Text classification for breast cancer patient identity from Weibo posts

Zhouqing ZHANG†, Kongmeng LIEW†, Wan-Jou SHE†, Shuntaro YADA†, Shoko WAKAMIYA†, and Eiji ARAMAKI†

† Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma, Nara 630–0192, Japan
E-mail: †{zhang.zhouqing.yx6,liew.kongmeng,she.wanjou,s-yada,wakamiya,aramaki}@is.naist.jp

Abstract As one of the most common cancers among female patients, breast cancer patients find social media a handy platform to share their feelings after diagnosis, disease episodes, and experiences of coping with cancer. Furthermore, some close others of the breast cancer patients also often utilize the platform to share their experiences related to breast cancer. In this study, we aim to identify social media users who tend to initiate or engage in breast cancer discussions. To do so, we applied a transformer-based language model from Chinese Weibo posts to identify the author's relationship with the breast cancer patient. We have identified six categories according to the types of relationship with the cancer patients: post_user, family_members, friends_relatives, acquaintances, heard_relation, and no_patient. By classifying different patient in the post, researchers can develop a more in-depth understanding on individuals' opinions and emotional investment about breast cancer discussion. For instance, when we research on breast cancer patients themselves, which can allow us to know and meet patients' needs.

Key words Text classification, machine learning, natural language processing, breast cancer, social media

1 Introduction

In the modern era, breast cancer is a globally prevalent disease. With more than 2 million people diagnosed with breast cancer around the world in 2020 (statistics releases by World Health Organization (WHO)), breast cancer is a familiar disease to almost all populations worldwide. Breast cancer is a chronic disease with a high mortality rate, which poses a serious threat to human life [1]. For this reason, people often have a negative view about breast cancer, such as sadness, fears, depression [2]. In recent decades, the number of newly diagnosed patients continues to increase each year, despite continuous improvements in medical technology worldwide [1]. In China, breast cancer is a major disease that plagues common people, more than 400,000 people were diagnosed with breast cancer, and more than 100,000 people died from breast cancer in 2020 [注1]. A large number of new confirmed cases emerge every year; behind this, there are numerous stories intertwined with breast cancer [3]. Almost every female individual can be at risk of developing breast cancer, and they can be our family members, friends, relatives, neighbors, school fellows, or celebrities. As a result, breast cancer-related topics have become an inseparable part in our everyday social lives. Discussion on breast cancer is to express different opinions by the powerful storytelling way [4, 5]. Such materials are extremely valuable to focus on.

Nowadays, social media is an indispensable part in our daily lives. Almost everyone is using at least one type of social media platforms [6]. People share their life events frequently on social media. As a social tool, it can smoothly interact and communicate with their friends and family in time [7, 8]. Sina Weibo[注2] is one of the most useful and popular social platforms in China, it is a Chinese counterpart of Twitter [9]. The number of Weibo's monthly active users reached 511 million by 2020. In other words, Weibo is known by almost everyone in China, according to the statistics released by Sina Corp. [10]. People can discuss all kinds of topics on Weibo, and this certainly including the topic of breast cancer. With the larger number of users developing Weibo, the large number of people discussing breast cancer-related topics on Weibo, correspondingly, the large amount of data generated makes Weibo data a valuable corpus for research.

In recent years, machine learning is a hot research branch in computer science. Researchers focusing on machine learning have opened up a lot of outstanding work, no matter whether it is in computer vision or natural language processing (NLP). In NLP, machine learning has shifted from

[注1]：https://www.who.int/

[注2]：https://weibo.com

conventional neural networks to a transformer-based framework, resulting in cutting-edge pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) [11], RoBERTa [12], and GPT-3 [13], which have greatly improved the use of downstream tasks (e.g., sentiment analysis, text classification, named entity recognition, and text generation). This has provided a massive boost to the development of NLP techniques [14].

To turn our attention to the topic of breast cancer, we aim to study how Chinese netizens discuss breast cancer-related topics on social media, and we believe that there should exist some reasons when people engage in breast cancer-related discussion, for example, family members, friends, relatives, as patients might be the reason why people want to speak with breast cancer, or hearing some stories make people want to focus on. In this study, we use a pre-trained language model as a research tool. We choose Chinese Weibo as a corpus resource to explore.

This study aims to identify whether a post is talking about breast cancer patient and what is the relationship between the mentioned breast cancer patient and the post user. To achieve these goals, we set six categories according to the distance score between the breast cancer patient mentioned in the post and the author of the post, they are **post_user**, **family_members**, **friends_relatives**, **acquaintances**, **heard_relation**, and **no_patient**, respectively.

2 Related work

In recent past, since social media contains a lot of valuable corpus, which attracts lots of researchers to take advantage of it. As Twitter[注3] and Reddit[注4] have a wide range of users, many researchers took Twitter and Reddit as corpus resources to explore various social issues [15, 16]. Researchers incorporated NLP and machine learning techniques to focus on social media, which made out excellent results.

In the earlier period, Orabi et al. [15] used conventional neural network (CNN)-based machine learning model, like CNNs, SVMs to detect depression from Twitter. Sekulic et al. [17] used CNN to detect a wide range of mental health issues. In recent years, machine learning has shifted from CNNs to a transformer-based framework, which made a lot of researchers started to use transformer-based model to study social media. Murarka et al. [7] used RoBERTa to study mental illnesses on social media. Ammer et al. [18] used transformer-based model to study mental illness classification on social media.

