Improved Text Classification of Long-term Care Materials

范姜頤 Yi Fan Chiang
國立政治大學語言學研究所
Graduate Institute of Linguistics
National Chengchi University
vianne6@gmail.com

李季陵 Chi-Ling Lee
國立政治大學華語文教學碩博士學位學程
Master's & Doctor's Program in Teaching Chinese as a Second Language
National Chengchi University
19940818jillian@gmail.com

廖恒佳 Heng-Chia Liao
國立政治大學華語文教學碩士學位學程
Master's Program in Teaching Chinese as a Second Language
National Chengchi University
106161011@nccu.edu.tw

蔡宜庭 Yi-Ting Tsai
國立政治大學語言學研究所
Graduate Institute of Linguistics
National Chengchi University
109555005@nccu.edu.tw

張瑜芸 Yu-Yun Chang
國立政治大學語言學研究所
Graduate Institute of Linguistics
National Chengchi University
yuyun@nccu.edu.tw

Abstract

Aging populations have posed a challenge to many countries including Taiwan, and with them come the issue of long-term care. Given the current context, the aim of this study was to explore the hotly-discussed subtopics in the field of long-term care, and identify its features through NLP. Texts from forums and websites were utilized for data collection and analysis. The study applied TF-IDF, the logistic regression model, and the naive Bayes classifier to process data. In sum, the results showed that it reached a F1-score of 0.92 in identification, and a best accuracy of 0.71 in classification. Results of the study found that apart from TF-IDF features, certain words could be elicited as favorable features in classification. The results of this study could be used as a reference for future long-term care related applications.

Keywords: long-term care, natural language processing (NLP), text classification, Chinese

1 Introduction

Long-term care, by the definition of Ministry of Health and Welfare[1], refers to "the living support, assistance, social participation, care and relevant healthcare services in accordance with the needs of any individual whose mental or physical incapacity has lasted or is expected to last for six months or longer".

As for long-term care services, according to Harris-Kojetin et al. (2019), include assistance with activities of daily living (abbreviated as ADLs, which includes activities such as dressing, bathing, and toileting), instrumental activities of daily living (abbreviated as IADLs, which includes activities such as medication management and housework), and health maintenance tasks. Long-term care services assist people to improve or maintain an optimal level of physical functioning and quality of life, which can include help from other people and special equipment or assistive devices.

According to National Development Council (2020)[2], driven by a low birth rate of 1.2 and an all-time high life expectancy in 2020, senior citizens (people aged 65 or older) will account for over 20 percent of Taiwan's total population by 2025, indicating Taiwan will step into a super-aged society. Within the context, how long-term care services can meet the escalating demand is absolutely critical.

While extensive long-term care studies have been conducted, few of them are from the perspective of caregivers, especially family caregivers. Chen (2013) pointed out that long-term care system in Taiwan lacked support services for the caregivers.

For most family caregivers, they provided ADLs and IADLs for their family members. Sometimes, they might get stuck when being confronted with several care problems. Lu (2005) showed that their needs included respite care services, psychological and educational support programs, and financial subsidies. Among psychological and educational support programs, caring skills and consulting services were what the caregivers desperately wanted while taking care of their family members.

In the course of caring, emergencies could happen. Caregivers might need timely help but lack sufficient time to search and browse – for an appropriate answer to their question. Besides, caregivers might be in need when it was late at night and had no one to turn to. Despite the abundant resources on the Internet, not every piece of information is suitable for caregivers, let alone those irrelevant discussions. Still, they had to filter.

To bridge the gap, this study aimed to explore features useful to identify (1) long-term care and unrelated texts (2) long-term care topics that generate intensive discussion, so that the application could be designed more user-friendly, and caregivers could get the needed information more efficiently.

At present, despite the fact that there are many online long-term care platforms, their manner and principles of classification are not based on caregivers' needs. This study intended to fill this gap by collecting authentic materials, adjusting its categories manually and processing them with caregiver-oriented topics.

