results and its trend	234 (3.3)
12.1.1.1.	Anxiety about the possibility of recurrence or metastasis	225 (3.2)
9.1.2.2.	Difficulty in asking questions or expressing concerns to the doctor	137 (2.0)
12.3.2.3.	I can't stop thinking about cancer	111 (1.6)
9.1.1.1.	Doctorś words and attitude	93 (1.3)
12.3.2.2.	Linking illness to cancer	82 (1.2)
12.5.2.4.	Depression (other)	75 (1.1)
3.1.1.1.	Uncertainty about treatment options	70 (1.0)
13.3.2.1.	Fluctuations, loss and change in femininity	68 (1.0)
3.1.1.6.	Concerns about hesitation when deciding on a treatment method and difficulty in selecting a treatment (other)	57 (0.8)
12.3.2.1.	Anxiety increases while waiting for test results	46 (0.7)
14.1.4.2.	Future anxiety due to inability to take out or renew private insurance	46 (0.7)
3.1.5.1.	Use of folk remedies and health foods	46 (0.7)
3.2.1.6.	Concerns about undergoing tests (other)	46 (0.7)
3.1.3.11.	I am worried about the effect of anticancer drugs	45 (0.6)

Figure 2.1: Classification using the D+E-based method (X-axis) and its frequency (Y-axis; top 30 categories). D+E: description and example combination-based method.

2.3.4 Similarity of Distribution between Manual Classification and the Proposed Method

We evaluated whether the frequency distribution of the proposed method was close to that of the manual method. Comparing the classification results with the manual classification results (Table 2.3), we found that 7 out of 10 categories with high frequency were the same, and the first and the second category in both cases were "worrying about cancer with subjective symptoms" and "cancer screening."

The frequency distribution of the CPC, including the low-frequency part, was compared between the proposed and manual methods. The top 30 categories' frequency distributions in the D+E-based method were used for visual and numerical evaluation of all categories. Figure 2.2 shows the distribution of the classification results using the D+E-based method and manual classification. The distributions were similar. In addition, we calculated the Jensen-Shannon divergence [39] for all categories to measure the distance between the distributions. Values closer to zero indicated higher degrees of similarity in distribution.

The value of the Jensen-Shannon divergence for the distribution of the manual classification and D+E-based classification result is 0.105, which shows that the two distributions are similar. Although the individual accuracy was low, the CPC distribution obtained by the proposed method was closer to the correct one.

Therefore, it is possible that the proposed method can be used to conduct a large-scale survey of patient concerns automatically.

Figure 2.2: Distribution of the classification results using the D+E-based method and manual classification.

2.3.5 Consistency of the Actual Questions and Results

Here we discuss the consistency of the actual questions and results with high and low cosine similarity, respectively. The questions are translated into English from Japanese.

For the question

> Please tell me the constitution that is susceptible to breast cancer!

which had the highest cosine similarity, the top three classification categories with cosine similarity are shown in Table 2.6. It is unclear whether a cancer patient asked the first question, but it appears to express concern about the possibility of developing cancer.

Table 2.6: The CPC code and name of question with the highest cosine similarity.

Cosine similarity	CPC code	CPC name
0.714	15.1.1.5.	I was told that I am cancer constitution
0.206	12.2.5.1.	Suspecting or worrying about another type of cancer
0.204	11.3.1.1.	Current health condition

For the question

> I had one of my breasts removed due to breast cancer. I did not have simultaneous reconstruction. My breasts are small, to begin with, so when I was asked about simultaneous reconstruction, I didn't think much about it and told my doctor that I would think about it after the surgery. In my 40s, I was admitted to the hospital, but most people my age had simultaneous reconstruction and expanders. I wondered if I had made the wrong choice. Since if I have to do it later, I'll have to have one more surgery, I think it's okay as is. I heard that it takes quite a few days to reconstruct. It needs one year at the earliest. Moreover, I heard that nipple and areola surgeries are different. That's a long time. But I still think I want to have reconstruction.

> If you have reconstructed, if you haven't reconstructed, if you have reconstructed in another way, if it's not covered by insurance, etc., please give me some advice! I'd like to hear about your experiences. It will be two months until my next visit to the hospital. I want to ask my doctor many questions, so if you could tell me anything, I would be very happy. Please give me some advice. Moreover, it seems that the implants and expanders for reconstruction have been discontinued because they are carcinogenic. I don't think I will be able to have reconstruction for a while, but please advise me.

with the second highest cosine similarity, the top three classification categories with cosine similarity are shown in Table 2.7. The second question was about breast reconstruction.

Table 2.7: The CPC code and name of question with the second highest cosine similarity.

Cosine similarity	CPC code	CPC name
0.682	13.3.1.6.	Matters related to breast reconstruction
0.264	3.1.1.1.	Uncertainty about treatment options
0.255	3.1.3.5.	The received treatment (choice), whether it is correct

For the question

> I am undergoing treatment for breast cancer, and my white blood cell count has dropped due to side effects, so my immune system is not high. I don't want to go to the birthday party at my parents-in-law's because I'm worried that I might get infected with the coronavirus. My mother-in-law and father-in-law know that I am undergoing treatment and my immunity is low, but they don't want to cancel the party because it's their adorable grandchild's birthday. It's hard for me to tell them. I don't want my husband to go either, but he doesn't seem to mind at all. Is there any way to avoid attending the party?

with the second highest cosine similarity, the top three classification categories with cosine similarity are shown in Table 2.8. The third was a concern about coronavirus (COVID-19).

Table 2.8: The CPC code and name of question with the third highest cosine similarity.

Cosine similarity	CPC code	CPC name
0.657	11.1.2.3.	Persistent side effects of anticancer drugs (other)
0.515	11.1.1.8.	Symptoms of side effects from anticancer drugs (other)
0.417	15.2.16.1.	Relationship with family (other)

We also discuss questions that could not be correctly classified. The reason for the inability to classify the questions with the highest cosine similarity correctly can be the use of the cosine similarity between the word vectors in the bag of words, and the context could not be taken into account. More specifically, since the question included the word "constitution," it was considered to be classified in the category that included the word "constitution." Similarly, the top 2 worries about breast reconstruction could be classified in the CPC category, which includes the phrase "breast reconstruction." The top 3 problems are related to COVID-19, which is not included in the current CPC category, and therefore must be newly defined.

For the question

> Please provide a Japanese translation.
> Survival rates are improving in the U.S. and the UK because of national mammography screening programs and vigorous campaigns to educate the public about breast cancer

with the lowest cosine similarity, the top three classification categories with cosine similarity are shown in Table 2.9.

Table 2.9: The CPC code and name of question with the lowest cosine similarity.

Cosine similarity	CPC code	CPC name
0.018	3.1.3.11.	Worried about the effect of anticancer drugs
0.017	12.3.2.4.	Anxiety about superimposing oneself on a (familiar) person with the same disease.
0.016	12.3.2.3.	Can't stop thinking about cancer

For the question

> Please provide a Japanese translation.
> Most people were all surprised to learn that Angelina Jolie one of the world's most celebrated actresses, had opted to undergo a preventive double mastectomy after being diagnosed with an 87% chance of breast cancer because she was carrying a high-risk gene, BRCA1 There are some risk factors that can increase a person's chances of getting breast cancer: aging, excessive alcohol consumption, or obesity. Among the risk factors, heredity is responsible for 5% of breast cancer cases. Thanks to advances in genetics, I now know for sure that the BRCA1 and BRCA2 genes can cause breast cancer.

with the second lowest cosine similarity, the top three classification categories with cosine similarity are shown in Table 2.10.

Table 2.10: The CPC code and name of question with the second lowest cosine similarity.

Cosine similarity	CPC code	CPC name
0.020	3.1.2.1.	Before treatment: Concerns about future treatment
0.018	11.1.24.4.	Irregular/changing menstruation due to hormone therapy etc.
0.015	12.2.1.7.	Concerns about other diseases

For the question

Can AI prevent missed breast cancer?

with the third lowest cosine similarity, the top three classification categories with cosine similarity are shown in Table 2.11.