To our knowledge, however, there is not too much studies to focus on Chinese Weibo. Meanwhile, although there is not been much such work same with us in prior research, this is a real widespread social problem for the topic of breast cancer, so we broke precedent. We would like to use transformer-based language model to do text classification from Weibo posts.

3 Dataset

3.1 Data Crawling

Since Sina Weibo does not release an official application programming interface (API) to the public, we made a web crawler programming to request posts. Our web crawler is to simulate a user visit to Weibo official website with some adjustable parameters, and search for relevant posts. Through this approach, each web search request can get up to 50 posts, while asking for a new search request, then new posts will be available. Due to the strong anti-crawler mechanism, between every two search requests, an interval time must be set; otherwise, the request will be recognized by the Weibo server as an illegal access and the programming will be forced to terminate. In our crawler programming, the adjustable parameters include keywords, publish date, location, and interval time of two search requests.

In our data collection process, we conducted two searches with different keyword pairs: ("乳腺癌 (breast cancer)" and "悲伤 (sadness)") and ("乳腺癌 (breast cancer)" and "记录 (record)") in Chinese characters, and the posting time was set to January 1st, 2018 to December 31st, 2021 in both data collection processes, the interval time was 15 seconds, the location parameter was blank. Finally, for the two searches with different keywords, we obtained 160,182 posts and 144,125 posts, respectively, each posts includes user id, user name, user type, publish time, post text, location, number of comments, number of likes, and number of reposts.

3.2 Data Cleaning

In the data cleaning phase, we aim to extract posts concerning breast cancer by individual users. For this phase, we first combined the collected data with the two keyword pairs ("乳腺癌 (breast cancer)" and "悲伤 (sadness)") and ("乳腺癌 (breast cancer)" and "记录 (record)") together. To ensure that the post text is related to the breast cancer patient, we removed duplicated data, advertising content. Then, to ensure that the posts are from personal users, we checked user type and recognized that a post is from an official account or an individual account. According to this rule, official posts, institutional posts, or posts from non-personal users were excluded. Finally, the remaining data are not only related to breast cancer, but also from individual users, we obtained 10,322 posts in total.

[注3]：`https://twitter.com/`

[注4]：`https://www.reddit.com/`

3.3 Annotation

We define a distance score from 0 to 5, which refers intimacy level between people. We set up six categories based on the distance score from 0 (the most intimate level) to 5 (the least intimate level) between the breast cancer patient mentioned in the post and the post user (author) as follows.

post_user When the breast cancer patient mentioned in the post is the post user themselves, the breast cancer patient and the post user are the same, so the distance score is 0.

family_members With distance score 1, it indicated that the patient mentioned in the post is a family member (e.g., wife, mother, grandmother, sister), who have a very intimate relationship with the post user.

friends_relatives With distance score 2, it is considered that the patient mentioned in the post is a relative (e.g., cousins and aunts) or friends of the post user.

acquaintances Distance score 3 is assigned when the mentioned patient is an acquaintance of the post user (e.g., a colleague and a neighbor) or just an acquaintance who knows each other.

heard_relation Distance score 4 was assigned when the patient mentioned in the post is only a heard relationship with the post user, or a very distant relationship (e.g., someone who met shortly in a location, a celebrity, a character on a TV show, a public figure or someone who heard about the patient from someone around).

no_patient Distance score 5 is referred to as no patient mentioned in the post.

We randomly annotated about 30% of the data (here is 3,000 Weibo posts) based on the above classification criteria, each piece of data was assigned a label from six categories. In the process of labeling, first step is to determine if there is a patient in post or not, which is a binary classification. Then we continued to work on if a patient exists in post, give a label to the post according to the relationship between the mentioned patient and the post author (the proposed classification criteria), which was a multiclass classification. All data labeling was done by one of the authors who was a Chinese native speaker. See annotation results in Figure 1. Samples for the annotations were shown in Table 1.

In order to verify that our annotations were factually labeled and free of personal subjective bias, we implemented an agreement verification by Cohen's kappa [19]. 20% of the annotated data (600 Weibo posts) were randomly selected, given the same classification criteria, another native Chinese speaker repeated the same annotation work. Eventually, we calculated these two data labeled by different people, to the six categories. Cohen's kappa value is 0.672288, indicating substantial agreement between both annotators.

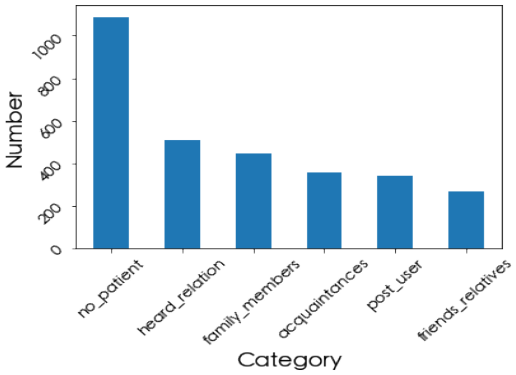


Figure 1 The distribution of the annotated posts

4 Methodology

4.1 Data Processing

This study used Chinese Roberta[注5] [20] as our classification baseline model. To improve the accuracy of multiclass text classification, we made two classifiers: a binary classifier and a multiclass classifier. Since the pre-trained language model Chinese Roberta has a limited input character length of 512, 522 posts in our dataset are longer than 512, therefore, we did a step to extract text abstract to ensure the text length is within 512. We took SnowNLP as a summarization tool [21] to extract abstracts for the post longer than the intended length. The abstract extraction mechanism is based on the weight of each sentence in each paragraph. By setting a number parameter, the corresponding number of sentences is output accordingly, and so the text summary is achieved.

