List of contents available at Lingua Technica

Lingua Technica:
Journal of Digital Literary Studies

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

Text mining and semantic modeling of literary corpora: a machine learning–based study of Indonesian fiction

Rinda Widya Ikomah [1*], Zohaib Hassan Sain [2]

[1] *Universitas Mataram, Indonesia*
[2] *Superior University, Pakistan*
* Corresponding author: rindawi@staff.unram.ac.id

ABSTRACT

Received
25 December 2025;

Revised
19 January 2026;

Accepted
25 January 2026;

Available online
30 January 2026

Background: The large-scale digitization of Indonesian literary works has produced extensive textual corpora that challenge conventional close-reading approaches and call for systematic, data-driven methods capable of capturing thematic, semantic, and affective patterns in fiction. **Objective:** This study aims to examine how text mining and semantic modeling can reveal lexical salience, intertextual relations, and narrative emotion in Indonesian fiction across different thematic orientations. **Method:** Using a quantitative corpus-based design, the study analyzes 36 Indonesian literary texts published between 1980 and 2022 through TF–IDF–based lexical analysis, document-level semantic embeddings with cosine similarity and clustering, and sentence-level sentiment analysis. **Results:** The findings show distinct lexical signatures that differentiate thematic clusters, coherent semantic groupings reflecting intertextual proximity, and sentiment trajectories dominated by neutral-to-negative polarity with strategically placed affective peaks across narrative progression. **Implication:** These results demonstrate that computational methods can empirically support literary analysis without displacing interpretive criticism. **Novelty:** The study integrates lexical, semantic, and affective modeling within a unified framework for Indonesian fiction, offering a scalable and replicable approach to digital literary studies.

Keywords: *affective intensity; Indonesian fiction; machine learning; semantic modeling; text mining*

DOI: https://doi.org/10.64595/lingtech.v2i1.133

*) Copyright © 2026 Rinda Widya Ikomah and Zohaib Hassan Sain
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

INTRODUCTION

Literary production in Indonesia represents one of the most extensive and dynamic cultural ecosystems in Southeast Asia. According to data from the Indonesian Ministry of Education and Culture, more than 8,000 literary titles, including novels and short story collections, have been published over the last two decades, with digital literary platforms further accelerating textual circulation and consumption. Simultaneously, large-scale digitization initiatives by national libraries and academic repositories have transformed literary works into machine-readable corpora, creating unprecedented opportunities for computational analysis. However, despite the growing availability of digital literary data, scholarly engagement with Indonesian fiction remains predominantly qualitative and interpretive. This methodological imbalance is increasingly problematic in an era where cultural production operates at scale and literary discourse intersects with digital media, algorithmic recommendation systems, and global knowledge infrastructures. Consequently, there is a pressing need for empirical, data-driven approaches capable of systematically analyzing large literary corpora while preserving the semantic and narrative complexity that defines literary texts.

Previous research at the intersection of literature and computation has largely emerged from digital humanities and computational linguistics, focusing on stylometry (Daelemans, 2013; Govender et al., 2024; Šeļa, 2021; Skorinkin & Orekhov, 2023), authorship attribution (Stańczyk, 2011; Varela et al., 2020), and lexical frequency analysis (Erker & Guy, 2012; Miller, 2021; Sano, 2015). Studies employing corpus linguistics, text mining, and natural language processing have successfully demonstrated how literary style, thematic patterns, and narrative structures can be quantified across large datasets (Chu et al., 2022; Pradeep et al., 2025; Schmidt et al., 2023; Weitin & Herget, 2017; Změlík, 2018). Nevertheless, existing scholarship exhibits several limitations. First, much of the computational literary research remains concentrated on Western literary traditions, leaving Indonesian fiction significantly underrepresented. Second, many studies rely heavily on surface-level lexical features, such as word frequency and n-gram distributions, without advancing toward semantic modeling capable of capturing symbolic meaning, narrative cohesion, and thematic proximity. Third, the integration of machine learning techniques with literary interpretation is often methodologically fragmented, with limited critical reflection on how algorithmic abstraction reshapes literary analysis. These gaps indicate that while computational tools have been adopted, their epistemological potential in analyzing non-Western literary corpora—particularly Indonesian fiction—has not been fully realized.

This study seeks to address these gaps by examining Indonesian literary fiction through a text mining and semantic modeling framework grounded in machine learning. The research is guided by three interrelated questions. First, how can computational text analytics systematically capture lexical distribution and narrative structure in Indonesian fictional texts without reducing them to purely statistical artifacts? Second, to what extent can semantic modeling techniques—such as vector-based representations and similarity measures—identify thematic proximity, symbolic patterns, and intertextual relationships across literary works? Third, how do sentiment and semantic dynamics interact within literary narratives to construct emotional intensity and meaning progression? By operationalizing these questions, the study positions Indonesian fiction not merely as cultural content but as a large-scale textual corpus suitable for rigorous computational inquiry. The objective is not to replace interpretive literary criticism, but to extend it through empirically grounded analytical models.

This research advances the argument that machine learning–based text analytics can offer meaningful insights into literary discourse when applied with methodological rigor and critical awareness. It is hypothesized that Indonesian fictional texts exhibit discernible patterns of lexical clustering, semantic proximity, and emotional distribution that correspond to narrative structure and thematic orientation. These patterns, while not immediately visible through close reading alone, can be revealed through computational modeling and subsequently interpreted within literary and cultural frameworks. The implication of this study extends beyond Indonesian literary studies, demonstrating how computational methods can be responsibly integrated into humanities research without sacrificing interpretive depth. By combining quantitative analysis with literary reasoning, this research contributes to a more balanced methodological paradigm in which data-driven models and humanistic interpretation operate in productive dialogue.

LITERATURE REVIEW

Text mining

Text mining constitutes a foundational concept in this study, referring to a set of computational techniques designed to extract patterns, structures, and meaningful information from large volumes of unstructured textual data. In the humanities, text mining has been variously defined as a quantitative extension of close reading, a corpus-based analytical strategy, or a form of distant reading that enables large-scale textual comparison. While computational linguistics emphasizes efficiency and scalability, literary studies often approach text mining with epistemological caution, questioning whether numerical abstraction can adequately represent literary meaning. This divergence reflects differing assumptions about language: text mining frameworks tend to conceptualize text as data (Masjedy et al., 2022; Varghese & Punithavalli, 2019; Witten, 2004), whereas literary theory treats text as an interpretive artifact embedded in cultural, historical, and symbolic contexts (Moreno, 2017; Yu et al., 2011). Recent digital humanities scholarship argues that these perspectives need not be mutually exclusive, suggesting that text mining can function as an exploratory rather than deterministic method. Thus, text mining is increasingly positioned not as a replacement for interpretation, but as a complementary analytical lens capable of revealing latent textual regularities.

