TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/



# RECONCEPTUALIZING LITERARY GENRE THROUGH STYLOMETRIC ANALYSIS: A COMPUTATIONAL LINGUISTIC APPROACH TO TEXT CLASSIFICATION

DR. MAHWISH SHAMIM[1], MARIAM AKRAM[2], DR. IRAM RUBAB[3], AYESHA INAM[4], DR. MUHAMMAD ARFAN LODHI*[5]

[1]ASSISTANT PROFESSOR DEPARTMENT OF ENGLISH UNIVERSITY OF RASUL, MANDI BHAODIN, EMAIL: mahwish.shamim@putrasul.edu.pk
[2]LECTURER DEPARTMENT OF ENGLISH THE GOVERNMENT SADIQ COLLEGE WOMEN UNIVERSITY BAHAWALPUR, EMAIL: maryam.akram@gscwu.edu.pk
[3]ASSISTANT PROFESSOR DEPARTMENT OF ENGLISH GC WOMEN UNIVERSITY SIALKOT, EMAIL ID: iram.rubab@gcwus.edu.pk
[4]M.PHIL SCHOLAR DEPARTMENT OF ENGLISH NCBA&E ALHAMRA UNIVERSITY, BAHAWALPUR, EMAIL ID: ayeshaslam28@gmail.com
[5]HIGHER EDUCATION DEPARTMENT, PUNJAB, EMAIL: Samaritan_as@hotmail.com

**ABSTRACT**

**Background:** This study rethinks literary genre from a linguistic and computational perspective, challenging the traditional view of genre as a fixed and rule-bound classification system. Drawing on structuralist and post-structuralist insights, genre is conceptualized as both systematic and fluid. Within this context, stylometry is positioned as a linguistic methodology that quantifies stylistic variation through lexical, syntactic, semantic, structural, and phonological features, enabling a shift from interpretive to empirical analysis.

**Method:** The study adopts a theoretical–computational approach rather than empirical text analysis. It develops a stylometric framework for modeling genre across poetry, drama, and fiction by transforming texts into multidimensional feature vectors. Comparative feature taxonomy is constructed, and computational techniques including statistical methods, machine learning, and deep learning are critically evaluated. A hybrid modeling strategy is proposed to integrate interpretability with predictive capability.

**Discussion:** The findings demonstrate that genre distinctions emerge from configurations of linguistic features rather than isolated markers. Poetry is characterized by phonological and structural constraints, drama by dialogic and interactional patterns, and fiction by narrative and syntactic complexity. Based on these insights, the study proposes the **Multidimensional Stylometric Genre Model (MSGM)**, in which genres are represented as probabilistic clusters within a multidimensional linguistic feature space. The MSGM further incorporates a hybrid analytical mechanism that combines interpretable statistical features with machine learning-based pattern recognition, addressing the tension between explainability and accuracy.

**Conclusion:** The study concludes that genre should be understood as a dynamic, probabilistic linguistic construct rather than a rigid category. The MSGM offers a flexible, scalable, and theoretically grounded computational model that aligns with contemporary digital humanities practices and supports more nuanced genre analysis across literary forms.

**KEYWORDS:** Stylometry; Computational Linguistics; Genre Classification; Multidimensional Stylometric Genre Model (MSGM); Digital Humanities

## 1. INTRODUCTION

Genre has long functioned as a central organizing principle in literary studies, shaping interpretation, canon formation, and pedagogy. However, despite its importance, genre remains conceptually unstable. Traditional approaches often rely on implicit assumptions about form, content, and historical context, leading to classifications that are difficult to formalize or replicate. The rise of computational methods in the humanities has introduced new possibilities for addressing these challenges. Stylometry, in particular, offers a framework for quantifying literary style through measurable linguistic features (Holmes, 1998). By transforming textual data into structured representations, stylometry enables systematic comparison across texts and genres. While stylometry has been widely applied to authorship attribution; its potential for genre classification remains underexplored. This gap is significant because genre, like authorship, is reflected in stylistic patterns that can be measured and modeled. The challenge lies in identifying which features are most relevant and how they can be integrated into a coherent framework. This study

this challenge by proposing a computational model of genre grounded in stylometric analysis. It focuses on three major literary forms—poetry, drama, and fiction—and examines how their stylistic characteristics can be represented in a multidimensional feature space. The study does not analyze specific texts; instead, it develops a theoretical and methodological foundation for future research.

### 1.1 Stylometric Analysis

Stylometric analysis refers to the systematic, quantitative examination of linguistic patterns in a text to uncover stylistic regularities that may not be immediately visible through traditional reading. Emerging from the convergence of linguistics, statistics, and digital humanities, it treats writing style as a measurable construct rather than a purely aesthetic phenomenon. In practice, stylometric analysis operates by extracting recurring features—such as word frequencies, sentence structures, function words, punctuation habits, and even character-level distributions—and subjecting them to statistical modeling. These patterns are then compared across texts to identify similarities, differences, or anomalies. Logically, stylometric analysis proceeds in a structured sequence: first, a corpus of texts is selected and digitized; second, relevant stylistic features are defined and extracted; third, statistical or computational techniques (e.g., cluster analysis, machine learning classifiers) are applied; and finally, the results are interpreted to draw conclusions about stylistic affinity, variation, or distinctiveness. The strength of stylometric analysis lies in its objectivity and replicability, allowing scholars to move beyond impressionistic judgments toward evidence-based claims about textual style.

### 1.2 Authorial Signatures

The concept of authorial signatures refers to the distinctive, often unconscious linguistic patterns that characterize an individual writer's style. Much like a biometric fingerprint, an authorial signature is not typically constructed deliberately; rather, it emerges from habitual choices in vocabulary, syntax, rhythm, and discourse organization. These features tend to remain relatively stable across a writer's body of work, even when the subject matter or genre changes. Narratively, an authorial signature can be understood as the "trace" a writer leaves behind in language which is a subtle but persistent imprint of identity embedded in textual production. It is shaped by cognitive habits, educational background, cultural influences, and even psychological tendencies. Logically, the notion rests on the premise that while writers can consciously manipulate certain stylistic elements, they cannot entirely suppress the deeper structural patterns of their linguistic behavior. Stylometric analysis seeks to detect and quantify these signatures by focusing on features that are less susceptible to conscious control, such as function word frequency or syntactic preferences.

### 1.3 Authorship Attribution Programs

Authorship attribution programs are computational systems designed to identify or verify the authorship of a text by analyzing its stylistic features. These programs operationalize the principles of stylometry by automating the processes of feature extraction, pattern recognition, and statistical comparison. Common examples include tools like JGAAP (Java Graphical Authorship Attribution Program) and various implementations in programming environments such as Python or R.

From a narrative perspective, authorship attribution programs function as analytical mediators between raw textual data and interpretive conclusions. They enable researchers to handle large corpora and complex datasets that would be impractical to analyze manually. Logically, these programs follow a pipeline model: input texts are first preprocessed (tokenization, normalization), then transformed into numerical representations based on selected stylistic features; next, algorithms—ranging from simple distance measures to advanced machine learning models—compare these representations; finally, the system outputs probabilities or classifications indicating likely authorship.

