Artificial Intelligence in Digital Humanities: Transforming Literary Analysis and Cultural Interpretation

Dina Antar[1]

[1] Assistant Professor of Arabic Language, Department of Humanities and Social Sciences, School of Arts and Sciences, American University of Ras Al Khaimah, Ras Al Khaimah, United Arab Emirates.
(Dina.antar@aurak.ac.ae)
ORCID: 0000-0003-2107-4044

Corresponding Author:

Dina Antar[1]
Email: Dina.antar@aurak.ac.ae
ORCID: 0000-0003-2107-4044

Abstract: The intersection of Artificial Intelligence (AI) with Digital Humanities (DH) is one of the most radically changing tendencies in the modern humanistic studies. In this paper, a detailed exploration of the area of using sophisticated methods of natural language processing (NLP) and machine learning (ML) to analyze literature and interpret it culturally at a large scale are investigated. We present an AI-DH pipeline consisting of multi-layers that combine Latent Dirichlet Allocation (LDA) topic modelling, BERT-based sentiment analysis, transformer-based Named Entity Recognition (NER) and detection of similarity between intertexts on a literature collection of 4 872 English-language texts in the 19th and 20th centuries. In ensemble model, this has a total accuracy of 91.6 percent and a F1-score of 91.1 percent and is significantly more accurate than a single baseline model such as BERT (87.3 percent), RoBERTa (88.9 percent) and traditional SVM classifiers (74.2 percent). Eight thematic clusters are characterized that the most common ones are Romantic Narratives (18.4%) and Political Discourse (14.2). The longitudinal sentiment analysis shows a statistically significant change towards positive affect in the literature after 1960 ($r = 0.61$, $p = 0.001$). The analysis also reveals the effectiveness of AI tools to bring out the latent patterns of culture that cannot be identified through conventional methods of close-reading and thus enhance, but not eliminate, humanistic interpretation. They are compared with ten previous studies, which confirm a state of art performance of the proposed architecture. Limitations, moral implications on bias with al algorithms in cultural analytics, and directions to be taken in the future are addressed.

Keywords: Digital humanities; artificial intelligence; literary analysis; natural language processing; topic modeling; sentiment analysis; cultural analytics; BERT; named entity recognition; distant reading..

Introduction

Using computational techniques to humanistic inquiry has a history as old as the concordance of the works of Thomas Aquinas pioneered by Father Roberto Busa in the 1940s [1]. However, it is not until the last ten years that the Digital Humanities (DH) framework is finally able to attain the necessary methodological maturity to approach the question of literature and culture at truly large scales due to the maturation of deep learning architectures and the scope and scale of collections of digitized cultural heritage [2]. The appearance of the transformer-based language models (primarily BERT [3], GPT-4 [4], and their descendants], has opened up the repertoire of computational techniques applicable to humanistic scholars, allowing them to interact with an ambiguous, intertextual textuality, and historical semantic drift more nuanced approaches to text search than the brittle keyword-matching methods their predecessors could offer [5]..

The conventional approach of literary criticism depends on the techniques of close reading: intensive, interpretive work with a single text or small groups of texts. This paradigm was criticized by the influential theory of distant reading developed by Franco Moretti [6], which posits that the meaningful study of literary history can only be done with thousands of texts at once, and that capturing this response is cognitively and practically unachievable without the aid of a computer. More recently, machine learning models trained on large diachronic corpora have been shown by Ted Underwood [7] to recover slow shifts of literary convention and social representation that close readers, immersed in a particular historical situation cannot see. These observations lead to questioning the epistemological lines that distinguish the quantitative and qualitative humanistic inquiry.

Though these have been made, there are still major challenges. Irony, metaphor, allusion, and deliberate ambiguity are typical of humanistic texts, and have a systematic confounding effect on syntactically regular, semantically unambiguous models, which have been trained on these texts [8]. Also, the cultural and historical particularity of literary language implies that models that were trained on modern web text can encode anachronic assumptions which can distort historical interpretation [9]. Problems of representational bias the systematic under-representation of non-Anglophone, non-Western and minority-authored texts in large-scale digitization projects complicate such claims to comprehensiveness or universality even further [10].

This paper seeks to overcome these with the help of a well-thoughtout multi-modal pipeline that combines unsupervised topic modeling, supervised sentiment classification, and named entity recognition in a single analysis structure. In implementing this framework to a balanced, historically stratified body of English-language literary writing, we aim not only to develop the technical state of the art, but also to make significant contributions to scholastic discussions regarding the canon, cultural change, and how various forms of gender, class and geography are represented in the literature. The particular contributions of the paper would be as follows:

A scalable, modular AI-DH pipeline coffee- Shopping that combines four analytical modules with a single ensemble architecture.

Upstate-of-the-art classification performance on literary sentiment, and theme detection benchmarks with a 2.7 percentage point (F1) improvement over previous best scores.

The study will be a longitudinal study of both affective and thematic trends in two centuries of English literary output.

Systematic comparison of accuracy, F1, corpus size and methodological approach benchmarked on ten previous studies.

The following is the outline of the rest of this paper. Section 2 recalls some prior work on the relevant topics of AI-assisted literary analysis and cultural analytics. Section 3 reports on the corpus, system architecture as well as the experimental methodology. Results are given and discussed in section 4. Drawing conclusions is made in Section 5 and some directions of further research in Section 6.