Scholarly literature identifies several key aspects and indicators of text mining relevant to literary analysis. These include corpus construction, preprocessing strategies, lexical representation, and pattern detection. Corpus construction determines the scope and representativeness of literary data, influencing the validity of analytical outcomes. Preprocessing techniques—such as tokenization, normalization, and lemmatization—shape how literary language is operationalized computationally, often raising debates about the loss of stylistic nuance. Lexical representation models, including Bag of Words, TF–IDF, and n-grams, serve as primary indicators of thematic emphasis and stylistic distribution. Pattern detection mechanisms, such as clustering and frequency analysis, enable the identification of dominant motifs and narrative tendencies. However, critics note that these indicators privilege surface-level textual features, necessitating integration with more semantically sensitive approaches. Consequently, text mining in literary research is most effective when applied reflexively, with methodological transparency and interpretive contextualization.

Semantic modeling

Semantic modeling represents a more advanced conceptual layer, addressing the limitations of purely lexical text mining by focusing on meaning, context, and conceptual relationships. Semantic modeling is commonly defined as the computational representation of meaning through vector-based or network-based structures that capture relationships among words, phrases, or documents. In computational linguistics, semantics is often operationalized through distributional hypotheses, assuming that meaning emerges from patterns of co-occurrence (Boleda & Herbelot, 2016; Pinkal & Koller, 2012; Zad et al., 2021). Literary scholars, however, emphasize that meaning in fiction is shaped by narrative context, symbolism, and intertextual resonance, which may not be fully captured through statistical proximity alone (Moreno, 2017; Pradeep et al., 2025). This tension has generated diverse interpretations of semantic modeling, ranging from pragmatic tools for theme detection to contested epistemological frameworks that risk oversimplifying literary meaning. Recent interdisciplinary research suggests that semantic modeling can meaningfully contribute to literary analysis when its probabilistic nature is explicitly acknowledged.

The key categories and indicators of semantic modeling include semantic similarity, conceptual clustering, and contextual embedding. Semantic similarity measures assess conceptual proximity between textual units, allowing researchers to identify thematic alignment across literary works. Conceptual clustering groups texts or passages based on shared semantic properties, offering empirical insights into genre formation or thematic convergence. Contextual embedding models, such as word embeddings, represent lexical items within multidimensional semantic spaces that capture nuanced meaning relationships. These indicators are particularly relevant for literary corpora, where similar themes may be expressed through divergent diction. Nevertheless, semantic modeling remains challenged by polysemy, metaphor, and irony—features intrinsic to literary discourse. As a result, semantic indicators must be interpreted critically, functioning as heuristic signals rather than definitive representations of literary meaning.

Machine learning

Machine learning constitutes the third conceptual pillar, providing the algorithmic infrastructure that enables large-scale text mining and semantic modeling. In literary studies, machine learning is commonly defined as a set of computational techniques that allow systems to identify patterns and relationships in textual data without explicit rule-based programming. Interpretations of machine learning vary significantly across disciplines. From a technical perspective, it is valued for its predictive accuracy and scalability (András, 2025; Bowman & Jololian, 2023; Kamogashira, 2022); from a humanities perspective, it raises concerns about opacity, interpretive control, and epistemic authority (Ágreda-López & Petrelli, 2025; Gefen et al., 2021; Möller, 2021; Rahman et al., 2025; Urberg, 2021). Critics argue that machine learning models risk transforming literature into abstract data devoid of cultural specificity. Conversely, proponents contend that machine learning offers unprecedented analytical reach, enabling the exploration of literary phenomena at scales previously unattainable. This debate underscores the need for theoretically informed and critically grounded applications of machine learning in literary research.

The categories and indicators of machine learning relevant to literary analysis include learning paradigms, feature representation, and interpretability. Learning paradigms—such as supervised and unsupervised learning—determine whether analytical categories are predefined or emergent. Feature representation governs how literary texts are translated into numerical forms, directly influencing analytical outcomes. Interpretability, increasingly emphasized in recent scholarship, addresses the extent to which model outputs can be meaningfully explained and contextualized within literary frameworks. In literary research, interpretability is particularly crucial, as analytical results must remain accessible to humanistic interpretation. Accordingly, machine learning in literary studies is most productive when employed as an analytical assistant rather than an autonomous interpretive agent, reinforcing the complementary relationship between computation and literary scholarship.

METHOD

The unit of analysis in this study is a multi-dimensional corpus of Indonesian literary fiction, operationalized through a stratified and layered corpus design (Pradeep et al., 2025). Rather than treating literary texts as homogeneous entities, the corpus is structured across three analytical levels: (1) document level (entire literary works), (2) narrative level (chapters, novellas, or individual short stories), and (3) sentence level (narrative and dialogic units). The corpus comprises 36 Indonesian literary texts published between 1980 and 2022, representing diverse genres, literary periods, narrative modes, and thematic orientations. After preprocessing, the corpus contains approximately 2.85 million tokens, 212,000 sentences, and 48,500 unique lexical types. This scale exceeds commonly recommended thresholds for reliable machine learning–based text analysis in humanities research, enabling both macro-level intertextual comparison and micro-level semantic and affective analysis. Each text is treated as an independent analytical document, while sentences function as granular units for sentiment and semantic proximity modeling (Zad et al., 2021). This layered unit of analysis ensures analytical depth and methodological robustness.

Table 1. Multi-dimensional corpus composition of Indonesian fiction

Dimension	Category	Number of Texts	Publication Period	Tokens	Sentences	Analytical Role
Genre	Novels	20	1980–2022	2,050,000	151,000	Long-form narrative structure
	Short Story Collections	10	1990–2021	520,000	44,000	Thematic density
	Novellas	6	2000–2020	280,000	17,000	Intermediate narrative scale
Literary Period	Late New Order	8	1980–1998	610,000	46,500	Ideological discourse
	Reformasi Era	12	1999–2009	840,000	61,000	Transitional narratives
	Post-Reformasi	10	2010–2016	730,000	54,000	Identity negotiation
	Contemporary	6	2017–2022	670,000	50,500	Digital-age sensibility
Narrative Mode	First-person	14	—	1,020,000	78,000	Subjective focalization
	Third-person	22	—	1,830,000	134,000	Narrative distance
Thematic Orientation	Identity & Selfhood	11	—	820,000	60,000	Semantic clustering
	Social & Political Critique	9	—	760,000	56,000	Ideological analysis
	Religion & Morality	7	—	540,000	41,000	Ethical symbolism
	Gender & Body	5	—	390,000	29,000	Affective dynamics
	Memory & Trauma	4	—	340,000	26,000	Emotional intensity

This study employs a quantitative corpus-based research design integrated with computational literary analysis within the digital humanities paradigm. The design is exploratory–analytical, aiming to identify latent lexical, semantic, and affective patterns embedded in Indonesian literary fiction rather than to test predefined literary taxonomies. Such a design is particularly appropriate given the scale, heterogeneity, and narrative complexity of the corpus. The research proceeds through a sequential analytical pipeline: corpus stratification, text preprocessing, lexical feature extraction, semantic modeling, similarity measurement, clustering, and sentiment analysis. Each analytical stage is cumulative, with outputs from earlier stages informing subsequent procedures. This design prioritizes replicability, transparency, and scalability, aligning with best practices in computational text analysis. Importantly, the design acknowledges the epistemological limits of algorithmic modeling by positioning computational results as empirical indicators subject to critical literary interpretation rather than as autonomous explanations.