Table 2.11: The CPC code and name of question with the third lowest cosine similarity.

Cosine similarity	CPC code	CPC name
0.023	12.2.4.1.	Anxiety due to lack of knowledge about cancer
0.022	12.5.2.2.	Depression due to anxiety about illness or death
0.021	14.1.2.14.	Medical expenses (other)

As for the results with the lowest to third-lowest cosine similarity, the question was a request for Japanese translation from English and was not in itself a question about breast cancer. This is because questions including the phrase “breast cancer” were also extracted when searching for “breast cancer.” Hence, the data acquisition method should be improved in the future.

3 Social Needs Focusing on COVID-19

3.1 Background and Related Work

In this chapter, we aimed to extract from Twitter data social needs related to COVID-19. We utilize data from Twitter because it can be a great source of information to summarize people's activities and thoughts. When the Japanese government announced a state of emergency, the phrase "Corona no-sei" was frequently tweeted on Twitter, followed by a sentence describing restricted actions. The words co-occurring with the phrase "Corona no-sei" are considered to represent disrupted events and actions. Therefore, we identify restrictions and extract social needs by tabulating co-occurring words from Twitter posts containing the word "Corona no-sei". Thus, we are able to explore and visualize the cancellation of events due to COVID-19 measures in Japan.

There are many studies on COVID-19 using social media data, such as Twitter. For example, Joanne, Eileen and Garving [40] investigated the topics and sentiments in the public COVID-19 vaccine-related discussion and Krittanawong et al. [41] investigated misinformation dissemination of COVID-19 on Twitter. Ferawati et al. [42] investigated how Twitter was used to report vaccine-related side effects and to compare the mentions of these side effects from 2 messenger RNA (mRNA) vaccine types developed by Pfizer and Moderna, in Japan and Indonesia. Gao et al. [43] investigated COVID-19 concerns in each Japanese prefecture. Uehara et al. [44] investigated attitudes toward vaccines or vaccination during the COVID-19 pandemic across different Japanese prefectures, using Yahoo! JAPAN search queries.

For research on citizen feedback, Ishida et al. [45] proposed a method using

social media data. They implemented a multitask learning framework to estimate associated viewpoints using a BERT model. However this method requires a lot of work to label the data.

The results of analyses need to be output in a timely manner for issues that have a strong social impact. When making decisions based on the results of data analysis, slower output of results slows down the decision-making. By responding promptly, it is expected that losses to society can be minimized. We propose that the setting up of appropriate search queries can enable the identification of social needs without using resource-intensive methods.

Therefore, we investigate the emergent social needs from COVID-19 using a simple approach on Twitter.

3.2 Materials and Methods

We utilized two types of data, one of which are Japanese tweets posted on Twitter between February 1st, 2020 and November 30th, 2021 and which are utilized to summarize the words that co-occurred with the phrase "Corona no-sei". The other data type are the daily COVID-19 case numbers from a special site for COVID-19 by Nippon Hoso Kyokai (Japan Broadcasting Corporation, abbreviated as NHK)[1]. Using this data, the association between the number of posted tweets and the number of COVID-19 cases is evaluated.

3.2.1 Materials

COVID-19 cases

For the daily increase in COVID-19 cases, we obtained the number of new COVID-19 positive cases by manual downloads from the special site for COVID-19 by NHK, to explore the relationship between the number of positive cases and the number of tweets. A total number of 1,726,943 COVID-19 positive cases were extracted during our target period.

[1] `https://www3.nhk.or.jp/news/special/coronavirus/data/`

Tweets

Another material for this study are the 300,778 samples of tweets that contained the Japanese phrase "Corona no-sei (meaning due to COVID-19, because of COVID, or considering COVID)."

The "Corona no-sei" phrase is frequently used by the general public in social media and everyday conversation to express (often negative) feelings when a Japanese citizens' activities or life plans gets interrupted by COVID-19 outbreak. The tweet data was provided by NTT DATA Corporation. In addition, we extracted co-occurring nouns and verbs from the obtained "Corona no-sei" tweets by applying dependency analysis implemented in the system developed by Yoshinaga et al. [46–48].

3.2.2 Needs Extraction Methods

The contexts following "Corona no-sei," which indicate a high level of negative concerns about COVID-19, frequently contain verbs in the negative form and nouns associated with them. Figure 3.1 shows an example of an actual tweet posted on Twitter. By aggregating these nouns and verbs, we extracted information on the restrictions imposed and the activities or plans cancelled due to the COVID-19 epidemic. These information allowed us to capture the potential social and psychological impact arising from disrupted life plans.

Although there are several expressions synonymous with "Corona no-sei" (e.g., "because of the new coronavirus" and "because of COVID-19"), we choose "Corona no-sei" as a casual expression used by the general public in social media and colloquial speech. The frequency of nouns and verbs in the tweets containing "Corona no-sei" is counted to identify the restrictions placed on people's lives.

3.3 Results

Figure 3.2 shows the time trend of "Corona no-sei" tweets (blue line) compared to the trend of positive cases (red line). There were three state of emergency announcements within our targeted period between February 1st, 2020 and November 30th, 2021, which are highlighted in gray color in the figure. The number of

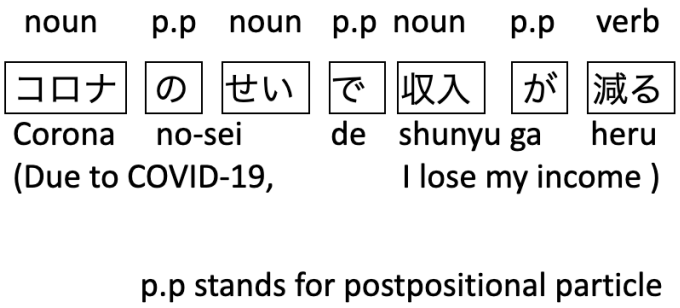

Figure 3.1: An example of an actual tweet posted on Twitter.

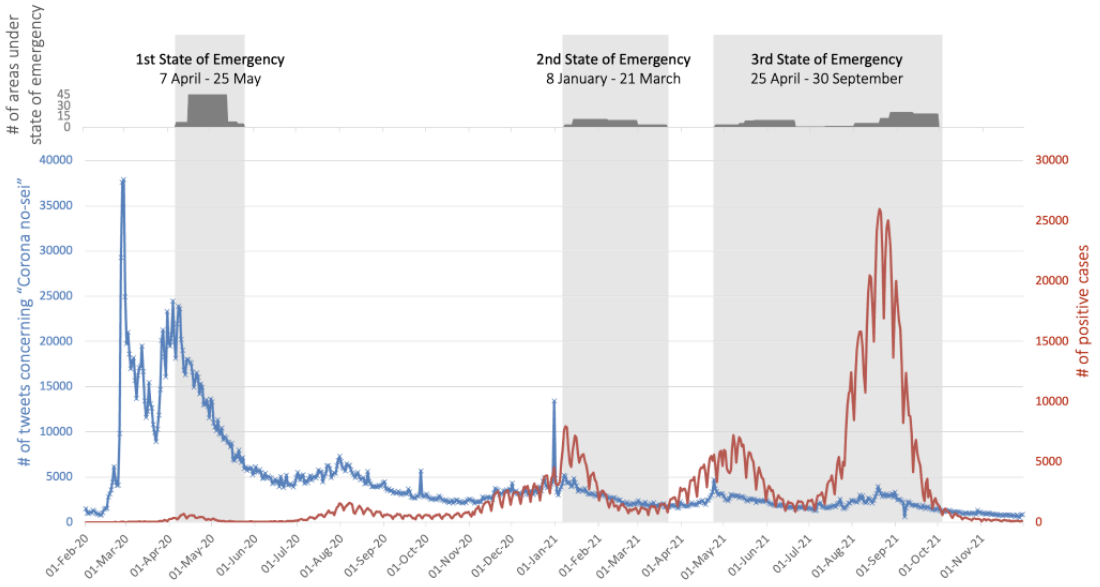

Figure 3.2: Trends in the number of "Corona no-sei" tweets and the number of corona-positive patients. The blue line indicates the number of "Corona no-sei" tweets, and the red line indicates the number of positive COVID-19 cases.

areas under states of emergency is indicated by the bar graph in the upper part of the figure because the target areas were changed during each state of emergency imposition. The periods in which the states of emergency were imposed roughly corresponded to the rise of COVID-19 case numbers. Interestingly, the announcements of states of emergency were very effective in suppressing case numbers. In the case of the spike caused by the Tokyo Olympics (which took place between July 23, 2021 and August 8, 2021), the case numbers quickly dropped to below 5000 per day within three months.