4.2 Classifier Training

Two classifiers were generated in this study. The first one was a binary classifier for classifying Patient and np_Patient. The second one was a multiclass classifier designed to group each post containing patient into one of the five categories (**post_user**, **family_members**, **friends_relatives**, **acquaintances**, and **heard_relation**).

4.2.1 Binary Classifier

To extract patient posts, we make a binary classifier and distinguish each post into patient class or no_patient class. The Chinese Roberta model was fine-tuned using the annotated data. We merged **post_user**, **family_members**, **friends_relatives**, **acquaintances**, and **heard_relation** into **patient**. 80% (2,400 Weibo posts) of the labeled posts were used to train the model and 20% (600 Weibo posts)

[注5]：https://huggingface.co/hfl/chinese-roberta-wwm-ext

Table 1　Weibo post examples with label

Chinese	English	Label（Distance score）
去让自己相信自己得了癌是一件漫长的事情，即使有两份确诊报告摆在我的面前, 当第二份报告我拿在手里，我心里默念三遍阿弥陀佛，打开看到结果和第一次结果一样的时候，我心里咯噔一下，我被确诊了乳腺癌！我能做的就是告诉自己，接受它，面对它，然后打败它！	It was a long time to convince myself that I had cancer, even though I had two reports in front of me. When I held the second report in my hand, I chanted Amitabha three times in my heart, and when I opened it and saw that the result was the same as the first result, my heart thumped, I was diagnosed with breast cancer! All I could do was tell myself to accept it, face it, and beat it!	**post_user** (0)
妈妈确诊乳腺癌的第三天，除了哭什么都做不了，不能替她生病，不能替她分担痛苦，确诊之前一家人高高兴兴的，我从来不奢求我的妈妈有多么的有钱，只想她平平安安健健康康的，可为什么那么难呢，妈妈一生多灾多难，所有的痛苦都承受了，我不怪老天不公，只怪老天为什么不放过她，如果可以，我希望生病的是我自己，如果能让妈妈健康，要我做什么都可以，只希望老天能善待我。	On the third day of my mom's breast cancer diagnosis, I couldn't do anything but cry, I could not share her sickness and pain, my family was happy before the diagnosis, I never wanted my mom to be rich, I just wanted her to be safe and healthy, but why was it so hard? I don't blame God for being unjust, I just blame God for not letting her go. If I could, I wish I was the one who was sick, if I could make my mom healthy, I could do anything, I just hope God would treat me well.	**family_members** (1)
难得有一会的空闲时间，想想人的一生做点自己喜欢的事真的挺难。昨天惊闻一个好朋友的了乳腺癌小女儿才两岁。好可怕！没有妈妈的孩子，想想就心疼。希望所有的孩子都有妈妈爱，妈妈们身体健康。希望人类早日攻克癌症！祝福我的好朋友早日康复！	It is rare to have a moment of free time to think about how hard it is to do something you love in your life. Yesterday, I was shocked to hear that a good friend had breast cancer and her little daughter was only two years old. How terrible! It hurts to think of a child without a mother. I hope all children have their mothers' love and mothers are in good health. I hope mankind will overcome cancer soon! Wish my good friend a speedy recovery!	**friends_relatives** (2)
今天突然得知以前的一个同事癌症到了最后的时期。三年前检查出来了乳腺癌，切的时候发现扩散到了淋巴一部分就一并切了，去年发现扩散到了骨头，这两天说是已经不认识人了。他的女儿才五年级，还有年迈的父母，不过还好她不是独生子女。有时候我在想这样和离婚失去父母一方，哪个对孩子的影响小一些?	Today I suddenly learned that a former colleague has reached the final stage of cancer. Three years ago, she was examined for breast cancer, and when she was cut, she found that it had spread to the lymphatic part and was cut together. Last year, she found that it had spread to the bones, and in the past two days, she said that she did not recognize people. His daughter is only in the fifth grade and has elderly parents, but fortunately she is not an only child. Sometimes I wonder which has less impact on the child, this or losing one parent to divorce?	**acquaintances** (3)
和同事闲聊，得知她同学经常生前男友的气，年纪轻轻就得了乳腺癌早期。感叹现在的癌症越来越年轻化以后更要好好爱自己的同时，暗自庆幸 2 年前的决定，一个男的时常让你生气，连你的心情都不顾及，还指望他以后善待你? 一段关系，开心舒适最重要！这年头，还是保命最要紧！顺带嘲讽一下牛哥最近可爱肉见涨，自带喜感。	I was chatting with a colleague and learned that her classmate was often angry with her ex-boyfriend and got early stage breast cancer at a young age. She sighed that cancer is getting younger and younger, and that she should love herself more, but she was glad that she made the decision 2 years ago. A relationship, happy and comfortable is the most important! These days, it's important to save your life! Incidentally, mock the cattle recently cute meat to see the rise, bring their own sense of comedy.	**heard_relation** (4)
结婚生小孩对我而言，真的是让我瞬间老了十岁的选择。每每一次郁结，我都觉得乳腺癌离我又近了一步。那些婆婆妈妈琐琐碎碎的生活，硬生生把我变成一个每天只会抱怨和唠唠叨叨黄脸婆。呵呵。只要不生活在同一个屋檐下面对屎尿屁的生活，谁在外面还他妈不是个人见人爱的小仙女呢。	Getting married and having children was really a choice that instantly aged me by ten years for me. Every time I got depressed, I felt that breast cancer was one step closer to me. The mother-in-law's trivial life has turned me into a yellow-faced woman who only complains and nags every day. Oh. As long as you don't live under the same roof to face the shit life, who is not a fucking fairy outside.	**no_patient** (5)

