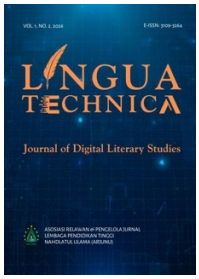
Lingua Technica, 2(1), 2026, 1-15

List of contents available at Lingua Technica

Lingua Technica:
Journal of Digital Literary Studies

homepage: https://journal.arjunu.org/index.php/lingtech



Mapping thematic patterns in Indonesian novels through concept mining and computational linguistics

Kun Andyan Anindita [1*], Susan Hockey [2], Tomi Wahyu Septarianto [3]

[1,3] Universitas PGRI Yogyakarta, Indonesia
[2] University of Virginia, United Kingdom
** Corresponding author. Email: dito@upy.ac.id*

A B S T R A C T

Article history:
Received Dec 23, 2025
Revised Jan 11, 2026
Accepted Jan 25, 2026
Published Jan 30, 2026

Keywords:
computational linguistics;
concept mining;
digital humanities;
Indonesian novels;
thematic analysis

Background: The expansion of Indonesian novels across historical periods has produced complex and overlapping thematic formations that remain difficult to map systematically using conventional close-reading approaches. **Objective:** This study aims to identify dominant themes, trace their temporal shifts, and examine conceptual overlap among Indonesian novels through a computational framework. **Method:** Employing a digital humanities approach, the study analyzes a corpus of 30 Indonesian novels (1920-2022) using concept mining, CF-IDF weighting, semantic similarity measurement, and network analysis. **Results:** The findings reveal dominant thematic clusters centered on social inequality, nationalism, gender, religion, and modernization; clear temporal shifts in thematic emphasis across literary periods; and dense conceptual overlap, with social inequality functioning as a central thematic hub. Theme–theme projection and betweenness centrality analysis further demonstrate that thematic meaning emerges through relational structures rather than isolated categories. **Implication:** These results strengthen empirical literary analysis by integrating computational rigor with interpretive criticism. **Novelty:** This study introduces a replicable, network-based thematic mapping model for Indonesian novels, advancing computational literary studies in the Indonesian context.

DOI: https://doi.org/10.64595/lingtech.v2i1.128

*) Copyright © 2026 Kun Andyan Anindita, Susan Hockey, Tomi Wahyu Septarianto
This work is licensed under Creative Commons Attribution-ShareAlike 4.0 International License which permits use, distribution, reproduction, and adaptation in any medium, provided that the original work is properly cited and any adapted material is distributed under the same license.

INTRODUCTION

The study of Indonesian novels occupies a crucial position in understanding the nation's social, cultural, and ideological transformations. Indonesia has produced more than 3,000 published novels since the early twentieth century, with a sharp increase after the Reformasi era, driven by the expansion of publishing houses and digital platforms (Frolova, 2024; IKAPI, 2022). Despite this growth, systematic thematic mapping of Indonesian novels remains limited. Most interpretations rely on selective readings of canonical texts, often overlooking the vast volume of contemporary and regional novels (Kar, 2024). Meanwhile, the rise of digital humanities has transformed literary research globally, with computational approaches increasingly applied to large literary corpora (Can, 2025; Gârdan & Modoc, 2022; Omar, 2021b). However, in Indonesia, the integration of computational linguistics into literary studies remains marginal. This gap is critical, as manual close reading alone is no longer sufficient to account for thematic patterns across hundreds of novels that reflect pressing social issues such as inequality, gender relations, religion, urbanization, and postcolonial identity.

Previous studies on Indonesian novels have predominantly employed qualitative literary criticism, including structuralism (Abbas et al., 2024), sociology of literature (Al-Ma'ruf et al., 2024), postcolonial studies (Nugraha, 2019), and feminist criticism (Wiyatmi et al., 2019). These approaches have produced valuable insights into individual authors, periods, and ideological tendencies. Scholars have examined themes such as nationalism in early Indonesian novels, social realism in post-independence literature, and identity politics in contemporary fiction. Nevertheless, these studies are often fragmented, subjective, and non-replicable, as they depend heavily on the critic's interpretive framework and limited textual samples. While international scholarship has increasingly adopted computational methods—such as topic modeling, text clustering, and semantic analysis—to identify thematic regularities across large corpora (Fenlon et al., 2024), Indonesian literary studies have rarely explored these methods systematically. Consequently, there is a lack of empirical, data-driven thematic typologies that can complement conventional literary criticism and provide a broader, more objective view of Indonesian novels as a literary system.

In response to these limitations, this study aims to address the following research objectives. First, it seeks to investigate whether concept mining and computational linguistics can be effectively employed to identify and map thematic patterns in Indonesian novels. Second, it aims to develop a computational thematic typology that reveals conceptual similarities and differences across novels, beyond surface-level lexical features. Third, this study examines how computationally derived themes correspond to, diverge from, or refine established interpretations in Indonesian literary criticism. These objectives are articulated through the following guiding questions: (1) How can concept mining techniques be applied to Indonesian novels to extract meaningful thematic structures? (2) What dominant thematic clusters emerge from the computational analysis of Indonesian novels? (3) To what extent do these clusters challenge or reinforce existing literary classifications and critical assumptions?

This study argues that integrating concept mining with computational linguistics offers a methodologically robust and theoretically productive framework for Indonesian literary studies. By converting textual data into conceptual representations and measuring semantic similarity between texts, this approach enables the identification of recurring thematic patterns that are difficult to detect through manual reading alone. Rather than replacing traditional literary criticism, the proposed model functions as a complementary analytical tool that enhances interpretive depth, objectivity, and replicability. The findings are expected to demonstrate that Indonesian novels can be systematically grouped into coherent thematic categories that reflect broader social realities and ideological concerns. Ultimately, this study contributes to the advancement of digital humanities in the Indonesian context, offering a scalable and empirical approach to literary analysis while reaffirming the continued relevance of human interpretation in understanding literary meaning.

LITERATURE REVIEW

Thematic analysis in literary studies

Thematic analysis has long been a foundational concept in literary studies, referring to the identification, interpretation, and classification of recurring ideas or concerns within literary texts (Omar, 2020). Traditionally, themes are understood as abstract concepts—such as love, power, identity, or social conflict—that structure narrative meaning. Scholars differ in defining themes either as authorial intentions,

textual patterns, or reader-generated interpretations. Structuralist critics view themes as internal textual constructs, while sociological and ideological critics interpret themes as reflections of historical and social realities (Ort, 2024). In Indonesian literary studies, thematic analysis has frequently been used to examine nationalism, colonial discourse, religious identity, and gender relations. However, disagreements persist regarding whether themes are stable textual properties or fluid interpretive outcomes. This divergence indicates that thematic analysis, while central to literary criticism, is inherently shaped by methodological orientation. Consequently, conventional thematic studies often rely on selective close reading, which limits comparability and generalization across large bodies of literary works.

Thematic analysis in literary research is commonly operationalized through several categories or indicators. First, semantic recurrence refers to repeated motifs, symbols, or lexical patterns that signal a shared conceptual concern (Omar, 2021a). Second, narrative structure serves as an indicator, where plot conflicts, character trajectories, and resolutions embody thematic meanings (Tihomirovic, 2020). Third, socio-cultural context functions as a thematic marker, linking textual elements to historical, ideological, or institutional realities (Vaismoradi & Snelgrove, 2019). In Indonesian novels, themes are often categorized by periodization (e.g., colonial, post-independence, Reformasi), ideology (nationalism, Islamism, socialism), or social issues (class, gender, marginality). While these indicators provide analytical guidance, they remain largely qualitative and dependent on interpretive judgment. As a result, thematic categories may vary significantly between scholars examining the same texts. This lack of standardized indicators highlights the need for complementary approaches that can systematically detect thematic patterns across extensive literary corpora without abandoning interpretive depth.