As indicated by the blue line, the "Corona no-sei" tweets peaked in March, 2020, roughly before the first state of emergency was announced and reached the second highest number at the time the first state of emergency was imposed. After the first announcement of a state of emergency, the number of "Corona no-sei" tweets showed a downward trend until the end of our data collection period. There were occasions of small increases of the "Corona no-sei" tweets before the second and third states of emergency announcements; however, in general, the number of reported plan disruptions never reached the level before the first state of emergency announcement. In comparison with COVID-19 case numbers, the number of "Corona no-sei" tweets showed a relative correlation with states of emergency announcements. Indeed, the scatter plot for case numbers and the numbers of "Corona no-sei" tweets is shown in Figure 3.3 with Pearson correlation coefficients of 0.86, 0.93 and 0.61 respectively for the first, second, and third states of emergency.

When compared against the number of "Corona no-sei" tweets during the entire period, the correlation between COVID-19 daily cases and "Corona no-sei" tweets was not very evident. We were able to observe a slight increase of "Corona no-sei" before the case numbers started rising, but the extent of case number increment was disproportional to that of the "Corona no-sei" tweets. Although the case numbers peaked in September, 2021 during the third state of emergency, there was only a slight increase in the number of "Corona no-sei" tweets, in comparison with the high amount of complains at the very beginning of the COVID-19 epidemic. This indicate the fact that Japanese residents might have adapted to the restrictions or disruptions caused by COVID-19 lockdown.

We, furthermore, examined the nouns and verbs in our sampled tweets. Table

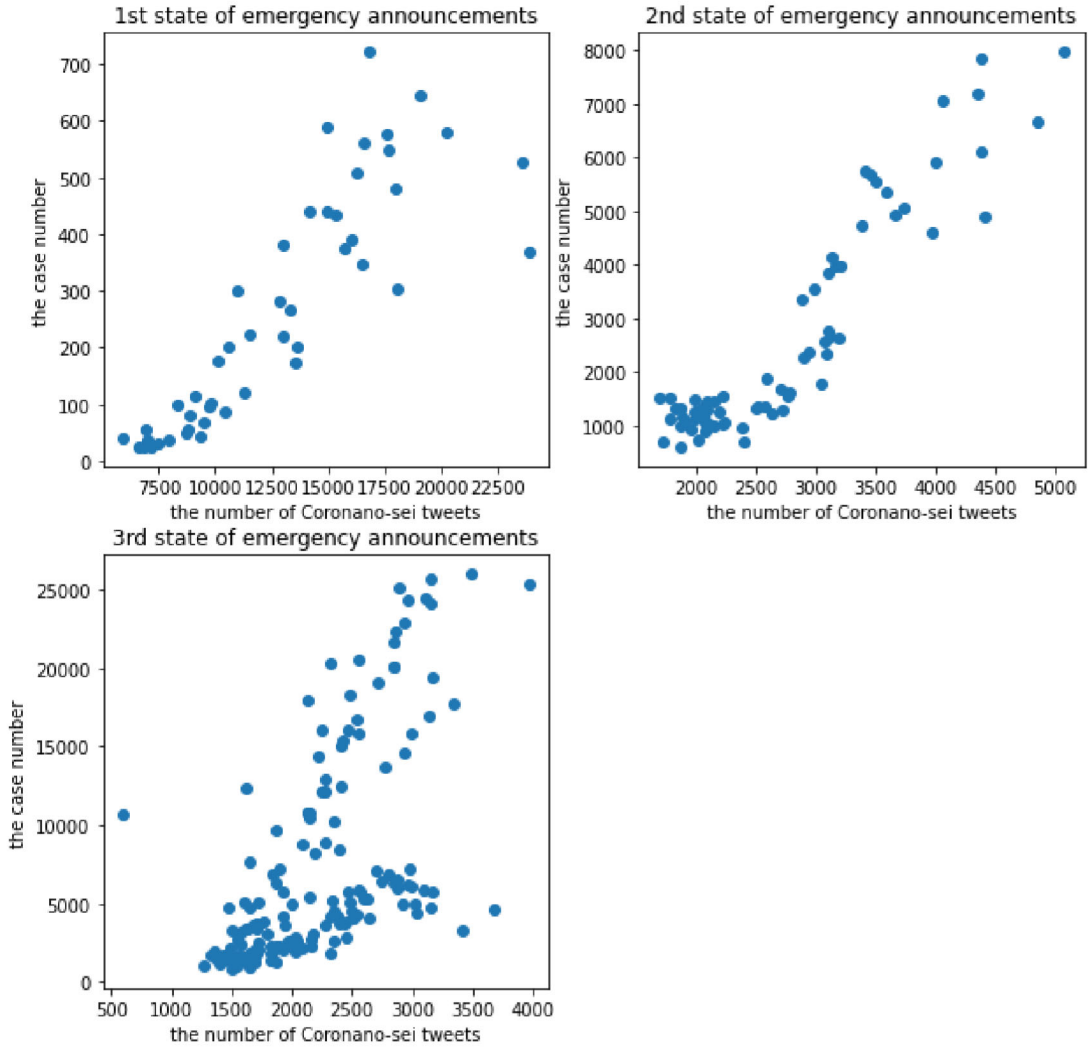

Figure 3.3: Scatter plot for COVID-19 case numbers, number of "Corona no-sei" tweets for 1st, 2nd and 3rd states of emergency announcements

3.1 and 3.2 shows the number of tweets of the top 20 nouns and verbs tweeted on February 28, 2020, when the tweet numbers reached their highest levels. Table 3.3 shows the top 20 words (nouns and verbs) that co-occurred with "Corona no-sei" tweets in descending order to highlight the most disrupted activities or plans during the period of our data aggregation. In considering nouns, "Corona" was excluded because it was included in the query and was evidently the most frequently detected word. For nouns, the top five most frequently mentioned words were "work," "abort," "home," "live," and "friends" after excluding the words that indicate the grammatical tense. These keywords indicated that over a longer period of time, Japanese citizens started developing concerns over their disrupted work and social life. For verbs, "go" was the most frequent; but in the actual tweets, it was sometimes used in the negative, and in the context, the verb was unlikely to be used in the affirmative. Therefore, the verb was likely used to indicate they "cannot go" even if it is in the affirmative in this paper (see Table 4.3). Hence, the top five means "go," "can go <negation>," "look," "meet <negation>," and "get out." The results show that there are restrictions on the actions of going, seeing, and meeting expressed in the verbs. In comparison with a single day result of February 28th 2020, the concern about work seemed to appear in The top five in both tables, indicating that Japanese citizens clearly emphasized their work routines. Moreover, the desire for live concerts also increased in the long run and placed "live concert" as the fourth most frequently mentioned keyword in Table 3.1. Coincidentally, concerns related to friends and the missing opportunities to meet up were also observed in both tables, showing the disruption of social relationships and recreational occasions. Both indicated that people regarded COVID-19 to be the main cause of their disrupted plans to hang out with friends or attend large public events. Besides activities, the keyword finding also reflected a concern about shortage of resources, such as toilet paper, masks and even money, which were considered to be critical in supporting daily lives or normal healthcare practices. We will discuss the implications of some key findings in the next chapter.

Table 3.1: The number of co-occurent noun on 28 February 2020.

