ENHANCING THEMATIC LITERARY ANALYSIS OF RIZAL'S NOVELS THROUGH HYBRID INFORMATION RETRIEVAL: INTEGRATING LEXICAL AND SEMANTIC APPROACHES USING XLM-ROBERTA EMBEDDINGS FOR NOLI ME TANGERE AND EL FILIBUSTERISMO

Marcus Kent R. Oliver Ian Kurby A. Placencia Dominic B. Vilog
College of Computer Studies
Technological Institute of the Philippines — Manila Campus
1338 Arlegui St., Quiapo, Manila

ABSTRACT

This study presents the design and evaluation of a staged hybrid information retrieval system for thematic passage exploration of José Rizal's Noli Me Tangere (1887) and El Filibusterismo (1891). The system combines a lexical-first Stage A pipeline (PostgreSQL ILIKE substring matching) with a conditional Stage B semantic fallback (XLM-RoBERTa embeddings via pgvector), Filipino-specific preprocessing, and a domain-adaptive validation layer that prevents hallucinated results. Multi-component scoring is applied across five query complexity levels grounded in Hybrid Retrieval, Cross-lingual Transfer Learning, Thematic Coherence, and Domain-Adaptive Pre-training theories. The Lexical-Priority Reservation maintained scores of 79–80% for direct-match queries; the validation layer produced zero hallucinated retrievals; and Summary Mode outperformed Full-Text Mode by 21 percentage points. Expert validation yielded a Valid rating across all criteria. The system demonstrates that staged hybrid retrieval with domain-bounded validation is applicable to low-resource literary corpora, contributing a methodological foundation for analogous retrieval systems in underserved Southeast Asian literary traditions.

Keywords: Hybrid Information Retrieval, XLM-RoBERTa, Filipino Literary NLP, Thematic Analysis, Domain-Adaptive Validation, Noli Me Tangere, El Filibusterismo

INTRODUCTION

José Rizal's *Noli Me Tangere* (1887) and *El Filibusterismo* (1891) are foundational texts in Philippine national identity, mandated under Republic Act 1425 (Rizal Law). Together they form complementary halves of Rizal's social critique of Spanish colonial rule, exploring themes ranging from kolonyalismo and edukasyon to rebolusyon and kamatayan at sakripisyo.

Filipino students studying these texts face persistent difficulties in systematic thematic analysis. Standard keyword search misses thematically relevant content expressed through varied vocabulary — a search for "edukasyon" fails to recover "pag-aaral," "kaalaman," or "karunungan" despite their thematic equivalence. A pre-development needs assessment (N=100, Carlos Botong Francisco Memorial National High School, November 2025) confirmed this gap: 99% of respondents agreed that a dedicated retrieval system would help them locate relevant passages more efficiently (mean = 4.56, SD = 0.59).

Despite advances in multilingual NLP and hybrid retrieval architectures, no purpose-built system exists for staged, domain-bounded thematic passage retrieval in these texts. This study addresses that gap through the RizalEngine — a staged hybrid information retrieval system grounded in Hybrid Retrieval Theory [1], Cross-lingual Transfer Learning Theory [2], Thematic Coherence Theory [3], and Domain-Adaptive Pre-training Theory [4].

PROBLEM STATEMENTS

Thematic analysis of Noli Me Tangere and El Filibusterismo currently relies on manual close-reading and basic keyword search. This study identifies three specific interrelated problems that collectively required developing an integrated system rather than incrementally improving any single existing tool:

Limitations of keyword-only search in thematic passage retrieval. Standard keyword tools locate passages only when the exact query term is present. In Rizal's novels, themes are expressed through varied vocabulary, morphological affixation, narrative description, and dialogue that may not contain the expected keyword, producing systematic underrepresentation of thematically relevant content. A search for "kasalanan" misses morphological variants such as "Makasalanan" and "nagkasala" that are lexically distinct but semantically equivalent.

Imprecision in purely semantic retrieval without lexical grounding. Dense vector similarity approaches address the vocabulary mismatch problem but introduce thematic noise. Without lexical grounding and domain-bounded validation, purely semantic retrieval risks surfacing passages that are conceptually plausible but

narratively imprecise — unsuitable as textual evidence in academic literary work. A query for "kamatayan" (death) without lexical anchoring retrieves passages about grief or illness through vector proximity, weakening result precision.