Table 2　Binary classifier's metrics report

	precision	recall	f1-score	support
no_patient	0.90	0.90	0.90	204
patient	**0.95**	**0.95**	**0.95**	396
macro avg	0.92	0.92	0.92	600
weighted avg	0.93	0.93	0.93	600

were used to test the fine-tuned model. The hyperparameters for the model training were as follows: batch_size = 16, learning rate = 10^{-5}, and training epochs = 5.

4. 2. 2　Multiclass Classifier

The target classifier focused on patient categorizing, assigning each post into one of the five categories based on the patient's info: **post_user**, **family_members**, **friends_relatives**, **acquaintances** and **heard_relation**. We removed **no_patient** category from the annotated data. 1,515 Weibo posts were used to fine-tune the Chinese Roberta model and 396 Weibo posts were used to test the fine-tuned model. The main parameters were the same as for the binary classifier.

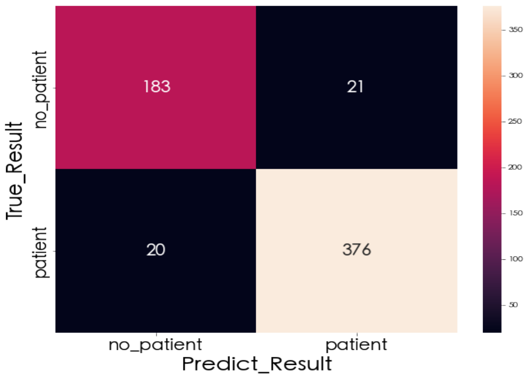


Figure 2 Binary classifier's confusion matrix

Table 3 Multiclass classifier's metrics report

	precision	recall	f1-score	support
acquaintances	0.76	0.67	0.71	75
family_members	**0.93**	0.90	**0.91**	86
friends_relatives	0.74	0.84	0.79	51
heard_relation	0.83	0.83	0.83	102
post_user	0.86	**0.91**	0.89	82
macro avg	0.82	0.83	0.83	396
weighted avg	0.82	0.83	0.83	396

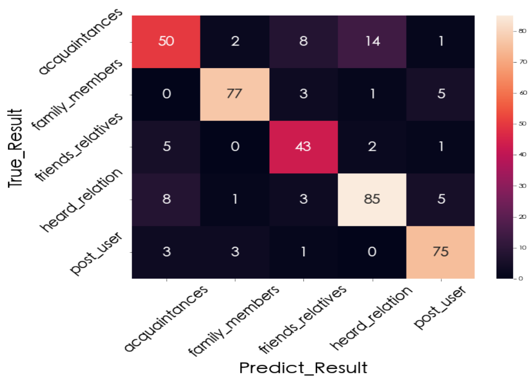


Figure 3 Multiclass classifier's confusion matrix

5 Results and Discussion

5. 1 Results

To evaluate the accuracy of the binary classifier, we used 600 Weibo posts as a test set. Table 2 shows the results of the binary classification and Figure 2 shows the confusion matrix of the binary classification.

To evaluate the accuracy of the multiclass classifier, we used 396 Weibo posts as a test set. Table 3 shows the results of the multiclass classification and Figure 3 shows the confusion matrix of the multiclass classification.

In addition to the annotated data, there were still 7,322 unlabeled posts. We classified them by applying the two classifiers. Firstly, the binary classifier was assigned to make the binary classification (**patient** and **no_patient**). Then among the post containing patient, the multiclass classifier was used to predict a label for each post. Finally, we obtained the predicted label for all data. Figure 4 shows the distribution of the predicted labels.

5. 2 Discussion

The prediction results were consistent with the manually labeled data in all categories' distribution in Figure 1. From them, we can easily know **no_patient** category occupies the first place in all categories, with a proportion of about one third. This indicated that about one third of the posts, users did not mention the breast cancer patient but only the keywords of breast cancer in their posts. In the remaining two thirds of the posts, according to our classification definition, the breast cancer patients most frequently mentioned in user's posts were the one most distantly related to the post user, and it was assigned to **heard_relation** category. Besides, **family_members**, **acquaintances**, **post_user**, and **friends_relatives** categories occupy the second, third, fourth, fifth, and sixth positions, respectively. Referring to our classification criteria, **heard_relation** was defined as the mentioned patient in the post is only a heard relationship with the post user, like celebrity, a character in a TV show. These easily resonate with people, and so that might be the reason **heard_relation** ranked the first among the five categories that contains the mentioned patient.

