LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

ARTIFICIAL INTELLIGENCE AND DIGITAL HUMANITIES FOR COMPUTATIONAL ANALYSIS OF LITERARY STYLE, NARRATIVE PATTERNS, AND THEMATIC EVOLUTION IN ENGLISH LITERATURE

Tahreem Sheikh*[1], Fakiha Umar[2], Tayyeba Ayyub[3], Kinza Saleem[4]

[1,2]*Department of English, Bahauddin Zakariya University, Multan, Pakistan*
[3]*Department of English, International Islamic University Islamabad, Islamabad, Pakistan*
[4]*Department of English Literature, Government College University Faisalabad, Faisalabad, Pakistan*

*[1]tahreemsheikh000@gmail.com, [2]fakihaumarumar@gmail.com, [3]tayyebaayyubrao@gmail.com, [4]kinza200645@gmail.com

DOI: https://doi.org/10.5281/zenodo.23591242

Keywords
digital humanities; computational stylistics; distant reading; natural language processing; topic modelling; narrative networks

Article History
Received: *15 July 2026*
Accepted: *06 October 2026*
Published: *22 October 2026*

Copyright @Author
Corresponding Author: *
Tahreem Sheikh

Abstract

The computational turn in literary scholarship has reorganised the terms on which English literature can be read, compared, and historicised, yet the field still lacks an integrated account of how artificial intelligence, natural language processing, and digital humanities methods jointly illuminate literary style, narrative form, and thematic change. This paper advances a conceptual and methodological framework that unites three analytical registers that are usually pursued in isolation: stylometric attribution and authorial fingerprinting, the structural modelling of narrative and character relations, and the diachronic tracing of theme across literary history. Building on the distant-reading tradition associated with Moretti (2013), the macroanalytic programme of Jockers (2013), the stylometric lineage descending from Burrows (2002), and the historically oriented machine learning of Underwood (2019), the study argues that transformer-based language models now permit a semantically sensitive reading of literary form that earlier bag-of-words approaches could not support. The proposed pipeline combines Python-based preprocessing (NLTK, spaCy), stylometric classification (Stylo), exploratory visualisation (Voyant), character-network analysis (Gephi), and topic modelling that moves from Latent Dirichlet Allocation to the embedding-based BERTopic. Illustrative readings of Austen, Dickens, and Woolf show how these instruments can be made to answer genuinely literary questions rather than merely to count words. The paper also engages the methodological critique of computational literary studies, arguing that interpretive accountability, not technical sophistication, is the discipline's central problem.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

1. Introduction

For most of its institutional history, the study of English literature has been organised around the close reading of a comparatively small canon. The interpretive protocols inherited from practical criticism and, later, from theory privileged the single text attended to with great care, and the value of a reading was measured by its attentiveness to nuance rather than by the breadth of its evidentiary base. This settlement produced extraordinary interpretive riches, but it also left a structural blind spot: the overwhelming majority of what was published in English was never read by literary scholars at all, and conclusions about periods, genres, and movements were routinely drawn from a canonical fraction that may be unrepresentative of the whole (Moretti, 2013). The digital humanities emerged, in part, as a response to this asymmetry between the scale of literary production and the scale of literary attention.

The intervening two decades have seen the gradual consolidation of a set of computational methods capable of operating on corpora far larger than any individual could read. Stylometry has matured into a reliable instrument for authorship attribution and for the quantitative description of style (Burrows, 2002; Eder, Rybicki, & Kestemont, 2016). Topic modelling has given scholars a tractable way to describe the thematic composition of thousands of documents at once (Blei, Ng, & Jordan, 2003). Network analysis has reframed the study of character and plot as a question about structure rather than about individual figures (Moretti, 2005). Most recently, the arrival of contextual language models has begun to close the long-standing gap between what computers can count and what literary critics actually care about, namely meaning in context (Devlin, Chang, Lee, & Toutanova, 2019). Taken together, these developments constitute not merely a new toolkit but a reconfiguration of the kinds of questions that literary history can ask.

Yet the field remains curiously fragmented. Stylometrists, narratologists, and intellectual historians each work with their own instruments, their own corpora, and their own evaluative norms, and the result is a body of work that is methodologically sophisticated but conceptually siloed. A study that fingerprints an author's prose rarely speaks to a study that maps a novel's character network, and neither tends to engage the diachronic question of how a theme rises and falls across a century. This compartmentalisation is intellectually costly, because style, structure, and theme are not separable features of a literary work; they are mutually constituting dimensions of the same artefact. The free indirect discourse that gives Austen her distinctive stylistic signature is also the mechanism through which her novels organise narrative sympathy, which is in turn inseparable from their thematic preoccupation with judgement and self-knowledge. An analysis that treats these as independent variables risks mistaking the convenience of its methods for the structure of its object.

