Towards a framework integrating peer networks and narratives related to Alzheimer's and Related Dementias in social media: a case study on Bluesky

Derun Xia[1[0000-0003-3980-7437]], Kayo Fujimoto[1[0000-0002-8445-2711]]

and Sahiti Myneni[2[0000-0003-3980-7437]]

[1] The University of Texas Health Science Center at Houston, School of Public Health, TX, US
Derun.Xia@uth.tmc.edu
Kayo.Fujimoto@uth.tmc.edu
[2] The University of Texas Health Science Center at Houston, McWilliams School of Biomedical Informatics, TX, USA
Sahiti.Myneni@uth.tmc.edu

Abstract. Alzheimer's Disease and Related Dementias (ADRD) pose urgent public health challenges that require targeted interventions for risk reduction and early detection. This study introduces an integrated analytical framework combining social structures with semantic clusters to examine ADRD-related discussions on the Bluesky platform, in which we combined natural language processing and network models along with manual coding. Our analysis of 26,981 posts resulted in a directed, weighted user interaction network with 133 distinct communities with high modularity (0.8811), indicating strong but fragmented structures and differences in interaction patterns. Our semantic similarity-based graph clustering revealed five largest dominant thematic clusters: delayed diagnosis, caregiver burden, early-onset dementia, fear of cognitive decline, and skepticism about medical diagnosis. Our study underscores the mechanisms through which we can leverage social networks at scale addressing knowledge barriers, care delays, and digital influence pathways on care seeking behaviors. Future implications for bipartite models and interventional targets are discussed.

Keywords: Alzheimer's Disease, Social Network Analysis, Semantic Network, Online Health Communication, Bluesky.

1 Introduction

Alzheimer's Disease and Related Dementias (ADRD) are an urgent global health priority with rising prevalence and significant societal impacts.[1]. An individual's social ties play an important role in the information seeking behaviors and care seeking behaviors that are pivotal for ADRD risk reduction and early detection. The role of online communities leads to further expansion of socially-influencing environments[2]. Social media platforms offer unique opportunities to study these underlying mechanisms by capturing digitized peer interactions as they reflect inter- and intra-personal factors and

sociobehavioral mechanisms on a scale. Prior studies have used Twitter and Facebook data to analyze dementia-related discourse and stigma with social network and semantic approaches[3–5]. These methods have allowed researchers to conduct secondary analyses of health communications and community dynamics. However, they are limited in terms of siloed perspectives on content and network context[2, 6]. By leveraging both social network structures and semantic content, researchers can take significant strides towards scalable network-based precision interventions for risky behavior modification and care coordination.

The objective of this paper is to develop and apply an integrated analytical framework that facilitates semantic awareness and network modeling to investigate ADRD-related peer discussions on the Bluesky platform, with the aim of identifying user communities, thematic topics, and their interconnections to inform public health strategies. For this purpose, we have leveraged open-source dataset from Bluesky[7], a newer decentralized platform, which offers a relatively less-curated and user-driven environment for studying ADRD conversations in a more natural context[8]. Our main contributions include, (a) development and application of a novel pipeline for transforming open-ended ("blue sky") online discourse into structured, disease-relevant social data cohorts, (b) construction of a bipartite topic-user network that links semantic framings with social contexts through user-topic mapping, and (c) identification of network-amenable targets for behavioral interventions. In the next sections of the paper, we will present our materials and methods with a detailed description of our analytical framework, followed by results, and thorough discussions on implications and limitations of our work.

2 Methods

2.1 Analytical Framework

In this paper, we present an analytical framework that integrates social network analysis and semantic analysis to examine ADRD on the Bluesky platform. The framework transforms raw, large-scale social media data into structured insights about user interactions and thematic content. We first extracted our ADRD-specific data cohort (explained below in section 2.2), constructed a directed, weighted user interaction network to capture relational patterns and identify user communities (section 2.3). In parallel, we performed semantic analysis on posts to explore latent topics and thematic clusters (section 2.4). Finally, we combined these social and semantic dimensions by building a user–post bipartite network in which users are assigned community attributes and posts carry topic attributes, enabling joint analysis of social structure and thematic content to inform precision public health technologies and behavior support interventions.

