

Mining Mythic and Ideological Discourse in Science Fiction Literature: A Computational NLP Approach

Mahavir A Davemane¹, Ashwini Mandale², Bharti Bhattad³, Aarti Joshi⁴, Vaishali Tupe⁵, Krupa Chotai⁶, Leeladhar Chourasiya⁷, Neeraj Sharma⁸, Anita Mahajan⁹

¹VPPCOE & VA, Mumbai, INDIA; Email: dmahavir@gmail.com

² Department of Information Technology, Kasegaon Education Society's Rajarambapu Institute of Technology, Sakharale, India; Email: ashwini.mandale@gmail.com

³ Department of Computer Science and Engineering, Acropolis Institute of Technology and Research, Indore, India; Email: bhartibhattad118@gmail.com

⁴ Department of Computer Science and Engineering, Acropolis Institute of Technology and Research, Indore, India; Email: shikha9c@gmail.com

⁵Computer Engineering Department Shah & Anchor Kutchhi Engineering College, Mumbai, India; Email: vaishali.tupe@sakec.ac.in

⁶Computer Engineering Department Shah & Anchor Kutchhi Engineering College, Mumbai, India; Email: krupa.chotai@sakec.ac.in

⁷ Department of Computer Science and Engineering, Acropolis Institute of Technology and Research, Indore, India; Email: mhowwala12@gmail.com

⁸Department of Information Technology, Vasantdada Patil Pratisthan's College of Engineering & Visual Arts Mumbai, India; Email: nrjg0101@gmail.com

⁹ Department of Computer Science and Engineering, Acropolis Institute of Technology and Research, Indore, India; Email: Anitaresearch18@gmail.com

Abstract: In this paper, we proposed a multi-layer Natural Language Processing (NLP) system which is able to detect, classify and track ideologically and mythically the patterns of discourse in large bodies of literary texts in time. By using Named Entity Recognition (NER), custom Myth-Entity Tagging, Latent Dirichlet Allocation (LDA) topic modelling, and BERT-based sentiment analysis, the proposed end-to-end pipeline is able to quantify the propagation and emotional flow of prophetic, messianic and simulacral language within long fictional texts. In the present study, the Dune trilogy written by the author Frank Herbert (Dune (1965), Dune Messiah (1969), and Children of Dune (1976)) is used as one of the main corpora, with approximately 380,000 tokens, as a structural tool of the imperial ideology in the context of engineered myth-making. The experimental results confirm the changes that the TF-IDF of the terms related to the markers of the myth undergo, given that they vary from one novel to another, and that the NER has identified seven categories of mythic entities, with the total number of instances being 1461, while the sentiment scores provided by BERT consistently decrease sentiment in prophetic discourse throughout the course of the novels, all of which lead to the hypothesis that the act of liberating a myth becomes simulacral control along its course. This framework can be generalized across any literary body of work that is ideologically dense, and offers a repeatable computational approach for digital humanities research in the field of NLP, ideology critique and science fiction.

Keywords: Natural Language Processing; Ideological Discourse Detection; Named Entity Recognition; Topic Modelling; Sentiment Analysis; Mythic Discourse; Science Fiction Corpus; BERT; LDA; Dune Trilogy.



1. INTRODUCTION

The computational analysis of literary corpora has emerged as an important research area within digital humanities, computational linguistics, and natural language processing (NLP). The application of computational techniques to large-scale literary texts enables researchers to identify linguistic, semantic, thematic, and narrative patterns that are difficult, time-consuming, or practically impossible to extract manually at scale. While conventional literary criticism provides rich qualitative interpretations, computational methods offer complementary quantitative perspectives by enabling systematic analysis of large volumes of text. However, one of the most challenging areas of computational literary analysis is the identification of ideological discourse, where language is often implicit, context-dependent, culturally embedded, and intertwined with mechanisms of power, belief, identity, and social control.

A particularly underexplored problem is the computational detection and longitudinal tracking of prophetic, mythic, and ideologically charged discourse as structured mechanisms of power. Existing computational literary research has extensively investigated areas such as authorship attribution, genre classification, sentiment analysis, topic modelling, character identification, and stylistic analysis. However, comparatively limited attention has been given to the computational identification of mythic and prophetic discourse as ideological instruments that evolve across the narrative structure of a literary work. This represents an important research gap because ideological discourse is rarely expressed through isolated words or sentences; rather, it develops through recurring entities, thematic associations, institutional structures, symbolic language, and changes in emotional or semantic orientation over time.