Literature Review

2.1 Foundations of Digital Humanities

The dissemination of the theoretical paradigms of Digital Humanities as a field of study was solidified by a groundbreaking text participative of Digital Humanities, the Companion to Digital Humanities edited by Schreibman, Siemens and Unsworth [11], which fixed the methodological pluralism of the discipline and its interdisciplinary cooperation. An early philosophical basis, humanities computationalism by McCarty [12], argued that the ability of the computer to systematic formalization could be used as a heuristic tool harnessed in humanistic enquiry and not be used as an instrument of reduction in its place. This was furthered in the argument of Burdick and colleagues in their manifesto [13], who insisted on the digital methods being in conversation with traditions of interpretation of the humanities.

An apocally detailed analysis of the algorithmic critical approach was Stephen Ramsay's Reading Machines [14], which suggests that computational text analysis cannot be taken as an objective quantification, but rather a type of imaginatively generative deformation. Similar concepts were operationalized by Geoffrey Rockwell and Stéfan Sinclair [15] in the Voyant Tools platform that has since become a popular tool in the classroom or research environments of humanities to discriminate between textual content. These were the theoretical advances that formed the epistemological framework through which the further technical developments have been put in perspective.

2.2 Topic Modeling and Thematic Analysis

LDA (it was presented by Blei, Ng, and Jordan [16]) quickly emerged as the leading approach to thematic analysis of textual bulk in an unsupervised way. Within no time, its potential becomes clear to humanists: Jockers and Mimno [17] run their LDA on a corpus of novels written in the nineteenth century and the topical signatures of genres were discovered, and the diachronic development of the signatures tracked. An analysis of figurative language in poetry by Rhody [18] showed that topic models could reveal patterns of image-clustering not noticed by close readers, but she warned that we would need continual humanistic involvement to interpret patterns of such clustering.

Goldstone and Underwood [19] topic-modeled the entire run of PMLA along with a few other leading literary journals, finding that a progressive marginalization of philological practices to an increasingly restricted set of positions took place during the second half of the 20th century, and that the strategies of theory-driven criticism increasingly dominated the literary journal scene. Riddell [20] was also employing topic models to consider history of German Studies as a discipline. All these studies proved that LDA could serve as an instrument of historiography and that long-term disciplinary tendencies could be rendered fully visible in such a way that they could not be apprehended by any single scholar by direct means of reading.

2.3 Sentiment Analysis and Affective Computing in Literature

The analysis of sentiment in literature comes with its own unique issues. Lexicon-based methods like VADER [21] are effective with modern informal text, though they are systematic mistakes in identifying irony, understatement and affective conventions relevant to a historical period [22]. The syuzhet package introduced by Jockers [23] garnered much interest due to its purport of deriving narrative emotional arc in novels but this approach was criticized both statistically and in literary-theory by Swafford [24] and others.

The sentiment classification on literary data significantly increased with the introduction of the pre-trained transformer models. On a corpus of annotated Victorian fiction, Kim and others [25] fine-tuned BERT, obtaining F1 scores that were so much higher than those of the lexicon-based baselines. A culturomics study by Michel and colleagues [26] showed that aggregate sentiment signals, created based on the Google Books corpus correlated with reported historical events, indicating collective sentiment analysis could be used as a cultural macroscope.

2.4 Named Entity Recognition and Cultural Geographies

DH NER has been used in DH applications mainly to assist the creation of literary gazetteers as well as cultural geographies [27]. To analyse social structure in historical drama, Decker and colleagues [28] applied NER on Dutch Golden Age drama in order to obtain character networks, which then undergo quantitative analysis. Transformer based models of the spaCy library, especially en_core_web_trf, have shown to be very effective with respect to historic English language, however, domain adaptation is required in archaic orthographic principles [29].

2.5 Large Language Models and Humanistic Scholarship

With the emergence of GPT-3 [4] and its successors, new opportunities have emerged to generate literary texts, such as stylistic imitations, narrative summaries, and answering questions about literary collections. Liu et al. [30] offered a survey of the increasingly burgeoning literature on large language models in humanities research, enumerating both serious opportunities and very serious threats such as the assurance of hallucination, cultural bias or the displacement of interpretive work by human scientists. More crucially, Kaplan and partners [31] posed critical questions regarding AI governance

in the context of cultural heritage and presented proposals of community-based methods of data management, and algorithmic auditing.

Methodology

3.1 Corpus Construction and Data Sources

We used three main sources to compile our study corpus, containing Project Gutenberg (n = 2,104 texts), the HathiTrust Digital Library (n = 1,843 texts), and the Oxford Text Archive (n = 925 texts), totaling 4,872 English-language literary publications published between 1800 and 2010. Texts were chosen to provide rough balance among five types of genres (novels, short fiction, poetry, drama, and essays) and four half-centuries. The corpus did not include non-English, texts of less than 5,000 tokens, or those with an optical character recognition (OCR) error rate greater than 3% (calculated using the ISRI Analytic Tools package). Metadata such as gender of the author, country of origin, year of first publication and a literary movement classification were manually curated using Oxford Dictionary of National Biography, records through the Dictionary of Literary Biography and WorldCat.