This study sought to answer the following research questions:

1. What are the topics that the caregivers have been hotly discussing?
2. How to provide caregiver-oriented information through NLP?

2 Literature Review

In terms of much-discussed topics, according to a global overview study by Fu et al. (2019), the simultaneous analysis of both references and keywords revealed that common long-term care hot topics included 'dementia care', 'quality of care', 'prevalence and risk factors', 'mortality', and 'randomized controlled trial'.

Back in Taiwan, Lee et al. (2019) had studied about what topics in an online blog would readers (including family caregivers or others) mainly follow. The study suggested that out of the eight categories, the most commonly read and discussed topics were 'family relationships', 'caregiving experiences', 'caregiving stress', 'physical, psychological, and social adaptation', and 'seniors care issues'. The left ones were 'long-term care

[1] https://www.mohw.gov.tw/mp-2.html

[2] National Development Council. 2020. Population Projection in ROC (from 2020 to 2070) (ISBN : 978-986-5457-22-8). Retrieved from: https://pop-proj.ndc.gov.tw/upload/download/中華民國人口推估(2020至2070年)報告.pdf

policies', 'the ups and downs in caring', and 'special caregiver groups' – such as male caregivers and former caregivers.

In order to find appropriate features, this paper applied TF-IDF, which was consulted from studies by Phetkrachang and Kittiphattanabawon (2019) and Paik (2013). They both applied TF-IDF weighting in their research.

3 Methodology

The procedure of this research could be presented as Figure 3.1.

Firstly, data from three platforms were collected and separated into caregiver-oriented ones and non-caregiver-oriented ones. Secondly, a task of annotation was done to appropriately classify relevant data into eight categories. Thirdly, by appling TF-IDF, the logistic regression model, and the naive Bayes classifier, features were extracted. Afterwards, manually obtained features were also taken into consideration. Lastly, with the model, it was hoped to be beneficial to different applications, such as chatbot development or website-building. It could help website designers to build a more caregiver-friendly platform.

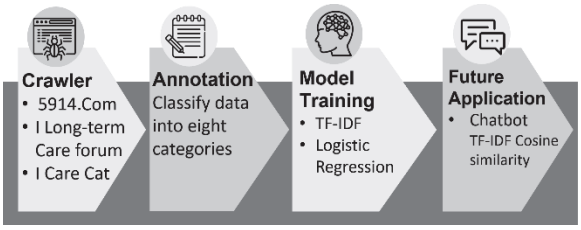

Figure3.1: Main procedure of this study

3.1 Data Collection

To investigate topics that spark widespread discussion in long-term care, 800 articles were collected from three online platforms, which included 長照喵 'long-term care cat'[3], 愛長照 'love long-term care'[4], and 呼叫醫師 'call for doctor'[5]. In the field of long-term care, all of these platforms are considered to be the most representative ones in Taiwan. First, 長照喵 'long-term care cat' is a website which mainly shares articles about long-term care information, health knowledge, and activities of long-term care. The website straightforwardly categorize these articles into two types: 找活動 'Activities' and 找知識 'Knowledge'. 找活動 'Activities' includes lectures or activity information about long-term care, whereas 找知識 'Knowledge' contains the practical knowledge or experience sharing on long-term care. In the research, we collected the articles of 找知識 'Knowledge', since the information of 找活動 'Activities' were time-sensitive – data collected went out-of-date or expired soon. Second, 愛長照 'love long-term care' is a website for caregivers to seek resources and advice. The platform also provides plenty of articles about long-term care. In addition, there is an online forum for users to ask questions or discuss long-term care issues. In the forum, the posts are classified into 16 categories, which are 居家服務 'home services', 照顧機構 'care institutions', 外籍看護 'foreign nursing workers', 我要找幫手 'find helpers', 照顧技巧 'care skills', 輔具 'assistive devices', 飲食營養 'diet and nutrition', 補助 'subsidy', 失智症 'dementia', 疾病 'diseases', 安寧 'palliative care', 照顧尊嚴 'dignity', 照顧苦水 'complaint', 健康保健 'health care', 生活/課程 'life and lessons', and 其他 'others'. This research obtained all posts and comments in this forum, and consulted the topic categorization to find the hot issues about long-term care. Last, 呼叫醫師 'call for doctor' is an online platform for people to talk to doctors directly. Doctors would professionally answer the questions posted online. The discussion field involves long-term care, which provides the data needed of the research.