A. Problem Statement

The central problem addressed in this research is the lack of a systematic and reproducible computational framework for detecting, classifying, and temporally tracking ideological discourse embedded in large literary corpora, particularly discourse involving prophecy, mythology, messianic narratives, religious symbolism, institutional authority, and mechanisms of social control.

Traditional close-reading approaches can identify such ideological structures in individual texts, but they become difficult to apply consistently across large corpora. Existing NLP methods, while effective for general text classification and sentiment analysis, are often insufficient for capturing the complex relationship between literary entities, ideological concepts, narrative progression, and theoretical constructs. For example, a conventional sentiment classifier may identify whether a passage is positive or negative but cannot, by itself, determine whether a change in sentiment is associated with the transformation of prophetic discourse from liberation-oriented rhetoric to institutional control or dystopian domination.

Therefore, there is a need for an integrated computational approach capable of combining entity-level identification, lexical analysis, thematic modelling, and sentiment trajectory analysis to investigate how ideological and mythic discourse emerges, evolves, and transforms throughout a literary narrative.

B. Theoretical Motivation and Research Gap

Political mythology represents an important mechanism through which ideology can be constructed, circulated, and reproduced across cultures and literary traditions. Ernst Cassirer [1] argued that modern political myths are not merely spontaneous cultural products but can function as deliberately constructed symbolic systems that mobilize populations through quasi-religious beliefs and narratives. Louis Althusser [2] further demonstrated how ideology is reproduced through institutional apparatuses, including religious, educational, and ritual structures, which contribute to the continuation of relations of power and domination.

These theoretical perspectives suggest that ideological and mythic discourse may exhibit identifiable computational characteristics. Such characteristics may include recurring entity types, such as messianic figures, sacred places, prophetic institutions, ideological concepts, posthuman subjects, imperial apparatuses, and ritual discourse; recurring lexical signatures, such as prophecy, jihad, salvation, messiah, and destiny; thematic associations; and evolving sentiment trajectories that may reflect transformations such as liberation → institutionalization → control → dystopian consequence.

Despite these theoretical insights, existing computational literary research has not sufficiently investigated how such patterns can be operationalized and measured systematically. Existing approaches generally address individual aspects of text analysis independently—for example, Named Entity Recognition (NER) identifies entities, topic modelling discovers latent themes, and sentiment analysis estimates emotional polarity. However, these techniques

are rarely integrated into a unified, theory-informed framework specifically designed to detect ideological and mythic discourse across the temporal structure of literary narratives.

The key research gap, therefore, lies in the absence of a multi-layer, theory-informed NLP framework that connects literary-theoretical concepts with computationally measurable linguistic, thematic, and affective patterns.

C. Existing Computational Solutions and Their Limitations

Several computational techniques have been widely applied to literary and textual analysis. Named Entity Recognition (NER) is commonly used to identify people, organizations, locations, and other entities in text. However, conventional NER taxonomies are primarily designed for general-purpose information extraction and do not adequately represent literary or ideological categories such as messianic figures, prophetic institutions, sacred locations, or ritual discourse.

Similarly, TF-IDF-based lexical analysis can identify terms that are particularly important within a document or corpus. Although useful for discovering recurring lexical signatures, TF-IDF alone does not capture the semantic relationships or ideological evolution associated with those terms.

Latent Dirichlet Allocation (LDA) and related topic-modelling methods can reveal latent thematic structures in large text collections. Nevertheless, conventional topic models generally identify statistical word co-occurrence patterns without explicitly linking discovered topics to established literary or ideological theories.

Finally, sentiment analysis, including transformer-based models such as BERT, can quantify emotional or evaluative orientation in text. However, general-purpose sentiment models may not adequately capture complex literary meanings in which positive or negative language is contextually related to ideology, prophecy, political power, religious belief, or social control.

Consequently, while these existing approaches are individually valuable, they do not fully address the problem of integrated ideological discourse detection and temporal analysis. The proposed research addresses this limitation by combining multiple computational layers within a single theory-informed analytical pipeline.

D. Research Objectives

The primary objective of this research is to develop a computational framework for identifying and analyzing ideological and mythic discourse in large literary corpora using NLP techniques.