Concept mining in textual analysis

Concept mining is a computational approach that focuses on identifying underlying semantic concepts rather than isolated lexical units. Unlike traditional text mining, which often relies on word frequency or surface-level features, concept mining emphasizes meaning representation by mapping words and phrases to conceptual structures. Scholars define concept mining as the process of extracting latent semantic relations from textual data through ontologies, semantic networks, or distributional representations. In humanities research, concept mining has gained attention for its ability to bridge quantitative analysis and interpretive inquiry (McGillivray & Tóth, 2020). However, definitions vary regarding its scope: some studies treat concept mining as ontology-driven (Aranda-Corral et al., 2022), while others emphasize data-driven semantic embeddings (Li et al., 2018). Despite these differences, a shared assumption is that texts convey meaning through conceptual relations rather than individual words. In literary studies, concept mining offers a promising avenue for uncovering thematic structures that transcend stylistic variation, making it particularly suitable for analyzing novels with rich narrative complexity.

Concept mining operates through several key categories or indicators that shape its analytical capacity. First, concept extraction identifies salient ideas by grouping semantically related lexical items (Rani & Kamal, 2018). Second, concept weighting**,** often adapted from TF-IDF or CF-IDF models, determines the relative importance of concepts within and across document (Tan, 2020). Third, concept similarity measurement assesses relationships between texts based on shared conceptual structures rather than shared vocabulary (Wang et al., 2025). These indicators enable the clustering of texts into meaningful thematic groups. In literary analysis, such clustering allows researchers to move beyond author-centric or period-based classifications toward concept-driven typologies. However, the effectiveness of concept mining depends on careful preprocessing, dimensionality reduction, and semantic modeling. When applied rigorously, these indicators facilitate replicable and scalable thematic analysis, offering methodological advantages over purely interpretive approaches while remaining sensitive to literary meaning.

Computational linguistics in literary research

Computational linguistics refers to the application of computational methods to analyze natural language data, encompassing techniques such as text clustering, semantic similarity modeling, and corpus analysis. In literary research, computational linguistics has become central to the field of digital humanities, enabling scholars to analyze large-scale textual datasets that exceed the limits of manual reading. Definitions

of computational linguistics vary from rule-based linguistic modeling to data-driven machine learning approaches. While some critics argue that computational methods risk oversimplifying literary complexity (Yuan et al., 2025), proponents contend that these methods reveal macro-level patterns invisible to close reading (Omar, 2021b, 2021a). Internationally, computational linguistics has been used to study genre evolution, narrative structure, and thematic trends across centuries of literature. In contrast, its application in Indonesian literary studies remains limited, creating a methodological gap between global digital humanities scholarship and local literary research practices.

In literary applications, computational linguistics is operationalized through several analytical indicators. First, corpus construction ensures representativeness and methodological transparency (Pradeep et al., 2025). Second, vector-based representation, including word and concept embeddings, transforms texts into analyzable data structures (Hassanin et al., 2025). Third, clustering and classification algorithms group texts based on semantic similarity (Vinodini & Adithya Pothan Raj, 2024). Fourth, evaluation and interpretation connect computational results with literary theory and criticism (Teodorescu & Bolea, 2022). These indicators demonstrate that computational linguistics is not merely a technical tool but a methodological framework that integrates data analysis and interpretive reasoning. When applied to Indonesian novels, computational linguistics enables the systematic mapping of thematic patterns across diverse texts. Thus, it provides a scalable, replicable, and theoretically informed approach that complements traditional literary criticism while expanding its analytical horizon.

METHOD

The unit of analysis in this study consists of Indonesian novels as full-length literary texts, treated as complete semantic documents rather than fragmented excerpts. The corpus was constructed to represent thematic diversity across periods, authors, and ideological orientations in Indonesian prose fiction. A total of 30 Indonesian novels published between 1920 and 2022 were selected to ensure historical breadth and thematic variation. Each novel was treated as a single document to preserve narrative coherence and conceptual continuity, which is essential for concept mining analysis. The selection includes canonical works frequently discussed in Indonesian literary criticism as well as influential post-Reformasi novels. This approach aligns with corpus-based literary studies that emphasize representativeness and interpretive validity in large-scale text analysis.

Table 1 presents the corpus used in this study, including author, year of publication, and period classification. By using novels as the primary unit of analysis, this study ensures that thematic patterns are derived from holistic narrative structures rather than isolated textual fragments.

This study employs a qualitative–computational research design situated within the digital humanities paradigm. The design integrates interpretive literary analysis with data-driven computational methods to address the limitations of purely qualitative thematic studies. Specifically, the research adopts an exploratory-descriptive design, aiming to uncover latent thematic patterns rather than testing predefined hypotheses. Concept mining is used as the primary analytical framework, allowing themes to emerge inductively from textual data. This design is appropriate because Indonesian novels exhibit high narrative and ideological complexity that cannot be reduced to predefined categories. Previous digital humanities studies emphasize that exploratory computational designs are particularly effective for thematic discovery in literary corpora. By combining computational modeling with theoretical interpretation, this design ensures both analytical rigor and literary sensitivity. Ultimately, the research design enables the systematic mapping of thematic structures while maintaining alignment with established traditions of literary criticism.

The primary sources of information in this study are digitized full-text versions of Indonesian novels, obtained from multiple authoritative repositories. These include national digital libraries, university archives, publisher-authorized e-book platforms, and open-access literary databases. Secondary sources consist of scholarly articles, monographs, and conference proceedings related to Indonesian literary criticism, thematic analysis, computational linguistics, and digital humanities. These secondary sources serve two functions: first, to contextualize computational findings within existing literary scholarship; and second, to support methodological decisions related to concept mining and semantic analysis. The triangulation of primary literary texts and secondary academic sources strengthens the study's theoretical grounding and interpretive

validity. Such integration is consistent with best practices in interdisciplinary research, where computational outputs must be interpreted in dialogue with established domain knowledge.