2.2 Material

The dataset used in this study comes from a publicly released Bluesky platform user behavior database[7]. Bluesky is a decentralized social platform that was originally proposed by Twitter in 2019 and began to operate independently in 2022[9]. In early 2023, Bluesky was opened to a wider range of users, and its open protocol design

attracted many users who wanted to control their own data and usage experience[9]. By the middle of the same year, the platform had established a relatively active community base. The original dataset is from the Zenodo database[7, 10]. The data were collected from mid-February 2023 to March 2024, with a total of ~ 235 million posts and about 4 million users. In addition to the content of the posts, the dataset also contains information about interactions between users, such as replies, reposts, quotes, and follows.

Only English posts are included in the analysis. In order to identify posts related to ADRD, we developed a set of 20 keywords based on online social listening, literature review, and insights from community advisory board. These keywords include common disease names (such as “Alzheimer” and "dementia"), symptom-related descriptions (such as "memory loss"), and terms related to drugs or biomarkers (such as "amyloid" and "Donepezil"). Some words (such as "apoe") are specially set to match the whole word to avoid misjudgment; other keywords are screened in a case-insensitive manner. In addition, posts containing political terms were excluded, as such content frequently appeared in unrelated or biased discussions, consistent with prior findings on the interference of political discourse in health-related social media analysis[11]. Our final ADRD data cohort consists of 26,981 posts from 13,110 users, totaling 18,830 engagement touchpoints. All original posts collectively received 869,145 likes.

2.3 Social network analysis

We constructed a directed, weighted social network to examine interaction patterns within ADRD discussions. Each node in the network represents a unique user, and each edge denotes an interaction between two users (i.e., reply, repost, or quote). To capture both the strength and type of interaction, each edge was assigned a weight reflecting the frequency of that specific interaction type. The network is directed, with edges pointing from the user who initiated the interaction (the sender) to the user whose post was engaged with (the receiver). Within this network, we analyzed the largest connect component (LCC) network by computing its node-level and network-level metrics. The node level centrality metrics includes degree, in-degree, out-degree, betweenness, closeness, eigenvector centrality, and structural constraint was also calculated to describe the redundancy in local connection[12, 13]. Besides node and edge counts, key network-level properties such as density, connectivity, average clustering coefficient, transitivity, and degree centralization are computed, along with other standard metrics.

We also applied the Clauset–Newman–Moore (CNM) algorithm to detect community structures within the largest connected component[14]. The CNM method is a hierarchical agglomerative algorithm that optimizes modularity by iteratively merging pairs of communities that result in the largest increase in modularity gain.

2.4 Semantic analysis

In parallel, we performed semantic analysis to extract and interpret the meaning of content exchanged in our ADRD data cohort. This allowed us to discover latent topics, sentiments, and relationships between concepts. We implemented a multi-step preparation pipeline (see Figure 1) through which non-English languages were removed, and

the remaining text was cleaned by removing hyperlinks, non-alphabetic characters, and stop words. We tokenized using Treebank tokenizer and then stemmed using WordNetLemmatizer.

We then used Sentence-BERT (SBERT) to generate sentence-level embeddings[15]. All-mpnet-base-v2 model was selected for our analysis, which is fine-tuned on over one billion sentence pairs and offers higher semantic accuracy than lighter models (e.g., MiniLM) [16]. We adopted a similarity threshold of 0.75 to identify semantically similar pairs of posts. This threshold was chosen based on existing literature[17] and was empirically tuned on a subset by inspecting labeled posts. We excluded edges below the threshold and removed isolated posts.

Using the resulting similarity graph, where each node represents a unique, non-duplicate post, we applied the Leiden algorithm for community detection[18]. This graph-based unsupervised clustering method groups semantically consistent posts by optimizing modularity, a measure of similarity within a community relative to separation between communities. To better understand the topics of each community, we computed a semantic centroid (average embedding) for each cluster and ranked all posts within that cluster based on their cosine similarity to the centroid. We then selected the most representative posts for each community as typical examples for manual interpretation.