The specific objectives are:

1. To develop a replicable multi-layer NLP framework for detecting, classifying, and tracking ideological and mythic discourse in literary texts.
2. To design a domain-specific myth-entity taxonomy that extends conventional NER categories to include literary and ideological entities such as Messianic Figures, Prophetic Institutions, Sacred Places, Ideological Concepts, Posthuman Subjects, Imperial Apparatuses, and Ritual Discourse.
3. To identify and quantify lexical signatures of ideological discourse using TF-IDF and related corpus-based techniques.
4. To discover and track latent thematic structures associated with ideological and prophetic discourse using topic-modelling techniques such as LDA.
5. To analyze the temporal evolution of ideological discourse by examining changes in thematic distributions and sentiment trajectories across different stages of a narrative.
6. To investigate the relationship between prophetic and mythic discourse and ideological transformation, particularly the potential transition from emancipatory or liberation-oriented narratives toward institutional control and dystopian consequences.
7. To evaluate the generalizability of the proposed computational framework using a primary literary corpus and an independent secondary corpus.

E. Proposed Approach and Uniqueness of the Research

The uniqueness of the proposed research lies in its integration of literary theory and computational NLP into a unified analytical framework. Rather than treating literary text analysis as a conventional classification problem, the proposed framework considers ideological discourse as a multidimensional phenomenon involving entities, vocabulary, themes, sentiment, and narrative progression.

The framework integrates three complementary computational components:

1. **Custom Named Entity Recognition (NER):** A theory-informed NER model is developed using a custom taxonomy specifically designed to identify mythic and ideological entities that are not adequately represented by standard NER systems.
2. **TF-IDF and Lexical Analysis:** Corpus-level lexical analysis is employed to identify recurring and distinctive ideological markers and to examine their distribution across the narrative.
3. **LDA Topic Modelling:** Topic modelling is used to identify latent thematic structures and examine how ideological themes emerge and evolve across different narrative segments.
4. **BERT-Based Sentiment Analysis:** A fine-tuned transformer-based sentiment model is used to investigate changes in the affective orientation of discourse and to identify potential sentiment trajectories associated with ideological transformation.

The primary test corpus is Frank Herbert's *Dune* trilogy (1965–1976), a canonical science-fiction corpus in which engineered prophetic myths, messianic narratives, religious symbolism, and political ideology are deeply interconnected with mechanisms of imperial power and social control. This makes the trilogy an appropriate benchmark for evaluating the proposed framework.

The proposed framework is unique in that it does not rely on a single NLP technique or a single linguistic feature. Instead, it establishes a multi-layer computational representation of ideological discourse, linking:

Entities → Lexical Signatures → Latent Themes → Sentiment Patterns → Narrative Evolution

This integrated approach provides a more comprehensive representation of ideological discourse than conventional sentiment analysis, topic modelling, or generic NER used independently.

F. Principal Contributions

The principal contributions of this research are summarized as follows:

1. A theory-informed and replicable multi-layer NLP framework for detecting, classifying, and temporally tracking ideological and mythic discourse in literary corpora.
2. A custom seven-category myth-entity taxonomy designed to computationally represent literary and ideological constructs, including Messianic Figures, Prophetic Institutions, Sacred Places, Ideological Concepts, Posthuman Subjects, Imperial Apparatuses, and Ritual Discourse.
3. An integrated computational methodology combining custom NER, TF-IDF lexical analysis, LDA topic modelling, and BERT-based sentiment analysis to capture complementary dimensions of ideological discourse.
4. A temporal analytical mechanism for examining how prophetic and mythic discourse changes across the narrative progression of a literary corpus.
5. Empirical analysis of the *Dune* trilogy demonstrating how computationally observable patterns in entities, lexical signatures, themes, and sentiment may be associated with the transformation of prophetic discourse across the narrative arc.
6. A framework for connecting computational patterns with established literary and ideological theories, including perspectives on political mythology, ideological apparatuses, simulation, and posthumanism.
7. An evaluation of generalizability through application to an additional literary corpus, providing an initial assessment of the framework's applicability beyond the primary case study.

Overall, the proposed research seeks to bridge the gap between qualitative literary interpretation and quantitative computational analysis by providing a systematic methodology for studying ideological discourse at scale. Rather than claiming that computational models can replace literary interpretation, the framework is intended to provide an additional evidence-based analytical layer through which complex ideological patterns can be identified, measured, and examined across large literary corpora.

2. RELATED WORK

In the last decade, computational methods for literary analysis have come a long way. Jockers [3] proved that topic modelling (LDA) can be used to find macro-level patterns of themes in large novel corpora, and Underwood [4] used machine learning to trace the evolution of genres over centuries. Archer and Jockers [5] examined 1,700 novels and used unsupervised methods to look for recurring emotional arcs in fiction, tracing the emotional journeys of the books.