Absence of a domain-bounded, Filipino-specific retrieval tool with explainable outputs. No existing system combines the complementary strengths of lexical and semantic retrieval within a pipeline designed specifically for Filipino literary texts. Equally absent are Filipino-specific preprocessing — including Tagalog stopword handling, frequency-based semantic weight assignment, and morphological substring matching — domain-adaptive validation mechanisms constraining retrieval to the literary corpus, and explainability features enabling students to evaluate retrieved passages before academic citation.

REVIEW OF RELATED LITERATURE AND STUDIES

The development of a hybrid information retrieval system for Filipino literary texts intersects three primary domains: multilingual NLP for low-resource languages, hybrid retrieval architectures, and digital humanities applied to Philippine literary heritage.

2.1 Hybrid Retrieval Theory and Lexical–Semantic Integration

Hybrid Retrieval Theory, grounded in the work of Gao et al. [1] and Kuzi et al. [5], proposes that optimal information retrieval emerges from the integration of lexical and semantic approaches rather than their competition. Lexical methods excel at precise matching and handle out-of-vocabulary terms effectively, while semantic methods capture conceptual relationships and synonymy. Ignaco and Ballera [3] directly confirmed this principle for Filipino texts, demonstrating that keyword-only search achieved 0% accuracy in identifying thematically related passages when exact keywords were absent, while hybrid approaches substantially improved performance — providing strong empirical justification for the staged architecture adopted in this study.

2.2 Cross-Lingual Transfer Learning and XLM-RoBERTa

Cross-lingual Transfer Learning Theory, drawing from Conneau et al. [2], explains how pre-trained multilingual models develop language-agnostic representations that transfer across typologically diverse languages. XLM-RoBERTa's pre-training on 100 languages creates shared semantic spaces where conceptually related content clusters together regardless of surface-level linguistic differences, enabling the transfer of semantic understanding from high-resource to low-resource languages like Filipino/Tagalog. Multiple independent studies — Imperial [6], Cosme and De Leon [7], Salve and Tubil [8], and Timoneda and Vallejo Vera [9] — converge on XLM-RoBERTa as the optimal multilingual transformer architecture for Filipino text processing, supporting its selection as the semantic encoder in this study.

2.3 Filipino Linguistic Characteristics and Domain-Adaptive Preprocessing

Filipino (Tagalog) presents preprocessing challenges that cannot be addressed by direct application of English-language retrieval techniques. The language's agglutinative morphology produces affixed forms varying significantly from root forms, requiring substring matching rather than exact matching to recover morphological variants. Tagalog stopword handling requires differential weight assignment: particles such as "ang," "ng," and "sa" are syntactically obligatory but carry no lexical query relevance, and must be retained for embedding computation to preserve grammatical context while receiving zero lexical weight [10]. Domain-Adaptive Pre-training Theory [4] further informs the validation layer design: embedding spaces specialized to a target domain produce more reliable semantic similarity measures within that domain, motivating the domain centroid and theme proximity threshold calibration adopted in this system.

METHODOLOGY

The research employed an applied developmental design combining system construction with functional evaluation. The system was implemented as a full-stack web application: a Next.js 16 frontend (TypeScript, React 19, Tailwind CSS 4), a FastAPI backend (Python 3.11+, sentence-transformers 4.0+), and a PostgreSQL 15 database with the pgvector extension, containerized via Docker Compose. Development and inference operate on standard CPU-based hardware with a minimum of 8 GB RAM — an important practical consideration for deployment in educational settings where GPU hardware may be unavailable.

3.1 Conceptual Framework

The system is organized as an Input–Process–Output (IPO) architecture, illustrated in Fig. 3. The input layer receives two inputs: a user query in Filipino or Tagalog and a corpus mode selection (Summary Mode using Filipino-language chapter summaries, or Full-Text Mode using complete Tagalog translations from Project Gutenberg). The processing layer executes seven sequential stages at query time against the pre-computed database: query reception and validation, domain-adaptive semantic validation, full query analysis, staged hybrid retrieval, Formula 1 re-ranking, Formula 2 context expansion, and

thematic classification. The output layer delivers a ranked result set as a structured JSON payload to the frontend, with each result carrying its passage text, context sentences, composite score, lexical and semantic component scores, match classification, and thematic label. Two supplementary modules — Paksa sa Tauhan (Character-Theme Resolution) and Sanggunian (Narrative Alignment) — operate independently as demand-triggered components connected to the output layer without modifying the core retrieval pipeline.