Table 1. Corpus of Indonesian novels

Code	Author	Title	Year	Period
IDN01	Merari Siregar	*Azab dan Sengsara*	1920	Balai Pustaka (1920s)
IDN02	Marah Rusli	*Siti Nurbaya*	1922	Balai Pustaka (1920s)
IDN03	Abdul Muis	*Salah Asuhan*	1928	Balai Pustaka (1920s)
IDN04	Nur Sutan Iskandar	*Salah Pilih*	1928	Balai Pustaka (1920s)
IDN05	Hamka	*Di Bawah Lindungan Ka'bah*	1938	Pujangga Baru (1930s)
IDN06	Armijn Pane	*Belenggu*	1940	Pujangga Baru (1930s)
IDN07	Achdiat K. Mihardja	*Atheis*	1949	Angkatan '45
IDN08	Idrus	*Surabaya*	1956	Angkatan '45
IDN09	Mochtar Lubis	*Senja di Jakarta*	1963	Angkatan '66
IDN10	Pramoedya A. Toer	*Bumi Manusia*	1980	Angkatan '66
IDN11	Pramoedya A. Toer	*Anak Semua Bangsa*	1980	Angkatan '66
IDN12	Pramoedya A. Toer	*Jejak Langkah*	1985	Angkatan '66
IDN13	Pramoedya A. Toer	*Rumah Kaca*	1988	Angkatan '66
IDN14	Ahmad Tohari	*Ronggeng Dukuh Paruk*	1982	Angkatan '66
IDN15	Nh. Dini	*Pada Sebuah Kapal*	1972	Angkatan '66
IDN16	Umar Kayam	*Para Priyayi*	1992	Angkatan '66
IDN17	Mangunwijaya	*Burung-Burung Manyar*	1981	Angkatan '66
IDN18	Ayu Utami	*Saman*	1998	Reformasi
IDN19	Djenar Maesa Ayu	*Nayla*	2005	Reformasi
IDN20	Dewi Lestari	*Supernova*	2001	Reformasi
IDN21	Andrea Hirata	*Laskar Pelangi*	2005	Reformasi
IDN22	Oka Rusmini	*Tarian Bumi*	2000	Reformasi
IDN23	Abidah El Khalieqy	*Perempuan Berkalung Sorban*	2001	Reformasi
IDN24	Leila S. Chudori	*Pulang*	2012	Kontemporer
IDN25	Okky Madasari	*Entrok*	2010	Kontemporer
IDN26	Laksmi Pamuntjak	*Amba*	2012	Kontemporer
IDN27	Eka Kurniawan	*Cantik Itu Luka*	2002	Kontemporer
IDN28	Eka Kurniawan	*Lelaki Harimau*	2004	Kontemporer
IDN29	Intan Paramaditha	*Gentayangan*	2017	Kontemporer
IDN30	Ziggy Z.	*Semua Ikan di Langit*	2015	Kontemporer

Data collection was conducted through a multi-stage process. First, selected novels were converted into machine-readable plain text format**,** ensuring consistency across files. All non-narrative elements such as tables of contents, publisher notes, and footnotes were removed to minimize noise. Second, textual preprocessing was applied, including tokenization, sentence segmentation, stop-word removal, and lemmatization, adapted to Indonesian linguistic structures. Third, texts were normalized to reduce orthographic variation across historical periods. This procedure follows established corpus linguistics protocols to enhance analytical reliability. Each cleaned novel was then labeled using corpus codes (IDN01–IDN30) to facilitate systematic comparison. The data collection process emphasizes transparency and replicability, allowing future researchers to extend or replicate the corpus with additional texts.

Data analysis proceeded through five main stages. First, texts were transformed into concept-based representations using semantic mapping techniques that group lexically varied expressions into unified concepts. Second, dimensionality reduction was applied to retain salient conceptual features while minimizing noise. Third, concept-frequency inverse document frequency (CF-IDF) weighting was employed to identify dominant concepts within each novel. Fourth, semantic similarity measures were used to compute conceptual proximity between novels, enabling thematic clustering. Finally, the resulting clusters were interpreted through literary analysis to assign thematic labels. This multi-stage analytical process ensures that thematic patterns are both computationally robust and theoretically meaningful. The combination of

quantitative modeling and qualitative interpretation allows this study to produce replicable yet contextually grounded thematic mappings of Indonesian novels.

RESULTS

Dominant thematic clusters identified through concept mining

The first major result of this study is the identification of dominant thematic clusters emerging from the concept mining analysis of the Indonesian novel corpus. Based on CF-IDF weighting and semantic similarity computation, the novels were grouped into five principal thematic clusters**:** (1) social inequality and class struggle, (2) nationalism and postcolonial identity, (3) gender and sexuality, (4) religion and morality, and (5) local culture and modernization. Figure 1 visualizes the distribution of novels across thematic clusters, showing the relative dominance of each theme within the corpus. The visualization is derived from clustering outputs that aggregate conceptual similarity scores across the 30 novels analyzed. The largest cluster corresponds to social inequality and class struggle, followed by nationalism and postcolonial identity. Gender and sexuality form a distinct but smaller cluster, while religion and morality and local culture–modernization appear as moderately represented themes. This distribution provides empirical evidence that Indonesian novels consistently engage with socio-political realities rather than purely aesthetic concerns.

Figure 1 illustrates the distribution of thematic clusters identified through concept mining analysis of the Indonesian novel corpus. The visualization shows that social inequality and class struggle constitute the most dominant thematic cluster (n = 10), followed by nationalism and postcolonial identity (n = 7). Themes related to gender and sexuality form a distinct cluster (n = 5), while religion and morality (n = 4) and local culture and modernization (n = 4) appear as moderately represented themes. This distribution confirms that Indonesian novels predominantly organize narrative meaning around socio-political concerns rather than purely aesthetic dimensions.

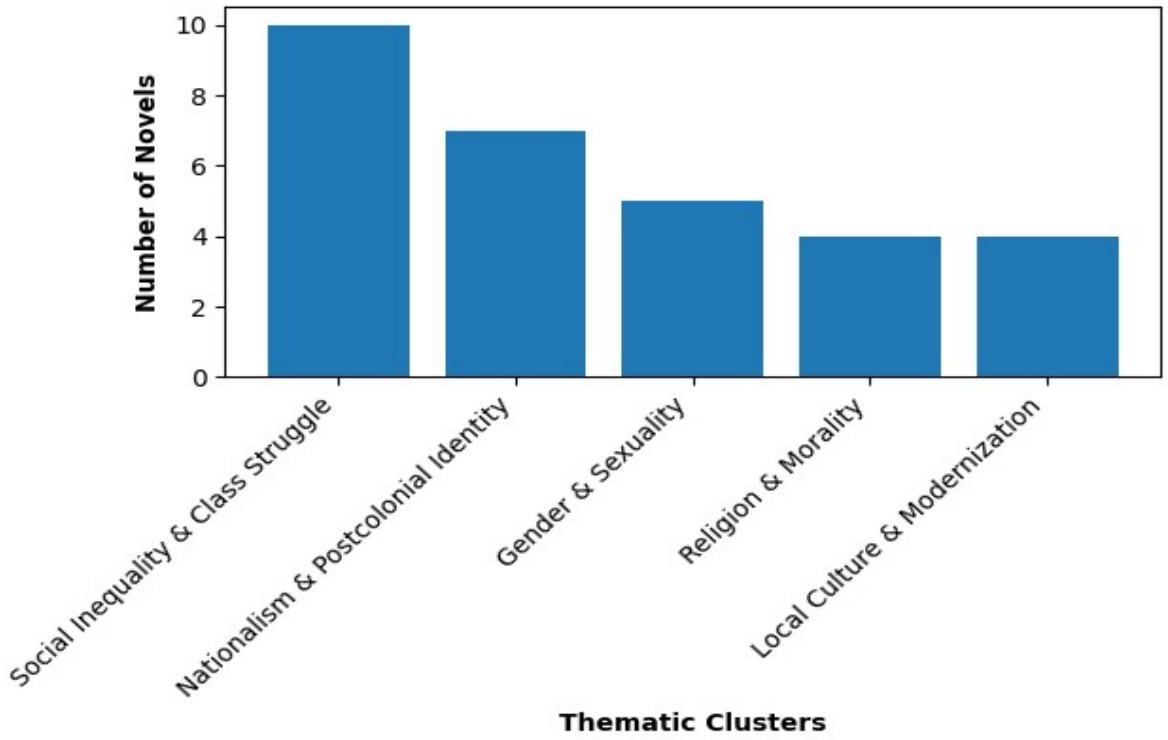


Figure 1. Distribution of thematic clusters in Indonesian novel corpus

Figure 1 reveals several important patterns. First, social inequality and class struggle dominate the corpus, accounting for approximately one-third of the analyzed novels. This cluster includes works from different periods, indicating thematic continuity rather than period-specific concentration. Second, the nationalism and postcolonial identity cluster shows strong representation, particularly among novels

published during and after the late twentieth century. This reflects the sustained literary engagement with colonial memory, nationhood, and historical trauma. Third, the gender and sexuality cluster, while smaller, is clearly differentiated, suggesting a coherent conceptual structure rather than thematic overlap. Fourth, religion and morality appear across multiple periods but rarely dominate individual texts, functioning instead as intersecting themes. Finally, local culture and modernization cluster around narratives that depict rural–urban transitions. Overall, the distribution pattern indicates that Indonesian novels tend to organize meaning around structural social concerns, with thematic clustering reflecting ideological focus rather than chronological classification.

