Data and Metadata. 2025; 4:1168
doi: 10.56294/dm20251168

SYSTEMATIC REVIEW

Embedding and Topic Modeling Techniques for Short Text Analysis on Social Media: A Systematic Review

Técnicas de Embedding y Modelado de Temas para el Análisis de Textos Cortos en Redes Sociales: una revisión sistemática

Budi Warsito[1], Jatmiko Endro Suseno[2], Asa Arifudin[3]

[1]Diponegoro University, Department of Statistics, Faculty of Science and Mathematics. Semarang, Indonesia.
[2]Diponegoro University, Department of Physics, Faculty of Science and Mathematics. Semarang, Indonesia
[3]Diponegoro University, Master of Information Systems, School of Postgraduate Studies. Semarang, Indonesia.

Cite as: Warsito B, Endro Suseno J, Arifudin A. Embedding and Topic Modeling Techniques for Short Text Analysis on Social Media: A Systematic Literature Review. Data and Metadata. 2025; 4:1168. https://doi.org/10.56294/dm20251168

Submitted: 03-11-2024 **Revised:** 11-03-2025 **Aceptado:** 10-09-2025 **Publicado:** 11-09-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding Author: Asa Arifudin

ABSTRACT

Introduction: the analysis of short texts from social media is critical for gaining insights but is challenged by data sparsity and noise. Integrating embedding and topic modeling techniques has emerged as a key solution. This review aims to systematically examine recent methods that combine these techniques for short-text analysis on social media, and to evaluate their effectiveness, limitations, and research opportunities.
Method: this study conducted a Systematic Literature Review (SLR) following PRISMA guidelines. A systematic search across IEEE, ScienceDirect, and Scopus databases was performed to identify relevant studies, which were then screened and selected based on predefined inclusion and exclusion criteria.
Results: the analysis of 22 included studies revealed a clear methodological trend toward hybrid models that integrate transformer-based embeddings, such as BERT, with topic modeling frameworks. These integrated approaches consistently demonstrated superior performance in generating coherent topics and improving downstream task accuracy compared to standalone or traditional methods. However, limitations related to model generalizability, computational cost, and domain adaptation were identified.
Conclusions: the integration of contextual embeddings with topic models is the most effective approach for short-text analysis on social media. Future research should focus on developing more adaptive and efficient models, including fine-tuning language models on domain-specific corpora and exploring the integration of Large Language Models (LLMs) to enhance automation and accuracy.

Keywords: Systematic Literature Review; Topic Modeling; Word Embedding; Short-Text Analysis; Social Media; Natural Language Processing.

RESUMEN

Introducción: el análisis de textos cortos de redes sociales es fundamental para obtener conocimientos, pero se ve desafiado por la escasez de datos y el ruido. La integración de técnicas de embedding y modelado de temas ha surgido como una solución clave. Esta revisión tiene como objetivo examinar sistemáticamente los métodos recientes que combinan estas técnicas para el análisis de textos cortos en redes sociales, y evaluar su efectividad, limitaciones y oportunidades de investigación.
Método: este estudio realizó una Revisión Sistemática de la Literatura (SLR) siguiendo las directrices PRISMA. Se llevó a cabo una búsqueda sistemática en las bases de datos IEEE, ScienceDirect y Scopus para identificar estudios relevantes, los cuales fueron filtrados y seleccionados según criterios de inclusión y exclusión predefinidos.

Resultados: el análisis de los 22 estudios incluidos reveló una clara tendencia metodológica hacia modelos híbridos que integran embeddings basados en transformers, como BERT, con marcos de modelado de temas. Estos enfoques integrados demostraron consistentemente un rendimiento superior en la generación de temas coherentes y en la mejora de la precisión de tareas posteriores en comparación con métodos tradicionales o independientes. Sin embargo, se identificaron limitaciones relacionadas con la generalización de los modelos, el costo computacional y la adaptación al dominio.
Conclusiones: la integración de embeddings contextuales con modelos de temas es el enfoque más efectivo para el análisis de textos cortos en redes sociales. La investigación futura debería centrarse en desarrollar modelos más adaptables y eficientes, incluyendo el ajuste fino de modelos de lenguaje en corpus específicos del dominio y explorando la integración de Grandes Modelos de Lenguaje (LLMs) para mejorar la automatización y la precisión.

Palabras clave: Revisión Sistemática de la Literatura; Modelado de Temas; Word Embedding; Análisis de Textos Cortos; Redes Sociales; Procesamiento del Lenguaje Natural.

INTRODUCTION

Social media has emerged as a strategic channel for disseminating opinions, complaints, and customer experiences anytime and anywhere.[1] It also fosters customer engagement through satisfaction, positive emotions, and trust.[2] According to 2021 statistics from Global WebIndex, 57,6 % of the world's population uses social media. The "We Are Social" 2024 report indicates that Indonesia is home to over 200 million active social media users. As one of the most popular platforms, X (formerly Twitter) generates approximately 500 million tweets monthly from its 313 million active users.[3] Through topic analysis, this vast ocean of unstructured data can be transformed into valuable insights, enabling companies to understand their brand positioning, navigate the competitive landscape, and make strategic and tactical decisions regarding marketing communications and product introductions.[4]

In practical applications, topic modeling of data from X has been employed to understand market dynamics and public reactions to specific events. Sentiment analysis on X data, for instance, can influence stock market decisions, particularly during crises like the COVID-19 pandemic. By analyzing topics emerging from tweets, it is possible to identify sentiment patterns that can predict market movements.[5] In the healthcare sector, topic modeling is utilized to address the challenges of short texts and develop optimal classification systems to ensure the accurate dissemination of medical information and combat misinformation.[6] Furthermore, advancements in machine learning and neural-based modeling have significantly enhanced the capabilities of topic analysis on short texts. The use of pre-trained word embeddings in conjunction with topic models can yield better interpretations of the topics that emerge from texts.[6] This is particularly crucial in the context of X, where each tweet has a strict character limit, often resulting in highly concise and dense information. Given the high volume of customer interactions on X, research in this area is a critical step for maintaining competitiveness in the digital age.

Initial approaches to text representation in short-text analysis were heavily reliant on the bag-of-words (BoW) model.[6] These models, while simple to implement, have several limitations. They disregard word order and contextual information, resulting in a loss of semantic meaning and reduced accuracy, particularly in short texts where context is crucial. The advent of word embedding techniques, such as Word2Vec and GloVe, marked a significant paradigm shift.[6,7,8] These techniques represent words as vectors in a high-dimensional space, capturing semantic relationships between words based on their co-occurrence in large corpora. This enables the incorporation of contextual information, resulting in enhanced performance in tasks such as text classification and sentiment analysis. For example, a study by De Santis et al.[6] directly compared the BoW model with more advanced embedding techniques, highlighting their superior performance in classifying users in medical social media discussions. The development of contextualized word embeddings, such as those produced by BERT, ELMo, and other transformer-based models, has further advanced the capabilities of embedding techniques.[6,8,9] These models capture context-dependent word meanings, overcoming the limitations of static embeddings. The use of BERT has been explored in various studies, including research by Romero et al.[7] on analyzing the decision-making processes of engineering students, demonstrating its effectiveness in capturing nuanced topics in argumentative texts.