To analyze the main topics discussed online, the language materials were first collected from websites mentioned above. Although articles were taken from long-term care websites, some of the materials were time-sensitive activities or advertisements. Therefore, the first task of the research was to identify the target articles. The relevant articles and the irrelevant ones were marked manually afterwards. Finally, 400 were annotated as long-term care articles and 400 articles were classified as irrelevant ones. This dataset was designed as the gold-standard data in the identification task.

[3] Long-term care cat: icarecat.com

[4] Love long-term care: https://www.ilong-termcare.com/

[5] Call for doctor: https://www.5914.com.tw/

In the course of inspecting the articles between caregiver-oriented and irrelevant articles, we found that the words used were quite similar. For instance, (3-1) is an article section from 'long-term care cat'; however, the content of it is about the requirements of long-term care worker, which is not suitable for long-term caregivers. This kind of articles contain the words which frequently show up in caregiver-oriented articles such as 照顧'look after' or 長照 'long-term care'. Therefore, the research first tried to build a model which identifies the articles between caregiver-oriented writings and irrelevant articles.

(3-1) Q1、成為 「照顧服務員」需要
Q1, become 'care worker' need

什麼 資格 ？
what qualification?

'Q1,What qualifications are needed to become a "care worker"? '

有 以下 資格 之 一
have below qualification of one

就 可以 成為 照顧服務員
and can become 'care worker'

'Having one of the following qualifications could become a care worker.'

受訓 參加 「照顧 服務員 專業
train participate 'care worker professional

訓練 課程」 取得 結業 證書
train course' obtain graduation certification

'participate the training of "Professional Training Course for Care Workers" and obtain a certification'

To find out the hot topics for long-term care discussion, the forum categories in the 'love long-term care' forum and the studies by Lee et al. (2019) and Fu et al. (2019) were consulted. Based on the data collected, categories in 'love long-term care' forum and the above-mentioned studies, the topics of our collected data were re-categorized manually into 8 categories. The topics of 'home services', 'care institutions', 'foreign nursing workers', 'find helpers' were merged into 'care manpower'. Besides, the topics of 'palliative care' and 'dignity' were merged into 'dignity' as one of the 'social issues'. The 'subsidy' posts were eliminated in this research. For one thing, articles in 'subsidy' included too many details about long-term care policies, statics (Lee et al., 2019) shown that caregivers were less interested in them. For another, most contents were time-sensitive. Besides, the original topic 'health care' overlapped issues of many topics. As a result, the articles under this topic were re-classified into topics of 'diet and nutrition' and 'care skills'. To sum up, the eight categories in the study are as follows – 失智症 'dementia', 疾病 'diseases', 照顧技巧 'care skills', 照顧苦水 'complaint', 照顧尊嚴 'dignity', 輔具 'assistive devices', 飲食營養 'diet and nutrition' and 照顧人力 'care manpower'. Figure 3.2 demonstrates the article counts of each category. The definitions of these categories are shown in Table 3.1. The second language model in the research would be employed to automatically characterize all the articles accordingly.