1.1 Problem Statement

The central problem this study addresses is the absence of an integrated computational framework through which literary style, narrative pattern, and thematic evolution can be analysed as interrelated rather than independent phenomena. Three difficulties follow from this absence. First, there is a methodological problem of fragmentation: the dominant techniques were developed within different disciplinary communities and have not been articulated into a common

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

pipeline that a literary researcher could apply to a single corpus with a single interpretive purpose. Second, there is a representational problem inherited from the field's reliance on bag-of-words models. Methods such as Latent Dirichlet Allocation treat a text as an unordered collection of tokens and are therefore constitutively blind to word order, syntax, irony, and the contextual modulation of meaning on which literary effect depends (Da, 2019). Third, there is an interpretive problem of accountability: computational outputs are frequently presented as findings without a sufficiently explicit account of how a cluster, a topic, or a network metric is to be translated into a claim about literature. This study treats these three difficulties as a single connected problem and proposes a framework designed to address them together.

2. Theoretical Framework

The framework proposed here rests on a single theoretical premise: that style, narrative structure, and theme are not three separable properties of a literary text but three scales of description of one underlying object. This premise has consequences for method. If the three dimensions are interdependent, then an analysis that isolates one of them will systematically distort the others, and the appropriate response is not to perfect each instrument in isolation but to coordinate the instruments so that their outputs can be read against one another. The theoretical scaffolding for such coordination is drawn from three sources that are rarely brought into contact.

The first is the systems-level conception of literary history that underwrites distant reading, in which the unit of analysis is not the work but the field of works, and in which formal features are understood as distributed properties of that field rather than as achievements of individual genius (Moretti, 2005, 2013). The second is the structuralist inheritance of narratology, which had long treated narrative as a system of relations, between characters, events, and points of view, before computation gave those relations a measurable form; network analysis can be understood as a formalisation of this structuralist intuition, rendering the relational logic of plot as a graph whose properties can be quantified. The third is the distributional hypothesis of modern linguistics, the proposition that the meaning of a word is a function of the company it keeps, which provides the theoretical justification for representing literary language as vectors in a semantic space and which finds its most powerful expression in contextual embedding models (Devlin et al., 2019).

Bringing these three traditions together yields a layered model of the literary text. At the lexical layer, the distributional frequencies of function words encode an authorial signature largely independent of conscious intention. At the relational layer, the co-occurrence of named entities across the span of a narrative encodes its structure of attention and sympathy. At the semantic layer, the contextual meaning of passages encodes the thematic preoccupations that rise and fall across literary history. The claim of this study is that these layers are legible to distinct computational instruments but intelligible only when their readings are integrated, and that the integration is itself an interpretive act requiring the judgement of a literary reader rather than the automatic output of a pipeline.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

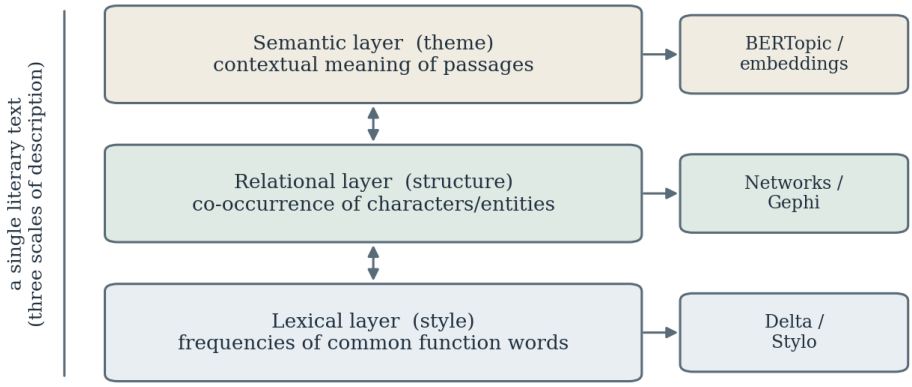


Figure 1. A layered model of the literary text. Style, structure, and theme are treated as three interdependent scales of description of a single work, each legible to a distinct computational instrument.

3. Literature Review

The scholarship relevant to this study can be organised around four intersecting lineages: the distant-reading programme, the macroanalytic tradition, computational stylistics, and the recent turn toward machine learning and historically oriented modelling. Each lineage has contributed a distinct conception of what it means to read literature at scale, and each has attracted its own critique.