The reliability of such programs depends on factors such as the size and representativeness of the training corpus, the choice of features, and the robustness of the statistical model. While not infallible, they significantly enhance the precision and scalability of authorship studies, making them indispensable in fields like forensic linguistics, literary scholarship, and digital humanities. Taken together, stylometric analysis provides the methodological foundation, authorial signatures supply the theoretical premise, and authorship attribution programs offer the technological implementation. The interplay among these three elements creates a coherent framework in which writing style is quantified, identity is inferred, and authorship is systematically evaluated.

## 2. THEORETICAL FOUNDATIONS

### 2.1 Classical and Structuralist Genre Theory

Structuralist approaches to genre emphasize the existence of underlying systems that govern literary forms. Tzvetan Todorov (1975) conceptualizes genre as a set of rules that shape both production and interpretation. From this perspective, genres are not merely descriptive categories but prescriptive frameworks that guide textual construction. Such models imply that genre distinctions are systematic and, therefore, potentially quantifiable. This assumption aligns with stylometric approaches, which seek to identify recurring patterns in textual data.

### 2.2 Post-Structuralist Perspectives

Post-structuralist theorists challenge the rigidity of structuralist models by emphasizing the fluidity and instability of genre. Jacques Derrida (1980) argues that texts inevitably transgress genre boundaries, participating in multiple categories simultaneously. This perspective complicates attempts at classification but also suggests the need for more flexible models. A stylometric approach can accommodate this fluidity by representing genre as a probabilistic rather than categorical phenomenon.

### 2.3 The Computational Turn

The integration of computational methods into literary studies marks a significant shift in methodology. Franco Moretti (2013) advocates for the analysis of large-scale patterns, arguing that traditional close reading cannot capture the full complexity of literary systems. Stylometry operationalizes this approach by providing tools for measuring and comparing stylistic features. It enables researchers to move beyond anecdotal evidence toward systematic analysis.

### 2.4 Stylometry as Methodological Framework

Stylometry is grounded in the assumption that style is quantifiable. Early work by Frederick Mosteller and Wallace (1964) demonstrated the feasibility of using statistical methods to analyze textual data. Subsequent research has expanded the range of features and techniques available, making stylometry a versatile tool for literary analysis.

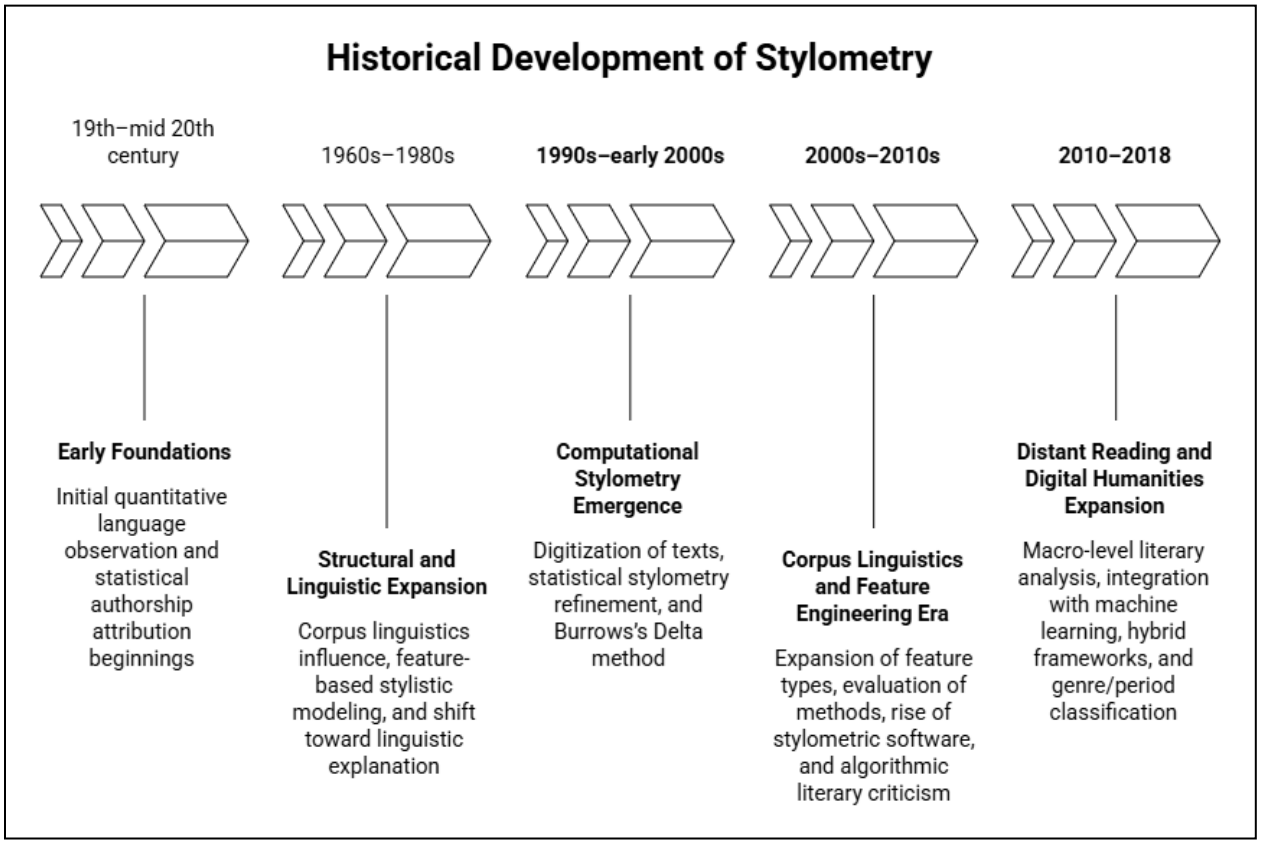


Figure 1. Stylometry across history

The development of stylometry begins in the nineteenth century and continues as a gradual shift from intuitive observation of stylistic differences to increasingly formalized and computational approaches. In its earliest phase, spanning the nineteenth century to the mid-twentieth century, attention was largely directed toward simple measurable aspects of language such as word length, sentence length, and basic frequency patterns. Style was not yet theorized as a structured system; instead, it was treated as a visible surface variation between writers. During this period, early attempts at authorship attribution introduced the idea that linguistic patterns could be used as evidence for distinguishing between writers, but these efforts remained limited by manual counting methods and small-scale textual comparison. A major turning point emerged with the work of Mosteller and Wallace (1964), who applied Bayesian statistical reasoning to disputed authorship in The Federalist Papers. Their analysis demonstrated that function words, which are often overlooked in interpretation, carry stable stylistic signals that can differentiate authors. This marked the beginning of stylometry as a statistically grounded discipline rather than a purely descriptive practice.

From the 1960s to the 1980s, stylometry began to develop more systematic linguistic foundations, influenced strongly by the rise of corpus linguistics. During this phase, researchers expanded their attention beyond simple frequency counts to more structured linguistic features, including syntactic patterns and grammatical variation. Style was increasingly understood as a multidimensional linguistic profile rather than a single measurable trait. This period also saw growing interest in distinguishing between spoken and written language, and in understanding how linguistic variation operates across different communicative contexts. Douglas Biber's work on variation between speech and writing provided an important foundation for this development by showing that registers of language can be systematically differentiated through distributional features. At the same time, stylometric analysis began to move closer to linguistic theory, as researchers attempted to explain stylistic variation rather than only measure it.