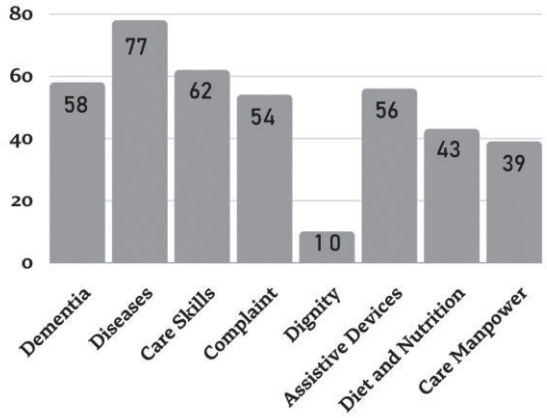

Figure 3.2: Counts on each topics of articles

category	content
失智症 'dementia'	Discussion about dementia
疾病 'diseases'	Discussion about diseases other than dementia
照顧技巧 'care skills'	Topics about caring skills and details
照顧苦水 'complaint'	Complaint about caring patients.
照顧尊嚴 'dignity'	Topics about how to treat patients with dignity, e.g., palliative care, good death
輔具 'assistive devices'	Discussion about Assistive devices, e.g., wheelchair
飲食營養 'diet and nutrition'	Details about patients' nutrition and diet
照顧人力 'care manpower'	Topics about foreign nursing workers and long-term care institutions

Table 3.1: Category definitions of topics on long-term care

3.2 Model

Before training, the data cleaning processes were implemented in advance. The punctuations and English characters were removed before training. Besides, the articles were segmented with *jieba*[6] package for the purpose of further data training.

The first model training is to identify the long-term care contents. We employ the trained TF-IDF model provided by jieba to help identify topic-related features for each category. To find the features of long-term care articles, TF-IDF scores were calculated to find the common and recurring keywords through the whole topics of the long-term care contexts. Each topic filtered out 20 keywords. Some of the keywords overlapped among categories. For finding the common features of long-term care, the words which repeated above three times were selected as common features. Moreover, for the purpose of gathering target articles more comprehensively, the top two keywords of each topic were added up as features for identifying the long-term care articles. All of the features for identifying long-term care contexts were 照顧 'look after', 因為 'because', 我們 'we', 照護 'care', 問題 'problem', 他們 'they', 治療 'treatment', 輪椅 'wheelchair',失智症 'dementia', 失智 'dementia', 藥物 'medicine', 醫療 'medical', 安寧 'palliative care', 輔具 'assistive devices', 服務 'service', 飲食 'diet', and 營養 'nutrition'[7]. With these features, the logistic regression model was applied to identify the long-term care articles. In the collected dataset, there were 400 articles with care-giver centered topics, and the other 400 articles discussed other topics. The entire dataset was split into 70% for training and 30% for testing.

The second step of model training was to classify the articles on long-term care into eight categories. The training model was also based on the logistic regression model.

4 Results

4.1 Model

Regarding the identification model based on the logistic regression model and TF-IDF features, the best f1-score was 0.72. The average was 0.65.

To further examine the model, 56 articles were collected from PTT 銀髮族板 'the Elderly board on PTT'[8]. This was the most relevant board to discuss long-term care on PTT. When the dataset was composed of PTT Elderly posts solely, the f1-score was 0.92.

In the identification phase, the features of the articles were extracted from the TF-IDF scores.

4.2 Text classification

First, to distinguish whether articles were related to long-term care, the TF-IDF weighting helped us in doing so. The features we got from the calculation of articles associated with long-term care were shown in table 4.2.1.

Category	Features
Dementia	失智症 'dementia'、失智 'dementia'、我們'we'、他們 'they'、認知'cognition'
Diseases	治療'treatment'、藥物 'medicine'、症狀'symptom'、用藥'medication'、障礙 'barrier'
Care Skills	我們'we'、治療'treatment'、訓練'training'、運動 'exercise'、活動'movement'
Complaint	我們'we'、悲傷'sadness'、自己'self'、一個'one'、他們 'they'
Dignity	醫療'medical care'、病人 'patient'、灌食'tube feeding'、痛苦'pain'、生命 'life'
Assistive Devices	輪椅'wheelchair'、輔具 'assistive devices'、我們 'we'、補助'subsidy'、支撐 'support'
Diet and Nutrition	飲食'diet'、營養'nutrition'、攝取'take in'、建議 'suggestion'、食物'food'
Care Manpower	服務'service'、單位 'affiliation'、居家'home'、照護'care'、機構'institution'

Table 4.2.1: Main TF-IDF features of topics on long-term care

[6] Jieba : https://github.com/fxsjy/jieba

[7] These features, as a part of the research outcome, would be further discussed in the result section.