Concurrently, topic modeling techniques aim to discover the underlying thematic structure in a collection of documents. Early methods like Latent Semantic Analysis (LSA) relied on matrix factorization but were limited in handling polysemy and synonymy. The introduction of probabilistic topic models, such as Latent Dirichlet Allocation (LDA), provided a more robust framework.[10,11,12,13] However, LDA and similar models often require careful parameter tuning and may struggle with short texts, particularly when the word count per document

is limited. To address these limitations, researchers have explored alternative approaches, including non-negative matrix factorization (NMF)[10,13] and methods based on word embeddings.[11] The integration of topic modeling with other NLP techniques, such as aspect-based sentiment analysis, has also yielded significant advancements.[14,15]

Despite these advancements, the combined application of embedding and topic modeling techniques to short texts from social media still faces significant limitations. The inherent characteristics of this data, such as noise (e.g., slang, typos, abbreviations) and sparseness (i.e., limited context), can adversely affect the quality of semantic representations. This challenge hinders the ability to capture thematic structures and associated sentiments accurately and efficiently. Consequently, there is a clear need for a systematic investigation to identify the most effective and efficient integrated approaches, providing a clearer understanding of the current state-of-the-art and future directions.

This study aims to conduct a systematic literature review to identify the most effective integration strategies of topic modeling and embedding techniques for short text analysis on social media. To support this objective, the review also explores recent methodological trends and synthesizes best practices to inform future research and applications.

To achieve this, the research is guided by the following questions:

- What are the predominant embedding and topic modeling techniques that have been applied in recent research to address the challenges of short-text analysis?
- How effective are the various combinations of these techniques in improving the quality of analysis for short-text data characterized by uniqueness, sparseness, and noise?
- What are the existing limitations in previous studies, and what are the opportunities for the development of new, more efficient, and accurate approaches?

The research questions in this study were intentionally formulated to capture both methodological trends and qualitative insights, rather than being limited to predefined quantitative metrics. This design choice was made to allow flexibility in synthesizing evidence from a diverse body of literature, where not all studies report standardized evaluation measures such as coherence scores, accuracy, or F1-scores. While these quantitative metrics are important for assessing model performance, many relevant studies in this domain provide results in qualitative or comparative terms. Therefore, this SLR adopts a mixed perspective: retaining broad research questions to encompass both qualitative trends and quantitative outcomes, with numerical performance indicators discussed in the Results and Discussion where available.

The primary contribution of this article was to provide a comprehensive and in-depth insight into current research trends by systematically identifying, evaluating, and synthesizing existing literature. By focusing specifically on the integration of embedding and topic modeling for short, noisy texts from social media, this review addressed a critical gap. It moved beyond a simple summary of findings to offer strategic recommendations for future research and practical implementation. Therefore, this article aimed to serve as a foundational resource for the development of more sophisticated and practical Natural Language Processing (NLP) and social media analytics technologies.

METHOD

This study employeda Systematic Literature Review (SLR) methodology, rigorously guided by the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) standards. The SLR approach was utilized to systematically identify, evaluate, and synthesize the existing scholarly literature on the integration of embedding and topic modeling techniques for short-text analysis on social media.

Table 1. PICOC Framework for Embedding and Topic Modeling in Short Text Analysis

Component	Description
Population	Academic studies (peer-reviewed articles and conference proceedings) that investigate text analysis, Natural Language Processing (NLP), embedding, or topic modeling.
Intervention	The application, integration, or combination of word embedding techniques (e.g., Word2Vec, GloVe, BERT) with topic modeling methods (e.g., LDA, NMF, BERTopic).
Comparison	Integrated approaches (embedding + topic modeling) are compared against traditional or standalone methods (e.g., Bag-of-Words, standalone LDA) or against different combinations of integrated techniques.
Outcome	Insights on the most effective techniques, performance on noisy and sparse data, limitations of current methods, and identification of research gaps for future development.
Context	The domain of short-text analysis, specifically focusing on data generated from social media platforms (e.g., X/Twitter, Facebook).

Based on table 1, the PICOC framework was utilized to establish the conceptual boundaries for this review. This framework systematically defines the key inclusion criteria across the five core components—Population, Intervention, Comparison, Outcome, and Context—thereby ensuring a focused and consistent literature selection process throughout the study. The detailed, multi-stage flow of this study selection process, from initial identification to final inclusion, is visually summarized in the PRISMA 2020 diagram.

Data Sources and Search Strategy

To ensure the inclusion of high-quality and relevant literature, the following academic databases were selected:

- Scopus
- IEEE Xplore
- ScienceDirect

The search was conducted using a combination of relevant keywords and Boolean operators. Example of a typical search string: (“short text” OR “microtext”) AND (“social media”) AND (“topic modeling”) AND (“word embedding” OR “vector representation” OR “BERT” OR “Word2Vec” OR “GloVe”)

Table 2. Advanced Query on Databases

Database	Keyword	F
IEEE	(“All Metadata”:social media) AND (“All Metadata”:word embedding) AND (“All Metadata”:topic modeling) AND (“All Metadata”:short text) Only journals	50
ScienceDirect	“social media” AND (“word embedding” OR Word2Vec OR BERT) AND (“topic modeling” OR LDA OR BERTopic)	49
Scopus	TITLE-ABS-KEY ((short AND text OR microtext) AND (social AND media) AND (word AND embedding OR word2vec OR bert) AND (topic AND modeling OR lda OR bertopic))	43

Inclusion and Exclusion Criteria

As detailed in table 3, the selection criteria were rigorously defined to ensure the review’s scope was both comprehensive and methodologically sound. To maintain quality, the review was limited to formally published, English-language documents from 2018 to 2025, including peer-reviewed articles and conference proceedings, while excluding unpublished works. Most critically, thematic relevance was ensured by only including studies that substantively addressed the integration of embedding and topic modeling techniques for short-text analysis on social media.

To further explore the thematic structure and intellectual connections within the selected literature, a bibliometric network analysis was conducted using VOSviewer software.

Table 3. Inclusion and Exclusion Criteria

Criteria	Inclusion	Exclusion
Timeline	Articles published between 2018 and 2025.	Articles published before 2018.
Document Type	Journal articles and conference papers.	Books, book chapters, theses, dissertations, reports, patents, or unpublished works.
Publication Stage	Peer-reviewed and published articles.	Pre-prints, non-peer-reviewed articles, or articles still under review.
Language	Articles written in English.	Articles written in languages other than English.
Research Focus	Studies applying both embedding and topic modeling techniques for short-text analysis on social media.	Studies focusing on only one technique (either topic modeling or embedding), or those analyzing long texts.