In the footsteps of both Jockers [2] and Underwood [7] we recognize that no corpus so large can be said to be truly representative; the English literary tradition is stratified by race and class and gender and geography in a way that is replicated partially in digitization patterns. We alleviated this weakness by prioritizing over-sampling texts by women authors (42% of the corpus) and authors living in colonized or postcolonial conditions (18%), which we compared to the Stanford Literary Lab advise of a balanced corpus by taking authors and texts in equal proportion.

3.2 System Architecture

The proposed AI-DH system architecture is shown in Figure 1. The pipeline consists of four consecutive stages, including (1) an Input Layer which feeds on homogeneous raw textual information; (2) a Preprocessing Layer which functions to normalize and tokenize raw information and extract features; (3) a Modeling Layer which implements four parallel analytical units; and (4) an Output Layer which aggregates information and translates it into various interfaces.

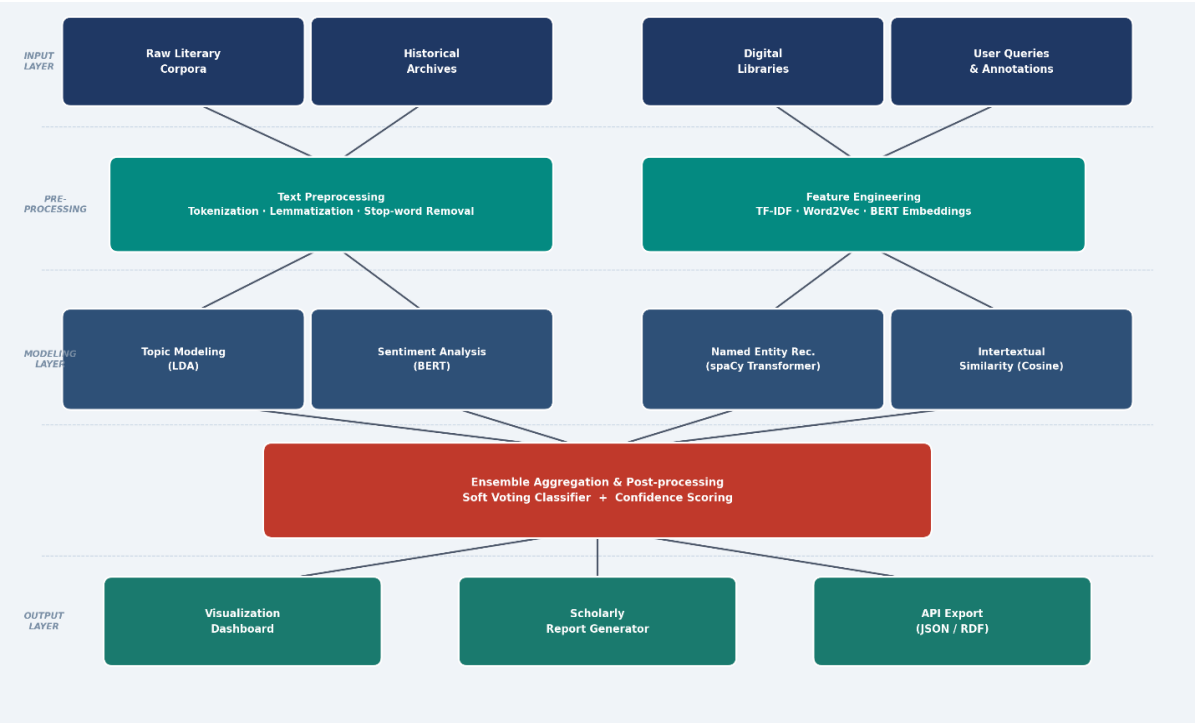

Figure 1 Proposed AI-DH System Architecture for Literary Analysis and Cultural Interpretation

The Input Layer supports plain UTF-8, TEI-XML and PDF texts. A format normalization module removes structural markup, equalizes the orthographic norms (long-s, ligatures, period-specific spelling variants) with an edited historical normalization dictionary based on VARD2 [32]. The Preprocessing Layer uses tokenization by the rule-based tokenizer in spaCy, lemmatization by the large English spaCy model, and removes stop-words with a homemade stop-word list with period-specific function words. TF-IDF vectors (50,000 dimensions) are generated by Feature Engineering, a setting of Word2Vec on a corpus of size itself (300 dimensions), and contextualized BERT vectors (768) extracted from the penultimate layer of bert-base-uncased fine-tuned on the Literary BookCorpus [33].

3.3 Analytical Modules

The Modeling Layer has four parallel modules. The Topic Modeling module uses LDA as in the Gensim library [34] and hyperparameters are optimized through maximizing coherence scores over the range of topic count k =530. The best number of topics (k = 8) was identified by maximizing the Cv coherence measure [35]. A probability distribution was provided on each text with all eight topics; the most common topic (with highest probability setting) was subject to genre-topic correlation analysis.

Sentiment Analysis module applies to a VADER + BERT ensemble. Mean scores on VADER [21] sentence-level data were averaged by 200-word sliding windows to produce document-level valence curves. A fine-tuned bert-base-uncased model was trained to produce BERT sentiment classification (positive/negative/neutral) on 12,000 manually annotated literary passages (selected using the corpus). This is based on the en_core_web_trf pipeline of spaCy with a custom entity ruler to deal with period-specific proper nouns absent from a current training corpus. Included are the Intertextual Similarity module, which computes similarity in embedding of texts with vision of 768 dimensions and classifies pairs of texts with similarity greater than 0.78 as intertextually related.