From the confusion matrix of the binary classification in Figure 2, we can see some cases were mistakenly classified into a different category. Examples of error cases are shown in Table 4. One common reason of the classification error is that the main character in a post is unclear and needs to be inferred through semantic understanding. For example, in (1), the post does not specify who the breast cancer patient is, but after reading it, we can infer that the breast cancer patient is the post user. In (2), the post mentions a wife's breast cancer, but does not say whose wife it is. After reading the post, from semantic understanding, we can infer there is nobody, but generally refers to the wife of a certain person. Under such situation, where semantic inference needs to be combined, our classifier is error-prone.

In Table 3 and Figure 3, we can find that the highest classification accuracy is **family_members** among the 5 categories, which is beyond 90% in F1 score. The lowest classification accuracy is **acquaintances**, which is around 70% in F1 score. For the 5 categories, the classification accuracy is beyond 80% in F1 score on average. Table 5 lists some examples of error cases in multiclass classification. In (3), the true label as **post_user** was predicted as **acquaintances**, we found that there was a "colleague" appearing in the post. Since it was not clearly stated that the breast cancer is "ME" (We inferred the breast cancer patient should be post user from semantic understanding), and according to our classification definition, "colleague" should be grouped

Table 4 Examples of error cases in binary classification

ID	True label	Predict label	Post
(1)	**patient**	**no_patient**	人与人之间的边界感真的太重要了，至少对我来说。今天开完会一个同事顺路带我回去，聊起来我一个人住妈妈会不会担心的问题。结果没讲两句就跳到：你妈妈现在还催婚吗（意思是得了乳腺癌后）? 你有没有准备找男朋友? 就很突然，我觉得自己够机智了，直接说我因为生病被人抛弃了，这不话都在里面了吗? 结果人家居然能够继续问：那对方现在结婚了吗? 我真的当场懵逼，真想回一句“关你屁事”或者“关我屁事”，人要是为了我不结婚了，从一开始就不会抛弃我这不难懂吧? 既然都说抛弃了我他结不结婚我还要去管? 不累么? 好好儿开心地活着不好么? 非要膈应自己? 哦，对了，这次是个女的。 The sense of boundaries between people is really too important, at least for me. Today after the meeting a colleague stopped by to take me back and chatted about whether my mom would be worried if I lived alone. I didn't talk for two sentences before I jumped to: Is your mom still pushing for marriage (meaning after having breast cancer)? Are you planning to find a boyfriend? I thought I was smart enough to say that I had been abandoned because I was sick. The result was that people were able to continue to ask: Is the other party now married? I really confused on the spot, I really want to return a "none of your business" or "none of my business", if people do not get married for me, from the beginning will not abandon me this is not difficult to understand it? Since they say abandoned me he married or not I have to care? Do not tired? What's wrong with living happily? Must be a pain in the ass? Oh yes, this time it is a woman.
(2)	**no_patient**	**patient**	有妻子乳腺癌术后化疗后 5 年复发转移，手机里 5 年的抽血、拍片的结果照片全在，用的什么化疗方案，什么反应，更改的方案全部记得的丈夫。也有几几年确诊，左肺还是右肺都不知道，跟他打电话让来看复查结果也不来院的丈夫。 There are husbands whose wives' breast cancer recurred and metastasized 5 years after chemotherapy, but they have all the photos of the 5 years' blood draws and film results in their cell phones, and they remember what chemotherapy regimen they used, what reaction they had, and what regimen they changed. There are also husbands who don't know whether their left or right lung was diagnosed several years ago, and who don't come to the hospital even though they called him to see the results of the review.

into **acquaintances** category. So this might be the reason (3) was mistakenly classified into other category. In (4), the reasons for misclassification are same as (3). Referring to our classification criteria, the breast cancer patient had a distant relationship with the post author in (4) from semantic understanding, so it should be **heard_relation**. However, there was "colleague" appearing at the beginning of the post, so this might be the reason that was mistakenly classified into **acquaintances** category. In (5), the breast cancer patient was post author regarded as **post_user** category, but in the post there were multiple family-related appellations nouns, such father, baby, son, daughter-in-law, granddaughter, grandma, so this would be the reason that lead to misclassification into **family_members** category. In (6), the breast cancer patient was the cousin of the post author, referring to our **friends_relatives** category, but it was classified into **acquaintances** category by our classifier. However, we could not explain the reason for this case from the point of view of the classification criteria.

Based on the above analysis, we believed that whether or not the post clearly indicated who the breast cancer patient was or if there was a breast cancer patient in the post would be a factor in the classification accuracy. In the post, multiple different appellations also will affect the classification.

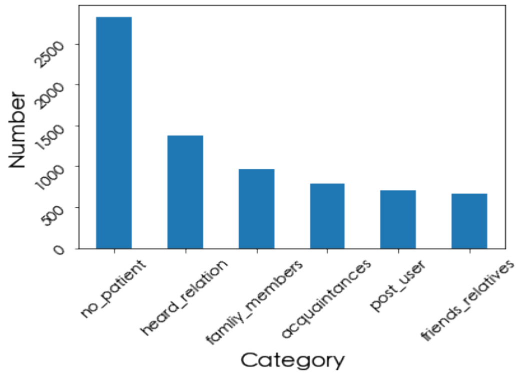


Figure 4 The distribution of the predicted posts

6 Case study

Among our classification results, one category is postuser, which refers to the mentioned patient in the post is the post user (author). That is, the breast cancer patients were talking about themselves, something related to their disease, such as treatment experiences, needs as patients, social relationships, and life changes. In this case study, we aimed to understand the needs of Chinese breast cancer patients. In addition, we compared the needs of Chinese patients and Japanese patients.