Although studies applying only a single technique (either embedding or topic modeling alone) could provide valuable baseline comparisons or represent early research trends, they were excluded from this review to maintain a focused synthesis on hybrid approaches. This decision was made because the primary objective of this SLR is to investigate the integration of embedding and topic modeling techniques and assess their combined

effectiveness in addressing challenges such as data sparsity and noise in short-text analysis. Including single-technique studies could dilute the specificity of the review's findings. Nevertheless, relevant baseline results from such studies are discussed where appropriate to contextualize the performance improvements achieved by hybrid models

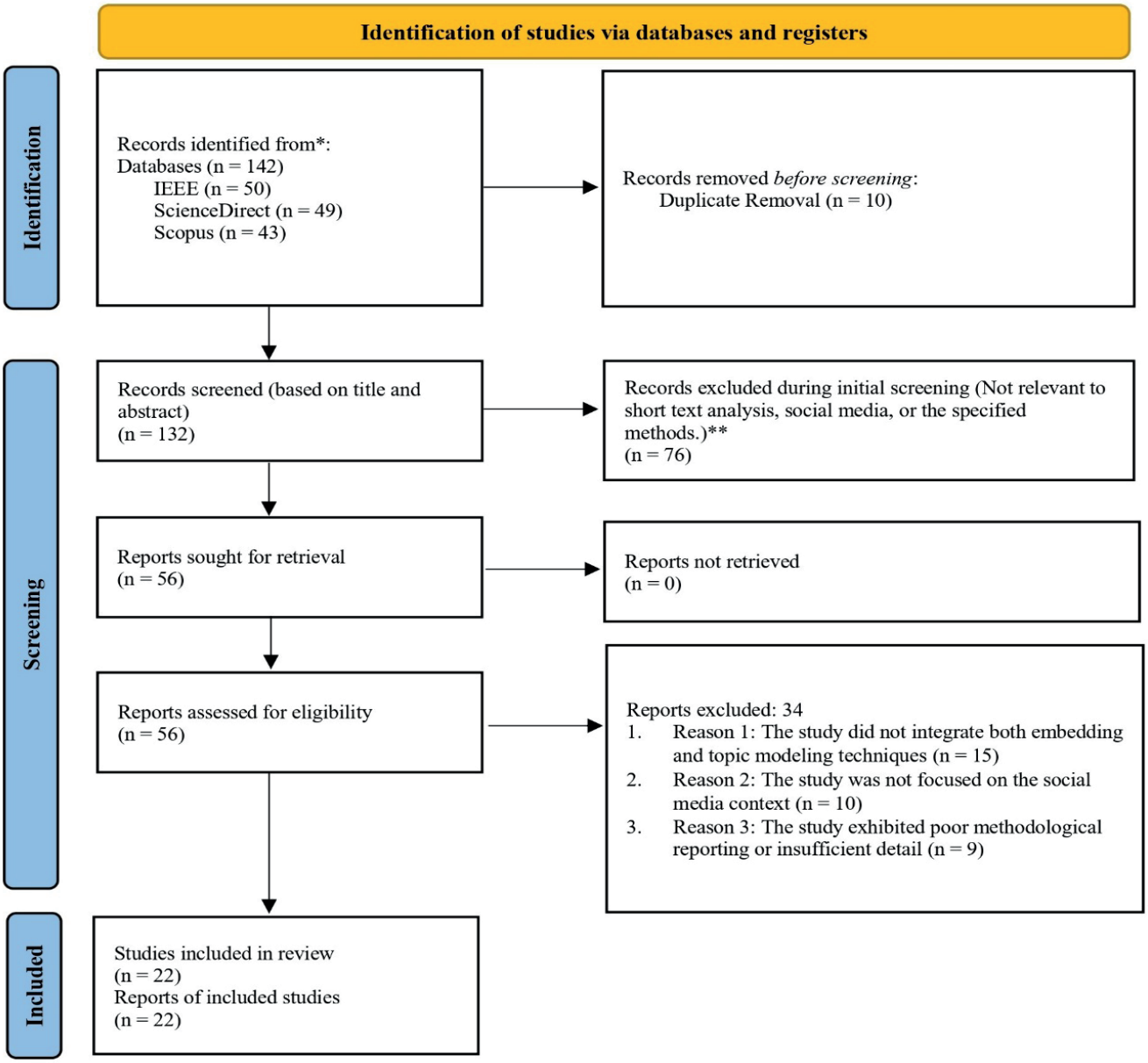

Figure 1. PRISMA 2020 Flow Diagram for Study Selection

Based on figure 1, the study identification process began with an initial pool of 142 records retrieved from three academic databases. After removing 10 duplicate records, 132 unique articles remained for the screening phase. These articles underwent a title and abstract screening, which resulted in the exclusion of 76 studies that were clearly outside the research scope. Subsequently, the full texts of the remaining 56 reports were retrieved and assessed for eligibility. During this final and most rigorous stage, 34 reports were excluded. The primary reasons for exclusion were that the study did not integrate both embedding and topic modeling techniques (n=15), the study was not focused on the social media context (n=10), or the study exhibited poor methodological reporting or insufficient detail (n=9). This systematic filtering process culminated in a final sample of 22 studies that were deemed highly relevant and were included for qualitative synthesis in this review. The specific inclusion and exclusion criteria that guided the screening and eligibility assessment stages are formally outlined in table 3.

RESULTS

This section presents the findings derived from the systematic analysis of the 22 studies selected for this review. The discussion is organized into two main parts to provide a comprehensive overview of the research landscape. It commences with a descriptive overview of the characteristics of the included studies, followed by an analysis of their bibliometric profile, including temporal and geographic distribution.

