doi:10.3233/SHTI220704

Comparing Themes Extracted via Topic Modeling and Manual Content Analysis: Korean-Language Discussions of Dementia on Twitter

Haeyoung LEE[a], Sun Joo JANG[a], Frederick F SUN[b], Peter BROADWELL[c] and Sunmoo YOON[d,e,1]
[a] *Department of Nursing, Chung-Ang University, South Korea*
[b] *Department of Rehabilitation and Regenerative Medicine, Columbia University, USA*
[c] *Center for Interdisciplinary Digital Research, Stanford University, USA*
[d] *General Medicine, Department of Medicine, Columbia University, USA*
[e] *Data Science Institute, Columbia University, USA*

Abstract. We randomly examined Korean-language Tweets mentioning dementia/Alzheimer's disease (n= 12,413) posted from November 28 to December 9, 2020, without limiting geographical locations. We independently applied Latent Dirichlet Allocation (LDA) topic modeling and qualitative content analysis to the texts of the Tweets. We compared the themes extracted by LDA topic modeling to those identified via manual coding methods. A total of 16 themes were detected from manual coding, with inter-rater reliability (Cohen's kappa) of 0.842. The proportions of the most prominent themes were: burdens of family caregiving (48.50%), reports of wandering/missing family members with dementia (18.12%), stigma (13.64%), prevention strategies (5.07%), risk factors (4.91%), healthcare policy (3.26%), and elder abuse/safety issues (1.75%). Seven themes whose contents were similar to themes derived from manual coding were extracted from the LDA topic modeling results (perplexity: -6.39, coherence score: 0.45). Our findings suggest that applying LDA topic modeling can be fairly effective at extracting themes from Korean Twitter discussions, in a manner analogous to qualitative coding, to gain insights regarding caregiving for family members with dementia, and our approach can be applied to other languages.

Keywords. Dementia caregiving, online intervention, social media, topic modeling

1. Introduction

Dementia is the seventh leading cause of death globally [1]. The psychological distress of family members of a person with dementia has been observed to increase during the COVID-19 pandemic [2]. Although recent qualitative studies have emphasized the burden experienced by family caregivers of persons with dementia during the pandemic [1,2], film, literature, and TV programs continue to associate negative attitudes and

[1] Corresponding Author, Sunmoo Yoon, General Medicine, Department of Medicine, Columbia University Irving Medical Center, 630W 168 Street, PH105, New York, NY, 10032, USA; E-mail: sy2102@cumc.columbia.edu.

images with dementia sufferers, and it remains unclear whether the positive or negative connotations of caregiving have gained greater prominence among Koreans.

Qualitative content analysis has been traditionally applied to manually extract themes from recorded human artifacts such as voice recordings or interview text data via independent coding personnel. Because this work depends on human efforts, human coding tasks are costly and require content expertise, coding experience, and time. Further, interpretation of the results of qualitative content analysis can be rather subjective and may introduce biases. By contrast, topic modeling applies statistical analyses to counts of words or word groups that co-occur within text documents to reveal latent word groupings with potential semantic significance across a collection of documents [3]. Compared to manual reading and keyword/keyphrase extraction algorithms (TF-IDF, TextRank), topic modeling offers the ability to find semantically meaningful groupings of multiple words across very large document collections, even when all of the related words do not appear together within any single document or section. This study aimed to compare themes extracted from Korean-language Tweets mentioning dementia/Alzheimer's via machine learning (LDA) to those detected with manual coding methods as a foundation for developing future Twitter-based interventions for family caregivers of persons with dementia among Koreans.

2. Methods

We applied qualitative content analysis using manual coding and machine learning (LDA topic modeling) to extract themes from a corpus of publicly available Korean-language Tweets (Tweet metadata "lang" field value: "ko") mentioning dementia/Alzheimer's disease, n= 12,413) from November 28 to December 9, 2020 [3] without limiting geographical locations. We used the NCapture software to collect this data. Two independent nursing researchers with content analysis expertise manually coded 12,413 total Tweets and extracted themes using qualitative coding methods. Inter-rater reliability was calculated on 100 randomly extracted Tweets (Cohen's Kappa: 0.842).

Another bilingual researcher with data science and domain expertise applied natural language processing (NLP) techniques to preprocess the text data (e.g., stop words removal) after translating the corpus [function =GOOGLETRANSLATE (cell, "ko," "en")] and then ran LDA topic modeling techniques to extract themes from the corpus. During the LDA topic modeling, an output set size of seven themes was selected after iteratively testing various numbers of themes from six to 21, comparing the perplexity and coherence scores of the resulting topic models and visually evaluating the overlapping of theme clusters for clinical meaningfulness and interpretability (perplexity: -6.39, coherence score: 0.45). Next, we compared the extracted themes derived from manual coding and machine learning (LDA). The larger study of which this was a part was approved by the Institutional Review Board (IRB). Resources, including analytic Python codes and de-identified data, are available on GitHub and OSF.io (https://osf.io/qruf3).

3. Results

Among a total of 12,413 Korean-language Tweets mentioning dementia/Alzheimer's disease, approximately one-third of the Tweets (N= 4,364, 35.16%) were identified via manual coding as non-relevant to dementia or Alzheimer's disease. Using manual content analysis, sixteen themes were extracted from the corpus (N= 8,048). The top themes from the manual coding include burden of family caregiving 48.50%, reports of wandering/missing family members with dementia 18.12%, stigma 13.64%, prevention strategies 5.07%, risk factors 4.91%, healthcare policy 3.26%, elder abuse and safety issues 1.75%, information on dementia caregiving, and financial support and warning symptoms 1.42%. Conversely, seven themes were extracted from the Tweet corpus from the LDA topic modeling approach. All themes from LDA were similar to those from manual coding (Table 1).