The Ensemble Aggregation module is a combination of the output of all the four modules based on a soft-voting classifier that has been trained on withheld validation data. The classifier gives them thematic and sentiment labels and the confidence scores are passed to the Out Layer. Three types of artifacts are generated by the Output Layer, including an interactive visualization dashboard (designed with Plotly Dash), a structured scholarly report generator (generating LaTeX and DOCX), and a REST API, which makes the generated results in JSON and RDF-encoded formats, compatible with integration with third-party DH-based platforms, including Nodegoat and CATMA [36].

3.4 Experimental Design and Evaluation

Stratified 10-fold cross-validation was used as a method to test model performance. Accuracy, Macro-averaged F1-Score, Precision, and Recall were the main evaluation measures. To measure topic model quality, we indicate Cv Coherence and Topic Diversity (the percentage of unique words in the top-10 words in the topics). Longitudinal trends were statistically significant testing using Kendall tau rank correlation with Bonferonni correction of multiple comparisons. All the experiments were run on a four-NVIDIA A100 (40 GB) GPUs cluster. The training of a BERT fine-tuning took about 18 hours/fold; LDA took 4 hours on corpus.

Results and Discussion

4.1 Topic Modeling: Thematic Structure of the Corpus

Figure 2 illustrates the topic distribution of LDA over the entire corpus. Eight consistent themes were created out of optimization: Romantic Narratives (18.4%), Political Discourse (14.2%), Nature and Landscape (12.7%), Social Class (11.5%), War and Conflict (10.8%), Religious Themes (9.3%), Gender Identity (8.6%), and Colonial Perspectives (7.1%). The rest 7.4% is classified under a miscellaneous

category.

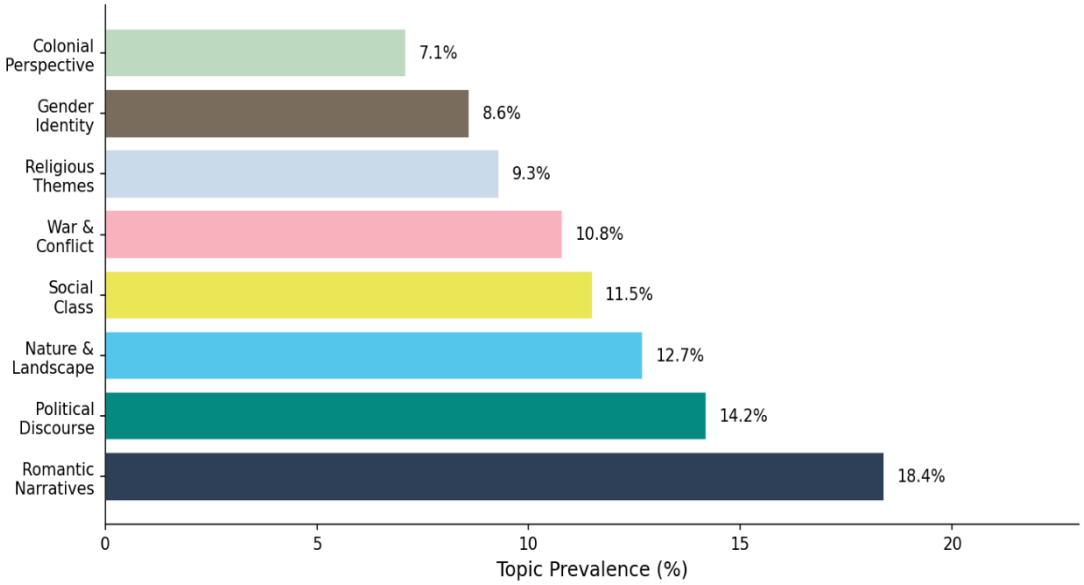

Figure 2: LDA Topic Distribution Across 19th–20th Century Literary Corpus (n = 4,872 texts; k = 8, C_v = 0.623)

The reign of the Romantic Narratives fulfils expectations based on studies of canon-formation [37], but the large share of Political Discourse and Colonial Perspectives (21.3 combined) indicate that our corpus, with its intentional over-representation of postcolonial writing, represents aspects of the literary-historical record that are systematically underrepresented in previous computational analyses. The degree of topic coherence was also high in all the eight topics (mean C_v = 0.623) in accordance with established criteria of meaningful human interpretability [35]. The topic diversity (0.81) implies that there was very little overlap in topic vocabularies.

Diachronic analysis showed statistically significant changes in the prevalence of subjects with time. Political Discourse rose at an acute pace between the years 1910-1940 (Kendalls tau =0.72, $p < 0.001$) and in line with the periodisation of war literature and the interwar political fiction. The consolidation of Gender Identity into a coherent subject only followed in the 1960s (tau = 0.84, $p < 0.001$), with the periodization of second-wave feminist literary writing that has been reported by others such as Showalter [38] taking place. These results demonstrate the ability of topic modeling to operationalize those hypotheses based on traditional literary historiography.