3.1 Distant Reading and the Macroanalytic Programme

The phrase that has come to name the entire enterprise is Franco Moretti's "distant reading," a deliberately provocative inversion of the close reading that had defined literary study. For Moretti (2013), distance is not a regrettable loss of textual intimacy but a positive condition of knowledge: certain objects of literary history, such as the life cycle of a genre or the diffusion of a form across national literatures, become visible only when the individual text recedes and the aggregate pattern comes into focus. His earlier Graphs, Maps, Trees (Moretti, 2005) had already proposed that quantitative graphs, geographical maps, and evolutionary trees could serve as abstract models for literary history, displacing interpretation of the single work in favour of the analysis of systemic form. The provocation was double-edged, and Moretti has been criticised both for the thinness of some of his quantitative demonstrations and for a rhetoric that seemed to disparage the close

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

reading it sought to supplement. Nonetheless, the reorientation he proposed, from the exceptional text to the literary system, remains the field's founding gesture.

Matthew Jockers gave this orientation a more rigorous empirical footing. Macroanalysis (Jockers, 2013) is at once a manifesto and a demonstration, applying stylistic and thematic classification to a corpus of some nineteenth-century novels in order to trace influence, nationality, and gendered style across literary history. Where Moretti often argued from secondary data and schematic models, Jockers worked directly from full text and from reproducible statistical procedures, and his later handbook (Jockers, 2014) did much to make these methods teachable to literary scholars without a computational background. The macroanalytic wager is that patterns invisible at the scale of the individual novel, such as the slow drift of thematic vocabulary across decades, are nonetheless real features of literary history that quantitative analysis can recover. Andrew Piper (2018) has extended this wager into a broader argument about the epistemology of literary data, insisting that quantification does not replace interpretation but reconfigures the evidence to which interpretation must answer.

3.2 Computational Stylistics and Authorship

Running parallel to distant reading is the older and, in some respects, more technically settled tradition of computational stylistics. Its decisive modern instrument is Burrows's Delta (Burrows, 2002), a measure of stylistic difference that compares texts on the standardised frequencies of their most common function words, those semantically inconspicuous items such as articles, prepositions, and pronouns whose frequencies an author deploys unconsciously and therefore cannot easily disguise. The counter-intuitive premise of the method is that an author's fingerprint lies not in the distinctive content words that a reader notices but in the fine-grained distribution of the words a reader ignores. Delta proved remarkably robust, and a substantial subsequent literature has tested, refined, and extended it (Hoover, 2004; Rybicki & Eder, 2011). Its consolidation into the freely available Stylo package (Eder et al., 2016) placed a rigorous stylometric toolkit within reach of literary scholars, enabling exploratory cluster analyses and attribution studies without bespoke programming. Stylometry thus offers the integrated framework its most mature quantitative description of style, even as it brackets the semantic and structural dimensions of the literary work.

3.3 Narrative Networks and Thematic Modelling

A third body of work treats the novel as a structure of relations rather than as a sequence of sentences. The application of network science to literature, prefigured in Moretti's (2005) network diagrams of Hamlet, reconceives character as a position within a web of interactions, so that protagonism becomes a measurable property, a matter of centrality within the graph, rather than a critical assumption. This reframing is particularly consequential for the multi-plot Victorian novel, whose large casts and interwoven storylines exceed the descriptive capacity of conventional close reading; the sprawling social architecture of a novel such as Bleak House becomes analytically tractable once it is represented as a network whose communities correspond to the novel's distinct

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

social worlds. Parallel to this structural line runs the thematic one. Probabilistic topic modelling, and above all Latent Dirichlet Allocation (Blei et al., 2003), gave literary scholars a method for inferring the latent thematic structure of a corpus from nothing but word co-occurrence, and the technique was rapidly adopted for the diachronic study of theme. Its limitations, however, are precisely those of the bag-of-words assumption on which it rests: because it ignores word order and context, it cannot distinguish between the literal and the ironic, nor register the semantic drift of a word across a century. The embedding-based successors to LDA, of which BERTopic (Grootendorst, 2022) is the most widely used, were developed in direct response to these limitations and form the thematic core of the present framework.

3.4 Machine Learning and the Historical Turn

The most consequential recent development is the migration of literary computation from counting to learning. Ted Underwood's Distant Horizons (Underwood, 2019) exemplifies this shift, using supervised machine learning not to attribute texts but to model literary change itself, asking how stable the markers of genre or of literary prestige remain across long historical spans. Underwood's striking finding, that certain distinctions persist with a continuity that neither periodising literary history nor theories of rupture would predict, is possible only because predictive modelling can quantify the gradualness of change in a way that episodic close reading cannot. The subsequent arrival of transformer architectures, and of contextual embeddings in particular (Devlin et al., 2019), promises to address the representational poverty that limited earlier methods. Where Latent Dirichlet Allocation (Blei et al., 2003) could model theme only as a distribution over an unordered vocabulary, embedding-based approaches such as BERTopic (Grootendorst, 2022) represent each passage as a dense, context-sensitive vector, allowing the clustering of semantically related material even where the surface vocabulary differs. For a discipline whose object is meaning in context, this is a difference of kind rather than of degree.