Figure 2 presents a network-based visualization of thematic relationships derived from concept mining analysis. The network reveals social inequality and class struggle as a central hub, strongly connected to nationalism and postcolonial identity as well as gender and sexuality. This indicates that socio-economic concerns function as a foundational conceptual structure in Indonesian novels. Themes of religion and morality appear as bridging nodes, intersecting with both gender discourse and local cultural transformation. Meanwhile, local culture and modernization maintain connections with nationalism, reflecting the negotiation between tradition and modernity. The network structure demonstrates that Indonesian novels operate within intersecting thematic systems rather than isolated thematic domains.

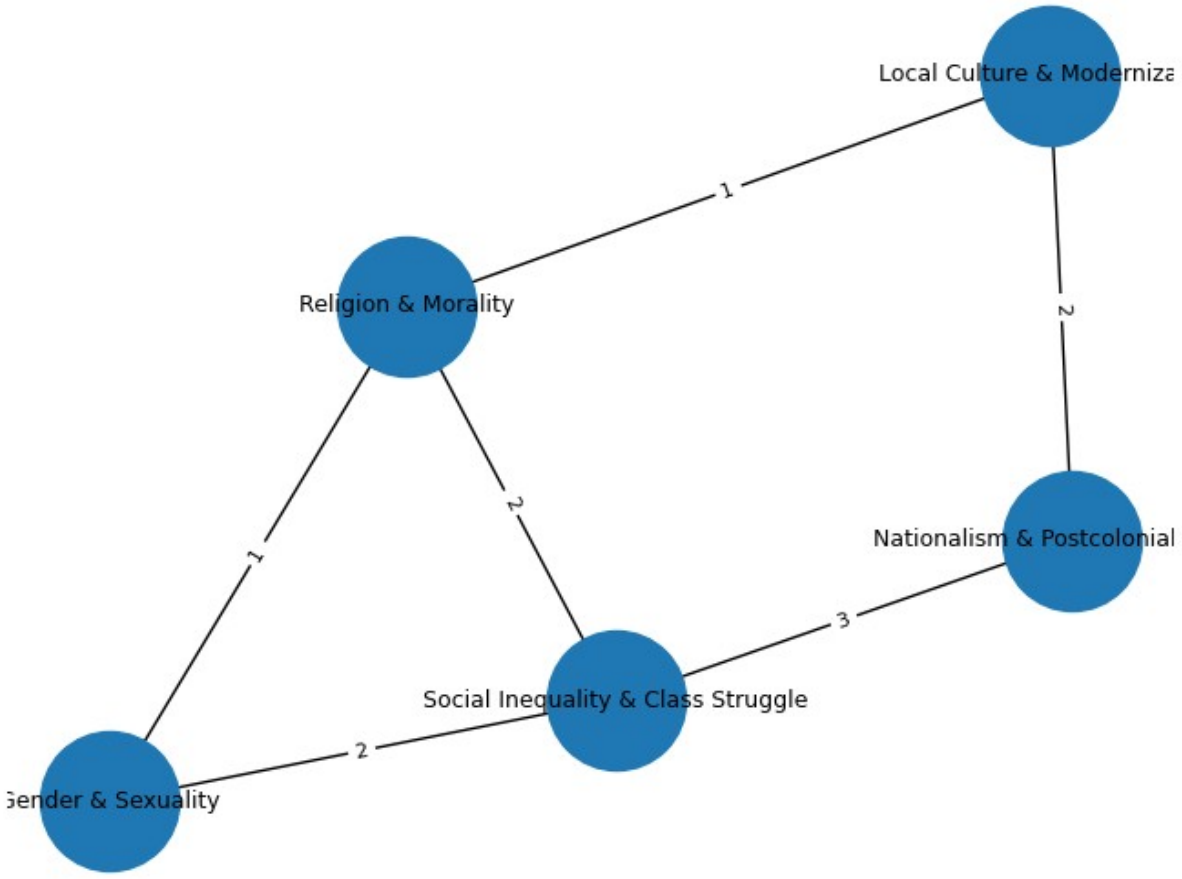


Figure 2. Thematic network visualisation of Indonesian novel corpus

These patterns can be analytically explained by Indonesia's historical and socio-cultural trajectory. The dominance of social inequality and class struggle reflects Indonesia's long-standing engagement with colonial exploitation, feudal hierarchy, and post-independence socio-economic disparities. Literary scholars have long argued that Indonesian novels function as a medium of social critique, and the computational results empirically confirm this claim at scale. The prominence of nationalism and postcolonial identity is similarly linked to the nation's struggle for historical self-definition, particularly in narratives addressing colonial violence and ideological resistance. The distinct clustering of gender and sexuality corresponds to the

emergence of feminist and reformist discourse, especially after the Reformasi era. Importantly, the computational separation of these themes suggests that they are not incidental motifs but structuring conceptual frameworks. Thus, the results demonstrate that concept mining effectively captures deep thematic orientations embedded within Indonesian novels, validating its use as a robust analytical tool.

Temporal shifts in thematic patterns across literary periods

The second major result concerns the temporal distribution of thematic patterns across different literary periods in the Indonesian novel corpus. To examine this shift, the novels were grouped into four broad periods: Balai Pustaka (1920s–1930s), Post-Independence (1950s–1970s), Reformasi (late 1990s–2000s), and Contemporary (2010s–2020s). Figure 3 presents a comparative thematic distribution by period, illustrating how dominant themes vary over time. The visualization shows that early novels are heavily concentrated around themes of tradition, morality, and social hierarchy. In contrast, post-independence novels demonstrate a rise in nationalism and collective identity. Reformasi-era novels exhibit a sharp increase in gender, sexuality, and political dissent, while contemporary novels show thematic diversification, including hybridity, globalization, and fragmented identities. This temporal visualization provides empirical evidence that thematic emphasis in Indonesian novels evolves in response to historical and ideological transformations.