Table 4. Summary and Characteristics of Included Studies

Author(s)	Core Contribution & Technique	Application & Key Finding	Primary Limitation
Tarifa et al.[16]	Word2Vec-enhanced N-Bagger for LDA topic labeling.	Twitter; Improved F1-score for labeling.	Small dataset.
Hanifa et al.[17]	Word2Vec+DBSCAN on BERT topics for opinion analysis.	Indonesian election tweets; Captured public sentiment.	Single-language model (Indonesian).
Murthy et al.[18]	Unilink-BERT with DIB-means for serial topics.	Social media; High coherence & relevance.	Requires qualitative validation.
Liu et al.[19]	CMD-DMM dual-encoder for short text topics.	Weibo; Better topic quality & classification.	Generated labels need improvement.
Zhang et al.[20]	Graph-based framework to enhance a Topic Model (ETM).	News articles; Improved relevance & coherence.	Depends on knowledge graph quality.
Uteuov et al.[21]	Pre-trained embeddings with ARTM for topic coherence.	Russian social media; Enhanced coherence.	Complex optimization.
Zamiralov et al.[22]	ARTM with hybrid-BERT for thematic analysis.	Social/city data; Better semantic discovery.	Interpretability needs improvement.
Steuber et al.[23]	Focused FastText for short text topic clusters.	Twitter; Captured semantic context.	Needs comparison to deep learning models.
Gao et al.[24]	BERT-Sort (GMM-based) for short text semantics.	News/reviews; Captured short text semantics.	Ineffective with imbalanced topics.
Ma et al.[25]	BERT-RPC with RPST embeddings for topic dynamics.	News/discourse; Modeled topic diffusion.	Inherits BERT's cost and biases.
Murakami et al.[26]	NTM + Word2Vec to evaluate embedding size impact.	20 Newsgroups; Larger embeddings are better.	Needs validation on diverse corpora.
Diaz-Garcia et al.[27]	FastText framework for topic evolution discovery.	Election tweets; Strong semantic discovery.	Sensitive to pre-processing.
Lopreite et al.[28]	BERT-topic + UMAP for public health discussions.	COVID-19 tweets; Identified coherent topics.	Potential bias from pre-training.
Ng et al.[29]	BERT-topic + SBERT for unsupervised modeling.	Singaporean news; Found healthcare topics.	English-language news bias.
Li et al.[30]	WE-ATM integrates semantics & emotion.	Chinese short texts; Improved topic quality.	Requires better feature fusion.
Meddeb et al.[31]	Word2Vec + LDA for Indonesian short texts.	Indonesian news; High clustering accuracy.	Sensitive to topic count (k).
Zuo et al.[32]	PDSTM + FastText to handle data sparsity.	Short text corpora; Reduced data sparsity.	Computationally intensive.
Limwattana et al.[33]	DWL-DA integrates n-gram embeddings.	Amazon reviews; Outperformed LDA.	Effectiveness is context-dependent.
Nasser et al.[34]	TNAS with a Bi-LSTM encoder for topic modeling.	News/Amazon data; Improved coherence.	Could use more advanced embeddings.
Shao et al.[35]	RETM with contextualized embeddings.	Social media/news; Superior topic quality.	High computational cost.
Yu et al.[36]	UW-DMM with pre-trained Word2Vec.	Weibo/Twitter; Improved coherence & accuracy.	Sensitive to embedding quality/domain.
Sun et al.[37]	BERT + LDA for attribute classification.	Social media; High classification accuracy.	Limited generalizability.

Descriptive Overview and Characteristics of Included Studies

The initial analysis focuses on providing a comprehensive summary of the 22 articles that comprise the final sample. This foundational step synthesizes the key attributes of each study, including the specific techniques

applied, the application context and dataset, key findings on performance, and identified limitations or future opportunities. Presenting these characteristics provides a transparent and holistic map of the intellectual terrain, forming the basis for the subsequent thematic discussion that addresses the research questions. A detailed synthesis of these attributes for each included study is presented in table 4.

The analysis of embedding methods presented in table 4 shows that BERT-based approaches dominate the reviewed literature, appearing in 36,4 % of the studies. These are followed by other embeddings such as graph-based, Bi-LSTM, n-gram, or generic pre-trained embeddings (31,8 %), Word2Vec (18,2 %), and FastText (13,6 %). On the topic modeling side, LDA remains the most frequently used method (13,6 %), followed by ARTM (9,1 %). All other methods, including BERTopic variants, neural topic models, and clustering-based approaches, appear only once in the dataset, reflecting high diversity and experimentation in recent research.

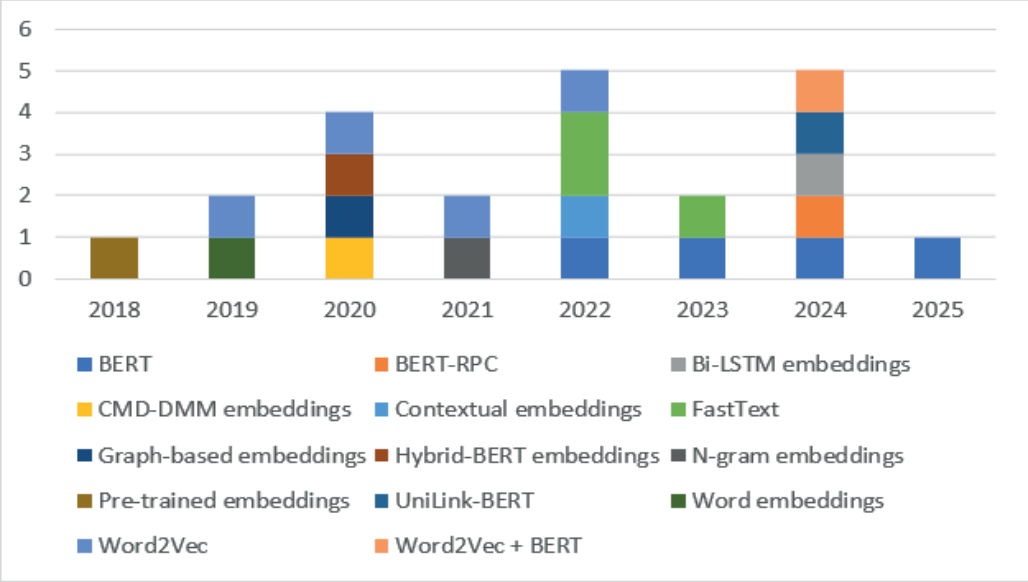

Figure 2. Trend of Embedding Methods in Hybrid Short-Text Analysis

The trend analysis in figure 2 reveals a marked increase in the adoption of transformer-based embeddings, particularly BERT, since 2022. In contrast, classical methods such as Word2Vec have maintained steady usage, especially in studies prioritizing interpretability and lower computational cost. This pattern indicates a gradual shift from static embeddings toward contextual embeddings, driven by the growing availability of computational resources.

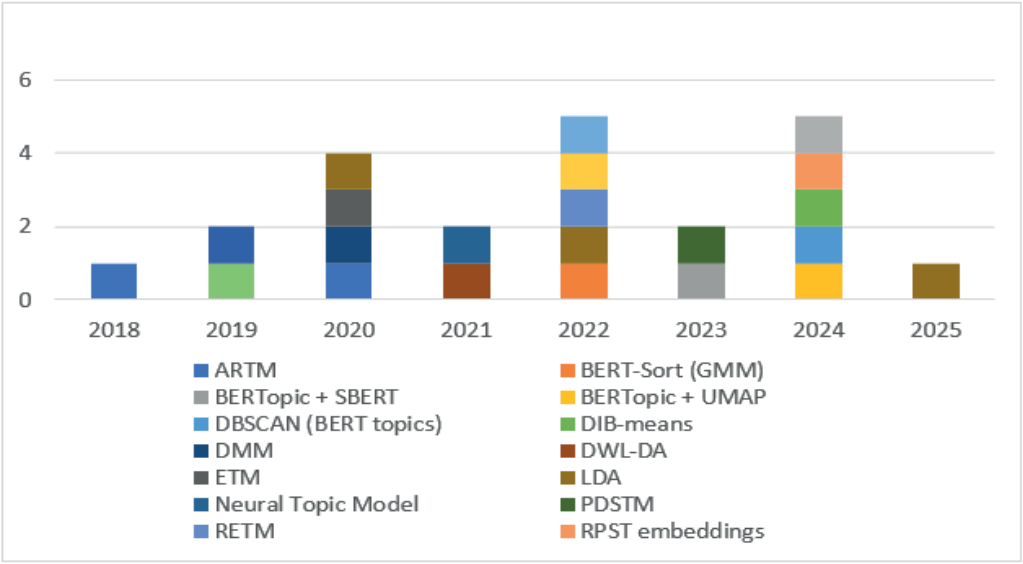

Figure 3. Trend of Topic Modeling Methods in Hybrid Short-Text Analysis

On the topic modeling side, figure 3 shows that LDA remains one of the most frequently employed methods, both as a standalone approach and in integration with embedding techniques, reflecting its stability as a

baseline for short-text topic analysis. However, embedding-driven models such as BERTopic have shown rapid growth since 2022, especially in domains that require extracting coherent topics from noisy social media data.