3.5 The Critique of Computational Literary Studies

No responsible framework can ignore the sustained critique that computational literary studies has attracted. The most forceful statement is Nan Z. Da's (2019) argument that the statistical reasoning in much published work is either trivial or unsound, that significant findings are frequently artefacts of method, and that the translation from numerical result to literary claim is too often performed by rhetorical sleight of hand rather than by argument. Whatever one makes of its polemical register, the critique names a real hazard: the ease of producing a cluster or a topic can substitute for the harder labour of justifying its interpretive significance. Stephen Ramsay (2011) had earlier offered a more constructive version of the same concern, proposing that computation in the humanities should be understood not as a hypothesis-testing science but as an "algorithmic criticism" whose purpose is to deform the text productively and provoke new readings. Michael Kirschenbaum (2010), writing on the institutional location of the digital humanities, and Willard McCarty (2005), writing on its methodological self-understanding, both insist that the field's legitimacy rests on its accountability to humanistic questions rather than on its technical

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

apparatus. This study takes these critiques as design constraints: the proposed framework is organised so that every quantitative step terminates in an explicit, contestable interpretive claim.

4. Research Objectives

The study pursues four objectives. The first is to develop an integrated analytical framework in which stylometric, structural, and thematic methods operate on a common corpus toward a unified interpretive end, rather than as isolated techniques. The second is to evaluate the extent to which transformer-based language models improve upon bag-of-words methods for the specifically literary tasks of detecting irony, free indirect discourse, and contextually modulated theme. The third is to demonstrate, through focused case studies of Austen, Dickens, and Woolf, how computational outputs can be translated into defensible claims about literary style, narrative organisation, and thematic change. The fourth is to formulate a protocol of interpretive accountability that specifies how a stylometric distance, a network metric, or a topic distribution is to be validated against qualitative reading before it is admitted as evidence.

5. Research Questions

These objectives are addressed through four research questions. First, to what extent can an integrated computational pipeline capture the interdependence of style, narrative structure, and theme within a single literary corpus? Second, how far do contextual embedding models outperform frequency-based models in identifying literary phenomena, such as free indirect discourse in Austen, that depend on contextual meaning rather than on lexical surface? Third, what do character-network metrics reveal about narrative organisation in the multi-plot Victorian novel, and how do these metrics correspond to established critical accounts of Dickensian form? Fourth, how does the thematic composition of the English novel evolve across the long nineteenth and early twentieth centuries, and does an embedding-based topic model recover transitions, such as the modernist turn inward associated with Woolf, that a frequency-based model obscures?

6. Methodology

The methodology is designed as a sequential pipeline in which each stage prepares the evidence for the next and in which every computational operation is paired with a mode of qualitative validation. The design is deliberately modular, so that the framework can be applied in whole or in part and so that each instrument can be interrogated independently.

6.1 Corpus Construction

The study draws on a purpose-built corpus of English prose fiction spanning roughly 1810 to 1930, assembled from public-domain full texts available through Project Gutenberg and the HathiTrust Digital Library. The temporal span is chosen to bracket the consolidation of the realist novel, its Victorian elaboration, and its modernist reorganisation, so that diachronic thematic analysis has a meaningful historical arc to describe. Three focal authors anchor the corpus and supply the qualitative case studies: Jane Austen, whose mastery of free indirect discourse poses the

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

hardest test for a model's sensitivity to contextual meaning; Charles Dickens, whose multi-plot novels and large casts make narrative structure tractable to network analysis; and Virginia Woolf, whose stream-of-consciousness technique marks the thematic and formal transition that the diachronic analysis must recover. Each text is cleaned of paratextual material, normalised for encoding, and annotated with bibliographic metadata, since reliable publication dates are a precondition of any diachronic claim.

Before the framework is applied across the full diachronic corpus, it is to be exercised on a deliberately small pilot of thirty texts, balanced across the three focal registers: ten works from the Austen–Brontë grouping, ten from the Dickens–Eliot grouping, and ten from the Woolf–Joyce grouping. The purpose of the pilot is not to generate substantive literary findings but to validate the pipeline end to end, to expose the sensitivity of each stage to preprocessing decisions, and to establish the annotation standards against which the later full-scale analysis will be measured. The pilot sample is small enough to permit exhaustive qualitative checking of every computational output, which is precisely what the full corpus will not allow, and it therefore functions as the calibration phase on which the accountability of the wider study depends.