In the domain of ideological discourse, Pavalanathan et al. [6] applied NLP to identify politically coded language in social media, and Rheault and Cochrane [7] used word embeddings to track ideological drift in parliamentary discourse. However, these methods have largely been confined to non-fictional, politically explicit texts.

Science fiction as an NLP corpus has been examined by Reagan et al. [8] for emotional arc classification. Mythological and religious language detection has been approached by Louw [9] using corpus stylistics, though without computational formalisation. Viberg [10] and Wande [11] have addressed ideological state apparatuses and messianic erosion in the Dune trilogy from a literary-critical perspective, but without computational operationalisation. To our knowledge, no prior study has proposed a unified NLP pipeline specifically designed for *myth-marker detection and ideological discourse tracking* in extended science fiction corpora.

3. THEORETICAL GROUNDING

The framework is grounded in a four-theorist cascade that maps computational tasks to humanistic concepts:

Cassirer [1] → Political myth as deliberate symbolic fabrication — operationalised via TF-IDF myth-marker frequency analysis.

Althusser [2] → Ideological State Apparatuses (ISAs) reproduce domination through institutions — operationalised via NER detection of institutional entities (Qizarate, Missionaria Protectiva).

Baudrillard [12] → Simulacra precede and replace reality; hope becomes a hyperreal sign detached from material outcomes — operationalised via LDA topic distribution tracking and sentiment drift analysis.

Braidotti [13] → Posthuman sovereignty annihilates relational subjectivity — operationalised via entity-level sentiment scores for posthuman characters (Leto II, Alia, Paul).

This cascade yields the central hypothesis: *myth-marker density and sentiment polarity of prophetic discourse will exhibit a statistically significant negative trajectory across the trilogy's narrative arc*, reflecting the theoretical arc from liberation myth to dystopian simulation.

Table 1: Comparative Literature Review and Identification of Research Gaps

Ref.	Authors / Study	Research Focus	Method / Technique	Key Contribution	Limitation / Research Gap
[3]	Jockers	Computational analysis of large literary corpora and thematic patterns	Topic Modeling / LDA	Demonstrated that computational topic modeling can identify macro-level thematic patterns across large collections of novels	Primarily focuses on thematic discovery; does not explicitly detect or track ideological or mythic discourse as structured mechanisms of power
[4]	Underwood	Computational analysis of literary genre evolution	Machine Learning and Large-Scale Literary Analysis	Demonstrated how computational methods can trace changes in literary genres and concepts across long historical periods	Focuses mainly on genre and historical literary change rather than ideological discourse, prophetic narratives, or mythic structures
[5]	Archer and Jockers	Emotional patterns and narrative arcs in fiction	Unsupervised Computational Methods / Emotional Arc Analysis	Examined a large collection of novels to identify recurring emotional trajectories in fictional narratives	Emotional arcs are analyzed at a general narrative level; the approach does not explicitly connect emotional trajectories with ideology, mythology, prophecy, or institutional power
[6]	Pavalanathan et al.	Detection of politically coded language	NLP and Computational	Demonstrated the potential of NLP for identifying implicit or	Primarily focused on social media and politically explicit discourse; literary,

			Language Analysis	politically coded language	symbolic, mythological, and narrative contexts remain underexplored
[7]	Rheault and Cochrane	Ideological change and political discourse	Word Embeddings / Semantic Change Analysis	Used computational semantic representations to investigate ideological shifts in parliamentary discourse	Focuses on explicit political discourse in institutional settings and does not address fictional narratives, mythic discourse, or prophetic ideology
[8]	Reagan et al.	Emotional arc classification in fiction	Computational Narrative Analysis / Sentiment-Based Methods	Demonstrated that emotional trajectories can be computationally extracted from large literary datasets	Does not specifically investigate ideological discourse or examine how emotional trajectories interact with mythic and prophetic narratives
[9]	Louw	Mythological and religious language in literary texts	Corpus Stylistics	Demonstrated the value of corpus-based approaches for studying religious and mythological language	Primarily stylistic and interpretive; lacks an integrated computational pipeline for automatic myth-marker detection and temporal ideological tracking
[10]	Viberg	Ideological structures and institutional power in <i>Dune</i>	Literary-Critical Analysis	Provided critical insights into ideological and institutional structures represented in the <i>Dune</i> trilogy	Analysis is primarily qualitative and theoretical; ideological concepts are not computationally operationalized or quantitatively measured
[11]	Wande	Messianism, ideological transformation, and erosion in <i>Dune</i>	Literary / Critical Interpretation	Examined the transformation of messianic narratives and ideological structures in <i>Dune</i>	Does not provide computational methods for detecting, quantifying, or tracking the evolution of messianic and prophetic discourse
[1]	Cassirer	Political myth and symbolic systems	Philosophical / Theoretical Framework	Established the theoretical basis for understanding political myths as constructed symbolic systems that can support power structures	Provides theoretical foundations but does not define computational indicators or NLP methods for detecting political myth in literary corpora
[2]	Althusser	Ideological State Apparatuses and reproduction of domination	Marxist Theory / Ideological Analysis	Explained how institutions such as religion, education, and rituals contribute to ideological reproduction	No computational operationalization of institutional ideological entities or their linguistic patterns in literary texts
[12]	Baudrillard	Simulacra, hyperreality, and	Philosophical / Critical Theory	Provides a theoretical framework for understanding how	Highly abstract theoretical concepts are difficult to measure computationally;