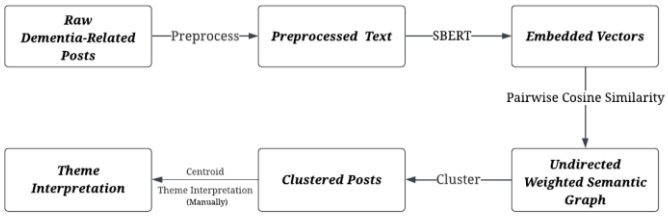

Fig. 1. Semantic Graph-Based Pipeline ADRD Social Media Posts

3 Results and Discussion

3.1 Data description

The most frequently matched keywords in our ADRD-specific Bluesky are “dementia” (43.26%), “Alzheimer” (24.08%), and “brain fog” (23.25%), which together account for over 90% of all matches. Among all selected posts, 18,830 (69.79%) were interaction posts, while the remaining 8,151 (30.21%) are individual, non-interaction posts. Among the interaction posts, replies made up the largest proportion (45.86%), followed by reposts (44.52%) and quotes (9.62%).

3.2 Social network analysis

Our analysis indicates the LCC has the most interaction with 8,438 nodes (64.37% of all nodes) and 13,302 edges (80.04% of all edges), compared to the original network's 13,109 unique nodes and 16,620 edges.

The average shortest path length is 5.83, and the network diameter is 19 in the LCC. It reflects that information can usually spread between users with only a few intermediaries, despite the sparse connectivity (density = 0.0002). The average clustering coefficient is 0.0898 and transitivity 0.0165. It indicates that users rarely form tightly connected groups. The assortativity coefficient is negative (–0.1192), implying that high-degree users tend to interact with low-degree users. The low reciprocity (0.0601) indicates that interactions are primarily unidirectional and have limited mutual engagement. These results indicate there is extensive room for capitalization of digital media space through targeted communication campaigns for ADRD risk reduction and early detection.

For key node-level centrality metrics, the average degree is 3.15, suggesting that most users have only a few connections, while a small number serve as highly connected hubs. A relatively high constraint score (mean = 0.73) implies that many users are embedded in redundant positions with limited bridging roles. A comprehensive summary of network-level and centrality metrics for the largest connected component is presented in Table 1. Based on the CNM algorithm, the modularity score is high (Modularity = 0.881), indicating a strong community structure within the network. We identify a total of 133 distinct communities. As shown in Figure 2, the top 20 largest communities account for 6,922 nodes, comprising 82.03% of all nodes in the LCC.

Table 1. Summary of Network-Level and Node-Level Metrics for the Largest Connected Component

Network-level Metric	**Value**	**Centrality Metric (mean ± std (25% – 75%))**	**Centrality Value**
Number of Nodes	8438	Degree	3.15 ± 8.07 (1.0 – 3.0)
Number of Edges	13302	In-degree	1.58 ± 7.57 (0.0 – 1.0)
Density	0.000187	Out-degree	1.58 ± 2.09 (1.0 – 2.0)
Diameter	19	Betweenness	0.0001 ± 0.0005 (0.0 – 0.0)
Radius	10	Closeness	0.0051 ± 0.0127 (0.0 – 0.0002)
Avg. Shortest Path Length	5.83	Eigenvector	0.0006 ± 0.0109 (0.0 – 0.0)
Avg. Clustering Coefficient	0.0898	Constraint	0.7283 ± 0.3287 (0.5 – 1.0)
Transitivity	0.0165	—	—
Self-Loop Proportion	0.0495	—	—
Modularity (CNM)	0.8811	—	—
Assortativity Coefficient	–0.1192	—	—
Reciprocity	0.0601	—	—
Degree Centralization	0.0286	—	—