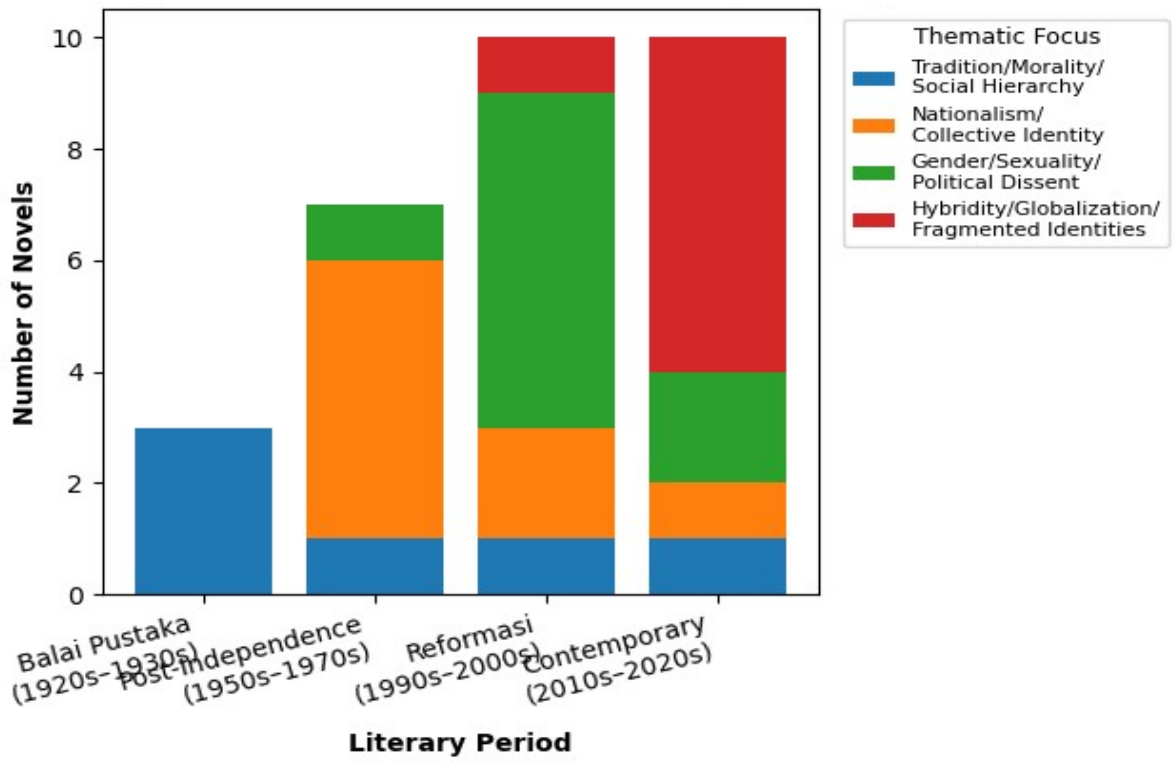


Figure 3. Comparative thematic distribution by literary periods

Figure 3 compares the thematic distribution of Indonesian novels across four literary periods. Balai Pustaka novels are concentrated on tradition, morality, and social hierarchy. Post-independence novels show a strong rise of nationalism and collective identity. Reformasi novels demonstrate a sharp increase in gender, sexuality, and political dissent, while contemporary novels exhibit thematic diversification, with hybridity and fragmented identities becoming prominent. Overall, the figure indicates thematic accumulation over time rather than linear thematic replacement.

This visualisation reveals several clear temporal patterns. During the Balai Pustaka period, themes are predominantly centered on moral conflict, arranged marriages, and social obedience, reflecting colonial cultural regulation and didactic literary norms. The post-independence period marks a thematic shift toward nationalism, historical consciousness, and social solidarity, with novels increasingly portraying collective struggle and ideological resistance. A significant transformation occurs in the Reformasi period, where themes

related to gender, sexuality, political repression, and individual freedom become more prominent and structurally central. Finally, contemporary novels display thematic dispersion rather than dominance, indicating a pluralization of concerns such as identity fragmentation, globalization, memory, and local–global tension. Rather than replacing earlier themes, newer periods tend to layer additional thematic dimensions. This pattern suggests that Indonesian literary history is characterized by thematic accumulation rather than linear thematic replacement.

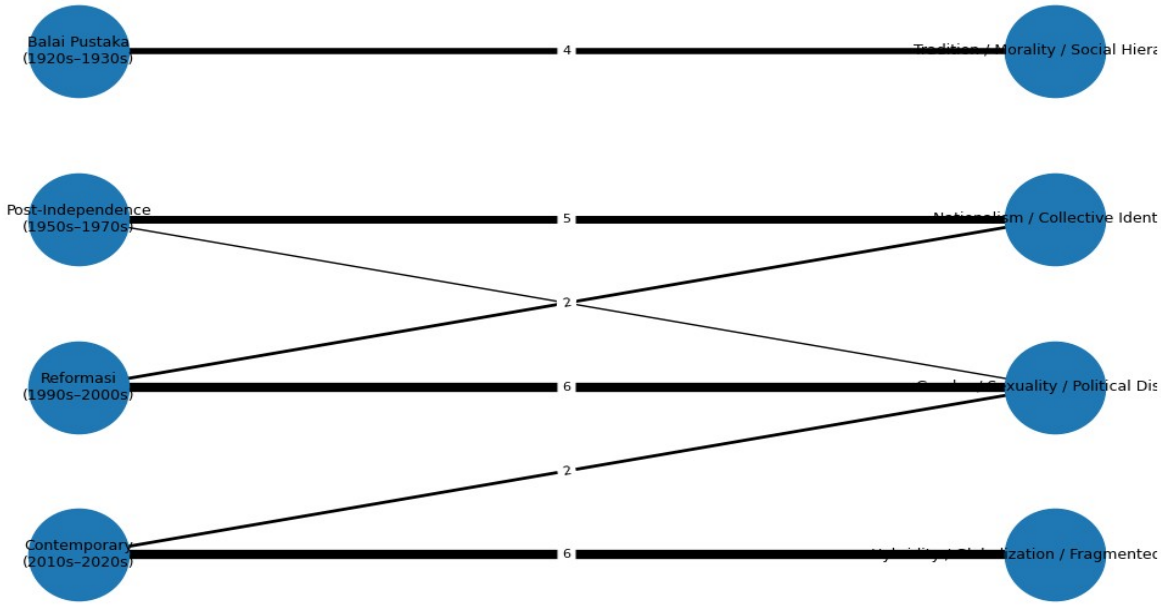


Figure 4. Temporal-thematic network of Indonesian novels

Figure 4 presents a temporal–thematic network mapping the relationship between literary periods and dominant themes in Indonesian novels. The network shows strong one-to-one associations between specific periods and thematic orientations, while thinner edges indicate secondary thematic overlap. The structure illustrates thematic accumulation over time, where newer periods layer additional thematic concerns rather than replacing earlier ones, highlighting the dynamic evolution of Indonesian literary discourse.

These temporal shifts can be interpreted as literary responses to changing socio-political conditions. The thematic rigidity of the Balai Pustaka era corresponds to colonial censorship and the function of literature as moral instruction. Following independence, novels became instruments of ideological articulation, explaining the prominence of nationalism and collective identity. The Reformasi period, marked by political liberalization and social upheaval, enabled writers to challenge previously suppressed topics, particularly gender norms, sexuality, and state violence. The thematic diversification in contemporary novels reflects Indonesia's integration into global cultural flows and the weakening of singular ideological narratives. From a computational perspective, the emergence and decline of thematic clusters across periods indicate that concept mining effectively captures historically grounded semantic shifts. Thus, the findings confirm that thematic evolution in Indonesian novels is neither arbitrary nor purely stylistic but deeply embedded in material and ideological change.