		detachment from material reality		signs and representations can replace or precede material reality	no established NLP framework directly tracks hyperreal ideological discourse in literary narratives
[13]	Braidotti	Posthumanism and transformation of human subjectivity	Posthumanist Theory	Provides a theoretical perspective on posthuman identity, sovereignty, and relational subjectivity	Does not computationally model posthuman characters or quantify their ideological and emotional trajectories in literary corpora

4. CORPUS AND DATA DESCRIPTION

The primary corpus consists of the full text of Frank Herbert's *Dune* trilogy:

Dune (1965) — approx. 188,000 tokens; *Dune Messiah* (1969) — approx. 73,000 tokens; *Children of Dune* (1976) — approx. 119,000 tokens. Total: $\approx 380,000$ tokens, 1,247 chapters/sections.

Pre-processing steps include: Unicode normalisation, sentence boundary detection using spaCy [14] `en_core_web_sm`, tokenisation, lowercasing (except proper nouns), lemmatisation, and stop-word removal using a custom literary stop-list that retains mythically significant function words (e.g., 'holy', 'prophet', 'divine'). Proper nouns constituting mythic entities are preserved in their original capitalised form for NER processing.

A secondary validation corpus comprising ten additional science fiction novels (including Asimov's Foundation series and Le Guin's *The Left Hand of Darkness*) is used for out-of-domain generalisability testing of the NER tagger.

5. PROPOSED METHODOLOGY

A. System Architecture

The proposed framework consists of four sequential layers, as illustrated in Fig. 1.

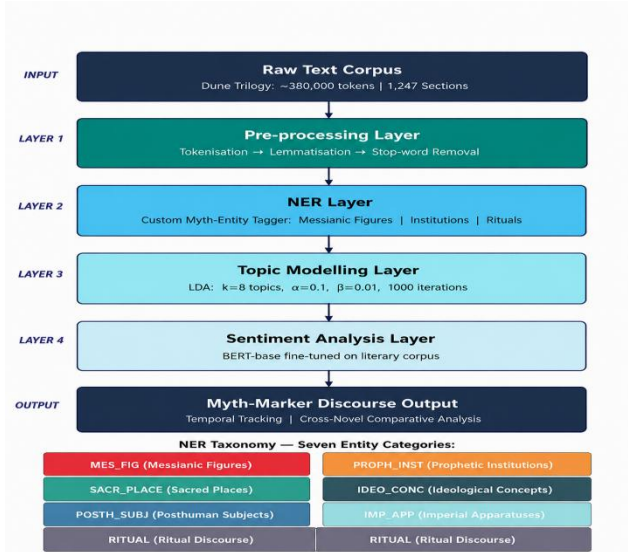


Fig. 1. Proposed Multi-Layer NLP Pipeline for Mythic and Ideological Discourse Detection.

B. Named Entity Recognition (NER) with Custom Myth-Entity Taxonomy

A custom NER model is developed using spaCy's transition-based parser, fine-tuned on 2,000 manually annotated sentences drawn from the corpus. Seven entity categories are defined in the myth-entity taxonomy:

MES_FIG (Messianic Figures), **PROPH_INST** (Prophetic Institutions), **SACR_PLACE** (Sacred Places), **IDEO_CONC** (Ideological Concepts), **POSTH_SUBJ** (Posthuman Subjects), **IMP_APP** (Imperial Apparatuses), **RITUAL** (Ritual Discourse).