The period from the 1990s to the early 2000s marks the beginning of computational stylometry in a more explicit sense. The digitization of texts made large-scale analysis possible, and statistical methods became more refined and computationally efficient. Researchers began to apply clustering techniques and multivariate statistics to literary corpora, allowing for more sophisticated comparisons across texts. A defining contribution of this period is Burrows's Delta method, which introduced a standardized way of measuring stylistic distance based on the distribution of function words. This method significantly strengthened the reliability of authorship attribution and demonstrated that consistent stylistic patterns could be captured through computational means. As a result, stylometry became more widely accepted as a method capable of producing reproducible results rather than isolated case studies.

In the 2000s to 2010s, stylometry expanded significantly under the influence of corpus linguistics and machine learning. Feature engineering became a central concern, with researchers systematically exploring lexical, syntactic, and semantic features to improve classification performance. This period also saw the emergence of specialized software tools that made stylometric analysis more accessible to humanities researchers. Tools such as Stylo and AntConc allowed users to perform clustering, frequency analysis, and comparative stylistic studies without requiring advanced programming knowledge. At the same time, computational approaches to literature began to develop under the broader framework of algorithmic literary criticism, where quantitative methods were used not only for classification but also for interpretation of literary patterns. Stylometry during this phase increasingly extended beyond authorship attribution to include genre classification and historical analysis of literary change.
Between 2010 and 2018, stylometry became closely connected to the field of digital humanities and the concept of distant reading. Large-scale literary analysis replaced close reading as a complementary method for understanding literary systems at scale. Franco Moretti's work played a key role in legitimizing this approach by arguing that literary history could be studied through patterns across large datasets rather than individual texts. During this time, machine learning methods such as support vector machines and random forests began to be widely used in stylometric classification tasks. These methods allowed researchers to model complex relationships between features and improve predictive accuracy in authorship and genre classification. Hybrid approaches also began to emerge, combining traditional statistical measures with machine learning techniques in order to balance interpretability and performance.
From 2018 onward, stylometry has entered a phase dominated by neural and transformer-based models. Deep learning systems, including architectures such as BERT, have enabled the automatic learning of linguistic representations without manual feature engineering. These models have significantly improved performance in tasks such as authorship attribution, genre classification, and detection of machine-generated text. However, this increase in predictive power has introduced a major limitation, as neural models often lack interpretability. Researchers can identify what a model predicts but not easily explain why it makes a particular classification. As a result, recent work has increasingly focused on hybrid models that combine deep learning with interpretable statistical features. This period also marks the expansion of stylometry into new domains, including forensic linguistics and artificial intelligence detection, where it is used to distinguish between human and machine-generated writing. At the same time, there is growing recognition that stylistic variation is not fixed but dynamic, and that genre and authorship interact in complex and sometimes overlapping ways.

**2.5 Review of the related literature**

Recent work on stylometry and genre classification shows a clear movement toward computational sophistication, yet this development raises important theoretical and methodological concerns. While newer models improve classification accuracy, they often do so at the expense of interpretability, which remains central to literary inquiry.

One major trend is the increasing reliance on deep learning models for stylometric tasks. Studies building on architectures such as BERT demonstrate that neural networks can capture complex stylistic patterns that extend beyond surface-level lexical features (Devlin et al., 2019). Reviews such as Sharma and Kumar (2024) argue that these models outperform traditional statistical techniques in authorship attribution and genre classification tasks. However, this claim requires careful qualification. High accuracy does not necessarily translate into meaningful literary interpretation. Neural models tend to obscure the relationship between input features and classification outcomes,

making it difficult to explain why a text is categorized in a particular way. This limitation is significant because genre theory depends on interpretive reasoning rather than prediction alone.

In contrast, earlier stylometric approaches retain analytical transparency. Methods based on word frequency and distribution, such as Burrows's Delta, continue to provide interpretable results even if their predictive performance is lower (Burrows, 2002). Recent discussions suggest that abandoning these methods entirely in favor of neural models may be premature. Gorman (2024) shows that morpho syntactic features, when carefully selected, can achieve competitive results while remaining interpretable. This suggests that feature engineering still plays a central role in stylometry, despite the popularity of automated representation learning. Another important development concerns the concept of authorial signature. Traditional stylometry assumes that each author has a stable stylistic fingerprint. Recent studies challenge this assumption by showing that stylistic patterns vary significantly across genres and contexts. He et al. (2024) note that features commonly used in authorship attribution may capture genre conventions rather than individual style. This raises a critical issue. If genre influences stylistic features more strongly than authorship, then models trained to detect authors may instead be detecting genre-specific patterns. This problem complicates both authorship attribution and genre classification, suggesting that the two tasks cannot be fully separated.

The expansion of stylometry into forensic and applied domains has further exposed its limitations. Cammarota et al. (2024) demonstrate that while stylometric techniques are increasingly used in legal and security contexts, their reliability decreases with shorter or more variable texts. This finding is particularly relevant for literary studies, where texts differ widely in length and structure. Poetry, for example, often provides limited data for statistical analysis, while drama introduces variability through multiple speakers. These differences make it difficult to apply a single model across genres without significant adaptation.

Recent research on multi-authored and internally variable texts adds another layer of complexity. Zamir et al. (2024) show that stylistic variation can occur within a single document, while reflecting shifts in authorship or narrative voice. This challenges the assumption that texts are stylistically uniform units. It also suggests that genre classification models must account for internal variation rather than treating texts as homogeneous entities. Without this consideration, models risk oversimplifying the structure of literary works.

The emergence of AI-generated texts has introduced a new dimension to stylometric research. Studies such as Przystalski et al. (2025) indicate that stylometric features can distinguish between human and machine-generated writing with high accuracy. While this finding appears promising, it raises conceptual questions about the nature of style. If computational systems produce identifiable stylistic patterns, then the notion of authorial signature must be expanded beyond human writers. At the same time, rapid improvements in text generation systems may reduce these distinctions, making detection more difficult over time. Across these studies, a consistent tension appears between generalization and specificity. Models that perform well on controlled datasets often fail to generalize across genres, languages, or contexts. Stamatatos (2009) had already identified this issue in earlier work, and recent research confirms that it remains unresolved. Genre classification models frequently rely on features that are sensitive to topic or corpus composition, which limits their broader applicability. This problem is particularly evident in multilingual settings, where linguistic differences complicate feature extraction and comparison.

Consequently, recent research supports a reconsideration of how genre is conceptualized in computational terms. The evidence suggests that genre cannot be treated as a fixed label derived from a stable set of features. Instead, it should be understood as a probabilistic configuration shaped by multiple interacting variables. This view aligns with theoretical arguments advanced by Jacques Derrida (1980), who maintains that texts participate in multiple genres simultaneously. Computational models that assign a single label to each text fail to capture this complexity. At the same time, the work of Franco Moretti (2013) provides a useful framework for interpreting these findings. By focusing on patterns across large datasets, distant reading makes it possible to identify genre tendencies without relying on rigid definitions. Stylometry extends this approach by offering measurable indicators of these tendencies. However, the success of this method depends on maintaining a balance between quantitative rigor and theoretical clarity.

In light of these considerations, it becomes difficult to support approaches that rely exclusively on either traditional stylometry or deep learning. Statistical models provide interpretability but lack flexibility, while neural models offer flexibility but obscure meaning. A more viable approach lies in combining these methods, allowing each to address the limitations of the other. Hybrid models can incorporate interpretable features while also capturing complex patterns, making them better suited to the study of genre.