Noun	Number of tweets
graduation ceremony	2154
cancel	1813
school	1498
work	1210
event	775
live concert	694
toilet paper	667
part-time job	639
school holiday	636
friends	616
Disney	605
Postponement	562
Home	553
Mask	531
New Corona	330
Test	328
Tissue	321
Hoax	318
Company	304
Next month	250

Table 3.2: The number of co-occurent verb on 28 February 2020.

Verb	Number of tweets
disappear	3265
go	1835
rest	1169
can go <negation>	1044
end	862
go out	754
look	734
work hard	711
buy	602
cry	576
vanish	556
can meet <negation>	524
play	505
spring rest	501
dies	499
be told	491
come	441
think	436
return	378
crush	319

Table 3.3: Words co-occurring with “Corona no-sei” in descending order.

Order	Noun	Verb
1	work	go
2	abort	can go <negation>
3	home	look
4	live concert	meet <negation>
5	friends	get out
6	event	lose
7	postponed	make effort
8	stress	increase
9	school	buy
10	part-time job	lose
11	company	end
12	new Corona	come
13	mask	think
14	graduation ceremony	meet
15	hospital	rest
16	one person	can go
17	opportunity	decrease
18	university	be told
19	family	meet
20	money	play

4 Proposals for the Identification of Needs and Issues

4.1 Proposals for the Extraction of the Medical Needs of Breast Cancer

In Chapter 2, we noted that this research could be effective for statistical surveys. In addition, there are other possible applications. In particular, we will examine the extraction of adverse drug events (signal detection) and the extraction of unmet needs. Here we blinded the proper noun because it is not relevant to extract the medical needs in this section.

4.1.1 Potential Application for Side Effect Signaling

This section shows the questions classified into category number 11 (extracted with high cosine similarity) to extract side-effect signaling. Extracting side effects from submitted questions could be very beneficial for pharmaceutical companies and patients because it allows them to collect significant information on drug safety. Specifically, because some questions classified under the overarching category of "symptoms, side effects, and sequelae" (category 11) of the CPC are considered to contain information on side effects, we can extract such information by applying intrinsic expression extraction to each question's text.

The first question

> I am undergoing treatment for breast cancer, and my white blood cell count has dropped due to side effects, so my immune system is not high. I don't want to go to the birthday party at my parents-in-law's because I'm worried that I might get infected with the coronavirus.

> My mother-in-law and father-in-law know that I am undergoing treatment and my immunity is low, but they don't want to cancel the party because it's their adorable grandchild's birthday. It's hard for me to tell them. I don't want my husband to go either, but he doesn't seem to mind at all. Is there any way to avoid attending the party?

contained information about a reduction in white blood cells count.

The second question

> Can I improve the numbness caused by the side effects of anticancer drug treatment? My sister is undergoing anticancer treatment for breast cancer, and she is suffering from numbness in her hands and feet. Is there anything she can do to relieve the numbness? Does she have to stop the anticancer treatment?

contained information about numbness in the hands and feet. However, we could not identify the drug that caused the side effect because there was no information about it. A third question below contained information about the side effects of hair loss due to fluorouracil, epirubicin, and cyclophosphamide (FEC) treatment, a type of chemotherapy.

However, the third question

> I am undergoing anticancer treatment, FEC treatment with infusions every 3 weeks, breast cancer. I have completed four courses, and I am about to start another one, and I have a question about hair loss. My hair still looks like a baby's, so I can say that I am losing hair. Although I have heard that other parts of my body, such as the eyelashes, eyebrows, shins, and lower hair, I am not losing other than my hair. My doctor said that you lose when I asked my doctor about it the second time. I'm worried that the medication might not be working correctly. If you have any experience with this or know anything about it, please advise me.

simply indicated that FEC treatment caused hair loss as side effect.

The fourth question

> I would like to know about mouth ulcers during anticancer treatment. I have breast cancer and will start anticancer treatment, but before that, I went to a dentist and was told that I should have my teeth treated. She told me that I would probably get many mouth ulcers from the anticancer treatment but that I should just go and see her. She told me that I should go in. If it's a common mouth ulcer, I'm sure they can treat it with ointment, but I'm not sure if the mouth ulcer will begin to heal before the anticancer drugs are finished? The side effects of the anticancer medicines make it hard to go to the dentist, and the thought of having to go stresses me out. If I can heal my mouth ulcers faster by going to the dentist, I'll do my best. However, if it doesn't make much difference, I don't want to push myself as much as possible because of the hair loss, fatigue, side effects, and other things. If I go to the hospital because of mouth ulcer, will it heal faster? If you have any experience or know of anyone who had mouth ulcers, please let me know.

contained information about stomatitis.

The fifth question

> My 66-year-old mother is undergoing anticancer treatment for lung adenocarcinoma. She takes Docetaxel plus Cyramza once every four weeks. She had numbness after the second dose and reduced the dosage for the third dose, but the numbness keeps getting worse. . . She's been taking the maximum daily dose of Lyrica to reduce the numbness, but she says it's not helping at all. She can't walk anymore, and it has become mentally painful for her, so I are hoping that I can alleviate her numbness. Can you tell me anything about how to deal with the numbness, herbal medicine, or anything else that might help reduce the numbness a bit? Thank you very much.

contained information about numbness.

For the above five questions, the code and name of categories classified by our model, drug name, and side effects are shown in Table 4.1.

Thus, although side effects can be extracted, the granularity of the drug information may be insufficient. Of the 6993 cases, 470 (6.7%) were classified under

Table 4.1: Questions that were classified into category 11

CPC code	CPC name	Drug name	Side effects
11.1.2.3.	Persistent side effects of anticancer drugs (other)	Anticancer drug for breast cancer	Leukopenia
11.1.1.2.	Nerve damage such as numbness and discomfort caused by anticancer drugs	Breast cancer drug	Numbness
11.1.1.1.	Hair loss due to anticancer drug treatment	FEC treatment	Hair loss
11.1.1.6.	Mucosal damage caused by anticancer drugs (stomatitis, etc)	Breast cancer anticancer drug	Stomatitis
11.1.1.2.	Nerve damage such as numbness and discomfort caused by anticancer drugs	Docetaxel + Thyramza	Numbness

category 11 using the D+E-based method, of which 100 (21.3%) cases were randomly sampled, and 15 (3.2%) had specific drug names.

4.1.2 Potential Application for Unmet Needs

In this section, we show some of the questions and their categorization into low-frequency categories, as well as "COVID" search to extract unmet needs.

Patients' unmet needs are becoming a major societal issue. In particular, the unmet needs of those who should provide answers have not yet been sufficiently addressed. Except for a few fee-based QA sites such as AskDocters[1] and Pearl[2], QA sites are generally answered by non-experts, whereas some questions should be answered by physicians.

Unmet needs are needs that are not addressed owing to a lack of services or resources, or that have never existed before. The former may be met by discussing high-frequency categories with medical workers, which may help identify needs that have been insufficiently addressed in the past, although many patients complain about them. The latter can be extracted by searching for low-frequency categories or words that have become popular in recent years (e.g., "COVID").

The question

> My mother has breast cancer with bone metastasis. I heard that bone metastasis has a high risk of fracture, so should I prevent her from driving a car in the future? Moreover, my 80-year-old grandmother is still driving. However, there are many accidents involving the elderly, and the risk of having an accident is probably higher than for younger people. If I assume the worst-case scenario, should I stop her from driving instead of saying, "It's a pity to take away her car?" If I ask her to quit driving, in what situation/venue should I tell her? Moreover, I have a driver's license, but I'm a driver on paper only. Should I go back to school to drive for my mother and grandmother when I go out with the family? I don't think I'll be able to drive on public roads since I have not driven for a long time...

[1] `https://www.askdoctors.jp/`
[2] `https://www.pearl.com/`

contained an unmet needs about car driving of a cancer patient with bone metastasis.