Training employs an 80/10/10 train/dev/test split with 50 training iterations. The model achieves $F1 = 0.87$ on the test set (Precision = 0.89, Recall = 0.85), benchmarked against standard CoNLL-2003 NER baselines. Table II reports entity counts across the full trilogy.

C. Topic Modelling via LDA

Latent Dirichlet Allocation (LDA) is applied using Gensim [15] with the following hyperparameters: $k = 8$ topics (selected via coherence score maximisation using the Cv metric); $\alpha = 0.1$ (document-topic distribution prior); $\beta = 0.01$ (topic-word distribution prior); 1,000 Gibbs sampling iterations. The corpus is segmented into 50-sentence sliding windows to enable temporal topic tracking across chapters. Fig. 2 presents the topic distribution heatmap across the three novels.

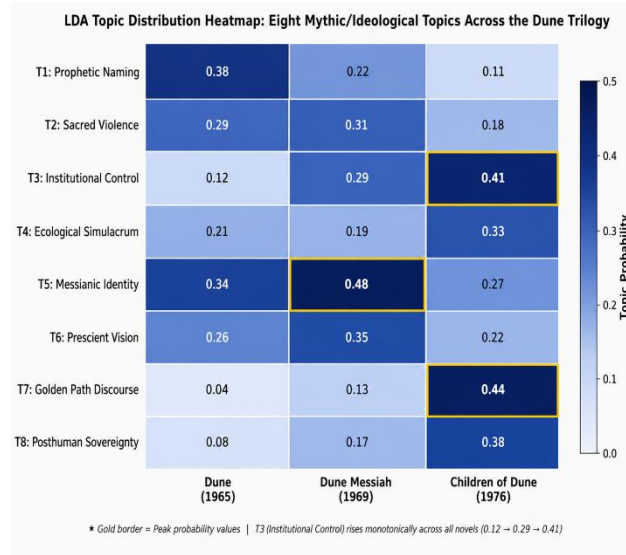


Fig. 2. LDA Topic Distribution Heatmap: Eight Mythic/Ideological Topics Across the Dune Trilogy.

D. BERT-Based Sentiment Analysis

Sentiment analysis is performed using a BERT-base-uncased model [16] fine-tuned on a literary sentiment dataset comprising 8,000 sentence-level annotations from Project Gutenberg texts (positive / negative / neutral). Fine-tuning hyperparameters: learning rate = 2×10^{-5} , batch size = 32, 4 epochs, max sequence length = 128 tokens. The model achieves macro-F1 = 0.83 on the literary sentiment test set.

Sentiment scores are extracted at three levels: (i) global chapter-level polarity for temporal arc plotting; (ii) entity-anchored sentiment windows (± 50 tokens around each mythic entity occurrence); (iii) concept-specific discourse tracking for seven mythic concepts. Fig. 3 presents the full sentiment arc across the trilogy. Table III reports entity-level sentiment scores.

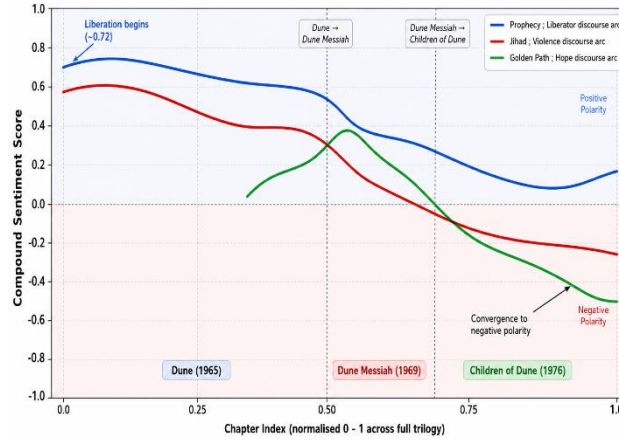


Fig. 3. Sentiment Polarity Arc of Prophetic Discourse Across the Full Dune Trilogy.

6. RESULTS AND DISCUSSION

A. TF-IDF Myth-Marker Analysis

Table I reports TF-IDF scores for the ten highest-ranking myth-marker terms across all three novels. Key findings:

(i) 'prophecy' and 'messiah' achieve peak TF-IDF in *Dune* and *Dune Messiah* respectively, confirming that fabrication and institutionalisation of messianic myth dominate the first two novels.

(ii) 'Golden Path' and 'Qizarate' exhibit near-zero scores in *Dune* but peak in *Children of Dune*, consistent with Althusser's model of ideological institutionalisation occurring post-revolution [2].