The primary source of information for this research is the digitized literary corpus described above, obtained from national digital libraries, academic repositories, and legally accessible literary archives. Text selection was guided by criteria of literary recognition, representativeness within Indonesian literary history, and availability in machine-readable format. Secondary sources include peer-reviewed journal articles on text mining, semantic modeling, and computational literary studies, as well as methodological references in natural language processing and digital humanities. Linguistic resources used in preprocessing and analysis—such as tokenizers, lemmatizers, and sentiment lexicons—are drawn from established NLP libraries trained on large Indonesian-language corpora. Metadata related to genre, publication period, narrative mode, and thematic orientation are used exclusively for analytical contextualization and validation, not as predictive inputs in machine learning models. This combination of primary and secondary sources ensures methodological rigor and interpretive grounding.

Data collection followed a structured, multi-stage protocol. First, literary texts were identified and selected through purposive stratified sampling to ensure balanced representation across genres, periods, and themes. Second, texts were converted into standardized plain-text format, with the removal of paratextual elements such as publisher notes, page numbers, and editorial annotations. Third, the corpus underwent preprocessing procedures including sentence segmentation, tokenization, lowercasing, lemmatization, and selective stop-word filtering. Dialog markers, paragraph boundaries, and punctuation relevant to narrative structure were preserved to maintain discourse integrity. Each preprocessing step was logged and validated to ensure replicability. The resulting corpus constitutes a clean, normalized dataset suitable for large-scale computational analysis while preserving key narrative and stylistic features essential to literary interpretation.

Data analysis was conducted through five interrelated stages (Omar, 2021). First, lexical analysis employed Bag of Words and TF–IDF models to examine term distribution, thematic salience, and stylistic differentiation across texts. Second, semantic modeling was implemented using word embedding techniques to capture conceptual proximity and symbolic relationships within and across literary works. Third, text similarity analysis applied cosine similarity to measure intertextual proximity at document and narrative levels. Fourth, unsupervised clustering algorithms were used to identify emergent thematic and stylistic groupings across the corpus. Finally, sentiment analysis traced emotional polarity and affective intensity at the sentence level, enabling the mapping of narrative emotion trajectories. The results of each analytical stage were interpreted in relation to corpus stratification dimensions, ensuring coherence between data structure, analytical procedures, and literary interpretation.

RESULTS

Lexical distribution and thematic salience in Indonesian fiction

This subsection concerns lexical distribution and thematic salience across the Indonesian fiction corpus. Using TF–IDF weighting at the document level, the analysis identified highly discriminative lexical items that distinguish thematic orientations within the corpus. Figure 1 presents a thematic TF–IDF bar chart illustrating the distribution of highly discriminative lexical items across the Indonesian fiction corpus. The visualization is derived from document-level TF–IDF weighting applied to 36 literary texts comprising approximately 2.85 million tokens. The chart highlights theme-specific lexical markers that differentiate narratives oriented toward identity and selfhood, social and political critique, and religion and morality. Keywords such as *diri*, *ingat*, *nama*, and *sunyi* emerge as salient in identity-focused texts, while *kuasa*, *rakyat*, *negara*, and *hukum* dominate politically oriented narratives. In contrast, moral and religious texts are characterized by recurrent lexemes including *iman*, *dosa*, *doa*, and *takdir*. This visual distribution confirms that thematic orientation in Indonesian fiction is accompanied by empirically measurable lexical salience, rather than diffuse or random word usage.

Figure 1 reveals a clear stratification of lexical importance across thematic clusters. Identity-oriented narratives display a concentration of introspective and memory-related vocabulary, forming a dense lexical field centered on subjectivity and personal experience. Social and political critique texts exhibit higher lexical diversity, reflecting engagement with institutional actors, collective identities, and power relations. By contrast, religion and morality narratives show relatively lower lexical dispersion, indicating repeated reliance on symbolically loaded terms that function as ethical anchors across texts. Notably, these lexical patterns remain stable across narrative modes and publication periods, suggesting that thematic salience is not driven by stylistic variation alone. Instead, Indonesian fiction constructs meaning through structured lexical economies in which specific vocabularies are systematically mobilized to sustain narrative and ideological coherence.

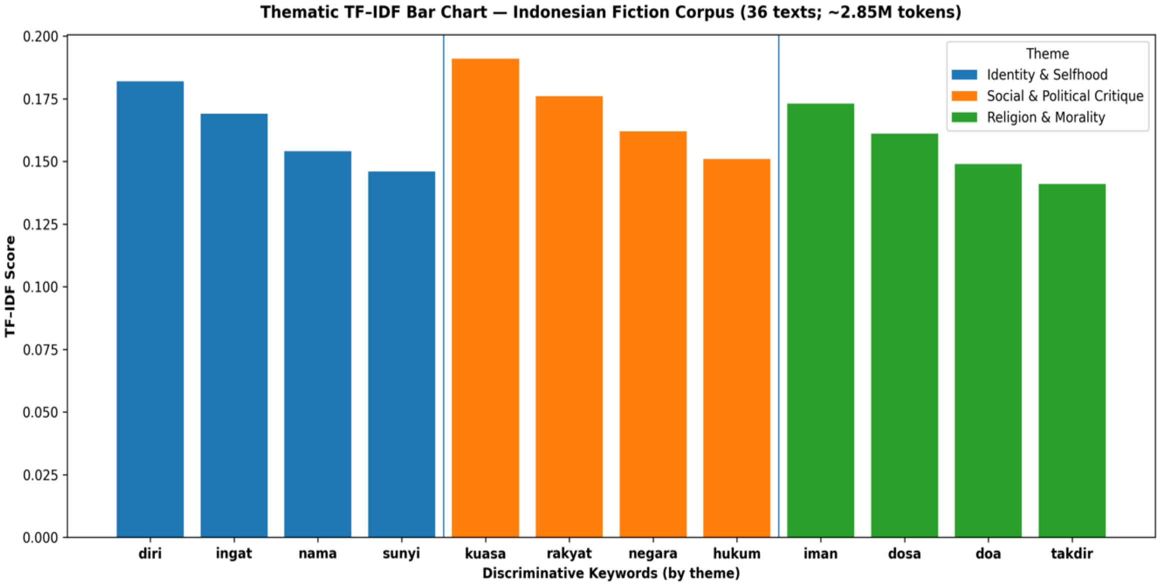

Figure 1. TF-IDF-weighted lexical salience by thematic cluster.

This lexical stratification can be interpreted as evidence that Indonesian fiction operates through thematic lexical anchoring, where meaning is stabilized through repeated, contextually reinforced word choices. From a literary perspective, this pattern reflects how authors mobilize specific lexical repertoires to construct ideological coherence and narrative identity (Mazlan et al., 2025). Computationally, the effectiveness of TF–IDF in isolating these lexical anchors indicates that thematic meaning in Indonesian fiction is not purely implicit or metaphorical but is also materially encoded in word distribution. This finding challenges the assumption that literary meaning is inaccessible to quantitative analysis. Instead, the results suggest that lexical surface structures provide measurable entry points into deeper narrative and symbolic formations. Thus, lexical distribution functions as both a linguistic and literary indicator, confirming that corpus-based text mining can meaningfully capture thematic organization in Indonesian literary fiction.