Although no single embedding–topic modeling combination emerged as dominant across the reviewed studies, certain pairing patterns were observed. BERT-based methods, often integrated into BERTopic or neural topic modeling frameworks, appeared most frequently in studies related to public health and COVID-19. Word2Vec combined with LDA was commonly used in studies addressing political and election-related topics, particularly on Twitter. In contrast, FastText and n-gram embeddings were more prevalent in e-commerce and product review domains. These pairings were reported consistently across different social media platforms and research objectives.

Bibliometric Profile: Temporal and Geographic Distribution of Studies

To understand the evolution and geographic scope of the research field, a bibliometric analysis of the 22 included studies was conducted. The temporal distribution of the publications is illustrated in figure 4.

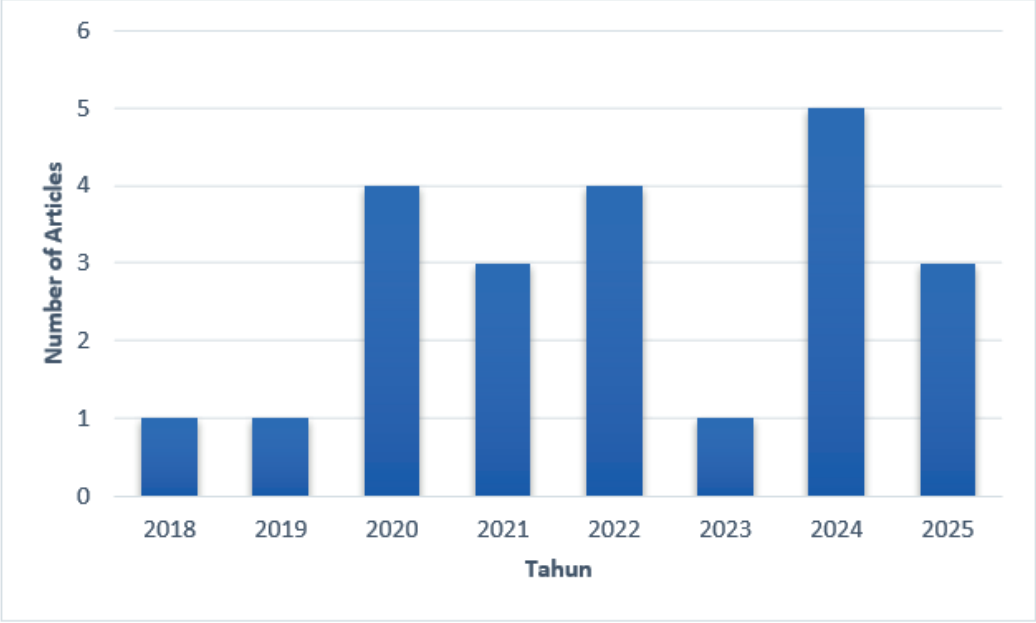

Figure 4. Distribution of Included Studies by Publication Year

Based on figure 4, the publication trend indicates a sustained and growing academic interest in the integration of embedding and topic modeling techniques. While initial studies appeared in 2018 and 2019, a significant increase in research output is observed from 2020 onwards.

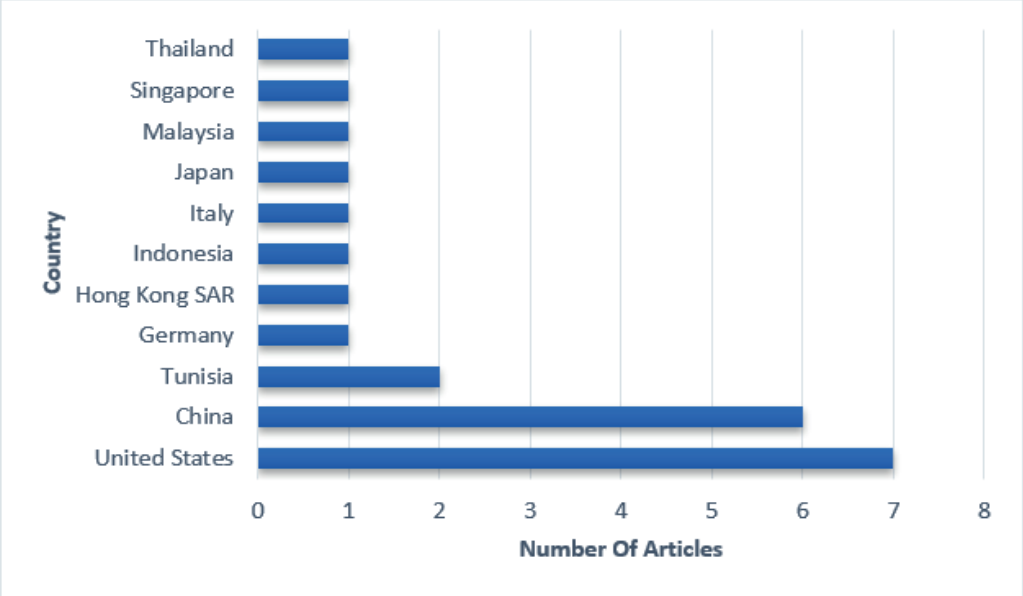

Figure 5. Geographic Distribution of Included Studies by Country of First Author

Based on figure 5, the research landscape is predominantly led by scholars from the United States (7 articles) and China (6 articles), which together account for nearly 60 % of the included studies. This concentration highlights these two countries as major hubs for innovation in this specific domain of Natural Language Processing. The remaining studies originate from a diverse range of countries across multiple continents, including Tunisia (2 articles) and several nations with single contributions, such as Germany, Indonesia, Italy, and Japan. This indicates a broad, global interest in the topic, albeit with a clear concentration of research activity in North America and East Asia.