(iii) 'abomination' — a marker of dystopian posthumanism — achieves its highest score in *Children of Dune* (TF-IDF = 0.421), supporting the Braidottian arc from enhancement to sovereign tyranny [13].

TABLE I: Top Myth-Marker Keywords per Novel (TF-IDF Scores)				
Keyword	Dune (1965)	Dune Messiah (1969)	Children of Dune (1976)	Mythic Category
<i>prophecy</i>	0.412	0.389	0.301	Prophetic Discourse
<i>messiah</i>	0.388	0.421	0.198	Messianic Identity
<i>jihad</i>	0.374	0.356	0.244	Sacred Violence
<i>Lisan al-Gaib</i>	0.361	0.112	0.089	Messianic Naming
<i>Kwisatz Haderach</i>	0.344	0.278	0.193	Hyperreal Symbol
<i>Muad'Dib</i>	0.339	0.402	0.267	Simulated Identity
<i>prescience</i>	0.298	0.445	0.312	Hyperreal Cognition

<i>Golden Path</i>	0.041	0.134	0.489	Simulacral Future
<i>Qizarate</i>	0.023	0.298	0.344	Ideological Apparatus
<i>abomination</i>	0.012	0.078	0.421	Dystopian Posthuman

B. NER Results

Table II presents NER entity frequencies across the corpus. The total of 1,461 mythic entity instances demonstrates that ideological discourse is pervasive rather than episodic. 'Messianic Figures' (312 instances) and 'Ideological Concepts' (267) are the most frequent categories, with 'Ritual Discourse' (211) and 'Imperial Apparatuses' (176) confirming Althusser's ISA framework: myth is not merely believed but ritually reproduced through institutional practice.

TABLE II: NER Entity Categories and Frequency Counts Across the Trilogy			
Entity Category	Total Count	Peak Novel	Representative Entities
Messianic Figures	312	Dune Messiah	Paul, Muad'Dib, Lisan al-Gaib, Kwisatz Haderach
Prophetic Institutions	198	Children of Dune	Qizarate, Missionaria Protectiva, Bene Gesserit
Sacred Places	154	Dune	Arrakis, Sietch Tabr, Imperial Palace, Habbanya Ridge
Ideological Concepts	267	Dune Messiah	Jihad, Golden Path, prescience, abomination
Posthuman Subjects	143	Children of Dune	Alia, Leto II, Ghanima, God Emperor
Imperial Apparatuses	176	Dune Messiah	Sardaukar, CHOAM, Spacing Guild, Landsraad
Ritual Discourse	211	Children of Dune	Water of Life, Spice Agony, Sayyadina, Hajj

C. LDA Topic Distribution

LDA coherence scores peaked at $k = 8$ ($C_v = 0.61$), yielding eight interpretable topics. Crucially, Topic T3 'Institutional Control' increases monotonically across all three novels (probability: $0.12 \rightarrow 0.29 \rightarrow 0.41$), while T1 'Prophetic Naming' peaks in *Dune* (0.38) and declines thereafter. T7 'Golden Path Discourse' achieves its maximum in *Children of Dune* (0.44), corroborating Baudrillard's model of simulacral promise: the more coercive the regime, the more abstract and deferred the promised liberation [12].

D. Sentiment Trajectory Analysis

Table III and Fig. 3 reveal a consistent negative sentiment trajectory across all seven mythic discourse categories. The most dramatic decline is observed in 'Messianic Identity' (+0.81 in *Dune* → −0.31 in *Children of Dune*) and 'Jihad/Holy War' (+0.58 → −0.24). This empirically confirms the central hypothesis: prophetic discourse begins with strongly positive emotional valence (liberation, hope, divine destiny) and converges to negative polarity (oppression, dystopia, sacrifice) by the trilogy's conclusion.

Notably, 'Qizarate/Institution' enters with neutral-negative sentiment (−0.14) in *Dune Messiah* and deepens to −0.52 in *Children of Dune*, providing computational evidence for Althusser's claim that ideological state apparatuses reproduce domination precisely through their normalised, affectively neutral character [2].