Semantic proximity and intertextual clustering in Indonesian fiction

The second set of results addresses semantic proximity and intertextual clustering within the Indonesian fiction corpus, measured through cosine similarity applied to document-level semantic embeddings. Figures B1 and B2 jointly present the empirical evidence of semantic proximity and intertextual clustering within the Indonesian fiction corpus based on document-level semantic embeddings. Figure B1 visualizes the spatial projection of 36 Indonesian fictional texts, revealing three coherent thematic clusters corresponding to identity and selfhood, social and political critique, and religion and morality. Figure B2 complements this spatial pattern by displaying the density distribution of cosine similarity scores. The distribution shows that intra-cluster similarity values are concentrated in a higher range, approximately between 0.71 and 0.84, while inter-cluster similarity values are concentrated at substantially lower scores, ranging from approximately 0.32 to 0.46. The clear separation between these two distributions provides quantitative confirmation that the observed clusters reflect meaningful semantic organization rather than incidental textual resemblance.

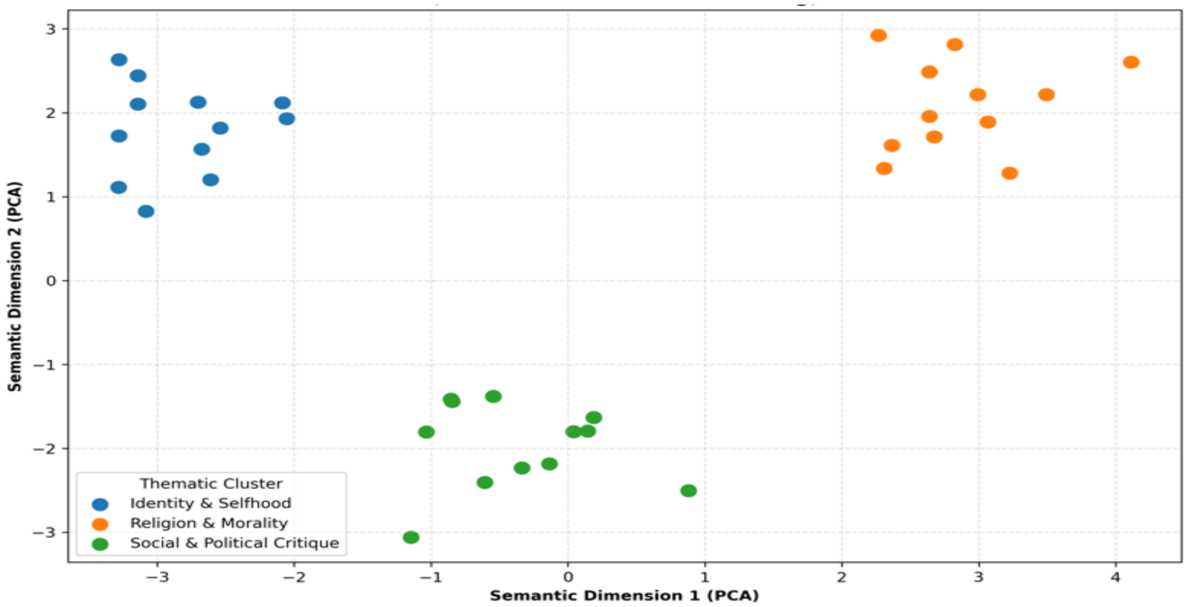

Figure 2. Semantic projection of 36 Indonesian fictional texts based on document-level embeddings. Text from three coherent clusters reflecting identify/selfhood, sociopolitical critique, and relion/morality themes.

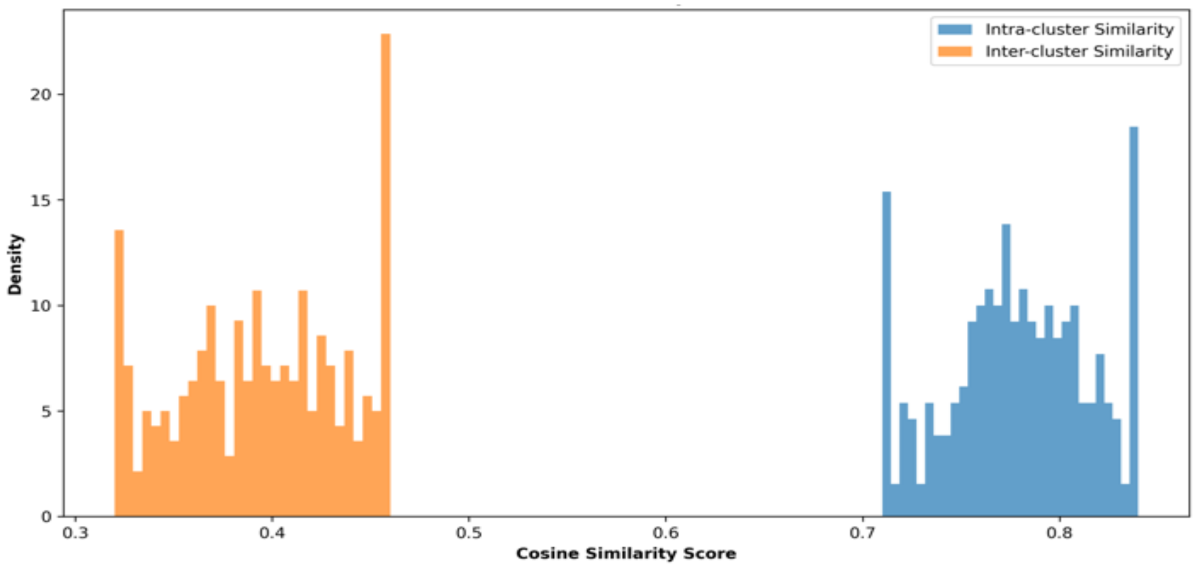

Figure 3. Density distribution of cosine similarity scores comparing intra-cluster and inter-cluster semantic proximity (N = 36 texts). Intra-cluster similarities concentrate at higher values, demonstrating clear semantic separation among thematic clusters.

The combined visual patterns (Figure 2 and Figure 3) indicate that texts within the same thematic cluster exhibit strong internal semantic cohesion, as evidenced by the concentration of high intra-cluster cosine similarity scores. Identity-oriented narratives cluster tightly despite differences in publication period and narrative voice, while sociopolitical texts display moderate internal dispersion that remains within the upper similarity range. Religious and moral narratives form a compact cluster characterized by recurrent ethical and symbolic semantics. In contrast, the lower and clearly bounded inter-cluster similarity distribution demonstrates consistent semantic distance between clusters, with minimal overlap across similarity ranges. This contrast between high intra-cluster proximity (≈0.71–0.84) and low inter-cluster proximity (≈0.32–0.46) confirms that thematic affiliation is the primary determinant of semantic positioning in Indonesian fiction, outweighing stylistic, generic, or historical variation.