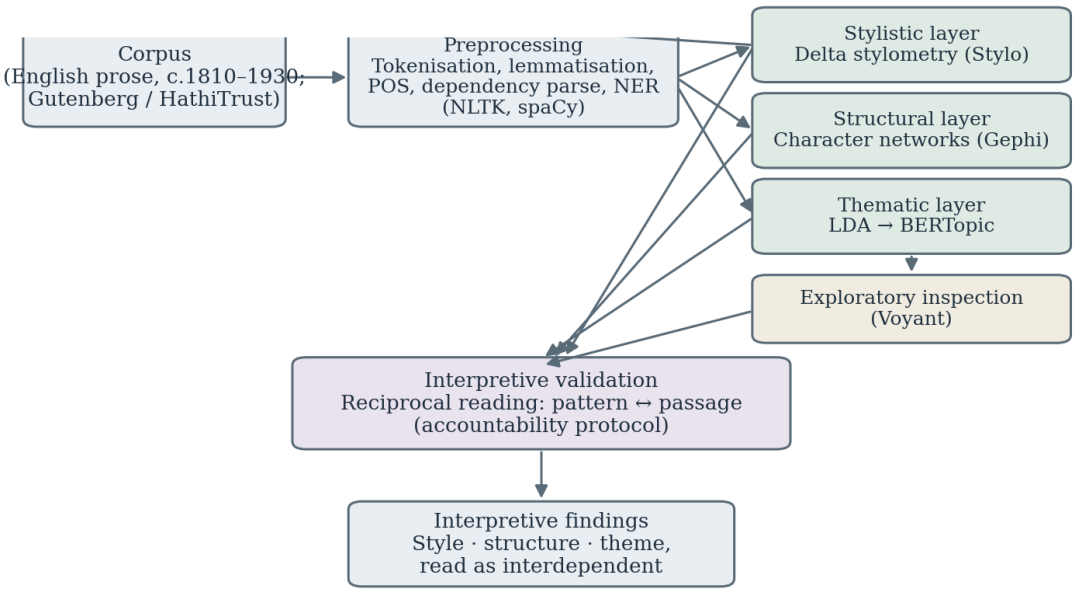


Figure 2. The integrated computational pipeline. Preprocessing feeds three parallel analytical layers (stylistic, structural, thematic), whose outputs converge on an interpretive-validation stage before any pattern is admitted as a finding.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

6.2 Tools and Their Literary Rationale

Preprocessing is performed in Python using the Natural Language Toolkit (Bird, Klein, & Loper, 2009) and spaCy (Honnibal, Montani, Van Landeghem, & Boyd, 2020). NLTK supplies tokenisation, lemmatisation, and part-of-speech tagging, operations that may appear merely clerical but that carry interpretive weight: the decision to lemmatise, for instance, collapses the inflectional variation on which some stylistic effects depend and must therefore be applied selectively rather than by default. spaCy contributes dependency parsing and named-entity recognition, the latter providing the character mentions from which narrative networks are later constructed. Stylometric analysis uses the Stylo package (Eder et al., 2016), which implements Burrows's Delta and its variants together with the cluster and consensus visualisations that make authorial grouping legible. Voyant Tools (Sinclair & Rockwell, 2016) supports rapid exploratory inspection, allowing the researcher to move between the aggregate view and the individual passage, a movement essential to interpretive accountability. Character networks are built and analysed in Gephi (Bastian, Heymann, & Jacomy, 2009), whose centrality and community-detection algorithms translate the sprawling casts of Dickensian fiction into quantifiable structures of narrative attention. Thematic analysis proceeds in two passes, first with Latent Dirichlet Allocation (Blei et al., 2003) as an interpretable baseline and then with BERTopic (Grootendorst, 2022), whose transformer embeddings (Devlin et al., 2019) permit the semantically sensitive clustering that the literary object requires.

6.3 Analysis Pipeline

The analytical workflow proceeds through four linked stages. In the stylistic stage, Delta-based stylometry establishes authorial signatures and tests them against known attributions, after which the function-word profiles that drive the classification are examined qualitatively to determine what, if anything, they reveal about style beyond the fact of authorship. In the contextual stage, a fine-tuned transformer classifier is trained to identify passages of free indirect discourse in Austen, and its predictions are compared against a manually annotated gold standard, so that the claimed superiority of contextual models over frequency-based ones is measured rather than merely asserted. In the structural stage, named-entity recognition and co-occurrence within a sliding window generate character networks for the Dickens subcorpus, whose centrality and modularity metrics are then read against the critical literature on the multi-plot novel. In the thematic stage, LDA and BERTopic are applied across the diachronic corpus, and their outputs are aligned on the publication timeline to trace the rise and decline of thematic clusters, with particular attention to whether the embedding model recovers the modernist interiorisation that culminates in Woolf.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