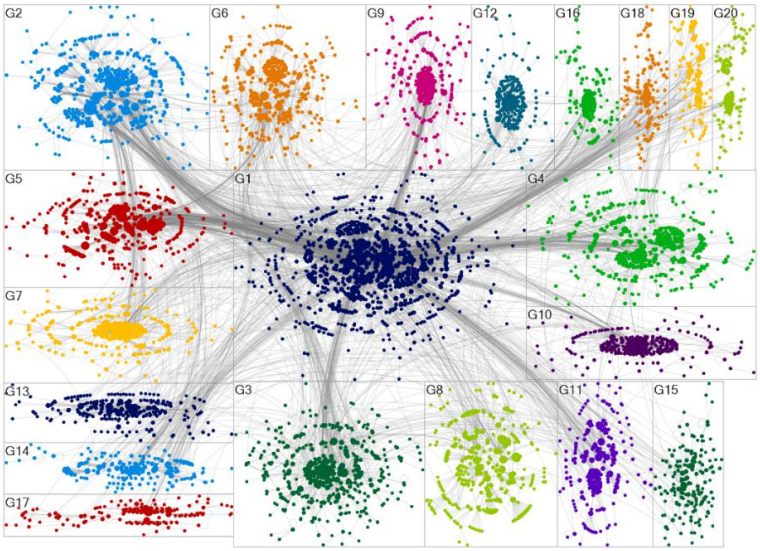

Fig. 2. Visualization of the Top 20 Largest Communities in the Largest Connected Component (LCC) of the Interaction Network

We also found clear differences in interaction types across community relationship categories. Inter-community interactions were mostly reposted (62.7%), suggesting users tend to amplify content from outside their own communities. In contrast, intra-community interactions showed more replies (40.7%) and quotes (10.8%), indicating greater conversational engagement. Outside the largest connected component (LCC), replies dominated (75.1%), while reposts (18.5%) and quotes (6.4%) are less common. The chi-squared test showed a significant difference ($p < 0.0001$) across community levels. This means users tend to communicate differently depending on which community they belong to.

3.3 Semantic analysis

In this analysis, after text cleaning and deduplication, a total of 9,985 unique and non-empty posts remained from the original 26,981 posts. The average post length was approximately 17.7 words, with lengths ranging from 1 to 56. The interquartile range was 14 words, spanning from the 25th percentile (11 words) to the 75th percentile (25 words). These results indicate moderate variability in the amount of content per post, with most posts falling within a concise length range suitable for sentence-level semantic encoding.

The modularity score of 0.76 indicated that the semantic similarity network was clearly partitioned into distinct communities. Of the 4,458 posts retained in the graph, the majority, over 64%, were concentrated within just five clusters. This pattern points to a thematic focus in the discourse, with a relatively small number of dominant topics driving much of the conversation.

Figure 3 provides a visualization the five largest semantic clusters in the network. Each node represented a unique post, and edges represented semantic similarity above the threshold. The network exhibited clear topical separation, as indicated by the sparse connections between clusters. In contrast, the dense ties within clusters pointed to tightly focused and thematically consistent discussions. For the top 5 largest clusters, Table 2 summarizes key posts and keywords selected based on similarity to each cluster's semantic centroid, reflecting themes like delayed diagnosis, caregiver burden, early-onset dementia, fear of cognitive decline, and doubt about medical labeling. These clusters revealed the many ways people voice their worries, frustrations, and doubts when talking about dementia online, highlighting the need for multi-level interventions that facilitate bridging, bonding, and linking social capital.