In Japan, the Shizuoka Cancer Center has made a cancer problem classification (we call the Shizuoka Cancer Classifi-

Table 5 Examples of error cases in multiclass classification

ID	True label	Predict label	Post
(3)	**post_user**	**acquaintances**	人与人之间的边界感真的太重要了，至少对我来说。今天开完会一个同事顺路带我回去，聊起来我一个人住妈妈会不会担心的问题。结果没讲两句就跳到：你妈妈现在还催婚吗（意思是得了乳腺癌后）？你有没有准备找男朋友？就很突然，我觉得自己够机智了，直接说我因为生病被人抛弃了，这不话都在里面了吗？结果人家居然能够继续问：那对方现在结婚了吗？我真的当场懵逼，真想回一句“关你屁事”或者“关我屁事”，人要是为了我不结婚了，从一开始就不会抛弃我这不难懂吧？既然都说抛弃了我他结不结婚我还要去管？不累么？好好儿开心地活着不好么？非要膈应自己？哦，对了，这次是个女的。 The sense of boundaries between people is really too important, at least for me. Today after the meeting a colleague stopped by to take me back and chatted about whether my mom would be worried if I lived alone. I didn't talk for two sentences before I jumped to: Is your mom still pushing for marriage (meaning after having breast cancer)? Are you planning to find a boyfriend? I thought I was smart enough to say that I had been abandoned because I was sick. The result was that people were able to continue to ask: Is the other party now married? I really confused on the spot, I really want to return a “none of your business” or “none of my business”, if people do not get married for me, from the beginning will not abandon me this is not difficult to understand it? Since they say abandoned me he married or not I have to care? Do not tired? What's wrong with living happily? Must be a pain in the ass? Oh yes, this time it is a woman.
(4)	**heard_relation**	**acquaintances**	和同事闲聊，得知她同学经常生前男友的气，年纪轻轻就得了乳腺癌早期。。。感叹现在的癌症越来越年轻化以后更要好好爱自己的同时，暗自庆幸2年前的决定，一个男的时常让你生气，连你的心情都不顾及，还指望他以后善待你？一段关系，开心舒适最重要！这年头，还是保命最要紧！顺带嘲讽一下牛哥最近可爱肉见涨，自带喜感。 I was chatting with a colleague. I learned that her classmate was often angry with her ex-boyfriend and got early stage of breast cancer at a young age. She sighed that cancer is getting younger and younger, and that she should love herself more, but she was glad that she made the decision 2 years ago. A relationship, happy and comfortable is the most important! These days, it's important to save your life! Incidentally, mock the cattle recently cute meat to see the rise, bring their own sense of comedy.
(5)	**post_user**	**family_members**	难忘的十一天就是我的这个宝贝在医院陪护我11天，谁也信不着爸爸要换她让她去上班不行。手术那几天整夜睡不好着，一会翻身，一会喝水，从来没有烦过，细心照顾着，营养餐天天换样，给我剪指甲，洗脚，真是无微不至，我虽然得乳腺癌是不兴的，但是幸运的是我养了个孝顺的好女儿，同时我还有个好儿子好儿媳妇，他们没有来陪我，是因为疫情期间不方便来，孙女十三岁了还得学习，我们来时太急了没告诉孙女，当视频知道奶奶生病了，就哭了说怎么不告诉我？唉想想这些我一定要坚强战胜病魔 The unforgettable eleven days is my baby in the hospital to accompany me 11 days, who can not believe that the father to replace her to let her go to work can not. I had breast cancer, but luckily I have a good daughter who is filial, and I also have a good son and daughter-in-law, they did not come to accompany me because it is not convenient to come during the epidemic, my granddaughter is 13 years old and still has to study. When the video knew that grandma was sick, she cried and said, “Why didn't you tell me? Thinking about all this I must be strong to overcome the disease.
(6)	**friends_relatives**	**acquaintances**	刚唠嗑唠到重疾险，说我堂姐一个女生做策划的，一直熬夜压力很大但混得贼好。（年薪200w左右，然后被检查出乳腺癌中期。切了乳腺之后本来人快恢复了，结果又突然恶化医生诊断可能最多也就是十月份了。我还感慨生命太无常了，我堂姐说就经常生闷气的人容易得吧，因为生闷气的话就会有很多脏东西出不来，郁结在胸口久而久之就成了肿瘤，淦，所以说做人嘛最重要的是开心，这句话一点都没错。不爽了就说。难受了就哭。压力大就吃喝玩乐。别搞得最后自己身心都受损。 Just natter natter to the critical illness insurance, said my cousin a girl doing planning, has been staying up all night stressful but mixed well. （The company's annual salary is about 200w, and then it was found to have mid-stage breast cancer.） After the mammary gland was cut, the person was almost recovered, and the result was suddenly deteriorated doctor's diagnosis may be up to October. I also lamented that life is too unpredictable, my cousin said that people who are often sulking are prone to it, because if you are sulking, there will be a lot of dirty things that can't come out and become tumors in the chest over time, Kam, so it is important to be happy, this is true. Not happy to say. Hard to cry. Stressed on eating, drinking and having fun. Don't make the last of their body and mind are damaged.

cation)[注6], which was established based on various consultations from cancer patients. There are 16 broad categories as shown in Table 6, multi subcategories were set under each major category, each major category has up to fourth level categories. The Shizuoka Cancer Classification is one of the most complete classification systems in Japan and the categories can cover almost all cancer patients' needs [22].