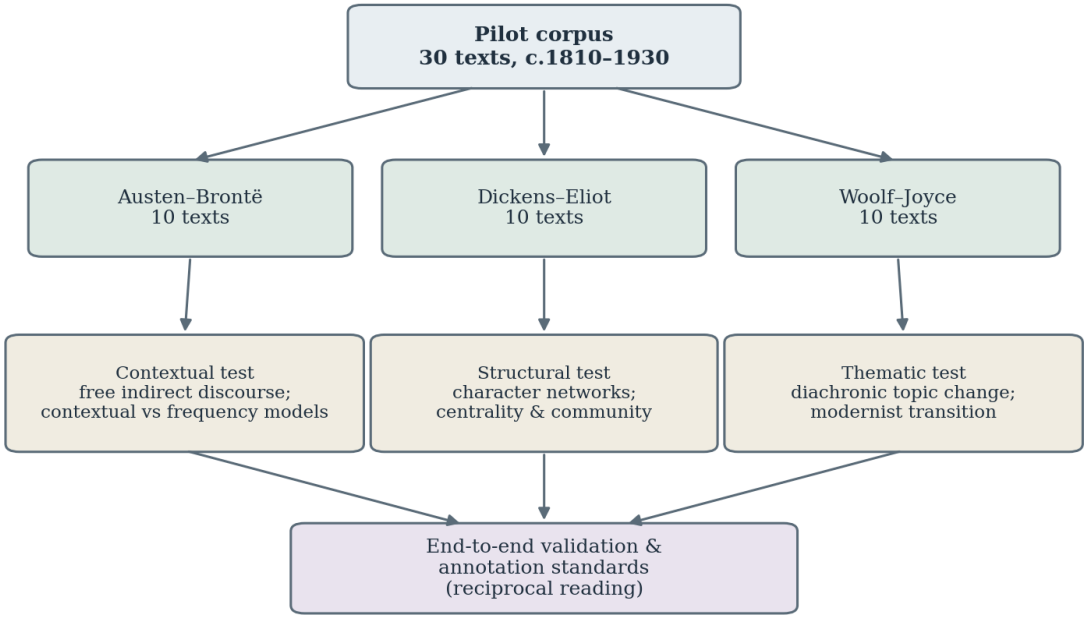


Figure 3. Research design. The thirty-text pilot is balanced across three author groupings, each mapped to the analytical test it is best suited to exercise, with all three converging on a shared validation stage.

6.4 Illustrative Case Readings

The abstract architecture of the pipeline is best clarified through the three cases that anchor it. The Austen case tests the contextual layer. Free indirect discourse, the technique by which a third-person narrator inflects its report with the idiom and judgement of a character, is the hinge of Austen's irony, and it is invisible to any method that treats words as context-free tokens, because the same sentence may be sincere in the narrator's voice and ironic in a character's. A frequency-based classifier, confronted with the famous opening proposition of Pride and Prejudice concerning a single man in possession of a good fortune, has no way to register that its apparent universality is being quietly disowned. A contextual model, by contrast, can be trained to detect the tonal dissonance between a clause and its narrative surroundings, and the case study measures how far it succeeds against a hand-annotated sample. The point is not that the machine reads the irony, but that it can be made to locate the sites where irony is at work, returning the critic to passages worth reading closely.

The Dickens case tests the structural layer. A novel such as Bleak House coordinates a double narrative and a vast cast across the institutions of law, aristocracy, and the urban poor, and critics have long described its form in terms of connection, the slow revelation that apparently unrelated lives are bound together. Network analysis gives this critical commonplace an operational form:

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

named-entity recognition extracts the characters, co-occurrence within a sliding window defines their ties, and community-detection algorithms partition the resulting graph into clusters. Where those clusters correspond to the novel's social worlds, and where the characters who bridge them register as high in betweenness centrality, the computational structure corroborates and sharpens the critical account; where it diverges, the divergence becomes a question for interpretation rather than an error to be corrected. The metric does not replace the reading of Dickensian form but disciplines it, forcing the critic to specify what connection means.

The Woolf case tests the thematic layer and the diachronic claim. The movement from the externally furnished world of the Victorian novel to the interior flux of modernist consciousness is among the most consequential transitions in the history of English fiction, and it is a transition of theme as much as of technique. A frequency-based topic model, keyed to vocabulary, tends to register this shift only crudely, since the interiorised novel does not necessarily introduce a new wordstock so much as redeploy a familiar one toward consciousness, time, and perception. An embedding-based model, sensitive to the contextual sense in which words such as light or moment are used, is better placed to detect the reorientation that culminates in a work such as To the Lighthouse, where the ostensible subject recedes before the registration of perception itself. Aligning the topic trajectories against the publication timeline allows the study to ask whether the modernist turn appears as a rupture or as the acceleration of a longer tendency, a question on which the choice of model may itself be decisive.