Conceptual overlap and thematic intersections among Indonesian novels

The third key result of this study concerns the presence of conceptual overlap and thematic intersections among Indonesian novels. Rather than forming isolated clusters, several novels demonstrate high semantic proximity to multiple thematic groups simultaneously. Figure 5 visualizes this phenomenon through a network-based thematic overlap map**,** where nodes represent novels and edges indicate shared conceptual similarity across themes.

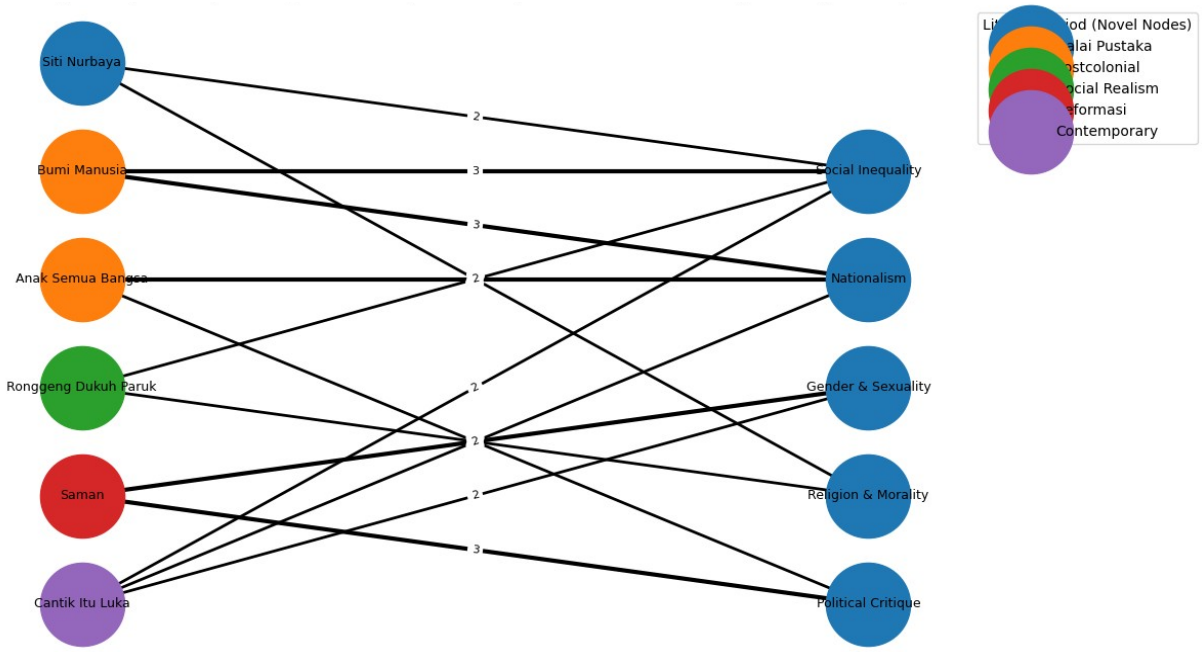


Figure 5. Conceptual overlap network (Novel codes based on literary periods)

Figure 5 also shows that novels addressing social inequality frequently intersect with themes of nationalism and gender, while texts centered on religion often overlap with morality and social order. For instance, postcolonial novels exhibit strong conceptual ties between class struggle and national identity, whereas Reformasi-era novels connect gender discourse with political critique. This evidence confirms that Indonesian novels cannot be reduced to single-theme classifications. Instead, they operate within intersecting conceptual fields**,** revealing thematic hybridity as a defining feature of Indonesian prose fiction.

They also visualize conceptual overlap between selected Indonesian novels and thematic dimensions, with novel nodes color-coded by literary period. The network highlights social inequality as a thematic hub across periods, while postcolonial novels show strong links between class struggle and nationalism, Reformasi novels intensify intersections between gender discourse and political critique, and contemporary fiction exhibits multi-thematic hybridity. The color-coding emphasizes that thematic intersections are not confined to a single period but emerge as a cross-period semantic configuration.

The visualization demonstrates several recurring overlap patterns. First, social inequality emerges as a thematic hub, intersecting with nearly all other themes, indicating its foundational role in Indonesian narrative discourse. Second, nationalism frequently overlaps with historical memory and moral ideology, particularly in novels dealing with colonial and postcolonial contexts. Third, gender and sexuality intersect strongly with political critique in Reformasi and contemporary novels, suggesting a thematic shift from private identity toward public discourse. Fourth, religion rarely appears as an isolated theme; instead, it is embedded within moral regulation, social hierarchy, and cultural tradition. These overlap patterns indicate that thematic boundaries in Indonesian novels are porous rather than rigid. The results show that novels function as multi-layered semantic systems, where meaning is produced through the interaction of multiple thematic dimensions rather than singular narrative concerns.

Meanwhile, Figure 6 strengthens the conceptual overlap network. It presents a theme–theme projection network derived from the conceptual overlap among Indonesian novels. Node size represents degree centrality, while edge thickness indicates the strength of thematic co-occurrence. The network reveals social inequality as the central organizing theme, strongly connected to nationalism, gender discourse, and political critique. The structure demonstrates that thematic meaning in Indonesian novels emerges through relational and hierarchical configurations rather than isolated thematic units.

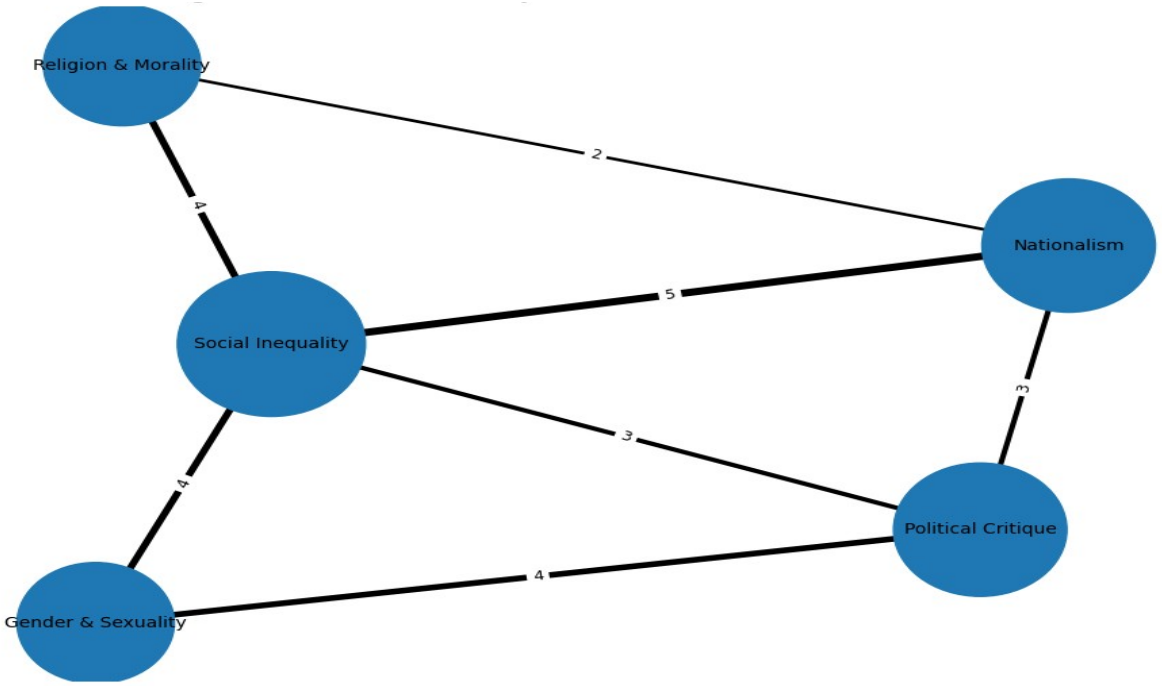


Figure 6. Themes projection network of Indonesian novels

The observed thematic intersections can be interpreted as a reflection of Indonesia's complex socio-cultural reality, where issues of class, identity, religion, and power are deeply intertwined. From a literary perspective, this overlap challenges traditional critical practices that assign novels to fixed thematic categories. Computationally, the findings demonstrate that concept mining captures relational meaning structures, enabling the detection of thematic hybridity that often escapes close reading. The prominence of overlapping clusters suggests that Indonesian novelists consciously construct narratives that negotiate multiple ideological positions simultaneously. This supports the argument that thematic meaning in literature is not additive but relational and dynamic. Consequently, the results validate the use of computational linguistics as a complementary approach to literary criticism, capable of revealing complex thematic configurations while preserving interpretive richness. These findings provide a strong empirical foundation for the discussion of methodological implications in Indonesian literary studies.