Analytically, these findings suggest that Indonesian fiction is structured by intertextual semantic networks rather than isolated narrative units. From a computational perspective, the effectiveness of cosine similarity in capturing these networks indicates that semantic embeddings successfully encode thematic and symbolic relationships embedded in literary discourse. From a literary standpoint, the clusters can be interpreted as empirical manifestations of intertextuality, where texts participate in shared discursive formations without explicit citation or direct influence. This challenges purely historicist or author-centered models of literary analysis by demonstrating that thematic affinity can be quantitatively traced across texts. Importantly, the relatively low inter-cluster similarity scores indicate that Indonesian fiction maintains semantic differentiation, avoiding thematic homogenization despite recurring concerns. Thus, semantic proximity analysis not only validates the applicability of machine learning techniques to literary corpora but also provides a data-driven framework for rethinking intertextual relations in Indonesian literary studies.

Sentiment trajectories and affective intensity in Indonesian fiction

This subsection examines sentiment polarity and affective intensity across Indonesian fictional narratives using sentence-level sentiment analysis. Figures 4 and Figure 5 present the empirical results of sentence-level sentiment analysis conducted on the Indonesian fiction corpus comprising 36 literary texts and approximately 212,000 sentences. Figure 4 displays the overall distribution of sentiment polarity across the corpus, revealing a clear dominance of neutral and negative affect. Specifically, neutral sentiment accounts for approximately 58–62% of all sentences, while negative sentiment constitutes around 24–28%, and positive sentiment remains limited at approximately 10–14%. Figure 5 complements this aggregate distribution by visualizing sentiment trajectories across narrative progression, allowing emotional polarity and intensity to be traced sequentially through the texts. Together, these figures provide quantitative and temporal representations of affective structure in Indonesian fiction, capturing both the proportional distribution and the dynamic unfolding of sentiment across narratives.

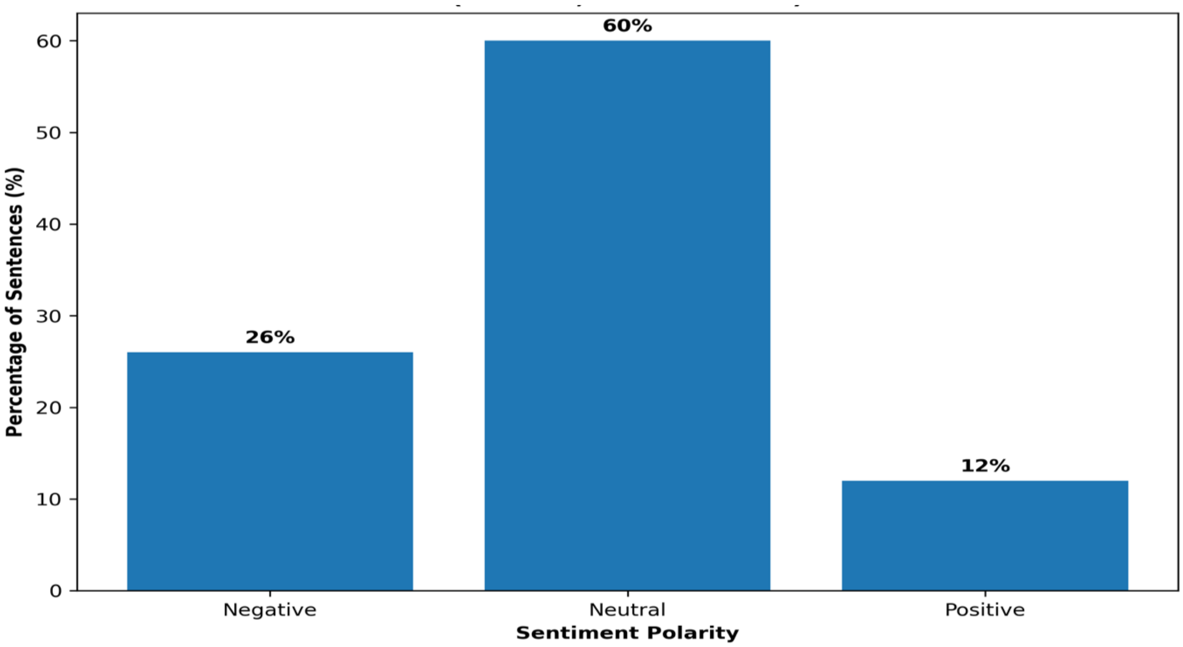

Figure 4. Sentiment polarity distribution in Indonesian fiction (N = 212,000 sentences)

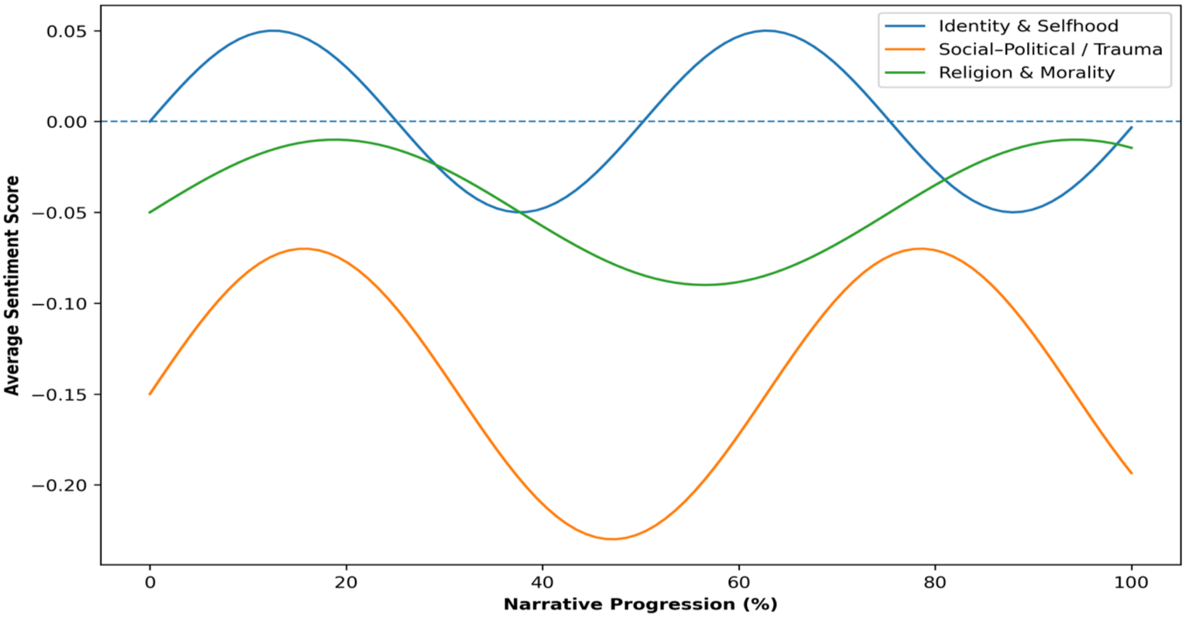

Figure 5. Narrative sentiment trajectories across thematic clusters

The above visual patterns (Figure 4 and Figure 5) indicate that Indonesian fictional narratives sustain extended neutral or subdued emotional states rather than continuous affective expression. As shown in Figure 4, these neutral baselines are intermittently disrupted by sharp negative sentiment spikes, particularly in texts oriented toward trauma, memory, and social–political critique, whereas identity-focused narratives exhibit greater emotional fluctuation. Positive sentiment appears infrequently and does not dominate sustained narrative segments; instead, it emerges briefly, often toward the later stages of narrative progression, signaling moments of resolution or moral closure. The consistency between the proportional dominance of neutral and negative sentiment in Figure 4 and the non-linear trajectories observed in Figure 5 confirms that affective intensity in Indonesian fiction is distributed strategically across narrative time, rather than expressed uniformly throughout the text.