Table 1. Meta-synthesis of the previous researches

| Year | Study | Focus | Methodological Approach | Key Findings | Conceptual Development |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **2009** | Stamatatos (2009) | Survey of authorship attribution methods | Statistical stylometry review | Identifies feature-based methods (n-grams, lexical stats) as dominant but notes poor generalization across domains | Establishes baseline limitation: models are sensitive to corpus and genre variation |
| **2019** | Devlin et al. | Transformer models for text representation | Deep learning (BERT) | Demonstrates strong performance in text classification and representation learning | Introduces contextual embeddings that reduce reliance on manual feature engineering |
| **2022** | Gupta et al. | Machine learning for authorship attribution | TF-IDF, n-grams, classical ML models | Shows high accuracy in controlled datasets but limited robustness across domains | Reinforces importance of feature selection in traditional stylometry |
| **2024** | Sharma & Kumar | Stylometry and deep learning review | Comparative literature review of neural models | Deep learning improves classification accuracy but reduces interpretability | Highlights tension between predictive performance and explainability |
| **2024** | He et al. | Authorship attribution survey | Comprehensive methodological survey | Identifies shift toward hybrid models combining lexical, syntactic, and semantic features | Suggests authorial signature is not stable across genres |
| **2024** | Gorman | Morphosyntactic stylometry | Linguistically structured feature engineering | Morphosyntactic features improve interpretability and classification performance | Reasserts value of linguistically grounded features over purely neural approaches |
| **2024** | Cammarota et al. | Stylometry in forensic applications | Applied computational stylometry review | Stylometry is effective in forensic contexts but weak for short or heterogeneous texts | Expands stylometry beyond literature into legal and security domains |
| **2024** | Zamir et al. | Multi-authored document analysis | Stylometric shift detection models | Successfully identifies intra-document style changes | Challenges assumption of stylistic uniformity in texts |
| **2025** | Przystalski et al. | AI vs human text detection | Stylometric classification of LLM-generated text | Stylometry can distinguish human vs machine text with high accuracy in controlled settings | Extends concept of “authorial signature” to AI systems |
| **2024** | Stamatatos (revisited in discussions) | General authorship and genre limitations | Meta-analysis across methods | Persistent issue: models fail to generalize across genres and languages | Reinforces need for genre-aware stylometric modeling |
| **2024** | Sharma & Kumar synthesis | Deep learning stylometry review | Cross-method comparison | Neural models outperform classical methods but lack inter | Pushes field toward hybrid explainable systems |

**3. Stylometric Feature Taxonomy**

A key step in developing a computational model of genre is identifying the features that distinguish different literary forms. These features can be grouped into several categories.

The table illustrates that genre distinctions emerge from combinations of features rather than single markers. Poetry, for instance, is not defined solely by rhyme but by the interaction of phonological, structural, and lexical elements. Drama’s defining feature is its dialogic structure, while fiction emphasizes narrative continuity and syntactic complexity. This multidimensionality supports the argument that genre should be modeled as a feature space.

Table 2: Comparative Stylometric Features Across Genres

| Feature Category | Poetry | Drama | Fiction |
|---|---|---|---|
| **Lexical** | High density, figurative language | Character-specific vocabulary | Broad vocabulary range |
| **Syntactic** | Fragmented, line-based | Short utterances | Complex sentence structures |
| **Semantic** | Symbolic, layered meanings | Context-dependent dialogue | Narrative coherence |
| **Structural** | Lineation, stanza forms | Acts and scenes | Chapters and paragraphs |
| **Phonological** | Meter, rhyme, stress | Limited but present in dialogue | Minimal emphasis |

## 4. Genre-Specific Stylometric Profiles

Stylometric analysis varies significantly across genres because each literary form encodes meaning through distinct structural, linguistic, and functional constraints. As argued by John Burrows (2002), stylistic signals are not uniform; rather, they are shaped by genre conventions, communicative purpose, and formal restrictions. Consequently, genre-sensitive stylometry enhances analytical precision by aligning feature selection with textual form.

### 4.1 Poetry

Poetry represents one of the most structurally constrained literary genres, particularly in traditional forms such as sonnets, ghazals, and odes. These constraints significantly influence stylometric patterns, requiring specialized analytical approaches. From a stylometric perspective, poetry foregrounds phonological and prosodic features. Meter (e.g., iambic pentameter), rhyme schemes (e.g., ABAB, AABB), and rhythmic regularities serve as key indicators of stylistic identity. Unlike prose, where lexical frequency dominates analysis, poetic stylometry often integrates sound-based patterns, including syllable counts, stress distribution, and phoneme repetition. Logically, this implies that conventional stylometric tools—primarily designed for prose—must be adapted or supplemented with phonological parsing algorithms. Studies demonstrate that features such as line length variation and enjambment frequency can also function as discriminative stylistic markers (Hoover, 2010).

Moreover, computational approaches have begun incorporating n-gram phonetic modeling to capture rhyme and alliteration patterns more accurately. Narratively, poetry can be understood as compressing meaning into patterned language; thus, stylometric analysis must account for both form and aesthetic constraint, not merely word usage.

### 4.2 Drama

Drama differs fundamentally from other genres due to its performative and dialogic nature. Rather than a single narrative voice, dramatic texts consist of multiple interacting voices, making stylometric analysis more complex and multidimensional.

- Stylometrically, drama emphasizes speaker-based variation. Key features include:
- Speaker distribution (frequency and prominence of characters)
- Turn-taking patterns (dialogue sequencing and interaction flow)
- Character-specific idiolects (distinct linguistic styles per character)

These features allow researchers to identify not only the overall authorial style but also how an author differentiates characters linguistically. For instance, function word usage and syntactic variation may differ systematically between protagonists and minor characters. A significant methodological advancement in dramatic stylometry is the use of network analysis, where characters are treated as nodes and interactions as edges. This approach enables the modeling of social and communicative structures within the text, revealing patterns of dominance, marginality, and relational dynamics (Moretti, 2013). Logically, drama requires a shift from text-centered to interaction-centered analysis, where meaning emerges through exchanges rather than isolated sentences. Narratively, this aligns with the theatrical essence of drama as a representation of human interaction rather than descriptive narration.

### 4.3 Fiction

Fiction, particularly the novel, presents the most expansive and structurally flexible genre, characterized by narrative depth, temporal shifts, and multiple discourse modes. This complexity allows for a broad range of stylometric features. In stylometric terms, fiction is often analyzed through:

- Sentence length and syntactic complexity
- Paragraph structure and narrative pacing
- Dialogue vs. description ratios

Unlike poetry, fiction relies less on formal constraints and more on discursive variation, making it well-suited for computational techniques such as topic modeling and discourse analysis. Topic modeling (e.g., Latent Dirichlet

Allocation) identifies thematic patterns across large corpora, while discourse analysis examines narrative voice, focalization, and cohesion. Research by Matthew L. Jockers (2013) highlights how large-scale stylometric analysis of novels can uncover macro-level literary trends, such as shifts in thematic focus or stylistic evolution across time. Additionally, fiction allows for the study of intra-author variation, where an author's style may shift across different works or narrative contexts. Logically, fiction demands multi-layered analysis, integrating lexical, syntactic, and semantic dimensions. Narratively, it reflects the fluidity of storytelling, where style adapts to character perspective, plot development, and thematic intent. Genre-specific stylometric profiling demonstrates that no single set of features is universally applicable. Poetry prioritizes phonological structure, drama foregrounds interactional dynamics, and fiction emphasizes narrative complexity. Therefore, effective stylometric analysis must adopt a genre-sensitive framework, selecting features and methods that align with the inherent properties of each literary form.