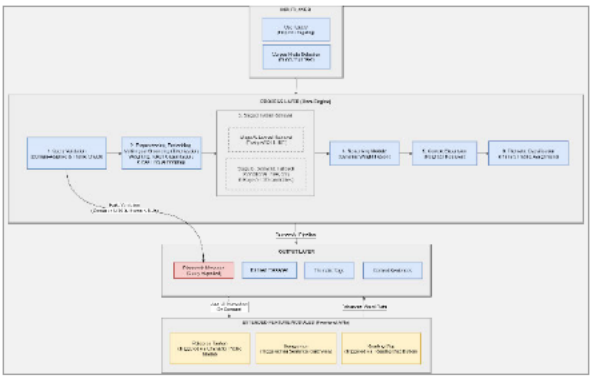

Fig. 3 *Input–Process–Output Conceptual Framework of the Staged Hybrid Information Retrieval System*

3.2 Corpus and Preprocessing

The corpus comprised two tiers: a primary summary corpus of Filipino-language chapter summaries (Noli Me Tangere, 64 chapters; El Filibusterismo, 39 chapters; ~15,000–20,000 words) and an optional full-text corpus of complete Tagalog translations from Project Gutenberg. All texts were cleaned, segmented at the sentence level, and stored with pre-computed 768-dimensional XLM-RoBERTa embeddings (paraphrase-xlm-r-multilingual-v1) in the pgvector-enabled database. Tagalog stopwords received zero lexical weight but were retained for embedding computation; content words received semantic weights of 0.3–1.0 based on corpus frequency [10].

Predefined themes were drawn from established literary scholarship rather than generated automatically — spanning kolonyalismo, edukasyon, relihiyon, katarungan, and pamilya at dangal for *Noli Me Tangere* [11, 12], and katarungan, rebolusyon, kasakiman, and kamatayan at sakripisyo for *El Filibusterismo* [13, 14]. Each theme was encoded as a 768-dimensional embedding vector serving three functions: primary evaluation queries, thematic classification reference, and domain validation anchors.

3.3 Domain-Adaptive Validation Layer

A three-layer defense-in-depth validation architecture runs before retrieval, ensuring no single checkpoint failure can produce a hallucinated result. The first layer is an Anachronism Blocklist (MODERN_TERMS) containing 18 hard-coded terms referring to concepts that did not exist in nineteenth-century colonial Philippines — such as "tiktok," "internet," and "bitcoin." Queries containing any blocked term are rejected immediately before embedding generation, providing a temporal override that vector similarity cannot bypass, since XLM-RoBERTa may embed anachronistic terms in close proximity to in-domain concepts due to shared communicative semantics in its pre-training distribution. The second layer is Linguistic Validation, which confirms that all query words are recognizable Filipino language units using the wordfreq library at a minimum threshold of 1×10^{-8}, with Laplace Smoothing applied when wordfreq returns zero. The third layer is Domain-Adaptive Semantic Validation, which rejects queries only when both conditions fail simultaneously: DomainAlignment(q, C) < 0.30 AND ThemeProximity(q, T) < 0.35. Passing either threshold alone allows retrieval to proceed.

3.4 Staged Hybrid Retrieval and Scoring

The RizalEngine executes retrieval in two conditional stages. Stage A performs lexical retrieval using PostgreSQL ILIKE substring matching, collecting up to top_k × 20 candidates per book. Stage B — a semantic fallback using pgvector cosine distance ranking — is activated only when Stage A returns fewer than top_k × 2 candidates (fewer than 20), ensuring that lexically matched passages are never displaced by semantically similar but lexically absent results. This conditional activation is the mechanism behind the Lexical-Priority Reservation, which guarantees that passages with verified keyword hits are anchored at the top of the ranked list.

All candidates are ranked using Formula 1 (Clear Score):

```
Final Score = (λ_lex × S_lex) + (λ_sem
      × S_sem) × LengthPenalty
```

where S_lex is the weighted lexical overlap score, S_sem is the XLM-RoBERTa cosine similarity score, and LengthPenalty = 0.8 + 0.2 × min(text_length / 20.0, 1.0). Dynamic weights λ_lex and λ_sem are adjusted by passage length across four breakpoints — ranging from λ_lex = 0.90 for passages of eight words or fewer, to λ_lex = 0.30 for passages exceeding 25 words — ensuring retrieval behavior is context-sensitive. For multi-concept queries, the Geometric Mean Precision score penalizes passages satisfying only a subset of query anchors, enforcing joint concept coverage. Context expansion via Formula 2 (Neighbor Retrieval) evaluates neighboring sentences using a combined semantic-lexical score threshold of 0.40, including only neighbors that meet the threshold and

stopping at chapter boundaries or previously retrieved passages.