The question

> Please tell me if I can sue for cancer misdiagnosis. Two years ago, I went to a hospital because a retest was required by mammography. Since there was something suspicious on the echo, I had cytology done on the spot. The cytology didn't give me any results due to a bad specimen, so I asked for histology. The doctor told me that I would have to stay overnight at another hospital for a mammotome biopsy, etc. I didn't want to spend a lot of time figuring out what was black and white, so I had a surgical biopsy, a definitive diagnosis that could be done at that hospital. As a result, I was diagnosed with "breast adenopathy" and told to visit the hospital regularly. But some of the results of the tissue examination were not convincing, so I had the examination done at another hospital. The result was breast cancer... how could they remove it from the definitive diagnosis... I would be horrified if they were convinced it was mammary gland disease and discovered it too late. I want to sue the doctor who is still examining and treating me as usual, but I heard that medical lawsuit are difficult. Is it possible to sue him? Do I have a chance to win?

contained an unmet need about misdiagnosis lawsuit.

The question

> I had breast cancer sparing surgery in early February and will start radiation treatment in April. However, I am going through a tough time with corona right now, and I feel anxious about going to the hospital every day. Is there anything else I can do except taking personal measures?

contained the unmet need about COVID-19 infection.

The question

> I had a breast cancer sparing surgery in February this year and was scheduled for radiation therapy, but it has been postponed due to the

> coronavirus. It will still take some time for the situation to improve, but should I avoid starting radiation therapy at this time? I am on hormone therapy, but I am getting anxious about not undergoing radiation therapy.

contained an unmet need about postponement of COVID-19 treatment.

The question

> My 88-year-old mother is in a special care facility and has a fever of 37.5. She has breast cancer, so I don't know if the fever is caused by breast cancer, corona, or a cold. What are the symptoms of a fever caused by breast cancer? Do I need to see my family doctor? If it is not caused by breast cancer, does the fact that I have a high fever in a special care facility mean that I have contracted the virus from a staff member?

contained an unmet need about inability to distinguish between cancer and COVID-19.

The question

> About 18 years ago, my mother was diagnosed with breast cancer. She had an operation and has been living a normal and healthy life since then. However, 2 years ago, she was told that the cancer had spread to her lungs. At present, she has difficulty breathing even when she moves a little, probably due to the accumulation of pleural effusion. When she was told that the cancer had spread, the doctor did not give her a life expectancy, but when she looked it up on the internet, she found all sorts of information that made her feel uneasy. Can you tell me whether she will live much longer or whether she may be able to live longer while coping with her illness? I'm getting married soon, and I was planning to show her my wedding dress next year. However, with the corona epidemic, that plan is now undecided. I want to show her my wedding dress at least. I'm not sure if this is practically possible.

contained an unmet need about wedding.

The question

> When I distrusted the female surgeon at the [omitted]b Hospital and applied for a second opinion (a letter of introduction was required), I was pressured to go to the hospital for a second opinion. The doctor there is a surgeon famous for his breast-conservation therapy, but he didn't listen to me very carefully and told me that he agreed with Dr. [omitted] (the doctor in charge at [omitted] Hospital) and that I should tell her that he agreed with her because doctors have a difficult relationship with each other. Is there such a thing? The book on breast cancer published by the [omitted] Hospital, famous for cancer treatment, claims it to be the "standard treatment," even though the treatment policy is different. Is there anyone who was notified that they had cancer and went for a second opinion and then were offered a different treatment plan? Do doctors always protect their doctors? I was amazed at the lecturers' pride in the national university hospital (even though they are quacks).

contained an unmet need about a second medical opinion.

In Table 4.2, we included the code and name of categories classified by our model and unmet needs that could be read from the text of the questions.

In this study, unmet needs were extracted by reading a questionnaire; however, constructing an automatic classification model for unmet needs is reserved as a future task.

4.2 Proposals for Social Needs of COVID-19

The previous section indicated that the most disrupted daily routines due to the COVID-19 pandemic in Japan were associated with work, education, social activities, and material shortages. The keyword findings from this study aligned with many other studies in different countries that had shown how COVID-19 had a wide impact on social life, the economy, public mental health and education [20]. Below we discuss key findings across the time span in four aspects: work routine disruption, public anxiety in reaction to (perceived) shortage of resources, social relationship concerns and interference with curricula.

Table 4.2: Questions and their classification categories considered as unmet needs.

CPC code	CPC name	Unmet needs
8.2.1.1.	Traffic conditions are bad	Driving a car with a displaced bone cancer patient
5.3.1.3.	Should I get a second opinion?	Lawsuits against misdiagnosis
8.2.1.3.	Frequent visits to the hospital are difficult	Worried about corona infection due to hospital
11.1.3.6.	Symptoms of radiation-related side effects (other)	Treatment postponed due to coronavirus
11.2.1.5.	Fever	I can't tell if it's cancer symptoms or corona symptoms
12.1.1.1.	Anxiety about the possibility of recurrence or metastasis	I want my mother to see me in my wedding dress, but she has been diagnosed with cancer
5.3.1.6.	Matters related to second opinions (other)	Not fulfilling the role of a second opinion

4.2.1 Top Concern

Work routine disruption rose to become the top concern for Japanese citizens. The labor market in Japan was undoubtedly severely challenged by COVID-19 as in many other countries. While the pandemic forced a change of work style and the introduction of remote collaboration, some types of Japanese workers (based on their employment contracts) appeared to be relatively vulnerable to the change in work style. The keyword "work" coexisted with "part-time," "abort," and "money" in our finding, suggesting that individuals who were concerned about work conditions could be struggling with job uncertainty due to the nature of their work being part-time, or the sudden drop in income. This finding was complementary to previous studies on the impact of COVID-19 on Japan's labor industry. As Kikuchi et al. described in their paper, contingent workers, female and low income workers were likely to be hurt the most owing to the change to telework and the uncertainty regarding long-term income during COVID-19 [49]. Fukai et al. also reported similar results in their large-scale government statistics analysis. Japanese citizens who worked part-time in the service industries, were forced to take a leave, or lost their job owing to the declaration of a state of emergency, were among the groups that were impacted most by COVID-19 [50]. While the introduction of part-time or contingent workers was a normal measure for Japanese companies to efficiently allocate their budget and resources, as work quickly rose to become the top concerning keyword when COVID-19 landed in Japan, researchers have warned that inequality could be exacerbated for vulnerable citizens without the government's active support [51]. Overall, our findings helped provide evidence for Japanese workers individual concerns about job disruption, employment disparity and lack of financial resilience. By failing to address these issues when announcing multiple states of emergencies, the Japanese government might end up severely compromising the equality of Japan's labor market.

4.2.2 Anxiety About Material Shortage

The shortage of some items, such as toilet paper, mask, and tissue as listed in Table 3.1 became a problem in Japan. Our findings was highly aligned with the

previous Twitter studies on the hoarding behavior over toilet paper [52]. While first observed in the United States, panic buying of all household goods quickly spread around the world. Among them, toilet paper was often the signaling product to be hoarded and stocked during a natural disaster [53, 54]. Although toilet paper stockpiling might seem irrational and has widely been ridiculed on social media, the harm caused by bulk purchasing was not as devastating and might be regarded by social scientists as a way of coping with natural disasters [54]. I n comparison to the frequently perceived over-hoarding of toilet paper, mask shortage was considered a more severe public health crisis and a direct threat to health and well-being. An agent-based simulation conducted by Tatapudi et al. in 2020 demonstrated that a universal use of masks could reduce infection by 20% [55]. At the time the study was conducted, the COVID-19 death toll across the world was 541 million, meaning that 108 million deaths could be spared if the universal use of masks could be implemented. Indeed, plenty of studies suggested that masks usage was negatively correlated to COVID-19 infection rate [56, 57]. In Japan, the situation was slightly different. The Japanese government received wide criticism about their slow reaction and realization of mask shortage since the pandemic was considered relatively "under control" in its early phase. As the mask crisis arose, many Japanese citizens were alarmed at their over-dependence on masks manufactured in foreign countries, prompting the government to take steps to boost mask production within the country. However, the over-promoted anxiety also resulted in what was commonly called "Abenomask," an incident that slammed the Japanese government for stockpiling over 82 million unused masks [58]. A tough lesson learnt from such an incident was that, while social media served as a critical channel to distribute news and raise public awareness, emotional contagious statements and over-promotion of a certain disaster could instead backfire, hampering the rational coping mechanism of citizens and the decision making of the government. As our findings and many other studies have highlighted, more work would need to be done to develop effective protocols to react to the widely contagious anxiety caused by sharing information about natural disasters on social media.