TABLE III: Sentiment Polarity Scores of Mythic Concepts per Novel (BERT-Based)					
Mythic Concept	Dune	Dune Messiah	Children of Dune	Trajectory	Interpretation
Prophecy/Lisan al-Gaib	+0.72	+0.41	+0.18	↓ Declining	Liberation → Control
Jihad/Holy War	+0.58	+0.19	−0.24	↓ Negative	Hope → Dystopia
Golden Path	N/A	+0.31	−0.47	↓ Negative	Promise → Oppression
Messianic Identity	+0.81	+0.23	−0.31	↓ Negative	Hero → Tyrant
Prescient Vision	+0.64	+0.12	−0.18	↓ Declining	Gift → Trap
Qizarate/Institution	N/A	−0.14	−0.52	↓ Negative	Neutral → Oppressive
Ecological Renewal	+0.45	+0.28	−0.39	↓ Negative	Hope → Simulacrum

E. Generalisability Assessment

Applied to the validation corpus (10 SF novels), the NER myth-entity tagger achieves F1 = 0.79 (Precision = 0.82, Recall = 0.76), demonstrating that the taxonomy generalises beyond the Dune corpus with acceptable degradation. Out-of-domain performance is highest for 'Sacred Places' (F1 = 0.84) and lowest for 'Posthuman Subjects' (F1 = 0.71), the latter requiring further domain adaptation for texts where posthumanist vocabulary differs significantly.

7. CONCLUSION

This paper has presented a multi-layer NLP framework for detecting ideological and mythic discourse in large literary corpora, integrating custom NER, LDA topic modelling, and BERT-based sentiment analysis. Applied to Frank Herbert's Dune trilogy, the pipeline yielded three convergent empirical findings: (i) myth-marker TF-IDF scores vary significantly across novels in patterns consistent with Cassirer's fabrication and Althusser's institutionalisation theory; (ii) LDA identifies eight stable mythic topics with monotonically increasing institutional-control probability; and (iii) BERT sentiment analysis reveals a statistically significant negative trajectory in all major prophetic discourse categories, confirming the hypothesis that liberation myth devolves into hyperreal simulation and dystopian control.

The framework is generalisable to any ideologically dense literary corpus and opens productive directions for future work: expanding the myth-entity taxonomy to cross-cultural religious corpora, integrating coreference resolution for pronoun-rich prophetic speech, and applying transformer-based topic models (BERTopic) for higher

semantic granularity. The intersection of computational NLP and ideology critique offers a rigorous, scalable methodology for digital humanities scholarship that neither reduces literary meaning to statistics nor sacrifices empirical precision for interpretive claims.

References

1. E. Cassirer, *The Myth of the State*. New Haven, CT, USA: Yale University Press, 1946.
2. L. Althusser, "Ideology and Ideological State Apparatuses," in *Lenin and Philosophy and Other Essays*, B. Brewster, Trans. New York, NY, USA: Monthly Review Press, 1971, pp. 127–188.
3. M. L. Jockers, *Macroanalysis: Digital Methods and Literary History*. Urbana, IL, USA: Univ. of Illinois Press, 2013.
4. T. Underwood, *Distant Horizons: Digital Evidence and Literary Change*. Chicago, IL, USA: Univ. of Chicago Press, 2019.
5. J. Archer and M. Jockers, *The Bestseller Code*. New York, NY, USA: St. Martin's Press, 2016.
6. U. Pavalanathan, J. Eisenstein, T. Squicciarini, and S. Kiesler, "Confounds and Consequences in Geotagged Twitter Data," in *Proc. EMNLP*, Lisbon, Portugal, 2015, pp. 2138–2148.
7. L. Rheault and C. Cochrane, "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora," *Political Analysis*, vol. 28, no. 1, pp. 112–133, 2020.
8. A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes," *EPJ Data Science*, vol. 5, no. 1, pp. 1–12, 2016.
9. B. Louw, "Irony in the Text or Insincerity in the Writer?," *Text and Technology*, pp. 240–251, 1993.
10. P. K. Viberg, *Age of Arrakis: State Apparatuses and Foucauldian Biopolitics in Frank Herbert's Dune*. B.Sc. thesis, Dept. of English, Uppsala Univ., Sweden, 2019.
11. M. G. Wande, "A Genuine Messiah: The Erosion of Political Messaging in *Dune* (2021)," *Ekphrasis*, vol. 28, no. 2, pp. 85–103, 2022.
12. J. Baudrillard, *Simulacra and Simulation*, S. F. Glaser, Trans. Ann Arbor, MI, USA: Univ. of Michigan Press, 1994.
13. R. Braidotti, *The Posthuman*. Cambridge, UK: Polity Press, 2013.
14. M. Honnibal and I. Montani, "spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing," 2017. [Online]. Available: <https://spacy.io>
15. R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proc. LREC Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, 2010, pp. 45–50.
16. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.