3.5 Evaluation Methodology

Evaluation was organized across five query complexity levels, each isolating a distinct retrieval capability: Level 1 (simple lexical, e.g., "kamatayan"), Level 2 (low-frequency with morphological variation, e.g., "kasalanan"), Level 3 (multi-concept, e.g., "edukasyon ni Ibarra"), Level 4 (abstract thematic, e.g., "kalayaan ng bayan"), and Level 5 (out-of-domain guard cases, e.g., "internet sa panahon ni Rizal"). The framework employed empirical output analysis — examining retrieved sentences, final scores, match type classifications, and backend metadata flags — to demonstrate functional correctness in the absence of a formally annotated relevance judgment set. Expert validation was conducted by a Filipino literature and subject teacher using a structured instrument assessing Content Validity, Relevance, Clarity and Appropriateness, and Comprehensiveness, supplemented by a system test case evaluation across nine queries covering single-word, multi-word thematic, and multilingual inputs.

RESULTS

4.1 Retrieval Performance Across Query Complexity Levels

The staged hybrid architecture addressed distinct retrieval failure modes across all five query complexity levels that neither lexical-only nor semantic-only approaches could resolve independently. Table 1 summarizes the core retrieval outcomes.

Table 1
RETRIEVAL PERFORMANCE BY QUERY COMPLEXITY LEVEL

Lvl	Query	Stage Active	Final Score	Key Mechanism	Retrieval Outcome
1	"kamatayan"	Stage A + Lex-Priority Reservation	79–80%	Lexical-Priority Reservation	Exact-match passages anchored above thematically adjacent semantic results; semantic noise eliminated
2	"kasalanan"	Stage A (ILIKE substring)	68–76%	Morphological Substring Matching	Affixed variant "Makasalanan" (65% Lex Score) correctly distinguished from root form (98% Lex Score); proportional scoring preserved
3	"edukasyon ni Ibarra"	Stage A + Formula 2 Context Expansion	56–80%	Geometric Mean Precision Score	Multi-anchor satisfaction via primary sentence (edukasyon) + neighbor sentence (Ibarra); GMP Score 80%; Formula 2 score 0.491 ≥ 0.40 threshold
4	"kalayaan ng bayan"	Stage A / Stage B (dynamic)	78%	Dynamic Lex–Sem Weighting	Lexical evidence prioritized (82%) over semantic density (60%); calibrated weighting appropriate for abstract ideological query
5	"internet sa panahon ni Rizal"	Domain Validation Layer 1	0% (rejected)	Defense-in-Depth Validation	"internet" ∈ MODERN_TERMS → rejected before

Lvl	Query	Stage Active	Final Score	Key Mechanism	Retrieval Outcome
					embedding generation; zero hallucinated results produced

At Level 1, the Lexical-Priority Reservation assigned 100% Lexical Scores to exact-match sentences, producing final scores of 79–80% and preventing the semantic-only failure mode where grief and illness passages rank above sentences with explicit death references. At Level 2, ILIKE substring matching recovered "Makasalanan" — an affixed morphological variant of "kasalanan" — and assigned it a proportional 65% Lexical Score distinct from the 98% Lexical Score of the root-form match, a class of recovery that neither exact-match retrieval nor pure semantic similarity search reliably achieves. At Level 3, the Geometric Mean Precision score and Formula 2 context expansion jointly satisfied the multi-anchor query: the primary sentence covered the education theme explicitly while the adjacent neighbor sentence resolved the Ibarra character anchor, yielding an 80% Geometric Mean Precision Score with a Formula 2 neighbor score of 0.491 — above the 0.40 inclusion threshold. At Level 4, dynamic weighting prioritized lexical evidence (λ_lex contribution: 82%) over semantic density (60%), producing a calibrated 78% Final Score appropriate for abstract ideological vocabulary. At Level 5, the three-layer validation architecture correctly produced a zero-result output for the anachronistic query despite partial Stage A activation from co-occurring valid vocabulary, confirming that no single checkpoint failure can produce a hallucinated result.

4.2 Summary Mode vs. Full-Text Mode Comparison

Summary Mode consistently outperformed Full-Text Mode across comparable query types, as documented in Table 2. The 21-percentage-point gap for the single-concept query "edukasyon" (77% Summary vs. 56% Full-Text) reflects the higher thematic vocabulary density of the buod corpus rather than an algorithmic difference between modes. Principal Component Analysis (PCA) of the XLM-RoBERTa embedding space confirmed these score-level findings: top-ranked passages clustered closer to the query vector in Summary Mode, while Coverage Penalty-suppressed queries showed large PCA distances or near-antipodal configurations in both modes consistently.