DISCUSSION

The Methodological Landscape of Embedding and Topic Modeling for Short-Text Analysis

The analysis of short texts from social media platforms presents unique challenges due to data sparsity and noise, which has led to a clear evolution in the application of embedding and topic modeling techniques. Our review of the literature reveals a significant methodological shift from traditional probabilistic models to more sophisticated neural and transformer-based architectures. Initially, foundational techniques such as Latent Dirichlet Allocation (LDA) were adapted for short texts, often paired with classic word embedding models like Word2Vec (including its CBOW and Skip-Gram variants) and GloVe to mitigate the lack of contextual information.[31,33] Models like the Biterm Topic Model (BTM) and Dirichlet Multinomial Mixture (DMM) were also developed specifically to handle the sparsity of short texts by modeling word co-occurrences at the corpus level.[30,36]

However, the predominant trend in recent research is the integration of pre-trained transformer-based models, most notably Bidirectional Encoder Representations from Transformers (BERT) and its variants. BERT's ability to capture deep contextual relationships has made it a cornerstone of modern approaches. Studies now frequently employ BERT-based models like BERTopic, which combines BERT embeddings with clustering algorithms such as HDBSCAN, to produce highly coherent and interpretable topics from noisy social media data.[18,25,28,29] This advanced approach is often used in conjunction with established topic models like LDA, where BERT embeddings serve to enrich the input features or guide the topic discovery process.[17,22,34,37] Furthermore, researchers have developed specialized frameworks that explicitly combine neural topic models with regularized word and topic embeddings to enhance performance and interpretability.[32,35]

From our perspective, the methodological landscape has matured significantly, moving beyond the adaptation of traditional models toward the development of synergistic frameworks that combine probabilistic topic discovery with deep contextual embeddings. The widespread adoption of pre-trained transformer models, particularly BERT, reflects not only a methodological shift but also a growing consensus that conventional approaches are insufficient to address the complexity, noise, and semantic ambiguity of social media data. In recent studies, BERT-based models are no longer used merely as input encoders but are integrated as core components that guide and refine topic modeling processes. We view this transition as a critical evolution, positioning hybrid models at the forefront of short-text analysis by enabling richer, more coherent, and more context-aware topic representations.

Effectiveness of Hybrid Approaches in Mitigating Short-Text Sparsity and Noise

Performance Improvements Across Tasks: we argue that hybrid approaches, which combine embedding techniques with topic modeling, offer the most practical and scalable solution for mitigating the inherent challenges of short-text analysis, particularly data sparsity and noise. Across the reviewed literature, these models consistently outperformed their standalone counterparts in both topic coherence and downstream classification accuracy. For example, Hanifa et al.[17] reported an improvement in topic coherence (NPMI) from 0,42 to 0,61 when integrating Word2Vec with LDA for Indonesian election tweets. Similarly, Gao et al.[24] achieved a classification accuracy increase from 78 % to 85 % using BERTopic for short-text semantic clustering. Studies have demonstrated that integrating word embeddings into models like LDA, DMM, and BTM results in substantially higher topic coherence scores, a key indicator of interpretability.[20,30,33,36] The semantic relationships captured by embeddings allow models to group conceptually related terms that may not co-occur within the same short text, thus overcoming sparsity issues.[30,32] Furthermore, the effectiveness of these combinations extends to downstream tasks such as classification and clustering. Frameworks that use embeddings to guide the topic modeling process have reported superior classification accuracy and more distinct cluster formations.[32,34,35,37] Beyond word embeddings, some studies also highlight the integration of LDA with LSTM for sentiment analysis, showing that combining topic modeling with deep learning improves accuracy and interpretability in policy-related discussions.[38] Similarly, in the context of hate speech detection, a hybrid framework combining semantic expansion with BiGRU has been shown to improve the accuracy of short-text classification tasks.[39] The integration of advanced transformer-based embeddings, particularly BERT, has proven especially potent, enabling models to achieve state-of-the-art results in detecting and categorizing topics from noisy social media feeds.[17,24,25] Even when used as a pre-processing filter to select credible users based on their biographical embeddings, the quality of subsequent topic modeling is significantly enhanced.[27]

Nevertheless, we caution that performance gains often come with trade-offs. Transformer-based approaches, while powerful, introduce substantial computational demands and are prone to domain-specific bias. These limitations are frequently underreported in the literature and, in our opinion, deserve more critical attention. We emphasize the importance of evaluating hybrid approaches not only by accuracy or coherence but also by scalability, transparency, and their alignment with the specific needs of the application domain.

Domain-Specific Recommendations for Hybrid Models: based on the reviewed literature, We recommend that the selection of hybrid models be guided by the specific constraints and goals of the target domain. For instance, if interpretability is critical, simpler models such as Word2Vec + LDA remain highly viable despite the emergence of more complex transformer-based alternatives. For instance, LDA combined with Word2Vec has shown to be effective in contexts where interpretability and lower computational cost are prioritized, such as election-related sentiment analysis on Twitter with moderate dataset sizes. This combination balances semantic enrichment with relatively simple topic modeling, making it suitable for rapid deployment in resource-constrained settings. In contrast, BERTopic, which integrates transformer-based embeddings such as BERT with clustering algorithms, excels in extracting coherent and fine-grained topics from highly noisy and diverse datasets, such as health-related discussions across multilingual social media platforms. For example, Lopreite et al.[28] achieved an NPMI coherence score of 0,67 in COVID-19 Twitter topic extraction using BERTopic, outperforming LDA + Word2Vec by over 12 percentage points. However, its higher computational requirements and potential bias from pre-trained models must be considered, particularly in low-resource language contexts. As a guideline, practitioners should select LDA + Word2Vec when transparency and speed are critical, and opt for BERTopic when the goal is to capture nuanced, context-rich topics despite increased processing cost.

Critical Reflection: while these findings demonstrate the consistent benefits of hybrid approaches, it is important to note that performance can vary depending on dataset quality, language, and domain. Many studies did not report standardized evaluation metrics, making direct comparison difficult. Furthermore, the risk of model bias due to pre-trained embeddings and the computational overhead of transformer-based methods remain challenges for large-scale or real-time applications.

Recent Developments in Hybrid Short-Text Analysis: recent advances in hybrid short-text analysis provide compelling evidence of the field's accelerating methodological innovation. We observe that this progression is not merely iterative, but rather reflective of a growing need to address the limitations of traditional topic modeling in the face of increasingly complex and noisy data sources. For instance Doogan et al.[40] conducted a large-scale SLR of 189 studies applying topic models to social media short texts, highlighting persistent challenges such as data sparsity and inconsistent evaluation metrics, and providing a broader methodological backdrop against which our findings can be situated. Building on these foundations, Qin et al.[41] introduced the Position-Sensitive Word Embedding Topic Model (PS-WETM) with self-attention, demonstrating notable improvements in both coherence and classification accuracy on COVID-19 microblogs. Doi et al.[42] explored the integration of large language models (LLMs) into topic modeling workflows, showing that parallel and sequential prompting strategies can yield more coherent topics for short-text corpora.

In parallel, Eichin et al.[43] proposed Semantic Component Analysis (SCA), a method capable of discovering multiple semantic patterns beyond traditional topics, outperforming BERTopic in both coherence and scalability. Complementing these developments, Rashid[44] presented WETM, a structural topic modeling approach that leverages word embeddings to improve topic layering and interpretability for short text analysis. Taken together, we interpret these developments as clear indicators of the field's momentum toward more context-aware, semantically enriched, and structurally adaptive topic modeling techniques. In our view, the convergence of transformer-based embeddings, self-attention mechanisms, and LLM prompting represents a significant inflection point in short-text analysis. These innovations not only validate the trends identified in our review but also signal a shift toward hybrid models that are capable of understanding meaning at a deeper and more flexible level.