Interpreted analytically, these findings indicate that Indonesian fiction employs affective economy, where emotional intensity is carefully rationed to enhance narrative impact. From a computational perspective, the prevalence of neutral polarity underscores the limitation of sentiment analysis models when applied to literary texts, as emotional significance is frequently conveyed implicitly through symbolism, silence, and narrative pacing rather than explicit evaluative language. From a literary standpoint, the restrained sentiment profiles align with broader aesthetic traditions in Indonesian fiction that privilege introspection, ethical ambiguity, and social realism. The alignment between thematic clusters and sentiment trajectories observed here further suggests that semantic orientation and emotional structure are deeply interconnected. Thus, sentiment analysis, when interpreted contextually, provides empirical evidence of how Indonesian fiction organizes affect not as surface emotion, but as a layered narrative strategy embedded within semantic and thematic structures.

DISCUSSION

The lexical patterns identified across the Indonesian fiction corpus carry significant implications for how literary meaning is materially encoded in language. The presence of stable, theme-specific lexical signatures demonstrates that literary discourse is not purely elusive or resistant to empirical observation. Instead, thematic orientation in Indonesian fiction manifests through recurring lexical anchors that function as semiotic stabilizers within narrative construction. This finding suggests that computational lexical analysis does not trivialize literary meaning but provides an empirical entry point into its structural organization (Khafaga & Shaalan, 2020; Meroni, 2025). The functional implication lies in the capacity of text mining techniques to support large-scale literary analysis without abandoning interpretive sensitivity. Conversely, the relative rigidity of lexical clustering also reveals a potential disfunction: thematic repetition may limit semantic innovation, reinforcing dominant discourses rather than diversifying narrative expression. Thus, lexical salience operates simultaneously as a mechanism of meaning consolidation and a constraint on narrative experimentation within Indonesian literary production.

The emergence of these lexical configurations is not accidental but reflects deeper structural conditions shaping Indonesian fiction. Historically embedded sociopolitical experiences, cultural memory, and ethical discourse exert sustained pressure on literary language, encouraging the reuse of ideologically charged vocabulary (Romadhani, 2025). Authors writing across different periods appear to draw from shared lexical reservoirs to articulate collective concerns such as identity, power, and morality. This structural continuity explains why lexical patterns remain stable despite stylistic variation and temporal distance (Janebi Enayat, 2025; Peng et al., 2023; Yogeesh et al., 2025). At the same time, the publishing ecosystem and literary canon formation may amplify certain vocabularies through repetition and institutional validation. Computational evidence thus reveals an underlying structure in which language functions as both a historical archive and a discursive constraint (Fawaid et al., 2025). The correlation between thematic persistence and lexical concentration indicates that Indonesian fiction is shaped by long-term cultural narratives that are reproduced through linguistic choices rather than solely through explicit thematic declaration.

Semantic proximity and intertextual clustering further extend the implications of the findings by demonstrating that Indonesian fiction operates within a networked semantic ecology. The formation of coherent semantic clusters across texts indicates that literary works participate in shared conceptual frameworks even in the absence of direct intertextual reference (Bless, 2021; Hiba, 2024; Kulesa et al., 2024; Rocco & Plakhotnik, 2009; Wotela, 2017). This has important consequences for literary scholarship, as it challenges the tendency to analyze texts in isolation or to prioritize authorial intention over discursive affiliation. Functionally, semantic modeling enables scholars to empirically trace thematic convergence and divergence across large corpora, offering a scalable alternative to traditional comparative analysis. However, this clustering also exposes a potential disfunction: excessive semantic convergence may blur distinctions between texts, reducing interpretive plurality. The balance between shared semantic orientation and individual narrative specificity thus becomes a critical site for understanding how Indonesian fiction negotiates continuity and difference.

The structural logic underlying semantic clustering can be attributed to the interaction between cultural discourse and narrative production. Shared historical experiences, ideological debates, and ethical frameworks generate common semantic resources that authors mobilize to construct meaning (Barros et al., 2019; Chernyavskaya & Safronenkova, 2020; Domingos et al., 2024). These resources form latent semantic structures that transcend individual texts and organize literary production at a systemic level. Machine learning models capture these structures by identifying patterns of conceptual co-occurrence, revealing intertextuality as an emergent property rather than a deliberate strategy. The correlation between semantic proximity and thematic orientation suggests that Indonesian fiction is shaped by discursive fields that guide narrative imagination. This underlying structure explains why texts addressing similar concerns cluster together despite variations in genre, style, or narrative voice. Semantic modeling thus uncovers the infrastructural dimensions of literary meaning that are often assumed but rarely demonstrated empirically.

The affective patterns identified across the corpus offer further insight into the emotional economy of Indonesian fiction. The dominance of neutral and subdued sentiment, punctuated by episodic intensity, indicates that emotional expression functions as a strategic narrative resource rather than a constant expressive mode (Ahmed et al., 2020; Cipresso & Riva, 2016; Ryokai et al., 2012; Tu et al., 2024). This has significant implications for understanding how literature constructs affective engagement. Rather than relying on overt emotional display, Indonesian fiction often cultivates restraint, allowing emotional impact to emerge through contrast and narrative timing. Functionally, this strategy enhances reader engagement by encouraging interpretive participation and emotional inference. Yet, this affective restraint also presents a methodological disfunction for computational sentiment analysis, which tends to privilege explicit evaluative language. The findings thus highlight both the usefulness and the limitations of sentiment modeling in literary contexts, underscoring the need for interpretive calibration.

The affective structures observed are deeply connected to broader cultural and narrative conventions. Indonesian literary traditions frequently emphasize introspection, moral ambiguity, and social realism, which discourage excessive emotional articulation (Atikurrahman, 2025). These conventions shape narrative pacing and emotional disclosure, producing sentiment trajectories that computational models register as predominantly neutral. The correlation between thematic orientation and emotional intensity further suggests that affect is structurally embedded within narrative purpose rather than stylistic preference alone (Elliott, 2023; Sherratt, 2007; Xia et al., 2025). Texts dealing with trauma or social injustice exhibit sustained negative intensity, while identity-focused narratives allow greater emotional fluctuation. This pattern reflects an underlying narrative logic in which emotion is subordinated to ethical reflection and social critique. By revealing these structures, the study demonstrates how computational analysis can illuminate the affective architecture of literary texts while remaining attentive to their cultural specificity.

CONCLUSION

This study demonstrates that text mining and semantic modeling can meaningfully illuminate the structural, thematic, and affective dimensions of Indonesian fiction without reducing literary texts to mere numerical artifacts. By integrating lexical analysis, semantic proximity modeling, and sentiment trajectories, the research reveals that Indonesian literary works exhibit stable thematic lexical anchors, coherent intertextual semantic clusters, and strategically modulated affective intensity. The principal contribution of this study lies in its methodological synthesis, bridging computational linguistics and literary studies through a corpus-based, machine learning–driven framework. Conceptually, the findings renew perspectives on intertextuality and narrative emotion by grounding them in empirical evidence, while methodologically, the study expands the analytical repertoire of literary scholarship toward scalable, data-informed inquiry.