4.2.3 About Social Relationships

Keywords related to relationships, social life, and collective events were frequently seen in our analysis. For instance, the top 20 frequent nouns associated with "Corona no-sei" included friends, family, live, events, and one person. The top 20 frequent verbs associated with "Corona no-sei" were "go," "can go <negation>," "meet <negation>," "buy," "meet," "can go," and "play." The example in Table 4.3 indicates how Japanese citizens associated "go" and "meet" to their socializing events. While it might seem that many tweets exhibit concerns about social relationships, those keywords showed how people vented their frustration of being unable to meet and conduct activities together rather than indicating the loss of relationships. Interestingly, a study conducted by Goodwin and Takahashi [59] also reflected similar findings. The majority of Japanese respondents to their survey regarding their perception of their relationship quality during COVID-19 indicated that there had been no noticeable changes in their perceived relationship quality, with only some reporting that their trust and relationship with communities declined when compared to the pre-pandemic era. There was also a report on students having less communication with friends, which became a risk factor for mental health problems [60]. These findings suggested that COVID-19, or similar pandemic events, might have caused individuals to experience higher anxiety and stress, and such an emotional response could result in a short term disruption to their social activities and coping mechanism against the trauma, but would not influence long term perceived relationship quality. Indeed, our example tweets showed how individuals were able to accept the fact that despite feeling frustrated, they looked forward to resuming their social activities post-pandemic. We, therefore, argue that concerns over relationship disruption might be temporary and instead was a positive signal for individuals in Japan to actively maintain their relationships and, as the study of Goodwin and Takahashi [59] suggested, in the case of romantic relationships, making more time for communication could further enhanced the quality of the relationship.

Table 4.3: Tweets examples.

Verb	Example
go	Due to Corona, the day I've been looking forward to going out with the guy I love has been postponed... I can't help it now and I'll accept it, but I was looking forward to it.
meet	It doesn't feel like April at all due to Corona, but I can't wait for it to end so that we can all meet, eat, and shop together comfortably. Six years already... I want to quit my job lol.

4.2.4 Concern for Education Discontinuation

The highest number of tweets occurred on February 28, 2020, when the government announced the simultaneous closure of all elementary, junior high, and senior high schools in Japan. Indeed, in the most frequent nouns and verbs shown in Table 3.1 and 3.2, the top words related to the simultaneous closure of schools were "graduation ceremony," "cancel," "lose," "rest," and "go <negation>," all of which reflected Japanese citizens' worries of their education discontinuing, the cancellation of the graduation ceremony and missing school classes. Note that the graduation ceremony is held in March and the new school/work year starts from April in Japan. Arguably, despite the Japanese government's desperate resort to curb COVID-19, as scientists questioned it, the closure of schools in Japan did not yield significant effect in preventing COVID-19 from spreading, but instead deprived learning and development opportunities for children [61]. Furthermore, due to the schools closure, the demand for digital education or a virtual learning platform soared, yet many schools and student households were severely underprepared for this makeshift education system. As Iwabuchi et al. [62] discussed in their in-depth analysis, the different resources allocated to each school in Japan further exacerbated the digital learning disparity caused by COVID-19 school closure. The well-funded private and prefecture-sponsored schools often already implemented or could quickly set up the necessary e-learning systems to cope with the lack of face-to-face lecturing. However, the majority of public schools were forced to send learning materials to students by mail, risking a huge learning gap between students in private and public schools. It is still unknown what kind

of impact to students' physical or mental development can be observed over the longer term because most of the schools were able to catch up after the lifting of the state of emergency. A study conducted by Nishimura et al. [63] on medical students clearly showed worsening subjective mental well-being and growing concerns about online alternatives failing to replace the much needed in-person learning and field practice in medical education. The various concerns reflected in education and associated keywords in both Tables 3.1 and 4.3 also indicate that most Japanese citizens shifted their focus from one-time events, such as "graduation ceremony" and "school holiday" to more long term mental and societal impacts, such as "opportunity," "stress," and "university." This shift implied that such long term impact would take time to emerge in comparison with the short term disruption of incidents (e.g., graduation ceremony), and more studies are required in the future to monitor and uncover the full picture of the disruption.

5 Conclusions

In this chapter, we provide an overall summary of our study, including its limitations, about the medical needs of breast cancer patients and social needs during the COVID-19 pandemic.

This thesis challenged to integrate informatics and sociology by approaching sociological issues using informatics technology to create new services using social media data such as Twitter and QA data. Although here are several academic disciplines that use data to study the science of society, such as computational social science and social informatics, our study differs from these studies in that our goal is to reach for developing real-world applications or services by utilizing research results and discoveries in social computing.

In this thesis, we approached sociological issues using informatics technology to extract the medical issues of breast cancer patients and social issues during COVID-19.

5.1 Medical Needs of Breast Cancer

We created three corpora and questions posted to the Yahoo Q&A service (YJQA) were classified into categories of the cancer problem classification (CPC). From the classified results, we showed potential applications for side effect signaling and unmet needs.

Because we used the Japanese text of questions found by the search phrase "breast cancer" as the training data of the method used in this study, the method may not apply to other cancer types and other countries. However, our method can be expanded to different cancer types and countries in cases where problem data are available. Here, the cancer problem categories specific for other countries are required because the ones used in this study were defined for Japanese citizens.

When expanding our method to other cancer types and countries, future work will have to focus on reproducibility. Therefore, it is necessary to reconstruct the training data from the questions found by searching for each cancer word to apply the method to other cancer types.

In addition, COVID-19 infections in Japan appeared in February 2020, and patients with cancer might experience COVID-19-related problems. Therefore, it is possible that the current CPC categories may not ensure proper classification. Thus, it is necessary to define new problem classification categories for patients with cancer after February 2020. Furthermore, because new topics, not limited to COVID-19, are always likely to occur, it is necessary to construct a model that could extract such uncommon topics.

The target of this study was question texts posted on QA services, and it may not be possible to classify other texts correctly. The fact that the accuracy of the D-based method was extremely poor among the 3 methods may be due to the difference between the questionnaire text used in the CPC and the text posted on the YJQA service. We also found that cancer patients' problems are not limited to questions posted on QA websites but include Twitter and blogs. It is necessary to broaden the training data of the classification method for these texts to classify the worries of cancer patients. In addition, there are many posts in which the content is unrelated to worries or contains too many emojis. Therefore, it is necessary to build a model to determine whether a post contains worries. Subsequently, two schemes are required to classify the blogs containing worries into CPC categories.

5.2 Social Needs for COVID-19

Overall, by adding the analysis on "Corona no-sei" to the conventional symptom-based monitoring, we were able to identify the underlying concerns at the peak of the disruption and across the whole time span of the three announcements of states of emergency. Our findings and comparison of the tweets against the COVID-19 case numbers yielded rich insights on people's short- and long-term concerns and potential societal impacts caused by the announcements of states of emergency. This analysis is expected to be useful for faster decision making

because it can produce results in a timely manner.

In addition, it can be extended to multiple languages by converting the query to other language. However, since the system in this thesis is designed to extract data from Japan, it is necessary to use the Twitter API to obtain data from overseas. Note that in such cases, the results will be limited because the API could not extract all tweet data [1].

It should be noted that bias exists on Twitter because the usage rate of Twitter among the elderly is lower than that among the young. To reduce such bias, it is necessary to remove the effect of age by stratified analysis. However, this system cannot obtain age data. Therefore, the results should be interpreted in consideration of the fact that the opinions of the elderly are limited.

