Mapping the Landscape of Social Media Mining for Applications in Healthcare: A Bibliometric Review

Mia Jiming Yang and Sabine Bohnet-Joschko

Chair of Management and Innovation in Healthcare
Faculty of Management, Economics and Society, Witten/Herdecke University, Witten

Awarded 3rd poster prize
Witten/Herdecke University
Research Day 2025

Introduction

By 2025, over 5 billion people worldwide use social media which host vast amounts of user-generated data[1]. Consequently, Social Media Mining (SMM) has become as an emerging field to get health information from social media data through data mining methods[2–4]. It enables real-time monitoring of health issues, serving as a supplement to traditional data analysis. Existing reviews mainly analyze isolated applications or methods, with limited exploration of correlations and temporal evolution of topics [3, 5–8]. This study addresses these gaps by a hybrid bibliometric analysis based on natural language processing (NLP) models and machine learning(ML) algorithms. The research workflow is illustrated in Figure 1.

Research Questions: What are themes of research of SMM in healthcare and their latent patterns? What are the developmental stages, strategic positioning, and future potential research themes?

We aim to provide researchers, funders, and policymakers with a systematic synthesis and strategic roadmap in SMM.

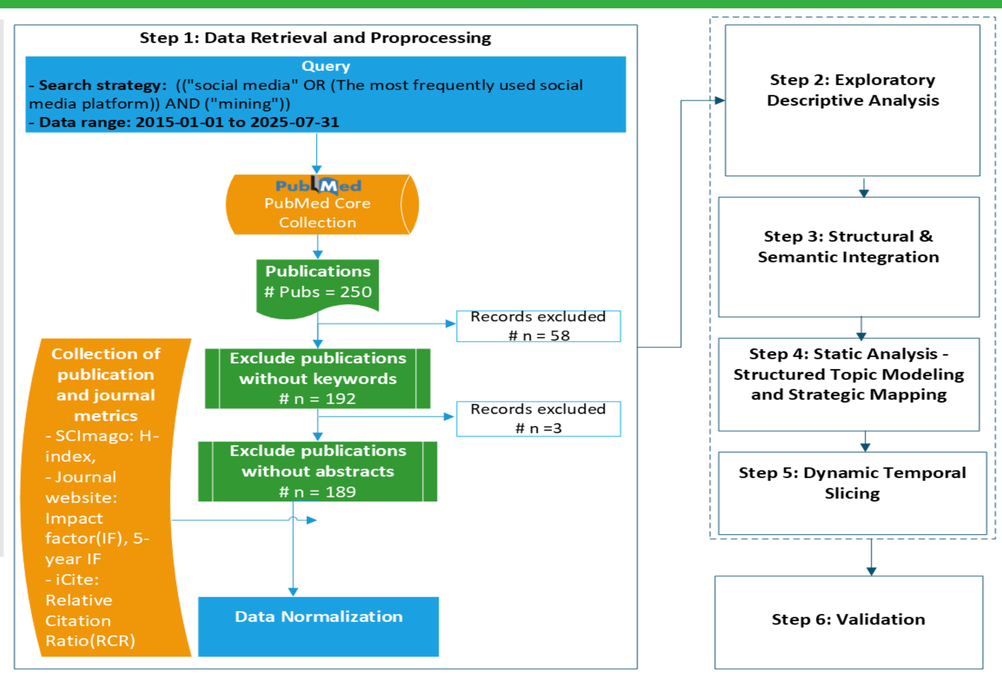


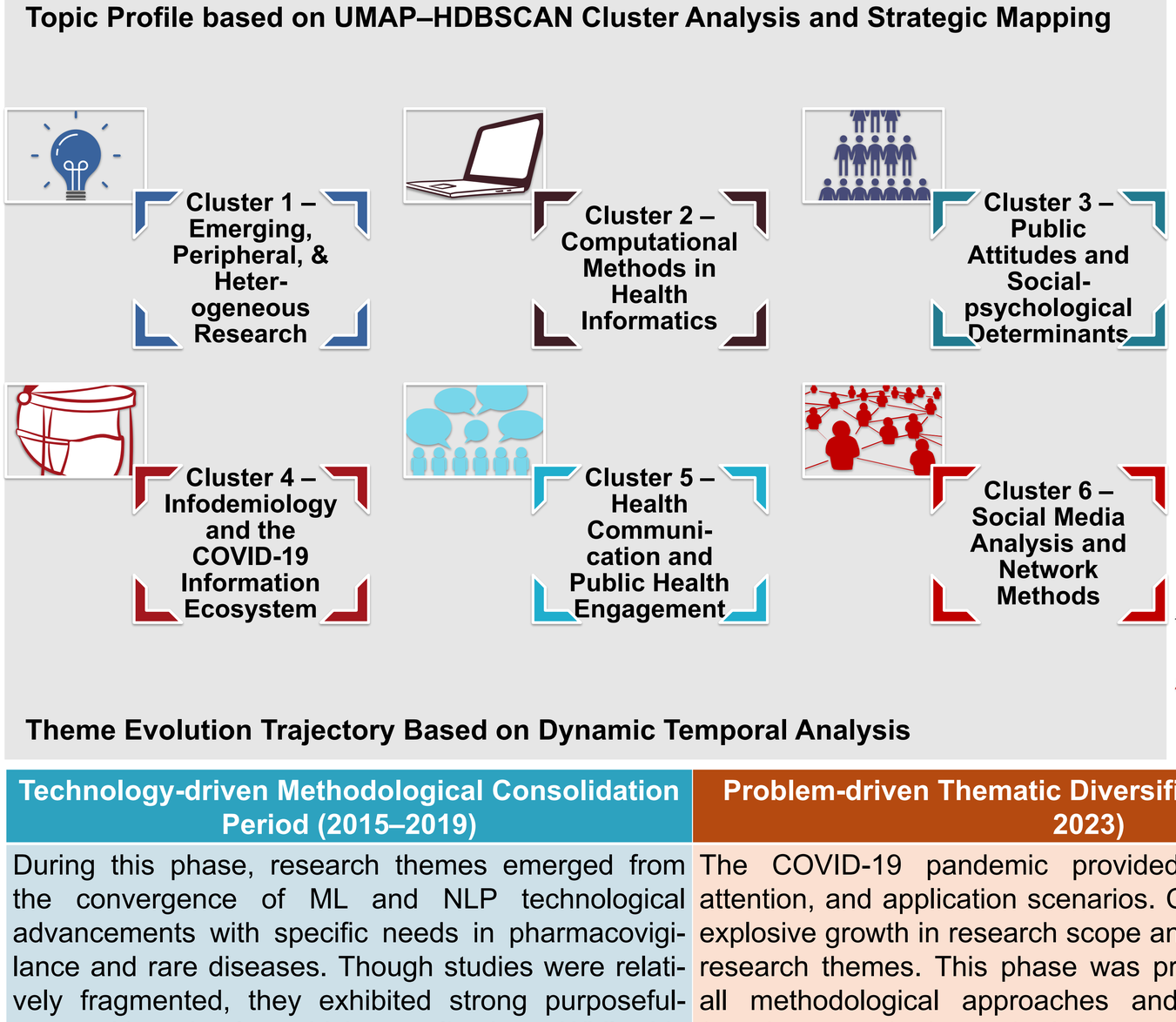
Figure. 1: Research Workflow

Results

Exploratory Descriptive Analysis

The field of social media mining for health has seen explosive growth since 2020, with publication volumes predicted to peak in 2025. The research landscape is concentrated, with the United States and China as primary hubs, and a small network of core scholars collaborating across key institutions. Foundational journals like the Journal of Medical Internet Research are central to the discourse. A keyword analysis confirms this dynamic, revealing a stable foundation of terms like ‘social media’, “NLP” and “ML” punctuated by event-driven bursts such as “COVID-19” and “vaccine hesitancy.”