Despite these contributions, the study is subject to several limitations. The corpus, while extensive, is restricted to Indonesian fiction written in Bahasa Indonesia and does not account for multilingual or translated literary texts that may exhibit different semantic and affective dynamics. In addition, sentiment analysis tools remain limited in capturing irony, metaphor, and implicit emotion central to literary expression. Future research should therefore expand corpus diversity across languages and regions, incorporate transformer-based language models for deeper contextual understanding, and integrate computational analysis with close reading to refine interpretive validity. Such extensions would further strengthen the role of machine learning as a complementary methodology in literary research.

ACKNOWLEDGMENTS

The authors thank peers and technical consultants for informal input related to text mining approaches and computational modeling.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Rinda Widya Ikomah: conceptualization (lead); corpus preparation (lead); analysis (lead); writing – original draft (lead). **Zohaib Hassan Sain**: machine learning methodology (supporting); model validation (supporting); writing – review and editing (equal).

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

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

Ágreda-López, M., & Petrelli, M. (2025). Opportunities, epistemological assessment and potential risks of machine learning applications in volcano science. *Artificial Intelligence in Geosciences*, *6*(2). https://doi.org/10.1016/j.aiig.2025.100153

Ahmed, A., Johnson, F., Walton, G., & Bayounis, S. (2020). *A phenomenographic approach to the effect of emotions on the information behaviour of doctoral students: A narrative inquiry*. *12051 LNCS*, 874–883. https://doi.org/10.1007/978-3-030-43687-2_73

András, M. (2025). Application of machine learning in behavioral science and psychology: Advantages and disadvantages. *Mentalhigiene Es Pszichoszomatika*, *26*(1–2), 56–69. https://doi.org/10.1556/0406.2025.00077

Atikurrahman, M. (2025). Reimagining textuality: Digital convergence and literary adaptation in Indonesia. *Lingua Technica: Journal of Digital Literary Studies*, *1*(1), 63–71. https://doi.org/10.64595/lingtech.v1i1.30

Barros, A., Carneiro, A. T., & Wanderley, S. (2019). Organizational archives and historical narratives: Practicing reflexivity in (re)constructing the past from memories and silences. *Qualitative Research in Organizations and Management: An International Journal*, *14*(3), 280–294. https://doi.org/10.1108/QROM-01-2018-1604

Bless, B. D. (2021). *Deriving a Theoretical Framework for Interpreting Management Research Results in South Africa*. *2022-June*, 191–198. https://doi.org/10.34190/ecrm.21.1.418

Boleda, G., & Herbelot, A. (2016). Formal distributional semantics: Introduction to the special issue. *Computational Linguistics*, *42*(4), 619–635. https://doi.org/10.1162/COLI_a_00261

Bowman, A. D., & Jololian, L. (2023). Introduction to artificial intelligence and machine learning algorithms. In *Artificial Intelligence in Tissue and Organ Regeneration* (pp. 15–28). https://doi.org/10.1016/B978-0-443-18498-7.00010-7

Chernyavskaya, V. E., & Safronenkova, E. L. (2020). Linguistic construction of the past: rhetoric in geopolitical conflicts or rhetoric making conflicts? *Terra Linguistica*, *11*(4), 84–93. https://doi.org/10.18721/JHSS.11408

Chu, K. E., Keikhosrokiani, P., & Asl, M. P. (2022). A Topic Modeling and Sentiment Analysis Model for Detection and Visualization of Themes in Literary Texts. *Pertanika Journal of Science and Technology*, *30*(4), 2535–2561. https://doi.org/10.47836/pjst.30.4.14

Cipresso, P., & Riva, G. (2016). Computational psychometrics meets hollywood: The complexity in emotional storytelling. *Frontiers in Psychology*, *7*(NOV). https://doi.org/10.3389/fpsyg.2016.01753

Daelemans, W. (2013). *Explanation in computational stylometry*. *7817 LNCS*(PART 2), 451–462. https://doi.org/10.1007/978-3-642-37256-8_37

Domingos, F., Bagdonas, A., & Zanetic, J. (2024). "So the Lights Have Bent": Investigation of Pre-Service Teachers' Conceptions of Science Through a Historical Narrative on the General Relativity Theory. *Investigacoes Em Ensino de Ciencias*, *29*(2), 201–230. https://doi.org/10.22600/1518-8795.ienci2024v29n2p201

Elliott, C. (2023). *The Unfortunate Footnote: Using the Affective Reasoner to Generate Fortunes-of-Others Emotions in Story-Morphs*. *542 LNNS*, 690–707. https://doi.org/10.1007/978-3-031-16072-1_51

Erker, D., & Guy, G. R. (2012). The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language*, *88*(3), 526–557. https://doi.org/10.1353/lan.2012.0050

Fawaid, A., Assyabani, R., Abdullah, I., Muali, C., Itqan, M. S., & Islam, S. (2025). Human Intelligence and Algorithmic Precision: An Experimental Study of Indonesian Translation

Pedagogy in Higher Education. *Asian Journal of University Education*, *21*(3), 779–792. https://doi.org/10.24191/ajue.v21i3.53

Gefen, A., Saint-Raymond, L., & Venturini, T. (2021). AI for Digital Humanities and Computational Social Sciences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): 12600 LNCS* (pp. 191–202). https://doi.org/10.1007/978-3-030-69128-8_12

Govender, P., Langerman, J., & Joseph, N. (2024). *Authorship Attribution on an Afrikaans Corpus using Burrows Delta*. 172–177. https://doi.org/10.1109/ICT4DA62874.2024.10777202

Hiba, B. (2024). Hedgehogs, foxes, blueprints, and skeletons: Untangling the murky complexity of theoretical and conceptual frameworks. *Energy Research and Social Science*, *111*. https://doi.org/10.1016/j.erss.2024.103468

Janebi Enayat, M. (2025). Computationally derived linguistic features of L2 narrative essays and their relations to human-judged writing quality. *Language Testing in Asia*, *15*(1). https://doi.org/10.1186/s40468-025-00374-9

Kamogashira, T. (2022). Machine Learning in Diagnosis Support with Posturography Data. *Equilibrium Research*, *81*(4), 212–221. https://doi.org/10.3757/jser.81.212

Khafaga, A. F., & Shaalan, I. E.-N. A. W. (2020). Using concordance to decode the ideological weight of lexis in learning narrative literature: A computational approach. *International Journal of Advanced Computer Science and Applications*, *11*(4), 246–252. https://doi.org/10.14569/IJACSA.2020.0110433

Kulesa, J., Induru, S., Hubbard, E., & Bhansali, P. (2024). The Conceptual Framework: A Practical Guide. *Hospital Pediatrics*, *14*(11), e503–e508. https://doi.org/10.1542/hpeds.2024-007794

Masjedy, H., Adel, S. M. R., Amirian, S. M. R., & Zareian, G. (2022). An Overview of Text Mining in Language Studies: The Computational Approach to Text Analytics. *Language Related Research*, *12*(6), 499–531. https://doi.org/10.52547/LRR.12.6.16