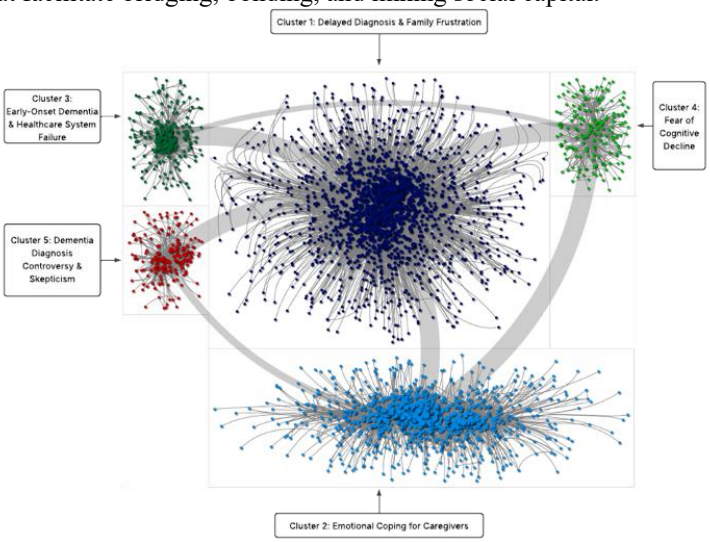

Fig. 3. Visualization of the Top 5 Clusters in ADRD discussions

Table 2. Identified Top 5 Thematic Communities in ADRD Discussions

Community	Representative Sentence	Sample Keywords
Delayed Diagnosis	*Dementia has now cruelly taken over my wee da. He just turned 70, …he couldn't get diagnosed in a timely manner.*	'delayed diagnosis', 'family', 'frustration'
Emotional Coping	*You can't change how she acts or her dementia, but you control how you handle your feelings.*	'caregiver', 'emotion','feelings'
Healthcare System Failure	*Dementia is an absolute nightmare, and it's only when someone close goes through it that you realise how utterly ineffectual and neglected our systems are.*	'early onset', 'healthcare system', 'neglect', 'failure'
Fear of Cognitive Decline	*Oh great, so I'll be developing early-onset Alzheimer's in 20 years...*	'fear', 'cognitive decline', 'future'
Dementia Diagnosis Controversy & Skepticism	*How will we know if he suffers cognitive decline?*	'skepticism', 'controversy', 'diagnosis',

3.4 Bipartite Network

Figure 4 illustrates the bipartite *user–post* network where users are depicted as blue circles while posts are shown as red squares. The overall network consists of 8,438 users and 13,550 posts, with the most active user engaging in 367 distinct posts. In the user–user projection, the highest-degree user shared posts with 244 others, with 88% being reposts. Among them, 13% of posts discussed *Healthcare System Failure*, 3–4% covered *Diagnosis Controversy* or *Fear of Cognitive Decline*, less than 1% addressed *Emotional Coping*. Meanwhile, the post–post projection reveals that the most connected post shared audiences with 466 other posts and exhibited a notably high average clustering coefficient of 0.824, indicating that posts tend to form tightly interconnected clusters characterized by substantial audience overlap.

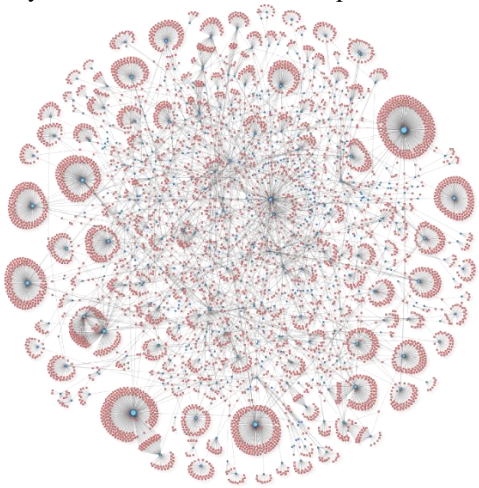

Fig. 4. Visualization of the bipartite user-post network

3.5 Findings and Discussion

Our analysis shows that combining social network and semantic approaches offers consolidated insights into large-scale ADRD discussions on the Bluesky platform. Our study focuses on the Bluesky platform, which has received very little attention in existing research, particularly within the public health domain due to its status as a nascent decentralized network with limited existing analyses and a very recent transition to public availability in 2024. Our social network analysis shows that user communities exhibit sparse interactions while remaining clearly differentiated. Several large communities form, but connections between them are limited, which may hinder broad information dissemination and create echo chambers. Similar patterns have been observed in Twitter discussions about dementia, although some studies have identified more centralized, influential communities there that can spread information more rapidly, highlighting possible platform-specific dynamics[19]. Our semantic analysis has allowed us to go beyond common surface terms like “dementia” and “Alzheimer,” allowing us to identify operational barriers related to delayed diagnosis concerns,

caregiver burden, early-onset challenges, and skepticism toward medical diagnosis. Importantly, our framework goes beyond analyzing user structures and discussion content separately by integrating them into a bipartite network. This design allows us to simultaneously capture user communities and the topics they discuss, providing richer insight into how discourse is structured and shared.

However, this study is limited to Bluesky, and expanding to other platforms could improve representativeness. Future work could also use interviews or surveys to capture keywords, themes and context that social media analyses might neglect. Additionally, our social network analysis has discounted users whose posts receive no interaction, however can still play a significant role in influence mechanisms online. For semantic analysis, we also need to consider how to further evaluate the recent advances in large language models, as well as integrate behavioral theory to develop robust and testable intervention targets.

4 Conclusions

In this work, we present a novel scalable approach to capture, analyze, and model peer interactions and their social and semantic structure. Our study highlights varied concerns, influence pathways, and experiences of individuals, indicating the complex needs that can be supported through a range of social capital interventions. Future works should focus on development and application of a targeted schema to map these insights to intervention priorities and strategies that can leverage emerging social media platforms and digital public engagement.