Topic Profile based on UMAP–HDBSCAN Cluster Analysis and Strategic Mapping



Theme Evolution Trajectory Based on Dynamic Temporal Analysis

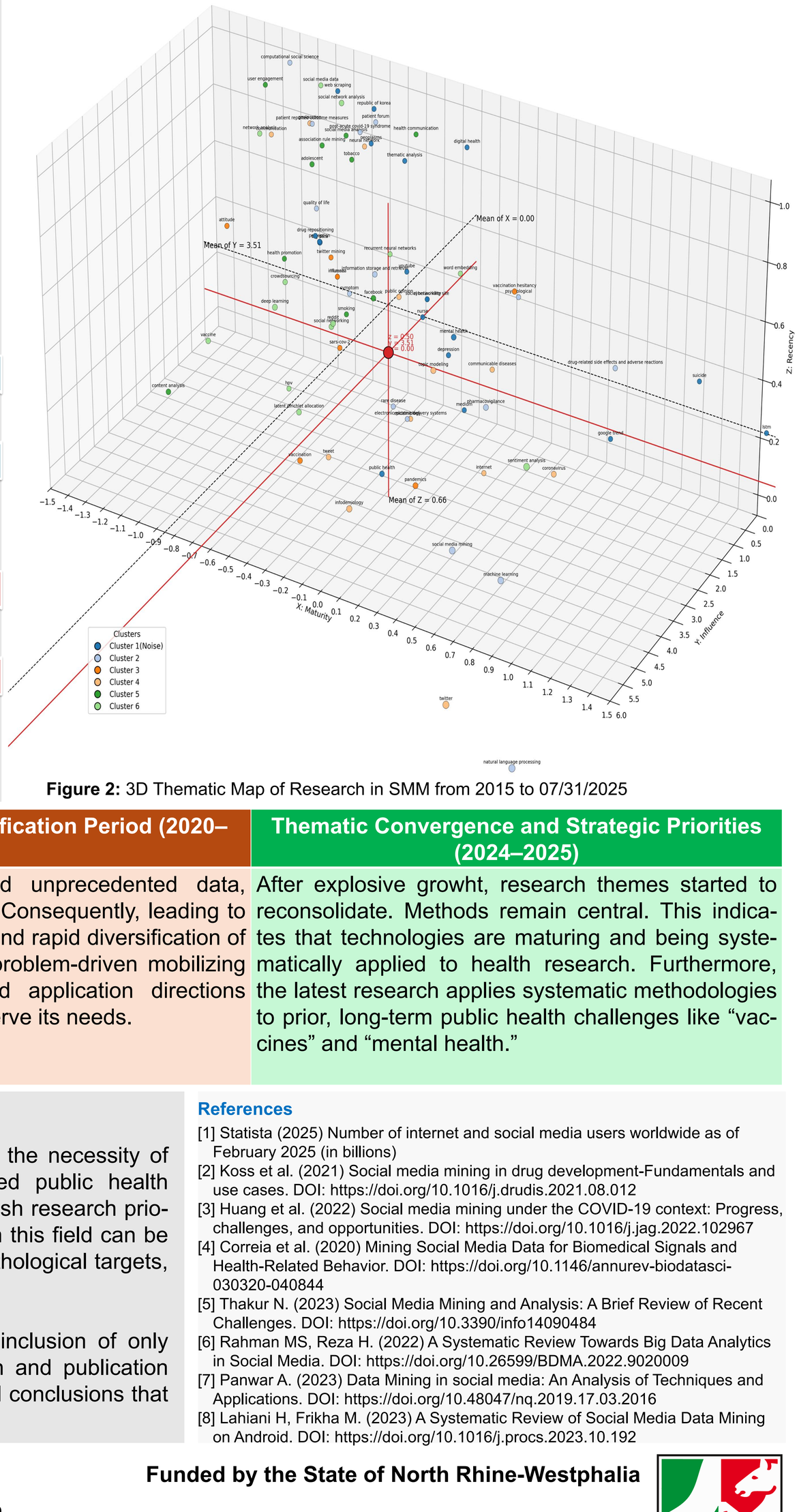


Figure 2: 3D Thematic Map of Research in SMM from 2015 to 07/31/2025

Technology-driven Methodological Consolidation Period (2015–2019)	Problem-driven Thematic Diversification Period (2020–2023)	Thematic Convergence and Strategic Priorities (2024–2025)
During this phase, research themes emerged from the convergence of ML and NLP technological advancements with specific needs in pharmacovigilance and rare diseases. Though studies were relatively fragmented, they exhibited strong purposefulness, with technology driving applications. While this marked only the foundational stage of methodology and emerging applications.	The COVID-19 pandemic provided unprecedented data, attention, and application scenarios. Consequently, leading to explosive growth in research scope and rapid diversification of research themes. This phase was problem-driven mobilizing all methodological approaches and application directions (vaccines, monitoring, attitudes) to serve its needs.	After explosive growht, research themes started to reconsolidate. Methods remain central. This indicates that technologies are maturing and being systematically applied to health research. Furthermore, the latest research applies systematic methodologies to prior, long-term public health challenges like “vaccines” and “mental health.”

Discussions

Our principal findings hold profound implications. For public health, the findings confirm the necessity of infodemiology and provide methodological support for developing audience-segmented public health communication strategies. For health research, the knowledge map can be used to establish research priorities and strategically optimize the allocation of resources. For healthcare, technologies in this field can be applied to develop real-world pharmacovigilance systems, assist in identifying potential pathological targets, and enhance understanding of patient journeys and improving medical services.

Limitations

This study has several limitations, including limited data comprehensibility due to the inclusion of only English-language literature from PubMed, time lag resulting from variations in research and publication dates, methodological reliance on keywords that may fail to capture minor distinctions, and conclusions that reveal correlations and trends rather than causal relationships.

References

[1] Statista (2025) Number of internet and social media users worldwide as of February 2025 (in billions)
[2] Koss et al. (2021) Social media mining in drug development-Fundamentals and use cases. DOI: https://doi.org/10.1016/j.drudis.2021.08.012
[3] Huang et al. (2022) Social media mining under the COVID-19 context: Progress, challenges, and opportunities. DOI: https://doi.org/10.1016/j.jag.2022.102967
[4] Correia et al. (2020) Mining Social Media Data for Biomedical Signals and Health-Related Behavior. DOI: https://doi.org/10.1146/annurev-biodatasci-030320-040844
[5] Thakur N. (2023) Social Media Mining and Analysis: A Brief Review of Recent Challenges. DOI: https://doi.org/10.3390/info14090484
[6] Rahman MS, Reza H. (2022) A Systematic Review Towards Big Data Analytics in Social Media. DOI: https://doi.org/10.26599/BDMA.2022.9020009
[7] Panwar A. (2023) Data Mining in social media: An Analysis of Techniques and Applications. DOI: https://doi.org/10.48047/nq.2019.17.03.2016
[8] Lahiani H, Frikha M. (2023) A Systematic Review of Social Media Data Mining on Android. DOI: https://doi.org/10.1016/j.procs.2023.10.192

ATLAS DIGITALE GESUNDHEITSWIRTSCHAFT

Contact: Mia Yang, mia.yang@uni-wh.de
Project lead: Prof. Dr. Sabine Bohnet-Joschko
www.atlas-digitale-gesundheitswirtschaft.de
www.uni-wh.de/mig

Funded by the State of North Rhine-Westphalia