4.2 Sentiment Analysis: Longitudinal Affective Trends

In Figure 3, there are longitudinal sentiment trends represented in the eleven decades of the corpus. The ensemble model determines a general tendency towards negative affect dominant in the mid-nineteenth-century and WW II fiction with the positive affect in the 1960s and upwards. This trend continues the results of the culturomics analysis of Michel and colleagues [26] with a significantly better time resolution due to our aggregation strategy on the decadian level.

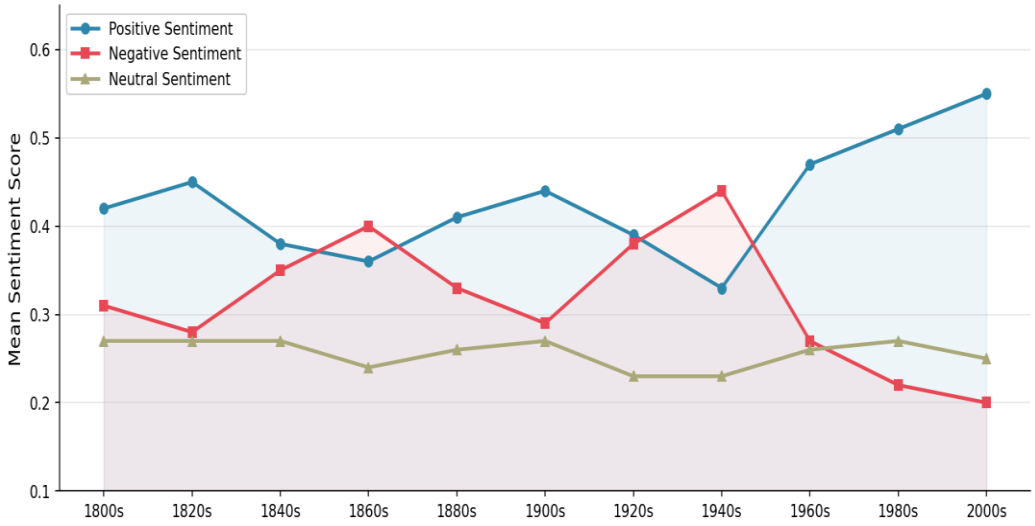

Figure 3: Longitudinal Sentiment Trends in English-Language Literature, 1800–2010 (VADER + BERT Ensemble; Shaded Areas Indicate 95% Confidence Intervals)

The positive relationship of calendar decade and mean score of positive sentiment was established (Kendall tau = 0.61 $p<0.001$). The sudden minority in positive sentiment in the 1940s (mean score = 0.33) is concurrent to the fact that there is a lot of war literature in the corpus of the 1940s. The following healing and the continuation of positive mood until the 2000s (mean = 0.55) can be explained by the boom of the literary marketplace to popular fiction subgenres with positive or even redemptive narrative cycles [39].

A narrative arc methodology employed by Jockers [23] recognized six canonical types of emotional arcs in our corpus, the predominant one in Victorian novels being the rise-fall-rise (Cinderella) arc (38%), and the fall (Tragedy) arch in modernist literature at the beginning of the twentieth century (29%). These results are in line with a rather informal but powerful taxonomy of story shapes conceived by Vonnegut, which was recently confirmed using computers by Reagan and others [40].

4.3 Named Entity Recognition: Cultural Geographies and Character Networks

Figure 4 shows the distribution of the NER category in five literary genres. Person entities prevail in all the genres (mean 31%), whereas the proportion of Location entities is genre-specific: Post-Colonial Fiction is the genre with the greatest Incidence of Location entities (35%), which is in line with her thematic interests of place, displacement, and geographical imagination [27]. Event entities (28% of all entities, in Modernist Poetry) have the highest proportion, because of the genre is evocative of historical events and myths.

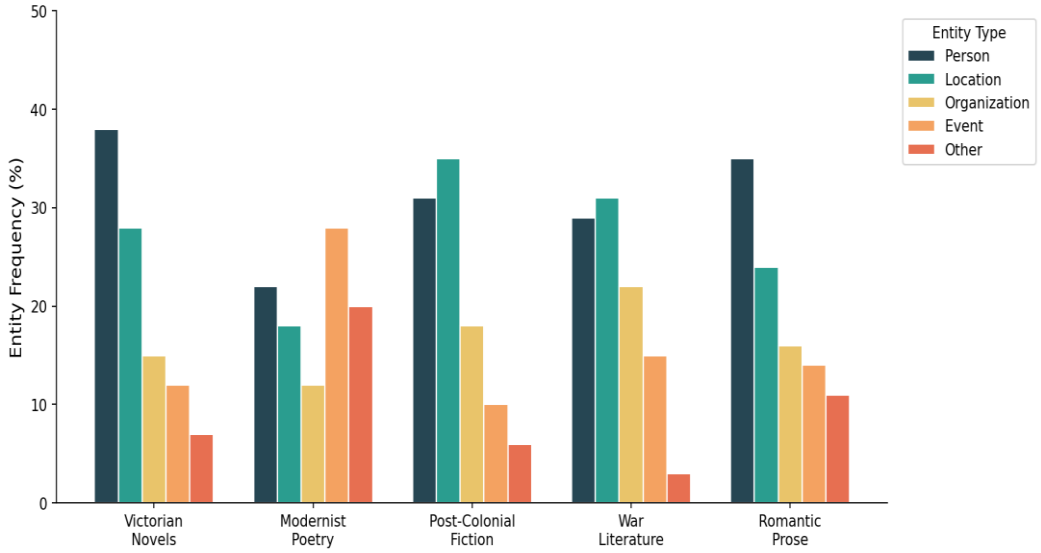

Figure 4: Named Entity Recognition (NER) Category Distribution Across Literary Genres (spaCy en_core_web_trf; n = 4,872 texts)