In [22], the Shizuoka Cancer Classification was used to classify questions by breast cancer patients from Yahoo! JAPAN question and answer service, Yahoo! Chiebukuro. They suc-

(注6)：https://www.scchr.jp/cancerqa/start_shizuoka.html

Table 6 The Shizuoka Cancer Classification categories

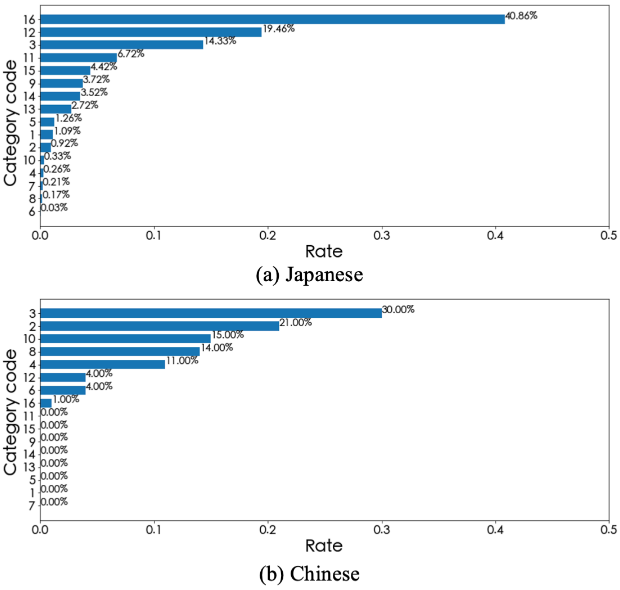
Code	Description
1	Outpatient
2	Hospitalization/discharge/transfer
3	Diagnosis/Treatment
4	Palliative care
5	Notice/Informed Consent/Second Opinion
6	Medical cooperation
7	Home care
8	Facilities/equipment/access
9	Relationship with medical staff (current hospital)
10	Relationship with medical staff (former hospital)
11	Symptoms/side effects/aftereffects
12	Mental problems such as anxiety
13	Way of life, purpose in life, sense of values
14	Employment/financial burden
15	Relationships with family and other people
16	Prevention of cancer, cancer screening, suspicion of cancer

cessfully classified 6,993 questions to fourth level categories, with a correct rate of more than 70%.

In this case study, we classify the needs of Chinese breast cancer patients into the the Shizuoka Cancer Classification categories. Then, we compare the results with the Japanese ones in [22] as a reference for Japanese patients with breast cancer. Specifically, we have randomly selected 100 Weibo posts from our classified **post_user** category. Taking the Shizuoka Cancer Classification categories as the classification criteria, we manually annotated each post with one label. As the Shizuoka Cancer Classification categories are too detailed, we only considered 16 broad classification criteria as shown in Table 6.

This annotation was completed by one of the authors who was a native Chinese speaker. To verify if this annotation work was fair and unbiased, another native Chinese speaking annotator repeated the same work based on the same standard. We adopted Cohen's Kappa [19] to verify agreement, resulting in 0.837581 (high agreement).

In order to keep the same pace with Chinese Weibo posts, we converted all Japanese data's predicted label into 16 broad categories as shown in Table 6. We visualized both distributions of Japanese patients' concerns and Chinese patients' concerns in Figure 5 based on the same categories criteria in Table 6.



(a) Japanese

(b) Chinese

Figure 5 Patient concerns' distribution. (a) Japanese case [22] and (b) Chinese case

In Figure 5(b), 100 posts were distributed into 8 of the 16 categories. Therefore, the bar chart of 8 other categories account for 0. Despite all this, we can draw a conclusion from a broad comprehensive horizon between Japanese breast cancer patients and Chinese breast cancer patients. In terms of the distribution ratio, for the Japanese patients, the top 5 categories are Code 16 (Prevention of cancer, cancer screening, suspicion of cancer), Code 12 (Mental problems such as anxiety), Code 3 (Diagnosis/Treatment), Code 11 (Symptoms/side effects/aftereffects), and Code 15 (Relationships with family and other people) and account for 40.86%, 19.46%, 14.33%, 6.72%, and 4.42%, respectively. For Chinese breast cancer patients, the top 5 categories are Code 3 (Diagnosis/Treatment), Code 2 (Hospitalization/discharge/transfer), Code 10 (Relationship with medical staff (former hospital)), Code 8 (Facilities/equipment/access), and Code 4 (Palliative care) and account for 30%, 21%, 15%, 14%, and 11%, respectively.

Even though Chinese posts were not large enough in this case study, we still could identify the tendency when breast cancer patients talk about their disease. Chinese patients were more interested in treatment, diagnosis, and treatment-related topics, whereas Japanese patients were more concerned with cancer prevention, emotional distress, and family relationships, that went beyond the treatment of breast cancer itself.