Although more studies from different fields would help to reveal the whole landscape of social and psychological impact caused by COVID-19, we believe that the keywords reflected in "Corona no-sei" tweets provided more nuanced descriptions of real-life problems that Japanese citizens faced during the COVID-19 and revealed the development of different concerns in response to the change of policies. For policy makers, particularly the Japanese government, such a study reflects the voice of the citizens and should be taken into consideration when reflecting on the effect and suitability of a policy and assessing further measures in supporting persons that were impacted during the pandemic.

5.3 Future Work

Although medical and social needs alone do not solve people's issues, a comprehensive understanding of the type and number of needs can help prioritize services to solve issues from the people's point of view. It is necessary to examine the services that could be provided in the future based on this information to solve medical and social issues.

[1] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

Acknowledgements

First of all, I would like to thank Professor Eiji Aramaki for providing with me the opportunity to study medical informatics, guiding me in my research, and providing valuable comments.

I thank Professor Yoshinobu Sato for being on my thesis committee and providing valuable comments.

I equally wish to thank Associate Professor Shoko Wakamiya, Assistant Professor Shuntaro Yada, and Assistant Professor Kongmeng Liew for guiding me in my research and providing valuable comments.

I thank all the students and staff of the social computing departmental laboratory.

Last but not the least, I would like to thank my dear wife for her understanding my study in college for my doctorate program.

Bibliography

[1] MedicationLife. Japan's healthcare issues - how doctors work. https://medionlife.jp/article3/ [accessed on November 30th, 2022].

[2] United Nations. Transforming our world: the 2030 agenda for sustainable development. https://www.mofa.go.jp/mofaj/files/000101401.pdf [accessed on November 30th, 2022].

[3] Patrick C Laurent A Nancy AD Elodie BA, Robert R. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiol Drug Saf*, 29, 2019.

[4] Wei H Richard F Tamar L, Ann WM. Methodologic approaches in studies using real-world data (rwd) to measure pediatric safety and effectiveness of vaccines administered to pregnant women: A scoping review. *Vaccine*, 39, 2021.

[5] Yasuhiko M Katsutoshi H, Annabel B. Current status, challenges, and future perspectives of real-world data and real-world evidence in japan. *Drugs – Real World Outcome*, 8, 2021.

[6] Research Group on the Sociology of Cancer. Voices of 7,885 people who have faced cancer (Report on the actual situation regarding the worries and burdens of cancer survivors, summary version). https://www.scchr.jp/cms/wp-content/uploads/2015/11/taiken_koe_jpn.pdf [accessed on November 30th, 2022], 2003.

[7] Research Group on the Sociology of Cancer. Voices of 4,054 people who faced cancer (Report on the actual situation survey on worries and burdens of cancer survivors). https://www.scchr.jp/cms/wp-content/uploads/2016/07/2013taikenkoe.pdf [accessed on November 30th, 2022], 2013.

[8] Research Group on the Sociology of Cancer. Voices of 1,275 people who have faced breast cancer (Report of a survey on the worries and burdens of cancer survivors). https://www.scchr.jp/cms/wp-content/uploads/2016/11/2013nyugan_taikenkoe.pdf [accessed on November 30th, 2022], 2013.

[9] Yom-Tov E Rosenblum S. Seeking web-based information about attention deficit hyperactivity disorder: Where, what, and when. *J Med Internet Res*, 19, 2017.

[10] Chen Z Oh S Bian J Park MS, He Z. Seeking web-based information about attention deficit hyperactivity disorder: Where, what, and when. *JMIR Med Inform*, 4, 2016.

[11] Gabrilovich E Yom-Tov E. Postmarket drug surveillance without trial costs: Discovery of adverse drug reactions through large-scale analysis of web search queries. *Med Internet Res*, 15, 2013.

[12] Tanaka A Narimatsu H Tsuya A, Sugawara Y. Do cancer patients tweet? examining the twitter use of cancer patients in japan. *J Med Internet Res*, 16, 2014.

[13] Evans R Wu H Hong Z, Deng Z. Patient questions and physician responses in a chinese health q&a website: Content analysis. *J Med Internet Res*, 22, 2020.

[14] Vigneux M Abetz L Arnould B Bayliss M et al Lasch KE, Marquis P. Pro development: rigorous qualitative research as the crucial foundation. *Qual Life Res*, 19:1087–1096.

[15] Black N. Patient reported outcome measures could help transform healthcare. *BMJ*, 2013.

[16] Maria Nicola, Zaid Alsafi, Catrin Sohrabi, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, Maliha Agha, and Riaz Agha. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International journal of surgery*, 78:185–193, 2020.

[17] Joseph A Lewnard and Nathan C Lo. Scientific and ethical basis for social-distancing interventions against COVID-19. *The Lancet infectious diseases*, 20(6):631–633, 2020.

[18] Meirui Qian and Jianli Jiang. COVID-19 and social distancing. *Journal of Public Health*, pages 1–3, 2020.

[19] Brendon Sen-Crowe, Mark McKenney, and Adel Elkbuli. Social distancing during the COVID-19 pandemic: Staying home save lives. *The American journal of emergency medicine*, 38(7):1519, 2020.

[20] Michèle Belot, Syngjoo Choi, Julian C Jamison, Nicholas W Papageorge, Egon Tripodi, and Eline Van den Broek-Altenburg. Six-country survey on COVID-19. *IZA Discussion paper*, 2020.

[21] LA. Andersen, TE. Hansen, N. Johannesen, and A Sheridan. Pandemic, Shutdown and Consumer Spending: Lessons from Scandinavian Policy Responses to COVID-19. *arXiv*, 2020.

[22] D. Alexander and E Karger. Do stay-at-home orders cause people to stay at home? Effects of stay-at-home orders on consumer behavior. *FRB of Chicago Working Paper*, 12:317–330, 2020.

[23] Avi J Hakim, Kerton R Victory, Jennifer R Chevinsky, Marisa A Hast, D Weikum, L Kazazian, S Mirza, R Bhatkoti, MM Schmitz, M Lynch, et al. Mitigation policies, community mobility, and COVID-19 case counts in Australia, Japan, Hong Kong, and Singapore. *Public Health*, 194:238–244, 2021.

[24] Toshikazu Kuniya. Evaluation of the effect of the state of emergency for the first wave of COVID-19 in Japan. *Infectious Disease Modelling*, 5:580–587, 2020.

[25] Kenji Karako, Peipei Song, Yu Chen, Wei Tang, and Norihiro Kokudo. Overview of the characteristics of and responses to the three waves of COVID-19 in Japan during 2020-2021. *Bioscience trends*, 2021.

[26] T Watanabe and T Yabu. Predicting intervention effect for COVID-19 in Japan: state space modeling approach. *PLOS ONE*, 20:317–330, 2021.

[27] Eiji Yamamura and Yoshiro Tsutsui. How does the impact of the COVID-19 state of emergency change? An analysis of preventive behaviors and mental health using panel data in Japan. *Journal of the Japanese and international economies*, 64:101194, 2022.

[28] K Katafuchi, Y. Kurita and S Managi. COVID-19 with Stigma: Theory and Evidence from Mobility Data. *Economics of Disasters and Climate Change*, 5:71–95, 2021.

[29] T Watanabe and T Yabu. Japan's voluntary lockdown. *PLOS ONE*, 16, 2021.

[30] T Watanabe and T Yabu. Japan's voluntary lockdown: further evidence based on age-specific mobile location data. *The Japanese Economic Review*, 72:333–370, 2021.

[31] T Mizuno, T. Ohnishi and T Watanabe. Visualizing Social and Behavior Change due to the Outbreak of COVID-19 Using Mobile Phone Location Data. *New Generation Computing*, 39:453–468, 2021.

[32] Swaroop Gowdra Shanthakumar, Anand Seetharam, and Arti Ramesh. Understanding the societal disruption due to covid-19 via user tweets. In *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 137–144, 2021.