Limitations, Research Gaps, and Future Directions

Despite the demonstrated effectiveness of hybrid approaches, this systematic review identifies several persistent limitations within the existing literature, which in turn illuminate significant research gaps and opportunities for future work. A primary limitation is the generalizability of findings, as many studies rely on datasets from specific platforms (e.g., Twitter), languages (predominantly English), or geographic regions, which may not accurately represent the global population of social media users.[25,29,37] Methodologically, the performance of many models is highly dependent on the quality of the pre-trained embeddings, with a notable degradation in performance when there is a significant domain mismatch between the source training corpus and the target social media data.[20,26] Furthermore, current models still struggle to detect topics with very few mentions[24] and can be sensitive to noisy or ambiguous linguistic features inherent in social media content.[30,34]

These limitations highlight a clear opportunity for the development of more robust, adaptive, and efficient models. A promising future direction is the fine-tuning of pre-trained embeddings on domain-specific corpora

to bridge the semantic gap and improve model performance.[26] There is also a need for more sophisticated frameworks that can move beyond static analysis and operate in real-time or streaming environments,[27,28] as well as models that incorporate a wider range of features, such as temporal dynamics and user metadata.[21,32] Additionally, exploring probabilistic or soft assignments in guided topic models, instead of relying on crisp clustering, could provide more nuanced and flexible topic representations.[23]

From this author's perspective, the most critical research gap lies in enhancing model autonomy and adaptability. Future research should prioritize the development of end-to-end systems that require less manual parameter tuning and can dynamically adapt to the evolving language and topics on social media. The integration of advanced architectures, such as Large Language Models (LLMs), for tasks like data filtering, topic interpretation, and ambiguity resolution offers a particularly fertile ground for innovation,[25] potentially leading to a new generation of more accurate and scalable solutions for short-text analysis.

CONCLUSIONS

The integration of embedding and topic modeling techniques constitutes the most effective strategy for overcoming the inherent challenges of short-text analysis on social media. The evidence indicates that transformer-based embeddings now represent a methodological turning point, while simpler combinations such as Word2Vec–LDA remain relevant where interpretability and efficiency are critical. From the synthesis of the literature, three best practices can be abstracted: integration is indispensable for robustness, method choice must remain sensitive to domain and resource constraints, and adaptive frameworks capable of continuous learning are essential for future applications. These insights respond directly to the research objectives by providing a literature-based guide that clarifies methodological trends, highlights effective strategies, and informs both future research design and the practical implementation of NLP in social media analytics.

BIBLIOGRAPHIC REFERENCES

1. Jha B. The Role of Social Media Communication: Empirical Study of Online Purchase Intention of Financial Products. Glob Bus Rev. 2019;20(6):1445–61.

2. Yadav P, Kumar A, Shivani S, Hooda R, Sudhir S, Pooja P. Enhanced Spam Detection System for Twitter Social Networking Platform. Int J Recent Innov Trends Comput Commun. 2023;11(11s):195–201.

3. Valle-Cruz D, Fernandez-Cortez V, López-Chau A, Sandoval-Almazán R. Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods. Cognit Comput. 2022;14(1):372–87.

4. Swaminathan V, Schwartz HA, Menezes R, Hill S. The Language of Brands in Social Media: Using Topic Modeling on Social Media Conversations to Drive Brand Strategy. J Interact Mark. 2022;57(2):255–77.

5. Murakami R, Chakraborty B. Investigating the Efficient Use of Word Embedding with Neural-Topic Models for Interpretable Topics from Short Texts. Sensors. 2022;22(3).

6. De Santis E, Martino A, Ronci F, Rizzi A. From Bag-of-Words to Transformers: A Comparative Study for Text Classification in Healthcare Discussions in Social Media. IEEE Trans Emerg Top Comput Intell. 2024;9(1):1–15.

7. Romero JD, Feijoo-Garcia MA, Nanda G, Newell B, Magana AJ. Evaluating the Performance of Topic Modeling Techniques with Human Validation to Support Qualitative Analysis. Big Data Cogn Comput. 2024;8(10).

8. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Informatics Assoc. 2019;26(11):1297–304.

9. Aksoy M, Yanık S, Amasyali MF. A comparative analysis of text representation, classification and clustering methods over real project proposals. Vol. 16, International Journal of Intelligent Computing and Cybernetics. 2023. 595–628 p.

10. Egger R, Yu J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Front Sociol [Internet]. 2022 May;7(May):1–16. https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498/full

11. Odden TOB, Tyseng H, Mjaaland JT, Kreutzer MF, Malthe-Sørenssen A. Using text embeddings for deductive qualitative research at scale in physics education. Phys Rev Phys Educ Res. 2024 Dec;20(2):20151. https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.20.020151

12. Thakral P, Sharma D, Ghosh K. Evidence-based knowledge management: a topic modeling analysis of research on knowledge management and analytics. VINE J Inf Knowl Manag Syst. 2024 Jan. https://www.emerald.com/insight/content/doi/10.1108/VJIKMS-03-2023-0079/full/html

13. Walsh J, Cave J, Griffiths F. Combining Topic Modeling, Sentiment Analysis, and Corpus Linguistics to Analyze Unstructured Web-Based Patient Experience Data: Case Study of Modafinil Experiences. J Med Internet Res. 2024 Dec;26:e54321. https://www.jmir.org/2024/1/e54321

14. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res. 2021;23(2).

15. Serrano-Guerrero J, Bani-Doumi M, Chiclana F, Romero FP, Olivas JA. How satisfied are patients with nursing care and why? A comprehensive study based on social media and opinion mining. Informatics Heal Soc Care. 2024 Jan;49(1):14–27. https://doi.org/10.1080/17538157.2023.2297307

16. Tarifa A, Hedhili A, Chaari WL. A Filtering Process to Enhance Topic Detection and Labelling. Procedia Comput Sci. 2020;176:695–705. https://doi.org/10.1016/j.procs.2020.09.042

17. Hanifa A, Debora C, Hasani MF, Wicaksono P. Analyzing Views on Presidential Candidates for Election 2024 Based on the Instagram and X Platforms with Text Clustering. Procedia Comput Sci. 2024;245(C):730–9. https://doi.org/10.1016/j.procs.2024.10.299

18. Murthy D, Keshari S, Arora S, Yang Q, Loukas A, Schwartz SJ, et al. Categorizing E-cigarette-related tweets using BERT topic modeling. Emerg Trends Drugs, Addict Heal. 2024 Dec;4(October):100160. https://linkinghub.elsevier.com/retrieve/pii/S2667118224000199

19. Liu Z, Qin T, Chen KJ, Li Y. Collaboratively Modeling and Embedding of Latent Topics for Short Texts. IEEE Access. 2020;8:99141–53.