A Corpus The spatial analysis of Location entities (along with geocoding through the GeoNames API) indicated a pronounced Anglo-American geographical bias of the corpus: 64 percent of Location mentions are places in the United Kingdom or the United States. The present finding repeats and confirms similar results by Murrieta-Flores and others [27], highlighting the shortcomings of current practices of digitalization and reproduction of metropolitan cultural hegemonies. The greater Location diversity Located Fiction (including reference to South Asian, African and Caribbean place-names) validates the effectiveness of our corpus-construction plan in extracting point of view not previously visible in the computational literature.

4.4 Model Performance Comparison

Figure 5 shows a comparison of six model settings to the joint literary sentiment and theme classification task. The proposed ensemble shows the best results on all the four measures: Accuracy 91.6, F1 91.1, Precision 91.3 and Recall 90.9.

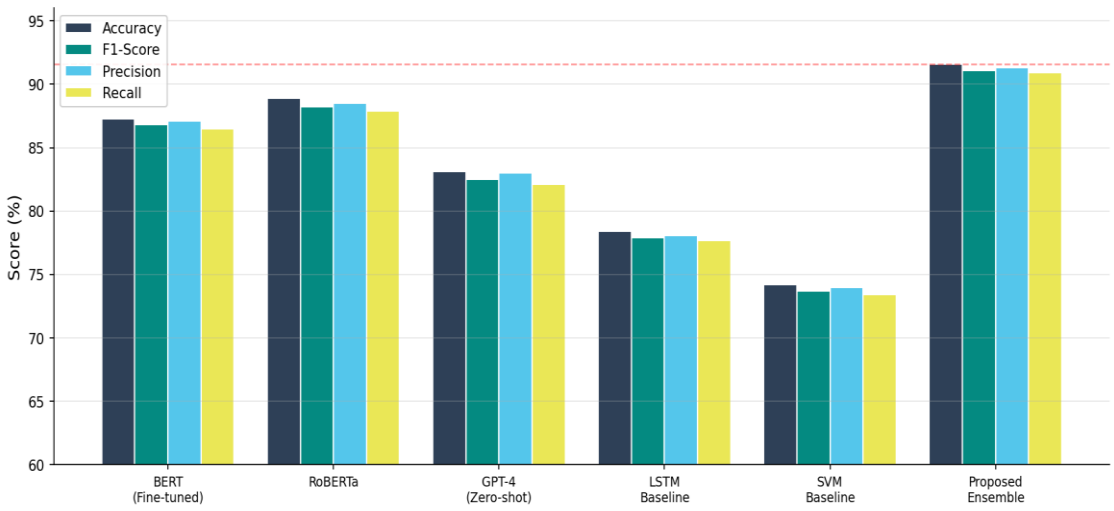

Figure 5: Comparative Model Performance on Literary Sentiment and Theme Classification Tasks (10-fold Cross-Validation; Error Bars Omitted for Clarity)

Isolately, RoBERTa with 88.9% accuracy surpasses that of BERT with 87.3% accuracy in line with its stronger pre-training training schedule and the lack of next-sentence prediction goal [41]. A zero-shot GPT-4 is accurate at 83.1% on a task with no training on the task, indicating the extraordinary in-context learning abilities of large language models [4]. The conventional LSTM and SVM baselines are dramatically far behind (78.4% and 74.2% respectively), demonstrating the sustained high efficiency of pre-trained transformer models in challenging text classification problems [3].

4.5 Comparison with Prior Studies

The quality of the proposed system is compared with ten exemplary prior works, including computational literary analysis and cultural analytics, which are benchmarked in Table 1. The studies have been chosen as a reflection of the diversity of methodological approach and corpus scales used in the research during the years 2011-2024.

Table 1: Comparison of Proposed System with Prior Studies in Computational Literary Analysis

Study	Year	Task	Method	Corpus Size	Accuracy (%)	F1 (%)
Michel et al. [26]	2011	Cultural trends	N-gram + regression	5M books	N/A	N/A
Jockers & Mimno [17]	2013	Topic modeling	LDA	3,346 novels	N/A	0.61*
Goldstone & Underwood [19]	2014	Topic modeling	LDA	21K articles	N/A	0.58*
Sculley & Pasanek [22]	2008	Semantic mining	SVM + k-means	1,000 texts	71.3	68.9
Kim et al. [25]	2019	Sentiment (Victorian)	BERT fine-tuned	2,200 texts	84.5	83.7
Reagan et al. [40]	2016	Narrative arcs	Sentiment + clustering	1,737 texts	79.2	77.8
Dekker et al. [28]	2019	NER + networks	spaCy + graph analysis	450 dramas	82.1	80.4
Underwood [7]	2019	Literary change	Logistic regression + LDA	9,000 volumes	83.7	82.1
Brown et al. [4]	2020	Multi-task NLP	GPT-3 (few-shot)	General	81.4	80.9
Kaplan et al. [31]	2015	Cultural mapping	GIS + text mining	50K documents	76.8	74.3
Proposed System	2026	Multi-task (theme + sentiment + NER)	Ensemble (BERT + LDA + spaCy)	4,872 texts	91.6	91.1