[33] Hocheol Lee, Eun Bi Noh, Sea Hwan Choi, Bo Zhao, and Eun Woo Nam. Determining public opinion of the COVID-19 pandemic in South Korea and Japan: social network mining on twitter. *Healthcare informatics research*, 26(4):335–343, 2020.

[34] Lynette Hammond Gerido Ying Cheng Ya Chen Lizhu Sun Hongru Lu, Juan Xie. Information needs of breast cancer patients: Theory-generating meta-synthesis. *Journal of Medical Internet Research*, 22, 2020.

[35] Anne McQueen Cristina García Vivar. Informational and emotional needs of long-term survivors of breast cancer. *Leading Global Nursing Research*, 51, 2005.

[36] Dominic Lunn Raja Sawhney Kelly Eu Rhea Liang Dean Vuksanovic, Jasotha Sanmugarajah. Unmet needs in breast cancer survivors are common, and multidisciplinary care is underutilised: the survivorship needs assessment project. *Breast Cancer*, 28, 2021.

[37] Jerome F Trevor H, Robert T. *The elements of statistical learning.* New York: Springer-Verlag. 2009.

[38] Nicole A. Lazara Ronald L. Wassersteina. The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, 70, 2016.

[39] Kevin P. M. *In: Machine Learning A Probabilistic Perspective.* Cambridge, Massachusetts. MIT Press, 2012.

[40] Garving K Luli Joanne Chen Lyu, Eileen Le Han. Covid-19 vaccine-related discussion on twitter: Topic modeling and sentiment analysis. *Journal of Medical Internet Research*, 23, 2021.

[41] Hafeez Ul Hassan Virk Harish Narasimhan Joshua Hahn Zhen Wang W H Wilson Tang Chayakrit Krittanawong, Bharat Narasimhan. Misinformation dissemination in twitter in the covid-19 era. *The American Journal of Medicine*, 133, 2020.

[42] Eiji Aramaki Shoko Wakamiya Kiki Ferawati, Kongmeng Liew. Monitoring mentions of covid-19 vaccine side effects on japanese and indonesian twitter: Infodemiological study. *Journal of Medical Internet Research Infodemiology*, 2, 2022.

[43] Nobuyuki Shimizu Kongmeng Liew Taichi Murayama Shuntaro Yada Shoko Wakamiya Eiji Aramaki Zhiwei Gao, Sumio Fujita. Measuring public concern about covid-19 in japanese internet users through search queries: Infodemiological study. *Journal of Medical Internet Research Public Health Surveill*, 7, 2021.

[44] Nobuyuki Shimizu Kongmeng Liew Shoko Wakamiya Eiji Aramaki Makoto Uehara, Sumio Fujita. Measuring concerns about the covid-19 vaccine among japanese internet users through search queries. *Scientific Reports*, 12, 2022.

[45] Wakako Kashino Noriko Kando Tetsuya Ishida, Yohei Seki. Extracting citizen feedback from social media by appraisal opinion type viewpoint. *Natural Language Processing (Japanese Paper)*, 29, 2022.

[46] N. Yoshinaga and M Kitsuregawa. Polynomial to linear: Efficient classification with conjunctive features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1542–1551, 2009.

[47] N. Yoshinaga and M Kitsuregawa. Kernel slicing: Scalable online training with conjunctive features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1245–1253, 2010.

[48] N. Yoshinaga and M Kitsuregawa. A self-adaptive classifier for effi-cient text-stream processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1091–1102, August 2014.

[49] Shinnosuke Kikuchi, Sagiri Kitao, and Minamo Mikoshiba. Who suffers from the COVID-19 shocks? Labor market heterogeneity and welfare consequences in Japan. *Journal of the Japanese and International Economies*, 59:101117, 2021.

[50] Taiyo Fukai, Hidehiko Ichimura, and Keisuke Kawata. Describing the impacts of COVID-19 on the labor market in Japan until June 2020. *The Japanese Economic Review*, 72(3):439–470, 2021.

[51] Shinnosuke Kikuchi, Sagiri Kitao, and Minamo Mikoshiba. Heterogeneous Vulnerability to the COVID-19 Crisis and Implications for Inequality in Japan. *Research Institute of Economy, Trade, and Industry (RIETI)*, 2020.

[52] Janni Leung, Jack Yiu Chak Chung, Calvert Tisdale, Vivian Chiu, Carmen CW Lim, and Gary Chan. Anxiety and panic buying behaviour during covid-19 pandemic—a qualitative analysis of toilet paper hoarding contents on twitter. *International journal of environmental research and public health*, 18(3):1127, 2021.

[53] Muneo Kaigo. Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake. *Keio Communication Review*, 34(1):19–35, 2012.

[54] Colleen P Kirk and Laura S Rifkin. I'll trade you diamonds for toilet paper: Consumer reacting, coping and adapting behaviors in the COVID-19 pandemic. *Journal of business research*, 117:124–131, 2020.

[55] Hanisha Tatapudi, Rachita Das, and Tapas K Das. Impact assessment of full and partial stay-at-home orders, face mask usage, and contact tracing: An agent-based simulation study of COVID-19 for an urban region. *Global epidemiology*, 2:100036, 2020.

[56] Adam Catching, Sara Capponi, Ming Te Yeh, Simone Bianco, and Raul Andino. Examining the interplay between face mask usage, asymptomatic transmission, and social distancing on the spread of covid-19. *Scientific reports*, 11(1):1–11, 2021.

[57] Dhaval Adjodah, Karthik Dinakar, Matteo Chinazzi, Samuel P Fraiberger, Alex Pentland, Samantha Bates, Kyle Staller, Alessandro Vespignani, and Deepak L Bhatt. Association between covid-19 outcomes and mask mandates, adherence, and attitudes. *PLOS ONE*, 16(6):e0252315, 2021.

[58] Tomohiro Osaki. Abenomask? Prime minister's 'two masks per household' policy spawns memes on social media. `https://www.japantimes.co.jp/news/2020/04/02/national/abe-two-masks-social-media/` [accessed on November 30th, 2022], 2020.

[59] Robin Goodwin and Masahito Takahashi. Anxiety, past trauma and changes in relationships in Japan during COVID-19. *Journal of psychiatric research*, 151:377–381, 2022.

[60] Masatoshi Tahara, Yuki Mashizume, and Kayoko Takahashi. Mental health crisis and stress coping among healthcare college students momentarily displaced from their campus community because of COVID-19 restrictions in Japan. *International Journal of Environmental Research and Public Health*, 18(14):7245, 2021.

[61] Kentaro Fukumoto, Charles T McClean, and Kuninori Nakagawa. No causal effect of school closures in Japan on the spread of COVID-19 in spring 2020. *Nature medicine*, 27(12):2111–2119, 2021.

[62] Kazuaki Iwabuchi, Kouki Hodama, Yutaka Onishi, Shota Miyazaki, Sae Nakae, and Kan Hiroshi Suzuki. Covid-19 and Education on the Front Lines in Japan: What Caused Learning Disparities and How Did the Government and Schools Take Initiative? In *Primary and Secondary Education During Covid-19*, pages 125–151. Springer, Cham, 2022.

[63] Yoshito Nishimura, Kanako Ochi, Kazuki Tokumasu, Mikako Obika, Hideharu Hagiya, Hitomi Kataoka, Fumio Otsuka, et al. Impact of the COVID-19 pandemic on the psychological distress of medical students in Japan: cross-sectional survey study. *Journal of Medical Internet Research*, 23(2):e25232, 2021.

List of Publications

Journals

[1] Masaru Kamba, Masae Manabe, Shoko Wakamiya, Shuntaro Yada, Eiji Aramaki, Satomi Odani, Isao Miyashiro, Medical Needs Extraction for Breast Cancer Patients from Question and Answer Services: Natural Language Processing-Based Approach, JMIR Cancer, Vol 7, No 4.

Domestic Conference

[1] Masaru Kamba, Masae Manabe, Shoko Wakamiya, Eiji Aramaki, Medical Needs for Breast Cancer Patients from WebQA Site Using NLP. In Proceedings of 23th Japan Association for Medical Informatics. 2020.