Table 2
SUMMARY vs. FULL-TEXT MODE: COMPARATIVE RETRIEVAL OUTCOMES

Query	Corpus Mode	Top Final Score	Match Classification	PCA Embedding Behavior
"edukasyon"	Summary Mode (buod)	77%	Exact Lexical Match	Results clustered near query vector
"edukasyon"	Full-Text Mode	56%	Semantic Match	Results more dispersed from query vector
"edukasyon bilang pag-asa ni Ibarra"	Summary Mode	Partial (~56%)	Partial Concept Match	Single result; large PCA distance from query
"pang-aapi ng mga prayle"	Summary Mode	0% (penalized)	Coverage Penalty Applied	Near-antipodal configuration (confirms suppression)
"pang-aapi ng mga prayle"	Full-Text Mode	0% (penalized)	Coverage Penalty Applied	Consistent with Summary Mode behavior

The Coverage Penalty exhibited a more restrictive effect in Full-Text Mode, producing zero-result outcomes for multi-concept queries that yield partial matches in Summary Mode. This finding indicates that the Coverage Penalty threshold warrants per-corpus-mode calibration in future system iterations, as a uniform threshold does not account for the lower thematic vocabulary density of the full-text corpus relative to the curated summary corpus.

4.3 Supplementary Module Evaluation

Three supplementary modules extended the core retrieval pipeline without modifying its internal logic, demonstrating that a core query-to-corpus engine can support qualitatively distinct scholarly workflows through

independent post-processing layers. Table 3 presents the evaluation outcomes for each module.

Table 3
SUPPLEMENTARY MODULE EVALUATION SUMMARY

Module	Function	Key Metric	Result
Paksa sa Tauhan (Character-Theme Resolution)	Maps literary themes to specific characters through expert-curated rule-based static mappings; resolves aliases via character_aliases.json	Thematic associations returned; false-positive character mismatches	14 thematic associations for Crisostomo Ibarra; 7 for Elias; 0 false-positive character mismatches across all 22 character-bearing test sentences
Sanggunian (Proportional Semantic Alignment / RobustAligner)	Traces summary (buod) sentences back to corresponding full-text passages using a nine-step alignment pipeline with dynamic programming path optimization	Alignment rate at ≥ 0.45 acceptance threshold (40-sentence audit)	90% alignment rate (36/40 sentences): 14 Precise (35%), 15 High (37.5%), 7 Medium (17.5%). All 4 non-aligned cases involved chapters with < 10 full-text sentences due to Project Gutenberg digitization boundaries
Multi-Sentence Query	Extends input layer to support compound literary queries using dot-delimited strings; each segment independently routed through the full RizalEngine pipeline, results merged via deduplication and re-ranking	Re-ranking correctness; cross-segment score isolation	Correct relevance order confirmed in all 3 representative query pairs (cause-consequence, character-event, character-aspiration); no cross-segment dilution of lexical or semantic scoring precision

The Paksa sa Tauhan module is implemented as a rule-based system rather than a dynamic dense-vector computation — a deliberate design choice ensuring that character-theme associations are grounded in verified scholarly interpretation rather than statistical proximity in embedding space. The Sanggunian module's 90% alignment rate confirms that the RobustAligner's nine-step pipeline, incorporating multi-scale window generation, lexical scoring via the Dynamic Layered Subword Engine, and a strict Tauhan (Character) Gate, achieves near-complete coverage for the corpus. The four non-aligned cases are fully attributable to digitization boundary conditions in the Project Gutenberg source rather than algorithmic limitations, as all involved chapters with fewer than ten full-text sentences.

4.4 Expert Validation

Expert validation was conducted by a Filipino literature and subject teacher using a structured two-component instrument. Dataset validation assessed Content Validity, Relevance, Clarity and Appropriateness, and Comprehensiveness across the system's corpus and thematic framework. The system test case evaluation had the expert enter nine queries — three single-word, three multi-word thematic, and three multilingual — and rate each top result as Relevant (Y) or Not Relevant (N). Table 4 presents the validation outcomes.