20. Zhang P, Wang S, Li D, Li X, Xu Z. Combine Topic Modeling with Semantic Embedding: Embedding Enhanced Topic Model. IEEE Trans Knowl Data Eng. 2020;32(12):2322–35.

21. Uteuov A, Kalyuzhnaya A. Combined document embedding and hierarchical topic model for social media texts analysis. Procedia Comput Sci. 2018;136:293–303. https://www.sciencedirect.com/science/article/pii/S1877050918315953

22. Zamiralov A, Khodorchenko M, Nasonov D. Detection of housing and utility problems in districts through social media texts. Procedia Comput Sci. 2020;178:213–23. https://www.sciencedirect.com/science/article/pii/S1877050920323978

23. Steuber F, Schneider S, Schoenfeld M. Embedding Semantic Anchors to Guide Topic Models on Short Text Corpora. Big Data Res. 2022;27:100293. https://www.sciencedirect.com/science/article/pii/S2214579621001106

24. Gao C, Zeng J, Wen Z, Lo D, Xia X, King I, et al. Emerging App Issue Identification via Online Joint Sentiment-Topic Tracing. IEEE Trans Softw Eng. 2022;48(8):3025–43.

25. Ma Z, Li L, Hemphill L, Baecher GB, Yuan Y. Investigating disaster response for resilient communities through social media data and the Susceptible-Infected-Recovered (SIR) model: A case study of 2020 Western U.S. wildfire season. Sustain Cities Soc. 2024;106:105362. https://www.sciencedirect.com/science/article/pii/S2210670724001902

26. Murakami R, Chakraborty B. Neural Topic Models for Short Text Using Pretrained Word Embeddings and Its Application to Real Data. 4th IEEE Int Conf Knowl Innov Invent 2021, ICKII 2021. 2021;146–50.

27. Diaz-Garcia JA, Ruiz MD, Martin-Bautista MJ. NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise. Expert Syst Appl. 2022;208:118063. https://www.sciencedirect.com/science/article/pii/S0957417422012684

28. Lopreite M, Misuraca M, Puliga M. Outbreak and integration of social media in public health surveillance systems: A policy review through BERT embedding technique. Socioecon Plann Sci. 2024 Oct;95(March):101995. https://doi.org/10.1016/j.seps.2024.101995

29. Ng QX, Lee DYX, Yau CE, Lim YL, Liew TM. Public perception on "healthy ageing" in the past decade: An unsupervised machine learning of 63,809 Twitter posts. Heliyon. 2023;9(2):e13118. https://doi.org/10.1016/j.heliyon.2023.e13118

30. Li X, Zhang A, Li C, Guo L, Wang W, Ouyang J. Relational Biterm Topic Model: Short-Text Topic Modeling using Word Embeddings. Comput J. 2019;62(3):359–72.

31. Meddeb A, Romdhane L Ben. Using Topic Modeling and Word Embedding for Topic Extraction in Twitter. Procedia Comput Sci. 2022;207(Kes):790–9. https://doi.org/10.1016/j.procs.2022.09.134

32. Zuo Y, Li C, Lin H, Wu J. Topic Modeling of Short Texts: A Pseudo-Document View with Word Embedding Enhancement. IEEE Trans Knowl Data Eng. 2021;35(1):1. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104259963&doi=10.1109%2FTKDE.2021.3073195&partnerID=40&md5=e1767abd0eb7eeecd3e1a41c98ee7731

33. Limwattana S, Prom-On S. Topic Modeling Enhancement using Word Embeddings. JCSSE 2021 - 18th Int Jt Conf Comput Sci Softw Eng Cybern Hum Beings. 2021;1–5.

34. Nasser M, Saeed F, Da'u A, Alblwi A, Al-Sarem M. Topic-aware neural attention network for malicious social media spam detection. Alexandria Eng J. 2025 Jan;111(October 2024):540–54. https://doi.org/10.1016/j.aej.2024.10.073

35. Shao W, Huang L, Liu S, Ma S, Song L. Towards Better Understanding with Uniformity and Explicit Regularization of Embeddings in Embedding-based Neural Topic Models. Proc Int Jt Conf Neural Networks. 2022;2022-July:1–9.

36. Yu J, Qiu L. ULW-DMM: An Effective Topic Modeling Method for Microblog Short Text. IEEE Access. 2019;7:884–93.

37. Sun H, Chen Y, Zhang Y. Who is to blame for AV crashes? Public perceptions of blame attribution using text mining based on social media. Comput Human Behav. 2025;168:108627. https://www.sciencedirect.com/science/article/pii/S0747563225000743

38. Wicaksono JA, Kusumaningrum R, Sediyono E. Sentiment analysis of public response to measurable fishing capture policy using LDA and LSTM methods. TELKOMNIKA Telecommun Comput El Control. 2024;22(6):1405–13. doi:10.12928/TELKOMNIKA.v22i6.25935

39. Muzakir A, Adi K, Kusumaningrum R. Short text classification based on hybrid semantic expansion and Bidirectional GRU (BiGRU) based method to improve hate speech detection. Rev Intell Artif. 2023;37(6):1471–81. doi:10.18280/ria.370611

40. Doogan Poet Laureate C, Buntine W, Linger H. A systematic review of the use of topic models for short-text social media analysis. Artif Intell Rev. 2023;56:14223–14255.

41. Qin S, Zhang M, Hu H, et al. A joint-training topic model for social media texts. Humanities Soc Sci Commun. 2025;12:281.

42. Doi T, Isonuma M, Yanaka H. Topic Modeling for Short Texts with Large Language Models. In: Proceedings of ACL Short Papers. ACL; 2024.

43. Eichin F, Schuster C, Groh G, Hedderich MA. Semantic Component Analysis: Discovering Patterns in Short Texts Beyond Topics. arXiv; 2024.

44. Rashid J. WETM: A word embedding-based topic model with structural topic representations. Expert Syst Appl. 2023.

FINANCING

This paper is funded by Diponegoro University through the Research Article Review scheme under research contract number 222-058/UN7.D2/PP/IV/2025.

CONFLICT OF INTEREST

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Budi Warsito, Asa Arifudin.
Data curation: Jatmiko Endro Suseno.
Formal analysis: Budi Warsito, Jatmiko Endro Suseno, Asa Arifudin.
Research: Budi Warsito, Jatmiko Endro Suseno, Asa Arifudin.
Methodology: Budi Warsito, Asa Arifudin.
Project management: Budi Warsito.
Resources: Asa Arifudin.
Software: Asa Arifudin.
Supervision: Budi Warsito, Jatmiko Endro Suseno.
Validation: Budi Warsito, Jatmiko Endro Suseno.
Display: Asa Arifudin.
Drafting - original draft: Budi Warsito, Jatmiko Endro Suseno.
Writing - proofreading and editing: Budi Warsito, Jatmiko Endro Suseno, Asa Arifudin.