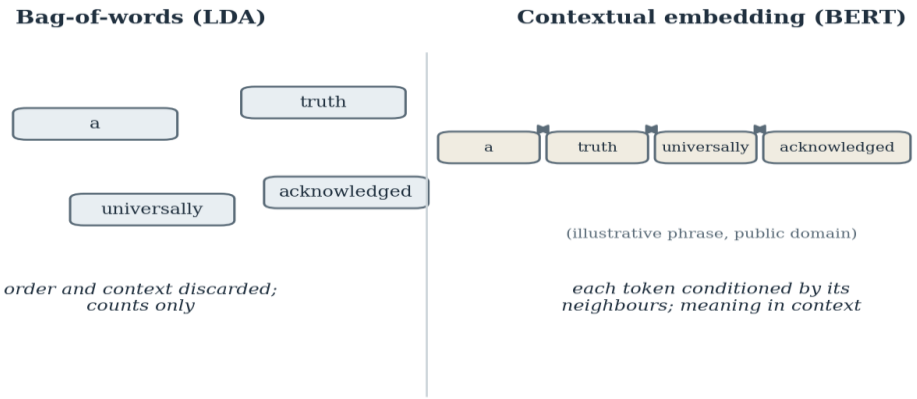


Figure 4. Why representation matters for literary reading. A bag-of-words model discards order and context, whereas a contextual embedding conditions each token on its neighbours, the property on which the detection of irony and free indirect discourse depends.

6.5 Interpretive Validation

Across all four stages, a single principle governs the transition from number to claim. No computational output is admitted as evidence until it has been subjected to what may be called a reciprocal reading, in which the aggregate pattern directs the researcher back to specific passages and those passages are read closely to confirm, qualify, or overturn the pattern. This protocol is the

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

study's direct response to the critique of computational literary studies (Da, 2019): it treats quantification not as a terminus but as a heuristic that earns its interpretive authority only through confrontation with the text. In this respect the framework aligns itself with Ramsay's (2011) conception of algorithmic criticism, in which the value of computation lies less in what it proves than in the readings it provokes.

6.6 Methodological Limitations

Three constraints specific to the method qualify every result the pipeline will produce. The first is corpus bias: the public-domain holdings of Project Gutenberg and comparable archives over-represent the male-authored, Western canon, so that any thematic trajectory the study reports describes the evolution of a particular curated archive rather than of English writing in general. The second is historical language drift: transformer models are trained overwhelmingly on contemporary English, and their embeddings may impose present-day senses on earlier usage, a hazard that domain adaptation can lessen but not remove. The third follows from the first two: because neither bias nor drift can be fully corrected computationally, the protocol of reciprocal reading is not an optional refinement but a necessary guard against algorithmic artefact, the point at which human interpretation must check what the models propose.

7. Significance of the Study

The significance of the study is at once methodological, disciplinary, and institutional. Methodologically, it offers one of the first explicitly integrated frameworks in which style, structure, and theme are analysed as interdependent dimensions of the literary work rather than as separable targets for separate tools, and it does so at the moment when transformer models make a semantically sensitive computational reading genuinely feasible. Disciplinarily, it responds to the most serious critique the field has faced by building interpretive accountability into the design of the pipeline rather than appending it as an afterthought, thereby modelling a form of computational literary study that remains answerable to the interpretive norms of literary criticism. The focus on canonical authors is strategic rather than conservative: because Austen, Dickens, and Woolf have been read with unusual intensity, they provide a demanding external check against which computational findings can be measured, so that the framework's claims can be falsified rather than merely illustrated.

Institutionally, the study has particular value for research environments in which large-scale computational infrastructure is limited, since the proposed pipeline relies entirely on open-source tools and public-domain corpora. This accessibility matters for the globalisation of the digital humanities, a field whose centre of gravity has remained in well-resourced Western institutions, and it positions the framework as a practicable basis for research in contexts where bespoke computational resources cannot be assumed. By demonstrating that rigorous computational analysis of English literature can be conducted with freely available instruments and an explicit interpretive protocol, the study aims to contribute both a method and a precedent.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

8.

Limitations and Ethical Considerations

Several limitations condition the claims of this study. The most fundamental is the dependence of every result on decisions taken during corpus construction and preprocessing. The confinement of the corpus to public-domain texts introduces a survivorship bias toward works that have remained in print and in copyright-free circulation, which are disproportionately canonical and anglophone; the diachronic findings must therefore be read as describing the evolution of a particular, institutionally shaped archive rather than of English literature in some neutral totality. Preprocessing choices compound this effect, since tokenisation, lemmatisation, and the handling of dialogue and paratext each silently shape the evidence before any analysis begins.