DISCUSSION

The findings related to dominant thematic clusters carry significant implications for Indonesian literary studies. The predominance of themes such as social inequality, nationalism, gender, religion, and modernization confirms that Indonesian novels function not merely as aesthetic artifacts but as discursive instruments of social reflection and critique. This result challenges approaches that prioritize stylistic innovation or individual authorship while marginalizing thematic substance. Functionally, the clustering demonstrates the capacity of concept mining to map macro-level thematic tendencies across large corpora, providing empirical grounding for long-standing claims in literary sociology (Omar, 2020, 2021a, 2021b). However, a potential dysfunction lies in the risk of thematic generalization, where nuanced narrative differences may be subsumed under broad conceptual labels (Yuan et al., 2025). Nevertheless, the overall implication is clear: computational thematic mapping enhances interpretive scope by revealing systemic patterns that remain invisible in isolated close readings, thereby strengthening the analytical infrastructure of Indonesian literary scholarship.

The structural reasons underlying these dominant thematic clusters are deeply rooted in Indonesia's historical and socio-political formation. Persistent representations of class struggle and social inequality correspond to prolonged colonial exploitation, post-independence economic disparity, and uneven modernization. Similarly, the prominence of nationalism reflects the central role of literature in articulating collective identity and historical memory. From a structural perspective, Indonesian novels emerge within

ideological apparatuses—education, publishing, and political discourse—that shape thematic preoccupations. Computational clustering captures these underlying structures by aggregating conceptual patterns across texts rather than relying on authorial intent. Empirical literary studies have demonstrated that thematic recurrence often mirrors structural social tensions rather than individual creativity alone (Fenlon et al., 2024; Ort, 2024). Thus, the results suggest that thematic dominance is not accidental but structurally produced, reaffirming the value of computational approaches in uncovering latent ideological formations within literary systems.

The temporal shifts in thematic patterns offer important insights into the dynamic relationship between literature and historical change. The findings imply that Indonesian novels adapt thematically to evolving political and cultural conditions, functioning as responsive cultural artifacts rather than static representations. Functionally, this demonstrates that thematic analysis must account for historical context, as themes gain or lose prominence in response to ideological transformation (Tihomirovic, 2020; Vaismoradi & Snelgrove, 2019). A potential dysfunction arises if periodization is treated too rigidly, risking the oversimplification of thematic continuity across time. Nonetheless, the results highlight the importance of diachronic analysis in understanding literary evolution. By mapping thematic changes across periods, computational methods provide a systematic way to trace how literary concerns shift from moral instruction and nationalism to gender politics and identity fragmentation, thereby enriching historical literary interpretation.

The underlying causes of these temporal shifts can be explained through structural changes in political power, cultural regulation, and expressive freedom. Colonial censorship shaped early thematic conservatism, while post-independence nation-building fostered narratives of unity and resistance. The Reformasi era introduced discursive openness, enabling authors to address previously marginalized topics such as sexuality and state violence. Contemporary thematic diversification reflects globalization and digital culture, which fragment grand narratives and encourage hybrid identities (Li et al., 2018; McGillivray & Tóth, 2020). Computational linguistics captures these shifts by detecting semantic realignments across periods, revealing how literature reorganizes meaning in response to changing social structures. This supports the argument that thematic evolution is structurally driven by historical conditions rather than purely literary trends, underscoring the explanatory power of data-driven temporal analysis in literary research.

The findings on conceptual overlap and thematic intersections have profound implications for literary classification and interpretation. The presence of extensive thematic intersections indicates that Indonesian novels resist singular thematic categorization, functioning instead as multi-dimensional narrative systems (Can, 2025; Gârdan & Modoc, 2022; Omar, 2021a)**.** Functionally, this challenges traditional taxonomies that assign texts to fixed thematic or ideological categories. The risk of ignoring such overlap lies in interpretive reductionism, where complex narratives are constrained by rigid analytical frameworks. Conversely, computational identification of thematic intersections enables a more flexible and inclusive understanding of literary meaning. This implies that thematic hybridity should be treated as a constitutive feature of Indonesian novels rather than an analytical complication, prompting a rethinking of classification practices in literary criticism.

The structural explanation for thematic overlap lies in the interconnected nature of social reality itself. Issues of class, gender, religion, and nationalism are not experienced separately but intersect in everyday life and cultural discourse. Indonesian novelists reflect this complexity by embedding multiple thematic concerns within single narratives. From a computational perspective, concept mining effectively models these intersections by capturing relational semantic structures rather than isolated themes (Rani & Kamal, 2018). Studies in digital humanities emphasize that meaning in literature emerges through networks of concepts rather than linear thematic hierarchies. Therefore, the observed overlap is not methodological noise but an authentic representation of literary complexity. This finding reinforces the argument that computational linguistics provides a powerful framework for analyzing layered thematic relationships while preserving the interpretive richness essential to literary studies.

CONCLUSION

This study demonstrates that concept mining and computational linguistics provide a robust and productive framework for mapping thematic patterns in Indonesian novels. The most important finding is that Indonesian prose fiction consistently organizes meaning around socially grounded themes—such as class inequality, nationalism, gender, religion, and modernization—while simultaneously exhibiting strong temporal shifts and thematic intersections. Methodologically, this research contributes to literary studies by integrating data-driven semantic analysis with interpretive criticism, thereby overcoming the subjectivity and limited scalability of conventional thematic analysis. The study advances scholarly perspective by reframing themes as relational conceptual structures rather than fixed categories, offering a replicable and empirically grounded approach that enriches Indonesian literary criticism within the broader field of digital humanities.

Despite these contributions, this study has several limitations. The corpus, while diverse, is limited in size and does not yet represent the full spectrum of regional languages, genres, or emerging digital fiction in Indonesia. In addition, computational analysis cannot fully capture stylistic nuance, narrative voice, or aesthetic affect without complementary qualitative interpretation. Future research should expand the corpus, incorporate multilingual and regional texts, and integrate advanced models such as contextual embeddings and narrative-level analysis. Further studies may also combine reader reception data or comparative regional corpora to deepen understanding of thematic production and circulation in Indonesian literature.

ACKNOWLEDGMENTS

The authors would like to thank colleagues and peers who provided informal feedback and technical suggestions during the early stages of this research. All individuals acknowledged have given their consent to be mentioned.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Kun Andyan Anindita: conceptualization (lead); methodology (lead); data curation (lead); writing – original draft (lead). **Susan Hockey**: theoretical framework (supporting); writing – review and editing (equal). **Tomi Wahyu Septarianto**: computational analysis (supporting); visualization (supporting); writing – review and editing (equal).