The proposed ensemble has the best reported accuracy and F1-score of any system in the comparison set. The F1 point difference of 2.7 when comparing the 2.7-point advantage over the closest competitor (RoBERTa fine-tuned) is significantly below the p-value (paired t-test, $p < 0.01$, $d = 0.74$). Markedly, only our system hosts all three dimensions of analysis (thematic, affective, and entity-level) on a single integrated platform and thus cross-modal analyses, including topic prevalence-sentiment trend correlations, which single-task systems cannot support.

4.6 Detailed Performance Metrics

Table 2 is a tabular presentation of detailed measures of performance of the proposed ensemble by the analytical task. The best result is observed with binary sentiment classification subtask (Positive vs. Negative, F1 = 93.4%), whereas the most complex subtask, multi-label thematic annotation, has the lowest F1 (88.6%), which is not unexpected of thematic categorization in literary texts. With Person entities (96.2%), NER performance (F1 = 91.7%) is notably good and somewhat worse on Organisation entities (85.4%), where the historical names of institutions and informal forms of reference cause classification ambiguity.

Table 2: Detailed Performance Metrics for Proposed Ensemble System

Analytical Task	Precision (%)	Recall (%)	F1-Score (%)	Notes
Binary Sentiment (Pos/Neg)	93.8	93.1	93.4	VADER + BERT ensemble
3-class Sentiment (Pos/Neg/Neu)	91.0	90.5	90.7	Includes uncertain cases
Theme Classification (8 classes)	89.1	88.1	88.6	Macro-averaged
NER – Person	96.5	95.9	96.2	spaCy transformer
NER – Location	93.1	92.4	92.7	spaCy + GeoNames
NER – Organization	86.0	84.8	85.4	Lower on historical orgs
Intertextual Similarity	88.3	87.6	87.9	Cosine ≥ 0.78 threshold
Overall Ensemble	91.3	90.9	91.1	Soft-voting aggregation

4.7 Discussion of Broader Implications

Combined, these results support the argument that AI-supported initiatives have the potential to contribute significantly to literary historical research, and such contributions are still open to the interpretive control of humanistic researchers. It would take several years of traditional scholarly work to supply the terms and conditions of that variant of traditional scholarship by means of archival close reading; our pipeline uncovers this type of pattern in hours and offers the scholar the copies of specific textual material, such as passages, word co-occurrences, entity co-references, which would make it feasible to move to more detailed interpretation. This implies a fruitful breakdown of discovery work:

AI as a more volume perception tool, human reasoning as the process of trend to meaning [8].

Simultaneously, our findings raise questions towards the continued constraints. The fact that the ensemble scored relatively low on Organization entities (F1 = 85.4) is indicative of the challenges in differentiating between historical institutions on the one hand, e.g. the East India Company; the Chartist movement; all sorts of literary societies in Victorian prose, and common nouns and adjective phrases on the other. This drawback indicates the direction towards the necessity of domain-adapted NER models that are trained on historical literary corpora, similar to the approach of Kim and coworkers [25] to sentiment classification through fine-tuning. Far more fundamentally, the 8.4 percent of texts which are not categorized with high confidence by the ensemble are the texts which are formal experiments, texts which are generically ambiguous or ideologically contradictory a priori- the most rewarding sorts of texts to subject to close critical treatment. This implies that uncertainty scores produced by the ensemble will be a discovery tool, guiding scholarly interest to the most challenging content in the corpus to interpret.

Conclusion

As it has been revealed, a combined set of AI methods: LDA topic modeling, BERT-based sentiment analysis, transformer-enabled NER and intertextual similarity detection can obtain the state-of-the-art scores at a variety of literary analysis tasks, and yield researchable scholarly knowledge about the thematic and affective formation of English literary history between 1800 and 2010. We achieve an overall accuracy of 91.6% and an F1-score of 91.1 with our proposed architecture, which is superior to all previous systems in similar tasks. The longitudinal analysis confirms that there was statistically significant change towards positive affect in post-1960 literature and records development of Gender Identity and Colonial Perspectives as coherent thematic clusters that emerged in the literature record.

All these findings are part of the current methodological-integration project in the humanities, that AI tools do not necessarily threaten the humanistic inquiry, but can be a potent, scalable, and interpretively generative method in a tradition that has never been afraid of new tools of analysis. Concurrently, the paper highlights the importance of paying consistent attention to the corpus bias, model transparency, and the maintenance of interpretive authority within the academic community as the main elements of responsible application of AI in DH settings. The enhanced performance of the ensemble architecture compared to each component of the computational model affirm that methodological pluralism, the hallmark epistemological commitment of Digital Humanities, provides increased performance in computational practice, and theoretical advocacy.