Acknowledgement:
Research reported in this publication was supported by the National Institute of Aging of the National Institutes of Health under award number 1R01AG089193. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Dementia, https://www.who.int/news-room/fact-sheets/detail/dementia, last accessed 2025/06/27.
2. Myneni, S., Fujimoto, K., Cohen, T.: Leveraging Social Media for Health Promotion and Behavior Change: Methods of Analysis and Opportunities for Intervention. In: Patel, V.L., Arocha, J.F., and Ancker, J.S. (eds.) Cognitive Informatics in Health and Biomedicine: Understanding and Modeling Health Behaviors. pp. 315–345. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-51732-2_15.
3. Chang, C.-Y., Hsu, H.-C.: Relationship Between Knowledge and Types of Attitudes Towards People Living with Dementia. IJERPH. 17, 3777 (2020). https://doi.org/10.3390/ijerph17113777.
4. Saha, K., Jain, Y., Liu, C., Kaliappan, S., Karkar, R.: AI vs. Humans for Online Support: Comparing the Language of Responses from LLMs and Online Communities of Alzheimer's

Disease. ACM Trans. Comput. Healthcare. 3709366 (2025). https://doi.org/10.1145/3709366.
5. Chahar, R., Dubey, A.K., Narang, S.K.: A review and meta-analysis of machine intelligence approaches for mental health issues and depression detection. IJATEE. 8, (2021). https://doi.org/10.19101/IJATEE.2021.874198.
6. Myneni Sahiti, Cobb Nathan K., Cohen Trevor: Finding Meaning in Social Media: Content-based Social Network Analysis of QuitNet to Identify New Opportunities for Health Promotion. In: Studies in Health Technology and Informatics. IOS Press (2013). https://doi.org/10.3233/978-1-61499-289-9-807.
7. Failla, A., Rossetti, G.: Bluesky Social Dataset, https://zenodo.org/doi/10.5281/zenodo.14669616, (2025). https://doi.org/10.5281/ZENODO.14669616.
8. Kleppmann, M., Frazee, P., Gold, J., Graber, J., Holmgren, D., Ivy, D., Johnson, J., Newbold, B., Volpert, J.: Bluesky and the AT Protocol: Usable Decentralized Social Media. In: Proceedings of the ACM Conext-2024 Workshop on the Decentralization of the Internet. pp. 1–7. ACM, Los Angeles CA USA (2024). https://doi.org/10.1145/3694809.3700740.
9. The AT Protocol, https://bsky.social/about/blog/10-18-2022-the-at-protocol, last accessed 2025/06/23.
10. Failla, A., Rossetti, G.: "I'm in the Bluesky Tonight": Insights from a year worth of social data. PLoS ONE. 19, e0310330 (2024). https://doi.org/10.1371/journal.pone.0310330.
11. Fuchs, J., Floor, F., Hargittai, E.: Engaging with COVID-19 content on social media in the United States: Does political affiliation matter? FM. (2023). https://doi.org/10.5210/fm.v28i11.13289.
12. Centrality in social networks conceptual clarification. Social Networks. 1, 215–239 (1978). https://doi.org/10.1016/0378-8733(78)90021-7.
13. Swedberg, R.: Review of Structural Holes: The Social Structure of Competition. Acta Sociologica. 37, 426–428 (1994).
14. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E. 70, 066111 (2004). https://doi.org/10.1103/PhysRevE.70.066111.
15. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3980–3990. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1410.
16. sentence-transformers/all-mpnet-base-v2 · Hugging Face, https://huggingface.co/sentence-transformers/all-mpnet-base-v2, last accessed 2025/06/23.
17. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, https://arxiv.org/abs/1908.10084, (2019). https://doi.org/10.48550/ARXIV.1908.10084.
18. Traag, V., Waltman, L., van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. (2018). https://doi.org/10.48550/ARXIV.1810.08473.
19. Alhayan, F., Pennington, D., Ayouni, S.: Twitter use by the dementia community during COVID-19: a user classification and social network analysis. OIR. 47, 41–58 (2023). https://doi.org/10.1108/OIR-04-2021-0208.