DOI: 10.1002/alz.093035

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

DEMENTIA CARE RESEARCH AND PSYCHOSOCIAL FACTORS

PODIUM PRESENTATION

DEMENTIA CARE RESEARCH (RESEARCH PROJECTS; NONPHARMACOLOGICAL)

Exploring Semantic Topics in Dementia Caregiver Tweets

Yanbo Feng[1] | Bojian Hou[1] | Ari Klein[1] | Karen O'Connor[1] | Jiong Chen[1] | Andrés Mondragón[1] | Shu Yang[1] | Graciela Gonzalez-Hernandez[2] | Li Shen[1]

[1]University of Pennsylvania, Philadelphia, PA, USA

[2]Cedars Sinai Medical Center, West Hollywood, CA, USA

Correspondence
Yanbo Feng, University of Pennsylvania, Philadelphia, PA, USA.
Email: Yanbo.Feng@Pennmedicine.upenn.edu

Abstract

Background: Caring for family caregivers of dementia patients has grown to an important topic. Social media platforms, like Twitter, provide great resources for studying the needs of caregivers. It would be beneficial to understand the caregivers' interested or concerned topics from their tweets. Meanwhile, topic modeling has become an efficient NLP tool for analyzing semantic themes within texts. Thus, in this study, we perform a novel topic model leveraging weighting strategies to analyze dementia-caregiver-related twitter data.

Method: The information about the twitter data collection is available at: https://aging.jmir.org/2022/3/e39547. In our analysis, we first performed quality control, as shown in Figure 1. Then, we deployed two word-frequency weighting strategies, *Log weighting* and *Balanced Distributional Concentration (BDC) weighting*, on the Dirichlet Multinomial Mixture (DMM) model to obtain 4 DMM variants and used them to mine semantic topics within the tweets. We compared the performance of DMM variants, LDA variants, and BERTopic variants. Topic numbers of 5, 10, 15, 20 were picked to achieve comprehensive comparison. The coherence score C_V was applied to evaluate the model performance. We then leveraged the GPT-4 to summarize the bag of words in each topic from the best model to generate interpretable themes.

Result: 224,862 tweets passed the quality control and were analyzed in our study. As shown in Figure 2, the performance evaluation demonstrated that the Log-BDC DMM has the best general performance among traditional and BERT-based models, with an averaged coherence score of 0.5648. The bags of words from the Log-BDC DMM with 10 topics were interpreted using GPT-4 due to the highest coherence score of 0.6054. The interpretation yielded 10 semantic themes, as shown in Figure 3.

Conclusion: Our investigation of the twitter data shows that the DMM model with Log and BDC strategies has promising power for mining meaningful semantic topics from short and noisy tweets. The model outperforms LDA models and the BERTopic models

while maintaining high model interpretability. This initial study on twitter data provides interesting findings that can provide better care to family caregivers and demonstrates the promise and interpretability of traditional topic modeling methods.

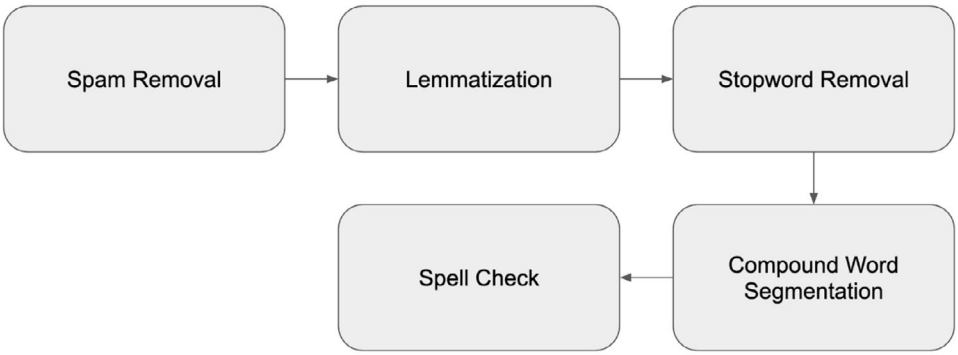

Figure 1. Quality control process of twitter data. The process is composed of spam tweet removal, lemmatization, stopword removal, compound word segmentation, and spell check. Spam Removal aims to remove junk tweets including repeated tweets, ad tweets, and web links. Lemmatization aims to inflect forms of a word so that the forms can be analyzed as a single term. Stopword removal aims to remove unimportant words from tweets. Compound word segmentation and spell check aim to decompose compound words and correct misspelled words to avoid uncommon words with low occurrence frequency.