Table 1. Extracted themes from Korean-language Tweets mentioning dementia/Alzheimer's disease using manual content analysis and LDA topic modeling

	Themes	Tweets	%	Manual	LDA
1	Caregiving burden of family members of a person with dementia	3,903	48.50	X	X
2	Report of missing and wandering dementia patients	1,458	18.12	X	
3	Stigma	1,098	13.64	X	X
4	Prevention strategies for dementia	408	5.07	X	X
5	Risk factors and causes of dementia	395	4.91	X	X
6	Healthcare policy for dementia patient care	262	3.26	X	X
7	Elder abuse and safety issues for dementia patients	141	1.75	X	
8	Information on dementia caregiving, financial support and warning symptoms	114	1.42	X	
9	Emotional distress in dementia patients and family members	61	0.76	X	
10	Dementia patient advocacy	58	0.72	X	
11	Treatment for dementia	53	0.66	X	
12	Diagnostic test for dementia	36	0.45	X	X
13	Coping with grief and loss	27	0.34	X	
14	Resilience	15	0.19	X	
15	Symptoms of dementia in early, middle and late stages	14	0.17	X	X
16	Prevalence of dementia in Korea	5	0.06	X	
	Total	8,048	100.0		

4. Discussion and Conclusion

This study explored differences between themes extracted from Tweets mentioning dementia using manual versus machine learning-based topic modeling approaches in the Korean language.

One of the major themes derived from the texts in both approaches was the burden caregiving places upon family caregivers, such as financial difficulties, emotional distress (e.g., depression), and family conflicts among Korean family caregivers of a person with dementia. Consistent with the existing literature [4], this study adds to the body of the knowledge that Twitter can be a platform for providing social and emotional support to the family caregivers of persons with dementia.

The substantial amount of Tweets inappropriately using dementia terms is concerning because such insensitive behaviors can hurt dementia patients and family members by promulgating stigma [2]. Providing appropriate counter interventions to promote a dementia-friendly, safe and supportive environment may be a priority for researchers. Moreover, it was remarkable to find themes focusing on non-evidence-based strategies to prevent dementia (e.g., turmeric, fine dust, gambling, drinking red wine).

Researchers and clinicians should be aware that the circulation of non-evidence-based information may delay timely diagnosis and treatment. We also found that family caregivers used Twitter to express their grief, which implies that Twitter is a potential space to support coping and recovery among Korean families after experiencing a loss.

Consistent with the findings from similar studies regarding the novel application of topic modeling for content analysis on qualitative data, we found that LDA was able to detect most of the major themes (present in more than 3% of the volume of the corpora) that were extracted via labor-intensive human coding, except one theme. At the same time, we also found that machine learning-based LDA missed several minor themes (cumulative volume = 5.89% of the corpus) identified via labor-intensive human coding. Consequently, it is relatively fair to point out that the quality of the results from the machine-driven LDA is not perfect compared to the human content analysis, given the following: 1) LDA missed major themes comprising approximately one-fifth of the corpus (18.12%); and 2) LDA also missed fine-grained, minor themes that were nonetheless clinically meaningful (e.g., elder abuse issues, financial support, emotional distress, advocacy, treatment, coping with grief and loss, resilience, prevalence).

Nevertheless, LDA is recommended for use to analyze a fairly large corpus despite the mid-level quality of its results due to its cost-effectiveness for projects with limited resources (e.g., time and budget). It is estimated that a maximum of two weeks were needed to apply LDA thoroughly with the tuning of various parameters (e.g., number of themes) to analyze a corpus size of 3,798 kilobytes of text (this consisted of 2,070 pages, 462,727 words in 11 point font size, double-spaced) while approximately three months were needed to apply qualitative content analysis (cost estimate: 20 hours X 1 analyst X 90 USD/hour = 1,800 USD per project for LDA application; vs. 120 hours X 2 coders X 30 USD/hour = 7,200 USD per project for manual content analysis).

In conclusion, the major themes detected from manual coding and LDA were broadly similar. Nevertheless, LDA lacked granularity in detecting minor themes in a Korean-language Tweet corpus mentioning dementia and Alzheimer's disease. These findings provide a foundation for developing future Twitter-based interventions for Korean dementia caregivers. Our approach can be applied to other languages.

Acknowledgments

This research was supported by the Chung-Ang University Research Grants in 2020 (PI: Lee), and US federal grant TweetS2 R01AG060929 (PI: Yoon).

References

[1] World Health Organization. Global action plan on the public health response to dementia. 2017.

[2] Brown EE, Kumar S, Rajji TK, Pollock BG, Mulsant BH. Anticipating and mitigating the impact of the COVID-19 pandemic on Alzheimer's disease and related dementias. The American Journal of Geriatric Psychiatry. 2020 (7):712-21.

[3] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. JML research. 2003;3:993-1022.

[4] Yoon S. What can we learn about mental health needs from Tweets mentioning dementia on World Alzheimer's Day?. Journal of the APNA. 2016, 22(6):498-503.