7 Conclusion

This study focused on Weibo posts about breast cancer-related topics. We used posts published by individual users on Chinese Weibo as a corpus, to classify posts by individual-centered categories based on proximity to close relationships. We used the latest NLP methods in machine learning to implement. Our basic binary classifier showed high accuracy (more than 90% F1 score), and for the further multiclass classification, our classifier also exceeded 80% in F1 score on average.

Based on our methods and results, when it comes to further study posts on topics related to breast cancer, we can understand how the post author's relationship distance with the cancer patient can potentially influence their opinions about and emotional responses to breast cancer itself. For example, when the cancer patient is post user herself, the content of the post may be more about her own treatment, relief, or personal concerns. When the cancer patient in the post is his or her mother, wife, or other family members, the content of the post may be about a kind of sadness, grief, or personal feelings or expectation. However, when the patient in the post is someone less close to the author, the content of the post may be more of a rational statement or a simple description of the story. As mentioned above, our research method provides a valid reference direction for further research on breast cancer-related topics in the future.

At the same time, there are some limitations in our study, for example, the amount of data is not large enough, so even the presence of small biases can have a significant impact on the final prediction results. Data labeling is done manually, so it is time-consuming and prone to human error. Since the data are collected from public social networks, the data structure is not uniform, which brings a lot of difficulties in the cleaning and analyzing process. Also, because the data structure is very different, too many different attempts are needed to select the classifier model, which makes the research more difficult.

In the future research, we can try to unify the data structure. The distance scores we currently define (0-5) might be too coarse, a more scientific and reasonable classification definition would be required. The data can be labeled and processed in a crowd-sourced manner to reduce human error. This will ensure that the data can be classified more accurately, which is good for subsequent in-depth research studies on breast cancer-related topics. We can get more specific valuable findings by focusing on each different category. This may provide useful references for medical health, human care, social services, and other relevant fields.

Acknowledgements

This work was supported by JST-Mirai Program Grant Number JPMJMI21J2 and JST CREST Grant Number JP-MJCR22N1, Japan.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] Nathan S Consedine, Carol Magai, Yulia S Krivoshekova, Lynn Ryzewicz, and Alfred I Neugut. Fear, anxiety, worry, and breast cancer screening behavior: a critical review. *Cancer Epidemiology Biomarkers & Prevention*, 13(4):501–510, 2004.

[3] Xinyan Zhao, Xiaohui Wang, Zexin Ma, and Rong Ma. Primacy effect of emotions in social stories: User engagement behaviors with breast cancer narratives on facebook. *Computers in Human Behavior*, 137:107405, 2022.

[4] Edel M Quinn, Mark A Corrigan, Seamus M McHugh, David Murphy, John O'Mullane, Arnold D Hill, and Henry Paul Redmond. Who's talking about breast cancer? analysis of daily breast cancer posts on the internet. *The Breast*, 22(1):24–27, 2013.

[5] Tânia Brandão, Rita Tavares, Marc S Schulz, and Paula Mena Matos. Measuring emotion regulation and emotional expression in breast cancer patients: A systematic review. *Clinical Psychology Review*, 43:114–127, 2016.

[6] Patti M. Valkenburg. Social media use and well-being: What we know and what we need to know. *Current Opinion in Psychology*, 45:101294, 2022.

[7] Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. Detection and classification of mental illnesses on social media using roberta. *arXiv preprint arXiv:2011.11226*, 2020.

[8] S Anand and K Narayana. Earthquake reporting system development by tweet analysis. *Int. J. Emerg. Eng. Res. Technol*, 2:96–106, 2014.

[9] Xinyue Ye, Shengwen Li, Xining Yang, and Chenglin Qin. Use of social media for the detection and analysis of infectious diseases in china. *ISPRS International Journal of Geo-Information*, 5(9):156, 2016.

[10] Binbin Ye, Padmaja Krishnan, and Shiguo Jia. Public concern about air pollution and related health outcomes on social media in china: An analysis of data from sina weibo (chinese twitter) and air monitoring stations. *International Journal of Environmental Research and Public Health*, 19(23):16115, 2022.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[13] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[14] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*, 2021.

[15] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 88–97, 2018.

[16] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[17] Ivan Sekulic and Michael Strube. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327, Hong Kong, China, November 2019. Association for Computational Linguistics.

[18] Iqra Ameer, Muhammad Arif, Grigori Sidorov, Helena Gòmez-Adorno, and Alexander Gelbukh. Mental illness classification on social media texts using deep learning and transfer learning. *arXiv preprint arXiv:2207.01012*, 2022.

[19] Ted Byrt et al. How good is that agreement? *Epidemiology*, 7(5):561, 1996.

[20] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*, 2020.

[21] Caixia Chen, Jue Chen, and Chun Shi. Research on credit evaluation model of online store based on snownlp. In *E3S Web of Conferences*, volume 53, page 03039. EDP Sciences, 2018.

[22] Masaru Kamba, Masae Manabe, Shoko Wakamiya, Shuntaro Yada, Eiji Aramaki, Satomi Odani, and Isao Miyashiro. Medical needs extraction for breast cancer patients from question and answer services: Natural language processing-based approach. *JMIR Cancer*, 7(4):e32005, Oct 2021.