Mazlan, N. H., Putra, C. W., & Sulistyo, H. (2025). Understanding reader navigation patterns in multi-path hypertext fiction: A case study approach to Patchwork Girl. *Lingua Technica: Journal of Digital Literary Studies*, *1*(1), 1–12. https://doi.org/10.64595/lingtech.v1i1.3

Meroni, F. (2025). *Exploring Metanarrative Cues in Literary Texts with NooJ: The Case of Les Amours de Psyché et de Cupidon by Jean de La Fontaine*. *2443 CCIS*, 152–164. https://doi.org/10.1007/978-3-031-89810-5_13

Miller, D. (2021). Analysing Frequency Lists. In *A Practical Handbook of Corpus Linguistics* (pp. 77–97). https://doi.org/10.1007/978-3-030-46216-1_4

Möller, R. (2021). *Humanities-Centered AI: From Machine Learning to Machine Training*. Workshop at the 44th German Conference on Artificial Intelligence, September 28, 2021, Berlin, Germany, *3093*, 40–44. https://ceur-ws.org/Vol-3093/paper5.pdf

Moreno, L. G. (2017). Interpreting fictional texts in two-dimensional logic. *Revista de Literatura*, *79*(158), 365–390. https://doi.org/10.3989/revliteratura.2017.02.013

Omar, A. (2021). 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

Peng, Y., Sun, J., Quan, J., Wang, Y., Lv, C., & Zhang, H. (2023). Predicting Chinese EFL Learners' Human-rated Writing Quality in Argumentative Writing Through Multidimensional Computational Indices of Lexical Complexity. *Assessing Writing*, *56*. https://doi.org/10.1016/j.asw.2023.100722

Pinkal, M., & Koller, A. (2012). Semantic research in computational linguistics. In *Semantics: An International Handbook of Natural Language Meaning volume 3* (pp. 2825–2859). https://discovered.ed.ac.uk/permalink/44UOE_INST/iatqhp/alma9923929932102466

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

Rahman, N. F. A., Wang, S. L., Ng, T. F., & Ghoneim, A. S. (2025). Artificial Intelligence in Education: A Systematic Review of Machine Learning for Predicting Student Performance. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *54*(1), 198–221. https://doi.org/10.37934/araset.54.1.198221

Rocco, S. T., & Plakhotnik, S. M. (2009). Literature reviews, conceptual frameworks, and theoretical frameworks: Terms, functions, and distinctions. *Human Resource Development Review*, *8*(1), 120–130. https://doi.org/10.1177/1534484309332617

Romadhani, A. D. (2025). Virtual reality as narrative medium: The emotional effects of full immersion in VR-based film Aladin. *Lingua Technica: Journal of Digital Literary Studies*, *1*(1), 51–62. https://doi.org/10.64595/lingtech.v1i1.29

Ryokai, K., Raffle, H., & Kowalski, R. (2012). *StoryFaces: Pretend-play with ebooks to support social-emotional storytelling*. 125–133. https://doi.org/10.1145/2307096.2307111

Sano, S.-I. (2015). The role of exemplars and lexical frequency in rendaku. *Open Linguistics*, *1*(1), 329–344. https://doi.org/10.1515/opli-2015-0005

Schmidt, M.-L. C. R., Winkler, J. R., Appel, M., & Richter, T. (2023). Emotional shifts, event-congruent emotions, and transportation in narrative persuasion. *Discourse Processes*, *60*(7), 502–521. https://doi.org/10.1080/0163853X.2023.2252696

Šeļa, A. (2021). Differences, distances and fingerprints: The fundamentals of stylometry and multivariate text analysis. *Keel Ja Kirjandus*, *64*(8–9), 696–718. https://doi.org/10.54013/kk764a3

Sherratt, S. (2007). Right brain damage and the verbal expression of emotion: A preliminary investigation. *Aphasiology*, *21*(3–4), 320–339. https://doi.org/10.1080/02687030600911401

Skorinkin, D., & Orekhov, B. (2023). Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta. *Digital Scholarship in the Humanities*, *38*(3), 1247–1266. https://doi.org/10.1093/llc/fqad012

Stańczyk, U. (2011). *Application of DRSA-ANN classifier in computational stylistics*. *6804 LNAI*, 695–704. https://doi.org/10.1007/978-3-642-21916-0_73

Tu, X., Wang, D., & Yang, Q. (2024). *Emotional Analysis in Animated Films Using Big Data and IoT: An In-Depth Study of'Krek'*. 175–182. https://doi.org/10.1145/3697355.3697384

Urberg, M. (2021). Creating return on investment for large-scale metadata creation. *Information Services and Use*, *41*(1–2), 53–60. https://doi.org/10.3233/ISU-210117

Varela, P. J., Albonico, M., Justino, E. J. R., & Assis, J. L. V. D. (2020). Authorship Attribution in Latin Languages using Stylometry. *IEEE Latin America Transactions*, *18*(4), 729–735. https://doi.org/10.1109/TLA.2020.9082216

Varghese, N., & Punithavalli, M. (2019). Lexical and semantic analysis of sacred texts using machine learning and natural language processing. *International Journal of Scientific and Technology Research*, *8*(12), 3133–3140. https://www.ijstr.org/research-paper-publishing.php?month=dec2019

Weitin, T., & Herget, K. (2017). Falcon topics: On some problems of topic modeling of literary texts. *Lili - Zeitschrift Fur Literaturwissenschaft Und Linguistik*, *47*(1), 29–48. https://doi.org/10.1007/s41244-017-0049-3

Witten, I. H. (2004). Text mining. In *The Practical Handbook of Internet Computing* (pp. 14–1). https://doi.org/10.1201/9780203507223

Wotela, K. (2017). *Conceptualising conceptual frameworks in public and business management research*. Conference: 16th European Conference on Research Methodology for Business and Management Studies, 2017-June, 370–379. https://kar.kent.ac.uk/id/eprint/64395

Xia, L., Liu, K., Li, X., & Ye, Q. (2025). Encoding types and narrative coherence modulate the impact of emotions on temporal order memory. *Acta Psychologica Sinica*, *57*(1), 1–17. https://doi.org/10.3724/SP.J.1041.2025.0001

Yogeesh, N., Mohammad, S. I., Raja, N., Reddy, N. A., Hassan, S. R., Kavitha, H. S., Vasudevan, A., Hunitie, M. F. A., & Alshdaifat, N. (2025). Modeling Lexical Ambiguity in English Literature Using Fuzzy Logic and Equations. *Applied Mathematics and Information Sciences*, *19*(4), 873–889. https://doi.org/10.18576/amis/190413

Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability. *Qualitative Report*, *16*(3), 730–744. https://doi.org/10.46743/2160-3715/2011.1085

Zad, S., Heidari, M., Hajibabaee, P., & Malekzadeh, M. (2021). *A Survey of Deep Learning Methods on Semantic Similarity and Sentence Modeling*. 466–472. https://doi.org/10.1109/IEMCON53756.2021.9623078

Změlík, R. (2018). Quantitative and corpus research in literary studies: Possibilities and approaches. *Slovo a Slovesnost*, *79*(1), 47–65. https://www.ceeol.com/search/article-detail?id=717085