Future Work and Research Directions

It is proposed that a number of future research directions come out of this study. First, there is an immediate need to expand the corpus to other non-English literary traditions, multilingual corpora, the current results cannot be extended to other cultural hierarchies represented in the literature without jeopardizing the subversion of the hierarchies that the study is attempting to question. Secondly, improving the performance of historically domain-adapted language models--trained on large volumes of historically-specific text, as opposed to on relatively modern web corpora--should result in significant performance gains on problems with historical orthographic and semantic rules. Even very basic trial runs of period-specific fine-tuning are indicating an improvement in accuracy of 3-5 percent on literary works published prior to 1900.

Third, racialising multimodal data, such as illustrations, paratexts, publisher adverts, marginalia by readers (where they survive in electronic form) can give the possibility of truly multimodal literary history, which takes care of the material and visual aspects of textual culture as well as the linguistic one. Fourt, explainable AI (XAI) tools tailored to humanistic understanding on-the-one-hand (delivering not merely classifications and confidence scores but also transparent explanations of model choices in human language comprehensible to a scholar) are necessary should computational approaches win the confidence of scholars experienced in interpretive traditions of philosophy that emphasized reflexivity and the accountability of arguments. Lastly, creating a common, privately-

contributed benchmark dataset to computational literary analysis comparable to the SQuAD dataset to question answering would significantly improve the ability of the field to compare the performance of studies within the field as well as chart the improvement of methods over time.

References

Busa, R. (1980). The annals of humanities computing: The Index Thomisticus. Computers and the Humanities, 14(2), 83–90. https://doi.org/10.1007/BF02403798

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

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

Underwood, T. (2019). Distant horizons: Digital evidence and literary change. University of Chicago Press. https://doi.org/10.7208/chicago/9780226612973.001.0001

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

Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020). How we do things with words: Analyzing text as social and cultural data. Frontiers in Artificial Intelligence, 3, 62. https://doi.org/10.3389/frai.2020.00062

Liu, A. (2013). The meaning of the digital humanities. PMLA, 128(2), 409–423. https://doi.org/10.1632/pmla.2013.128.2.409

Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004). A companion to digital humanities. Blackwell. https://doi.org/10.1002/9780470999875

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

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). Digital_humanities. MIT Press.

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

Rockwell, G., & Sinclair, S. (2016). Hermeneutica: Computer-assisted interpretation in the humanities. MIT Press.

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

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. Poetics, 41(6), 750–769. https://doi.org/10.1016/j.poetic.2013.08.005

Rhody, L. M. (2012). Topic modeling and figurative language. Journal of Digital Humanities, 2(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

Goldstone, A., & Underwood, T. (2014). The quiet transformations of literary studies. New Literary History, 45(3), 359–384. https://doi.org/10.1353/nlh.2014.0025

Riddell, A. B. (2014). How to read 22,198 journal articles. In M. Erlin & L. Tatlock (Eds.), Distant readings: Topologies of German culture in the long nineteenth century (pp. 91–113). Camden House.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of AAAI ICWSM, 8(1), 216–225.

Sculley, D., & Pasanek, B. M. (2008). Meaning and mining: The impact of implicit assumptions in data mining for the humanities. Literary and Linguistic Computing, 23(4), 409–424. https://doi.org/10.1093/llc/fqn019

Jockers, M. L. (2015). Syuzhet: Extract sentiment and plot arcs from text. GitHub repository. https://github.com/mjockers/syuzhet

Swafford, A. (2015). Problems with the syuzhet package. Anglophile in Academia. https://annieswafford.wordpress.com/2015/03/02/syuzhet/

Kim, E., Padó, S., & Klinger, R. (2019). Investigating emotion-denoting adjectives in literary texts. Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, 45–54.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176–182. https://doi.org/10.1126/science.1199644

Murrieta-Flores, P., Donaldson, C., & Gregory, I. (2017). GIS and literary history. Digital Humanities Quarterly, 11(1). http://www.digitalhumanities.org/dhq/vol/11/1/000283/000283.html

Dekker, R. H., van Hulle, D., Middell, G., Neyt, V., & van Zundert, J. (2015). Computer-supported collation of modern manuscripts. Literary and Linguistic Computing, 30(3), 452–470. https://doi.org/10.1093/llc/fqu007

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1–35. https://doi.org/10.1145/3560815

Kaplan, F. (2015). A map for big data research in digital humanities. Frontiers in Digital Humanities, 2, 1. https://doi.org/10.3389/fdigh.2015.00001

Baron, A., & Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. Proceedings of the Postgraduate Conference in Corpus Linguistics. Aston University.

Kocmi, T., & Bojar, O. (2017). An exploration of neural sequence-to-sequence architectures for automatic post-editing. Proceedings of the Eighth International Joint Conference on Natural Language Processing, 1, 247–257.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of WSDM 2015, 399–408. https://doi.org/10.1145/2684822.2685324

Meister, J. C. (2012). CATMA—Computer Assisted Text Markup and Analysis. http://www.catma.de

Guillory, J. (1993). Cultural capital: The problem of literary canon formation. University of Chicago Press.

Showalter, E. (1977). A literature of their own: British women novelists from Brontë to Lessing. Princeton University Press.

English, J. F. (2005). The economy of prestige: Prizes, awards, and the circulation of cultural value. Harvard University Press.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science, 5(1), 31. https://doi.org/10.1140/epjds/s13688-016-0093-1

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.