## 5. Computational Models for Genre Classification

### 5.1 Statistical Models

Statistical approaches provide a foundation for stylometric analysis. Burrows's Delta measures stylistic distance based on word frequency distributions (Burrows, 2002). PCA and clustering techniques enable visualization of genre groupings.

### 5.2 Machine Learning Models

Machine learning models extend these capabilities by learning patterns from data. Algorithms such as support vector machines and random forests can classify texts based on feature vectors.

### 5.3 Deep Learning Approaches

Transformer-based models such as BERT (Devlin et al., 2019) can capture complex patterns in textual data. However, their lack of interpretability poses challenges for literary analysis.

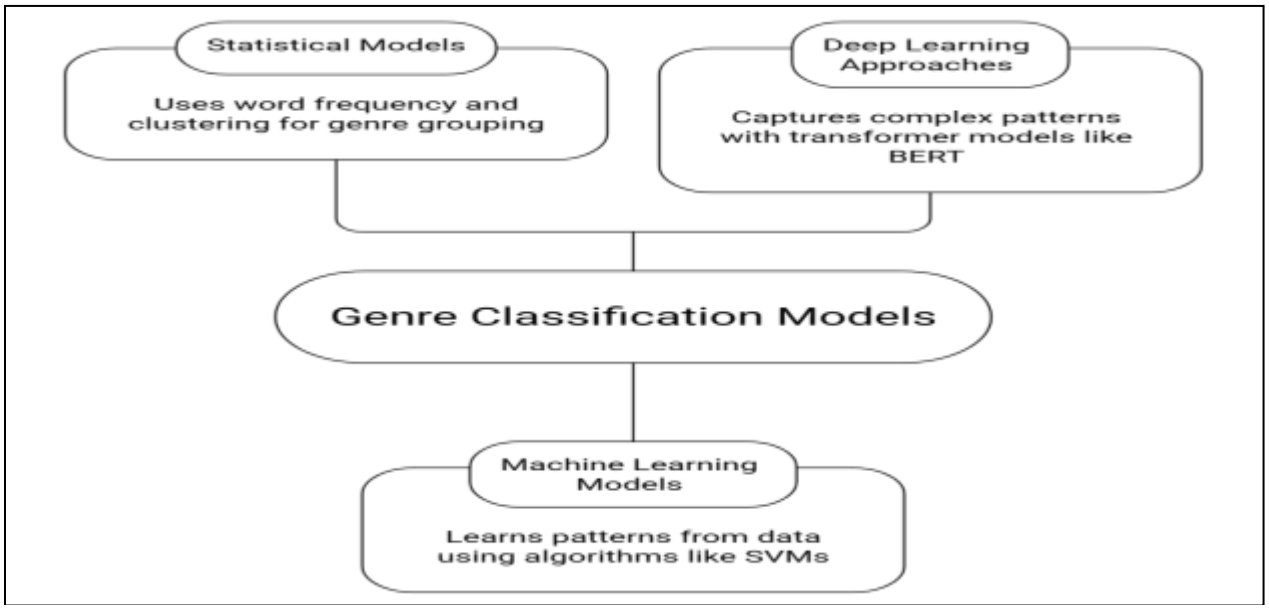


Figure 2. Models of genre analysis

## 6. Description of Stylometric Tools

### 6.1. Stylometric Tools for General Literary Analysis

These tools are widely used in digital humanities for corpus building, stylometry, keyword analysis, and stylistic comparison.

Table 3. Stylometric Tools for general literary analysis

| Tool | Function | Website | Research Use |
|---|---|---|---|
| **AntConc** | Concordance, keyword, frequency, n-grams | https://www.laurenceanthony.net/software/antconc/ | Core tool for lexical and frequency-based stylometry |

| | | | |
|---|---|---|---|
| **Voyant Tools** | Web-based text analysis and visualization | https://voyant-tools.org | Exploratory analysis of corpora and stylistic patterns |
| **Stylo (R package)** | Statistical stylometry and clustering | https://github.com/computationalstylistics/stylo | Authorship and genre classification |
| **WordSmith Tools** | Word lists, concordance, keyword analysis | https://www.lexically.net/wordsmith/ | Traditional corpus-based stylistic comparison |
| **JGAAP** | Authorship attribution system | https://github.com/evllabs/JGAAP | Machine learning-based authorship detection |
| **Signature** | Stylometric frequency and statistical analysis | https://www.philocomp.net/texts/signature.htm | Basic statistical stylistic profiling |
| **TXM** | Corpus + textometry platform | https://textometrie.ens-lyon.fr/ | Advanced corpus linguistic analysis |
| **MALLET** | Topic modeling toolkit | http://mallet.cs.umass.edu/ | Thematic structure analysis |
| **NLTK (Python)** | NLP toolkit | https://www.nltk.org/ | Feature extraction for stylometry |
| **spaCy** | Industrial NLP library | https://spacy.io/ | Syntactic and semantic feature extraction |

General tools primarily support feature extraction and statistical modeling. Systems like AntConc and Voyant Tools are preferred for early-stage analysis because they allow researchers to observe lexical patterns without programming knowledge. More advanced systems like Stylo or JGAAP enable classification and clustering, which are essential for genre and authorship modeling. However, these tools often treat genre as a secondary variable, which limits their ability to directly model genre theory. This creates a gap between computational output and literary interpretation.

### 6.2. Stylometric Tools for Poetry Analysis

Poetry requires tools that capture sound, rhythm, and structural compression, not only word frequency.

Table 4. Stylometric Tools for analysis of poetical text

| Tool | Function | Website | Research Use |
|---|---|---|---|
| **Rhyme Analyzer (Poetry tools)** | Detects rhyme patterns | https://github.com/ | Rhyme scheme detection |
| **Prosodic Toolkit** | Meter and stress analysis | https://github.com/ | Scansion of poetic lines |
| **Voyant Tools** | Word frequency + visualization | https://voyant-tools.org | Theme + lexical density in poetry |
| **AntConc** | Keyword and concordance | https://www.laurenceanthony.net/software/antconc/ | Repetition and lexical compression |
| **Python NLP (NLTK / spaCy)** | Custom feature extraction | https://www.nltk.org | Meter, syllable, and phonetic analysis |
| **Phonemizer tools** | Converts text to phonemes | https://github.com/ | Sound pattern modeling |
| **Stylo (R)** | Cluster poetic styles | https://github.com/computationalstylistics/stylo | Authorial / stylistic clustering in poetry |

Poetry-oriented tools emphasize phonological and rhythmic structure, which standard stylometric systems often ignore. Traditional tools like Voyant and AntConc capture repetition and lexical density but fail to represent meter or sound patterns directly. This creates a methodological gap: poetry requires integration of linguistic stylometry & phonological modeling, which is still underdeveloped in computational literary studies.

### 6.3. Stylometric Tools for Drama Analysis

Drama requires tools that analyze dialogue, speaker interaction, and structural segmentation.