[8] The Elderly board on PTT: https://www.ptt.cc/bbs/elderly/index.html

Then, some of the distinctive TD-IDF features of long-term care articles were picked, along with some manually obtained features. The adjustment was as follows. In sum, the best f1-score under such arrangement was 0.71. The average was 0.67.

TF-IDF features	**Manually obtained features**
Dementia 失智症	
失智症 'dementia'、失智 'dementia'、認知 'cognition'	阿滋海默症 'Alzheimer's disease'
Diseases 疾病	
治療 'treatment'、症狀 'symptom'、藥物 'medicine'、用藥 'medication'	疾病 'diseases'
Care Skills 照顧技巧	
訓練 'training'、運動 'exercise'、活動 'movement'	指甲 'nails'、技巧 'skills'、練習 'practices'
Complaint 照顧苦水	
悲傷 'sadness'、自己 'self'	媳婦 'daughter-in-law'、體會 'relate to'、加油 'hang in there'、累 'tired'、辛苦 'You've worked hard'、溝通 'communicate'
Dignity 照顧尊嚴	
灌食 'tube feeding'	鼻胃管 'nasogastric tube'、安寧 'palliative'、尊嚴 'dignity'、臨終 'hospice'、善終 'good death'
Assistive Devices 輔具	
輔具 'assistive devices'、輪椅 'wheelchair'、補助 'subsidy'	電動 'electric'、扶手 'handrail'、拐 'crutch'、杖 'walking stick'
Diet and Nutrition 飲食營養	
飲食 'diet'、營養 'nutrition'、攝取 'take in'	東西 'things'、牙口 'teeth'、維他命 'vitamin'
Care Manpower 照顧人力	
服務 'service'、單位 'affiliation'、居家 'home'、機構 'institution'	仲介 'agency'、雇主 'employer'

Table 4.2.2: Features of each category on long-term care

5 Discussion

In the previous section, several high-frequency features were retrieved from the dataset, including 照顧 'look after', 因為 'because', 治療 'treatment', 我們 'we', and 照護 'care'. All of these words occurred over three times among the 8 categories.

Among these features, 我們 'we' was the most impressive one. It was not a long-term care related word, but it had such a high weight compared with the other features that were directly associated with long-term care. One possible explanation for this is that articles related to long-term care topics sometimes offer information and suggestions with empathy. With this inclusive title, those readers as caregivers might feel a special closeness and feel understood.

Another finding was about a less frequent feature 他們 'they'. Both the type 失智症 'dementia' and 照顧苦水 'complaint' obtained the feature 他們 'they'. To explore this phenomenon, we compared where they occurred in posts of these two categories, and in posts of 疾病 'diseases'.

失智症 'dementia' and 疾病 'diseases' were both diseases. From our observation, when it came to diseases, narrators of the articles often talked about their symptoms and the corresponding treatments. However, when an article's topic is 失智症 'dementia', hardly would the narrator suffer from dementia. In general, narrators under this topic view people who have dementia as 'the constitutive other', and thus use the title 他們 'they' in particular.

On the whole, with these articles and features of various types, our observations could be roughly divided into two kinds. For the first kind, features indicated relationships, as mentioned above. Besides, features sometimes consisted of identities, such as 媳婦 'daughter-in-law'. These articles tended to express personal experience, and mostly were under topics such as 照顧苦水 'complaint'.

When articles with family titles were categorized into the complaint type, they often described care experiences from the perspective of caregivers themselves. Moreover, apart from 照顧苦水 'complaint', in articles under topics of 照顧技巧 'care skills' and 輔具 'assistive devices', the feature 我們'we' almost topped the TF-IDF list. Although articles of this kind were not solely

written from caregivers' perspectives, it reflected the fact that when offering advice from the clinical experiences, the doctors or specialists tended to use 我們'we' often.