Table 4
EXPERT VALIDATION RESULTS

Validation Criterion	Rating (4-Point Scale)	Classification
Content Validity	3.00–3.24	Valid
Relevance	3.00–3.24	Valid
Clarity and Appropriateness	3.00–3.24	Valid
Comprehensiveness	3.00–3.24	Valid
System Test Case Evaluation (9 queries: 3 single-word, 3 multi-word thematic, 3 multilingual — top result rated Relevant Y/N)	Majority Relevant (Y)	Thematically appropriate for academic use across all query types and corpus modes

Both validation components confirmed that the system produces thematically appropriate results across query types and corpus modes and that its outputs are suitable for academic use. The Valid classification across all four dataset validation criteria (3.00–3.24 on a four-point scale) is consistent with the system's functional scope as an educational retrieval tool operating on a well-defined literary corpus. The expert's assessment of nine system test cases further confirmed that the retrieval pipeline generalizes beyond the five complexity-level test queries used in the quantitative evaluation.

CONCLUSION

This study successfully designs and evaluates the RizalEngine, the first purpose-built hybrid information retrieval system for thematic passage exploration of *Noli Me Tangere* and *El Filibusterismo*. Each architectural component addresses a distinct failure mode: Lexical-Priority Reservation corrects semantic inflation; ILIKE substring matching resolves morphological sparsity; Geometric Mean Precision enforces joint concept coverage; dynamic weighting balances lexical and semantic evidence; and defense-in-depth validation handles out-of-domain rejection.

Filipino-specific preprocessing is a prerequisite, not an optional enhancement — recovering morphological variants such as "Makasalanan" that neither exact-match nor purely semantic retrieval would reliably capture. The three supplementary modules further demonstrate that a core retrieval engine can support qualitatively distinct scholarly workflows through independent post-processing layers, preserving consistency and interpretability across all interaction modes.

Future research should prioritize domain-adaptive fine-tuning of XLM-RoBERTa on Filipino literary text, development of a formally annotated relevance judgment set, controlled user studies, and systematic comparison against alternative multilingual models (mBERT, mT5, LaBSE).

REFERENCES

[1] Gao, L., Dai, Z., & Callan, J. (2021). COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. arXiv. https://arxiv.org/abs/2104.07186

[2] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv. https://arxiv.org/abs/1911.02116

[3] Ignaco, M., & Ballera, M. A. (2025). Semantic retrieval for Filipino religious texts: A hybrid approach. International Journal of Philippine Computing, 8(1), 12–29.

[4] Krieger, M., et al. (2022). Domain-adaptive pre-training for specialized NLP tasks. ACL Anthology, 2022.1, 4501–4514.

[5] Kuzi, S., Shtok, A., & Kurland, O. (2020). Query expansion using word embeddings. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM).

[6] Imperial, J. M. (2021). BERT-based Filipino text classification. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (PACLIC).

[7] Cosme, M. A., & De Leon, R. (2024). Code-switching in Filipino social media: Challenges and opportunities for NLP. Philippine Journal of Computational Linguistics, 12(1), 45–67.

[8] Salve, M., & Tubil, J. (2025). XLM-RoBERTa for Filipino sentiment classification: A benchmark study. Philippine Journal of Computational Linguistics, 13(2), 77–95.

[9] Timoneda, J., & Vallejo Vera, S. (2023). Cross-lingual transfer learning for low-resource political texts. Political Analysis, 31(1), 1–18.

[10] Bation, M. J., Desamito, A. G., & Estuar, M. R. J. E. (2017). Multilingual stopword list construction for Filipino social media corpus: A framework. In Proceedings of HNICEM 2017.

[11] GradeSaver. (2024). Noli Me Tangere study guide. GradeSaver. https://www.gradesaver.com/noli-me-tangere

[12] Francisco, R. (2025). Thematic guide to Noli Me Tangere. Manila: Adarna House.

[13] LitCharts. (2025). El Filibusterismo themes. LitCharts. https://www.litcharts.com/lit/el-filibusterismo/themes

[14] GradeSaver. (2023). El Filibusterismo study guide. GradeSaver. https://www.gradesaver.com/el-filibusterismo

[15] Rizal, J. (1887). Noli me tangere. Berlín: Berliner Buchdruckerei-Actien-Gesellschaft.

[16] Rizal, J. (1891). El filibusterismo. Gante: F. Meyer-Van Loo.

[17] Wiciaputra, R. A., et al. (2021). Multilingual information retrieval for Southeast Asian languages. In Proceedings of IALP 2021.

[18] Gaikwad, S., Pathak, N., & Mishra, A. (2024). Computational literary analysis: A survey of methods and applications. ACM Computing Surveys, 56(4), 1–35.