Table 5. Stylometric Tools for analysis of drama

| Tool | Function | Website | Research Use |
|---|---|---|---|
| **AntConc** | Dialogue extraction | https://www.laurenceanthony.net/software/antconc/ | Speaker-based lexical comparison |
| **Voyant Tools** | Text segmentation | https://voyant-tools.org | Scene-level analysis |
| **Gephi** | Network visualization | https://gephi.org | Character interaction networks |

| | | | |
|---|---|---|---|
| **Python NLP (spaCy)** | Speaker tagging | https://spacy.io | Dialogue parsing |
| **Stylo (R)** | Stylometric clustering | https://github.com/computationalstylistics/stylo | Authorial voice detection |
| **TXM Platform** | Corpus + structure analysis | https://textometrie.ens-lyon.fr/ | Structural segmentation of plays |
| **NLTK** | Sentence/dialogue processing | https://www.nltk.org | Turn-taking analysis |

Drama-based analysis depends heavily on interaction structure rather than lexical density alone. Network tools like Gephi allow researchers to model character relationships, which is crucial for understanding dramatic structure. However, most stylometric tools still treat drama as linear text, ignoring its performative and multi-speaker nature, which limits analytical precision.

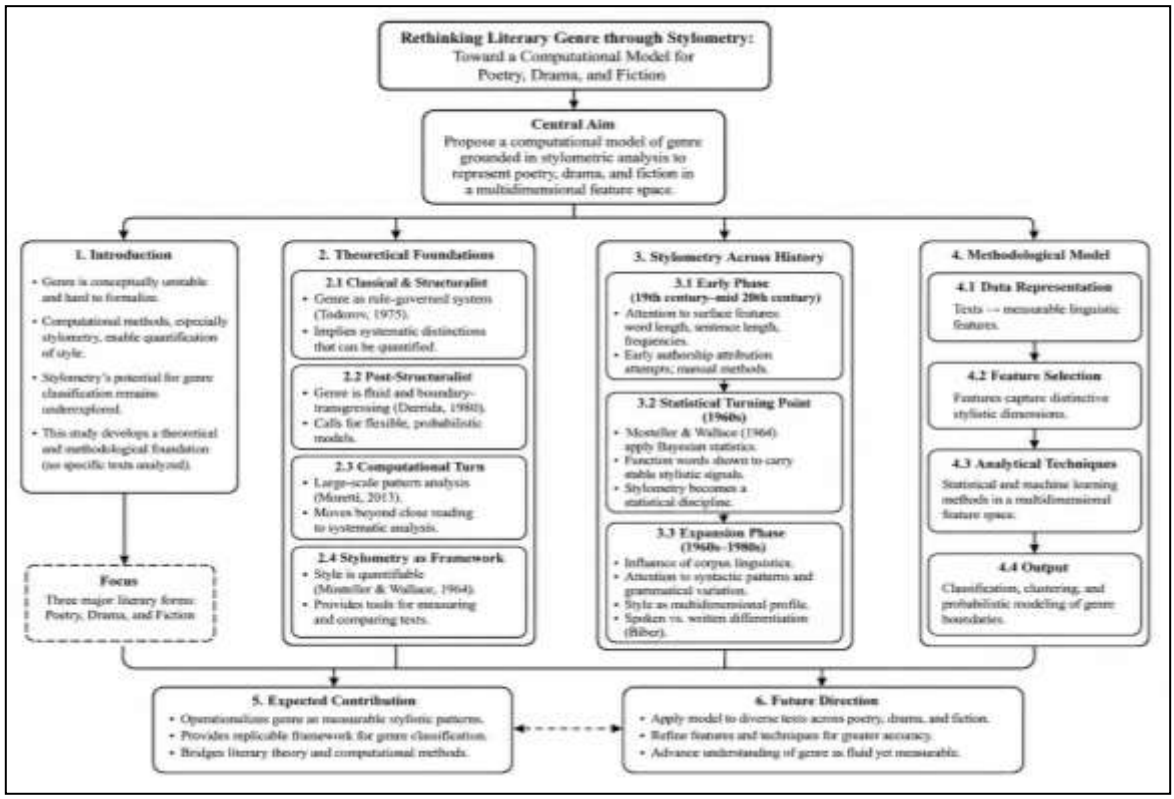

Figure 3. Conceptual framework

### 6.4. Stylometric Tools for Fiction / Prose Analysis

Fiction requires tools focused on narrative flow, discourse structure, and thematic evolution.

Table 6. Stylometric Tools for analysis of prose

| Tool | Function | Website | Research Use |
|---|---|---|---|
| **MALLET** | Topic modeling | http://mallet.cs.umass.edu/ | Theme detection across novels |
| **Voyant Tools** | Narrative visualization | https://voyant-tools.org | Word trends across chapters |
| **AntConc** | Frequency and collocation | https://www.laurenceanthony.net/software/antconc/ | Lexical variation in prose |
| **Stylo (R)** | Genre clustering | https://github.com/computationalstylistics/stylo | Fiction genre classification |
| **spaCy** | Named entity recognition | https://spacy.io | Character tracking |
| **Stanford NLP** | Syntax parsing | https://stanfordnlp.github.io | Narrative structure analysis |

| | | | |
|---|---|---|---|
| **Gensim** | Topic modeling | https://radimrehurek.com/gensim/ | Semantic evolution in novels |

Fiction analysis relies heavily on semantic and narrative-level modeling, especially topic modeling and named entity recognition. Unlike poetry and drama, fiction tools emphasize discourse continuity and thematic development. However, many tools still reduce novels to bag-of-words representations, which remove narrative structure. This limits the ability to fully model storytelling complexity.

## 6. CONCLUSION

The central argument of this study has been that literary genre, long treated as a stable classificatory system, is better understood as a dynamic, multidimensional construct that can be modeled computationally through stylometric analysis. Drawing on theoretical insights from Tzvetan Todorov and Jacques Derrida, the paper has demonstrated that genre is simultaneously structured and fluid—governed by conventions yet constantly transgressed in practice. This dual nature makes traditional rigid classification insufficient, particularly in the context of increasingly hybrid and experimental literary forms. By incorporating the computational paradigm advanced by Franco Moretti, the study reframes genre as a measurable phenomenon grounded in linguistic and stylistic features. Stylometry, as a methodological framework, enables the transformation of qualitative literary characteristics into quantifiable data. This shift does not replace interpretive criticism but complements it by introducing replicability, scalability, and empirical rigor. A key contribution of this study is the articulation of a proposed unified computational model of genre. In this model, texts are represented as vectors in a multidimensional feature space, where each dimension corresponds to a stylometric feature such as lexical frequency, syntactic complexity, or phonological patterning. Genres are not fixed categories but clusters within this space, characterized by probabilistic boundaries rather than absolute distinctions. This approach accommodates the reality of genre hybridity, allowing texts to occupy overlapping regions and exhibit mixed stylistic profiles.

The comparative analysis of poetry, drama, and fiction has further illustrated that genre differentiation emerges from configurations of features rather than isolated traits. Poetry emphasizes phonological and structural constraints, drama foregrounds dialogic interaction and character-specific language, and fiction prioritizes narrative continuity and syntactic complexity. These distinctions validate the feasibility of computational modeling while also highlighting the need for genre-sensitive feature selection. The evaluation of computational models—from statistical methods such as Burrows's Delta to machine learning and deep learning approaches like BERT—reveals a trade-off between interpretability and predictive power. While advanced models achieve high accuracy, their opacity limits their usefulness for theoretical inquiry. This reinforces the importance of hybrid approaches that balance computational efficiency with conceptual clarity.