The second kind of features belonged to objective things, instead of showing relationships and communications. They are closely related to symptoms and treatments, or associated affiliations and services on long-term care.

While features approximately reflect the content of every type of topic, certain features stood out for their good performance in discrimination.

The TF-IDF weighting indicated the feature 輔具 'assistive devices' was a favorable feature in identify long-term care articles. Besides, the feature 牙口 'teeth' also played an important role in distinguishing 飲食營養 'diet and nutrition' posts from others.

In the course of classification, we found articles under 照顧技巧 'care skills' were the hardest to be classified. This was because 'care skills' posts usually mention the diseases first, then the techniques and suggestions. The suggestions might contain reminders on diets, and if the situation went severe, the narrator might provide suggestions on corresponding assistive devices. These made the classification of 照顧技巧 'care skills' challenging.

Limitation for the present study was that it was a relatively small dataset for analyzing. More contents on long-term care may help refine the categories and enrich the features.

6 Conclusions and Future Work

The results of this study could have broad applications in the future. For example, chatbot, search engine, or Q&A keywords.

One possible application is a long-term care chatbot. The chatbot could be designed to classify the topic of the question or inputs, and give the answer accordingly. The TF-IDF Cosine similarity model of sklearn[9] could be applied to find the most appropriate answer to the input question.

Since our study focused on long-term caregivers, specialists could optimize their services to caregivers. Thus, caregivers would get suitable advice more efficiently. They could save time on browsing and filtering information on the Internet.

[9] Sklearn: https://scikit-learn.org/stable/

7 References

Zheng-Fen Chen. 2013. 我國長期照顧體系欠缺的一角：照顧者支持服務 (a missing piece of Taiwan's long-term care system: caregiver support services) [in Chinese]. *Community Dev J*, 141: 203-213.

Li-Ping Fu, Zhao-Hui Sun, Lan-Ping He, Feng Liu, and Xiao-Li Jing. 2019. Global long-term care research: A scientometric review. *International journal of environmental research and public health,* 16(12): 2077. https://doi.org/10.3390/ijerph16122077

Lauren Harris-Kojetin, Manisha Sengupta, Jessica P. Lendon, Vincent Rome, Roberto Valverde, and Christine Caffrey. 2019. Long-term care providers and services users in the United States, 2015-2016. *DHHS Publication No. 2019–1427, National center for health statistics.* https://www.cdc.gov/nchs/data/series/sr_03/sr03_43-508.pdf

Mun-Sim Lai, and An-Chi Tung. 2015. Who supports the elderly? The changing economic lifecycle reallocation in Taiwan, 1985 and 2005. *The Journal of the Economics of Ageing,* 5: 63-68. https://doi.org/10.1016/j.jeoa.2014.10.012

I Lee, Chii-Jun Chiou, and Yueh-Feng Lu. 2019. 應用網路部落格倡導家庭照顧者議題 (lessons learned from developing web-based educational support blogs for family caregivers in Taiwan) [in Chinese]. *The Journal of Long-Term Care in Taiwan,* 23(3): 217-230.

Pau-Ching Lu. 2005. 支持家庭照顧者的長期照護政策之構思 (toward a more family caregiver-responsive long-term care policy) [in Chinese]. *National Policy Quarterly,* 4(4): 25-40. https://doi.org/10.6407/NPQ.200512.0025

Jiaul H. Paik. 2013. A novel TF-IDF weighting scheme for effective ranking. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval,* pages 343-352.

Ketsara Phetkrachang and Nichnan Kittiphattanabawon. 2019. Fuzzy TF-IDF Weighting in Synonym for Diabetes Question and Answers. In *Proceedings of the 15th International Conference on Computing and Information Technology,* pages 59–68.

Hsiu-Hung Wang, and Shwn-Feng Tsay. Elderly and long-term care trends and policy in Taiwan: Challenges and opportunities for health care professionals. *The Kaohsiung journal of medical sciences,* 28(9): 465-469.