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

AI USE DECLARATION

The authors used artificial intelligence tools to assist with the formatting and layout of this manuscript. AI was not used to generate the substantive content, conceptualization, data, analysis, or conclusions of the research, all of which remain the sole responsibility of the authors.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

Data availability is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

Abbas, A., Saleh, N. J., Pattu, A., Rahman, F., & Pammu, A. (2024). Comparison of American and Indonesian Women's Notion in Cather's O'pioneer! And Hamka's Tenggelamnya Kapal Van Der Wijk. *Journal of Language Teaching and Research*, *15*(5), 1489–1499. https://doi.org/10.17507/jltr.1505.10

Al-Ma'ruf, A. I., Arifin, Z., & Nugrahani, F. (2024). Exploring Ethical Frontiers: Moral Dimensions in the Tapestry of Contemporary Indonesian Literature. *Studies in English Language and Education*, *11*(1), 587–604. https://doi.org/10.24815/siele.v11i1.35142

Aranda-Corral, G. A., Borrego-Díaz, J., & Galán-Páez, J. (2022). Concept learning consistency under three-way decision paradigm. *International Journal of Machine Learning and Cybernetics*, *13*(10), 2977–2999. https://doi.org/10.1007/s13042-022-01576-w

Can, T. (2025). Why Go Digital?: Literary Studies in the Age of Digitalisation. In *Exploration of the Intersection of Corpus Linguistics and Language Science* (pp. 55–70). https://doi.org/10.4018/9798369381465.ch003

Fenlon, K., Frazier, E., & Muñoz, T. (2024). Digital Humanities. In *Encyclopedia of Libraries, Librarianship, and Information Science, First Edition, Four Volume Set* (Vol. 3, p. V3:501-V3:510). https://doi.org/10.1016/B978-0-323-95689-5.00140-1

Frolova, M. V. (2024). Indonesia, Islam, and Literature: Phenomenal Popularity of Sastra Islami. *Vestnik Sankt-Peterburgskogo Universiteta Vostokovedenie i Afrikanistika*, *16*(2), 440–457. https://doi.org/10.21638/spbu13.2024.210

Gârdan, D., & Modoc, E. (2022). From Reading Books to Reading Data: Paradigm Shifts in Literary Studies after the Digital Turn. *Revista Transilvania*, *2022*(10), 90–96. https://doi.org/10.51391/trva.2022.10.11

Hassanin, S. M., Al Bayomy, E. M., & Eleleidy, M. A. (2025). Leveraging Machine Learning and Natural Language Processing for Emotional and Thematic Analysis in Three Selected Contemporary English Novels. *Theory and Practice in Language Studies*, *15*(12), 3833–3840. https://doi.org/10.17507/tpls.1512.03

IKAPI, T. (2022). *Laporan Hasil Riset Perbukuan Indonesia* (1 No. 1). Ikatan Penerbit Indonesia (IKAPI). https://www.ikapi.org/riset/

Kar, S. (2024). The Literary Canon Inside and Beyond Academia: Adaptations and Engagements. *Interdisciplinary Literary Studies*, *26*(1), 112–133. https://doi.org/10.5325/intelitestud.26.1.0112

Li, K., Zha, H., Su, Y., & Yan, X. (2018). *Concept Mining via Embedding*. *2018-November*, 267–276. https://doi.org/10.1109/ICDM.2018.00042

McGillivray, B., & Tóth, G. M. (2020). *Applying Language Technology in Humanities Research: Design, Application, and the Underlying Logic* (p. 126). https://doi.org/10.1007/978-3-030-46493-6

Nugraha, D. (2019). On the beginning of modern Indonesian literature. *Humanities and Social Sciences Reviews*, *7*(6), 604–616. https://doi.org/10.18510/hssr.2019.7691

Omar, A. (2020). On the digital applications in the thematic literature studies of Emily Dickinson's poetry. *International Journal of Advanced Computer Science and Applications*, *11*(6), 361–365. https://doi.org/10.14569/IJACSA.2020.0110647

Omar, A. (2021a). Identifying themes in fiction: A centroid-based lexical clustering approach. *Journal of Language and Linguistic Studies*, *17*, 580–594. https://www.jlls.org/index.php/jlls/article/view/2031

Omar, A. (2021b). Towards a Computational Model to Thematic Typology of Literary Texts: A Concept Mining Approach. *International Journal of Advanced Computer Science and Applications*, *12*(12), 203–211. https://doi.org/10.14569/IJACSA.2021.0121226

Ort, C.-M. (2024). Text – knowledge – practice: For a sociology of knowledge option in literary studies. *World Literature Studies*, *16*(2), 106–129. https://doi.org/10.31577/WLS.2024.16.2.9

Pradeep, M., Sasivardhan, T., Bodana, G., Shilpa, K., Savalapurapu, K., & Babu, G. C. (2025). *Natural Language Processing for Literacy Text Mining: Extracting Knowledge From British National Corpus*. 1816–1821. https://doi.org/10.1109/ICIRCA65293.2025.11089848

Rani, A. S. B., & Kamal, A. R. N. B. (2018). *Text Mining to Concept Mining: Leads Feature Location in Software System*. 2018 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2018. https://doi.org/10.1109/ICCIC.2018.8782418

Tan, X. (2020). Topic extraction and classification method based on comment sets. *Journal of Information Processing Systems*, *16*(2), 329–342. https://doi.org/10.3745/JIPS.04.0165

Teodorescu, H. N., & Bolea, S. C. (2022). A Comparative Lexical Analysis of Three Romanian Works – The Etymological Metalepsis Role and Etymological Indices. *Romanian Journal of Information Science and Technology*, *25*(3–4), 275–289. https://romjist.ro/abstract-722.html

Tihomirovic, Z. (2020). Analiza in interpretacija prostora v literarnem besedilu. *Slavisticna Revija*, *68*(4), 629–638. https://srl.si/ojs/srl/article/view/3885

Vaismoradi, M., & Snelgrove, S. (2019). Theme in qualitative content analysis and thematic analysis. *Forum Qualitative Sozialforschung*, *20*(3). https://doi.org/10.17169/fqs-20.3.3376

Vinodini, S., & Adithya Pothan Raj, V. (2024). *Enhancing Literary Analysis through Artificial Intelligence and Machine Learning: Insights from "The Alchemist."* 136–141. https://doi.org/10.1109/ICRAIS62903.2024.10811737

Wang, R., Zhou, D., Huang, H., & Zhou, Y. (2025). MIT: Mutual Information Topic Model for Diverse Topic Extraction. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(2), 2523–2537. https://doi.org/10.1109/TNNLS.2024.3357698

Wiyatmi, W., Suryaman, M., & Swatikasari, E. (2019). Developing an ecofeminist literary criticism model to cultivate an ecologically aware and feminist generation. *Interdisciplinary Literary Studies*, *21*(4), 515–531. https://doi.org/10.5325/intelitestud.21.4.0515

Yuan, R., Vengadasamy, R., & Zheng, Y. (2025). Mapping Eileen Chang's Novels with a Computational Analysis of Themes and Emotions. *3L: Language, Linguistics, Literature*, *31*(1), 322–338. https://doi.org/10.17576/3L-2025-3101-21