A second limitation concerns the instruments themselves. Transformer models are trained predominantly on contemporary text, and their representations of historical language may be anachronistic, imposing present-day semantic associations on words whose sense has shifted. The study mitigates this hazard through domain adaptation and qualitative validation, but it cannot eliminate it, and the risk of semantic anachronism remains a standing caution. A third limitation is interpretive: the framework's insistence on reciprocal reading, while principled, is labour-intensive and does not scale to the full corpus, so that the validation of any given pattern rests on a sample of passages whose selection is itself an interpretive act.

Ethical considerations, though less conspicuous in a study of long-dead authors, are not absent. The concentration of computational literary infrastructure in a small number of well-resourced institutions raises questions of equity that the field has begun to confront, and the deliberate reliance on open tools and open corpora in this study is a modest response to them. More broadly, the authority that quantification lends to interpretive claims carries a responsibility to represent uncertainty honestly, and the study treats the candid reporting of negative results and model disagreements as an ethical as well as a methodological obligation.

9. Expected Outcomes and Future Directions

The study is expected to yield three principal outcomes. The first is a documented and reproducible pipeline, released with its code and parameters, that integrates stylometric, structural, and thematic analysis on a common corpus and that other researchers can adapt to their own materials. The second is a set of substantive findings on the three focal authors, including a measured comparison of contextual and frequency-based models on the detection of free indirect discourse, a network description of Dickensian narrative structure tested against the critical literature, and a diachronic account of thematic change across the corpus that foregrounds the modernist transition. The third is methodological: an explicit protocol of interpretive accountability that specifies the conditions under which a computational output may be admitted as literary evidence.

Future work suggested by the framework extends in several directions. The pipeline could be applied beyond the anglophone canon to literatures whose computational study remains underdeveloped, provided the models are adequately adapted to their languages. It could be extended temporally, both backward into the early-modern period and forward into the

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

contemporary novel, where questions of style and theme intersect with the new problem of machine-generated text. And it invites methodological refinement, particularly in the development of models that are sensitive to historical language change rather than merely tolerant of it. These directions position the present study as a foundation rather than a conclusion.

10. Conclusion

This paper has argued that the analytical power of computational methods for English literature is currently constrained less by the sophistication of the available tools than by the fragmentation of their use and the underdevelopment of the interpretive protocols that connect number to meaning. By integrating stylometry, network analysis, and embedding-based topic modelling within a single accountable pipeline, and by testing that pipeline against the demanding cases of Austen, Dickens, and Woolf, the proposed framework seeks to show that artificial intelligence and the digital humanities can address recognisably literary questions about style, narrative, and theme. The wager throughout has been that the computational and the interpretive are not rivals but complements, and that the future of the field lies in the disciplined articulation of the two.

Author Contributions

Tahreem Sheikh conceived the study, developed the theoretical framework, and drafted the manuscript. Fakiha Umar contributed to the literature review, the design of the methodology, and the critical revision of the manuscript. Tayyeba Ayyub contributed to the literary analysis, the interpretation of findings, and the review and editing of the manuscript. Kinza Saleem contributed to corpus preparation, the thematic reading of the selected texts, and the proofreading of the manuscript. All authors read and approved the final version.

References

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. Proceedings of the International AAAI Conference on Web and Social Media, 3(1), 361–362.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: Analyzing text with the Natural Language Toolkit. O'Reilly Media.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3), 267–287.

Da, N. Z. (2019). The computational case against computational literary studies. Critical Inquiry, 45(3), 601–639.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 4171–4186). Association for Computational Linguistics.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 8 October, 2026

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. The R Journal, 8(1), 107–121.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. https://doi.org/10.48550/arXiv.2203.05794

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303

Hoover, D. L. (2004). Testing Burrows's Delta. Literary and Linguistic Computing, 19(4), 453–475.

Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.

Jockers, M. L. (2014). Text analysis with R for students of literature. Springer.

Kirschenbaum, M. G. (2010). What is digital humanities and what's it doing in English departments? ADE Bulletin, 150, 55–61.

McCarty, W. (2005). Humanities computing. Palgrave Macmillan.

Moretti, F. (2005). Graphs, maps, trees: Abstract models for a literary history. Verso.

Moretti, F. (2013). Distant reading. Verso.

Piper, A. (2018). Enumerations: Data and literary study. University of Chicago Press.

Ramsay, S. (2011). Reading machines: Toward an algorithmic criticism. University of Illinois Press.

Rybicki, J., & Eder, M. (2011). Deeper Delta across genres and languages: Do we really need the most frequent words? Literary and Linguistic Computing, 26(3), 315–321.

Sinclair, S., & Rockwell, G. (2016). Voyant Tools [Computer software]. https://voyant-tools.org

Underwood, T. (2019). Distant horizons: Digital evidence and literary change. University of Chicago Press.