Finally, the review of stylometric tools demonstrates that the field is supported by a robust ecosystem of software, including Stylo, AntConc, and Voyant Tools. However, no single tool fully captures the complexity of all genres, underscoring the need for integrated workflows and customized pipelines. In sum, this study advances the understanding of genre by bridging literary theory and computational methodology. It proposes a flexible, scalable, and theoretically grounded model that reflects the evolving nature of literary forms in the digital age.

### 7. Recommendations

1. Future research should begin by strengthening the connection between computational methods and established literary theory. Stylometry should not be treated as a purely technical procedure detached from interpretation. Instead, it can function as a testing ground for theoretical claims about genre. Derrida's idea that genre is unstable and always in a process of crossing boundaries can be examined through probabilistic models that allow texts to belong to more than one category at the same time. In this way, computational results can refine rather than replace theoretical arguments about literary form.
2. A central requirement moving forward is the development of more refined feature sets that are sensitive to genre differences. Current models often rely on generic linguistic features, but genre-specific modeling would improve both precision and interpretability. Poetry requires attention to phonological and rhythmic structures such as meter and rhyme, since these shape its identity more strongly than syntax alone. Drama depends heavily on interactional structures, especially speaker shifts and relational networks between characters. Fiction, on the other hand, is better captured through narrative progression and syntactic complexity, including sentence variation and discourse organization. A more systematic alignment between feature design and genre characteristics would strengthen computational modeling.
3. Future work should also move toward hybrid modeling frameworks rather than relying on a single methodological tradition. Statistical methods such as Burrows's Delta remain valuable because they are interpretable and theoretically grounded, while machine learning models such as support vector machines and random forests improve classification

performance. Neural approaches provide additional capacity for capturing high-dimensional patterns in text. When combined carefully, these methods can balance interpretability and predictive strength, allowing models to remain both analytically meaningful and computationally effective.

4. Another important direction concerns the development and integration of tools. At present, stylometric tools are often fragmented and designed for narrow purposes, which limits their usefulness in cross-genre research. There is a need for integrated systems that can handle poetry, drama, and fiction within a single analytical environment. Such systems should also include visualization features that make feature spaces and genre clusters easier to interpret. At the same time, accessibility should be improved so that researchers without advanced technical training can still engage with computational methods. Progress in this area will depend on collaboration between literary scholars and software developers.

5. Research must also expand beyond English-language corpora. Most existing models are trained on English texts, which restrict their applicability. Extending stylometric analysis to multilingual and cross-cultural datasets would allow researchers to test whether current models generalize across linguistic systems or whether genre behaves differently in different cultural contexts. This expansion is essential if computational genre theory is to become globally relevant rather than language-specific.

6. As models become more complex, interpretability remains a critical concern. It is not enough for a system to classify texts accurately; it must also allow researchers to understand why a classification occurs. Techniques from explainable artificial intelligence can help address this issue by making model decisions more transparent. Alongside this, visualization methods that map genre distributions in feature space can provide a clearer conceptual understanding of how texts relate to one another.

7. The quality of results in stylometric research also depends heavily on corpus design. Reliable modeling requires balanced and carefully constructed datasets that represent genres fairly and consistently. Researchers must document how corpora are built, including selection criteria and preprocessing decisions. Ethical considerations are also important, especially when dealing with digitized texts, copyright restrictions, and cultural representation.

8. Finally, the future of stylometric genre analysis depends on sustained interdisciplinary collaboration. Literary studies provide the theoretical grounding, computational linguistics offers methodological tools, and data science contributes modeling expertise. When these fields are combined, it becomes possible to develop more robust and theoretically informed models of genre. Such collaboration will be essential for moving from isolated experiments toward a coherent computational theory of literary form.

TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/



**Multidimensional Stylometric Genre Model (MSGM)**

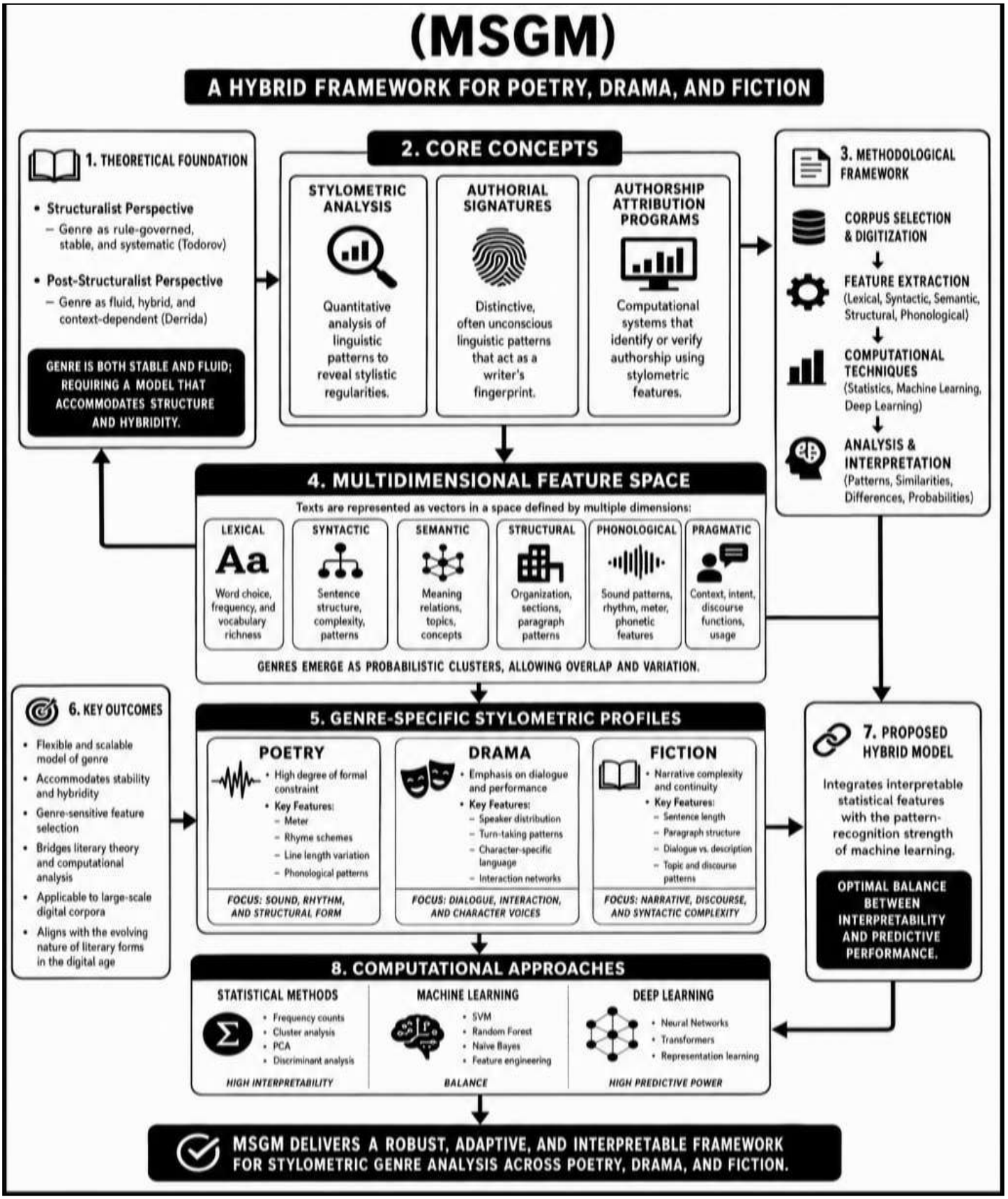


**Figure 4:** Multidimensional Stylometric Genre Model (MSGM)

**MSGM:** A Hybrid Framework for Poetry, Drama, and Fiction

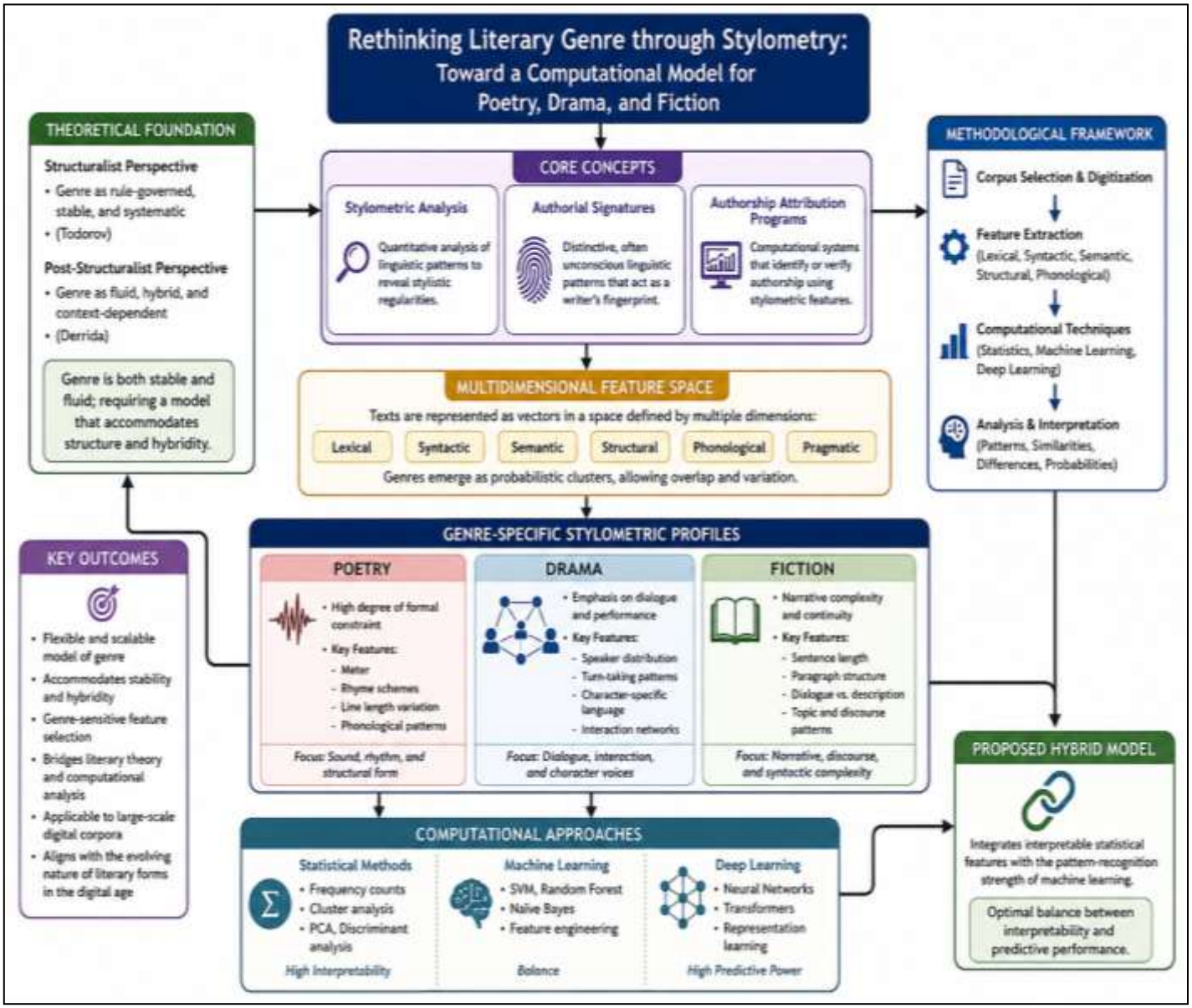


**Figure 5.** MSGM: A Hybrid Framework for Poetry, Drama, and Fiction

## REFERENCES

1. Anthony, L. (2020). *AntConc (Version 3.5.9) [Computer software]*. Waseda University. https://www.laurenceanthony.net/software/antconc/
2. Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024
3. Burrows, J. F. (2002). "Delta": A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267–287. https://doi.org/10.1093/llc/17.3.267
4. Cammarota, V., Bozza, S., Roten, C.-A., & Taroni, F. (2024). Stylometry and forensic science: A literature review. *Forensic Science International: Synergy, 9*, 100481. https://doi.org/10.1016/j.fsisyn.2024.100481
5. Daelemans, W. (2013). Explanation in computational stylometry. *Computational Linguistics, 39*(3), 491–494. https://doi.org/10.1162/COLI_a_00152
6. Derrida, J. (1980). The law of genre. *Critical Inquiry, 7*(1), 55–81. https://doi.org/10.1086/448093
7. 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 (NAACL-HLT)* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

8. Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *The R Journal, 8*(1), 107–121. https://doi.org/10.32614/RJ-2016-007
9. Gorman, R. (2024). Morphosyntactic annotation in literary stylometry. *Information, 15*(4), 211. https://doi.org/10.3390/info15040211
10. Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 251–270. https://doi.org/10.1093/llc/fqm020
11. He, X., Lashkari, A. H., Vombatkere, N., & Sharma, D. P. (2024). Authorship attribution methods, challenges, and future research directions: A comprehensive survey. *Information, 15*(3), 131. https://doi.org/10.3390/info15030131
12. Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing, 13*(3), 111–117. https://doi.org/10.1093/llc/13.3.111
13. Hoover, D. L. (2010). Quantitative analysis and literary studies. In *A companion to digital literary studies*. Blackwell.
14. Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
15. Kestemont, M. (2014). Function words in authorship attribution. *Literary and Linguistic Computing, 29*(2), 171–188. https://doi.org/10.1093/llc/fqt052
16. McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
17. Moretti, F. (2013). *Distant reading*. Verso.
18. Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley.
19. Przystalski, K., Argasiński, J. K., Grabska-Gradzińska, I., & Ochab, J. K. (2025). Stylometry recognizes human and LLM-generated texts in short samples. *arXiv preprint arXiv:2507.00838*.
20. Ramsay, S. (2011). *Reading machines: Toward an algorithmic criticism*. University of Illinois Press.
21. Sharma, N., & Kumar, A. (2024). Deep learning for stylometry and authorship attribution: A review of literature. *International Journal for Research in Applied Science and Engineering Technology, 12*(9), 212–215.
22. Sinclair, S., & Rockwell, G. (2016). *Voyant Tools* [Computer software]. https://voyant-tools.org/
23. Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538–556. https://doi.org/10.1002/asi.21001
24. Todorov, T. (1975). *The fantastic: A structural approach to a literary genre* (R. Howard, Trans.). Cornell University Press.
25. Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. University of Chicago Press.
26. Zamir, M. T., Ayub, M. A., Gul, A., Ahmad, N., & Ahmad, K. (2024). Stylometry analysis of multi-authored documents for authorship and style change detection. *arXiv preprint arXiv:2401.06752*.