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

	Topic Num = 5	Topic Num = 10	Topic Num = 15	Topic Num = 20
DMM	0.3423	0.3530	0.3689	0.3687
Log-DMM	0.4409	0.5266	0.5258	**0.5528**
BDC-DMM	0.4731	0.5297	0.5336	0.51621
Log-BDC-DMM	**0.5771**	**0.6054**	**0.5431**	0.5335
LDA	0.3194	0.3523	0.3713	0.3872
Log-LDA	0.3592	0.4527	0.4730	0.4579
BDC-LDA	0.3155	0.4093	0.4924	0.5199
Log-BDC-LDA	0.3138	0.4122	0.4584	0.4408
BERTopic-MiniLM	0.5743	0.4288	0.3809	0.4189
BERTopic-MPNET	0.3274	0.4385	0.4268	0.3677
BERTopic-Distl-RoBERTa	0.3009	0.3534	0.4172	0.3521

Figure 2. Topic model performance comparison on dementia-caregiver-related twitter data with number of topics equal to 5, 10, 15, 20. C_V coherence score is based on the concurrence of words and the pointwise mutual information between words. The higher coherence score represents the higher consistency of words within each topic and marks better performance of mining semantic themes in data. Generally, Log-BDC DMM overperforms traditional LDA variants, DMM variants, and BERTopic variants on mining semantic themes in the twitter data.

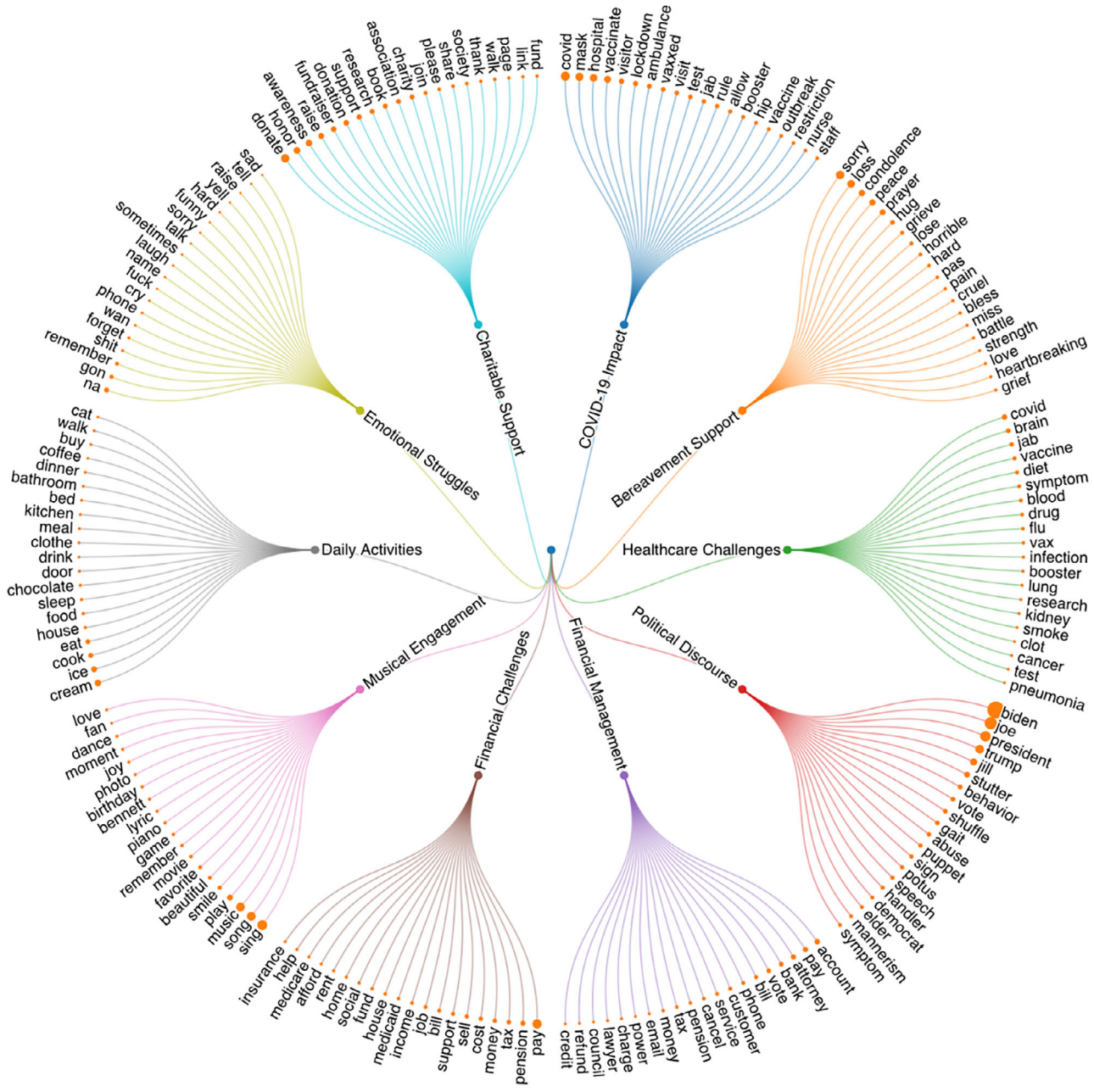

Figure 3. Semantic topic demonstration by Log-BDC DMM with number of topics equal to 10 and corresponding interpretation by GPT 4. The most important and frequent 20 words in each topic were picked for demonstration and theme interpretation and summarization. We used GPT-4 to generate a two-word summarization of each topic based on the top 20 words in each topic regarding the dementia-caregiver-related twitter data. Each cluster represents a different topic. The occurrence frequency or importance of words is marked by the size of the orange circle. The words in each topic are sorted in a descending order.