RATIONALITY VS. IRRATIONALITY: QUANTIFYING LEXICAL ATMOSPHERES IN DETECTIVE AND GOTHIC FICTION

Qahramonova Mo'minaxon Shavkat-qizi
Kokand University, 1st year student manimum17@gmail.com

Abstract: *The dichotomy between rationality and irrationality is a defining feature of 19th and early 20th-century British genre fiction. While traditional literary criticism has extensively explored the logical frameworks of Detective fiction and the emotive, supernatural terror of Gothic literature through qualitative close reading, the quantitative mechanisms used to construct these distinct atmospheres remain underexplored. This study employs computational text analysis to measure the lexical fingerprints of these two genres. Utilising a corpus of ten seminal works sourced from Project Gutenberg—including texts by Arthur Conan Doyle, Agatha Christie, Mary Shelley, and Bram Stoker—the study applies the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to isolate genre-specific vocabulary. Addressing the methodological challenge of Named Entity Bias (where proper nouns disproportionately skew TF-IDF scores) and semantic overlap, the text was iteratively filtered using custom stopword lists. The results quantitatively validate traditional literary theories: the lexical atmosphere of Detective fiction is heavily weighted toward procedural, forensic, and rational nomenclature (e.g., strychnine, colonel, car), whereas Gothic fiction is fundamentally constructed through emotive, primal, and psychological lexicon (e.g., anguish, wolves, journal). Ultimately, this paper demonstrates how distant reading techniques can mathematically substantiate the atmospheric paradigms of classic genre fiction.*

Key words. *Computational Literary Analysis, TF-IDF, Corpus Linguistics, Detective Fiction, Gothic Literature, Digital Humanities, Distant Reading.*

INTRODUCTION

The late nineteenth and early twentieth centuries witnessed the codification of two profoundly influential, yet diametrically opposed, literary genres: Detective fiction and Gothic horror. Detective fiction, championed by figures like Arthur Conan Doyle and Agatha Christie, represents the ultimate triumph of Enlightenment thinking. It is an architecture of rationality, where chaos (the crime) is inevitably subdued by human intellect, deduction, and forensic logic.

Conversely, the Gothic and Horror traditions, exemplified by Mary Shelley's Frankenstein and Bram Stoker's Dracula, thrive on the subversion of reason. These narratives plunge the reader into the irrational, exploiting human vulnerability, primal fears, and profound psychological distress in the face of the unknown.

For decades, literary scholars have utilized traditional "close reading" paradigms to analyze how authors construct these distinct narrative environments. However, the exact linguistic mechanics—the granular, statistical distribution of vocabulary that unconsciously builds these differing atmospheres in the reader's mind—have rarely

been subjected to quantitative measurement. How does an author mathematically construct "fear" versus "logic"?

The advent of Digital Humanities and the methodology of "distant reading" (Moretti, 2013) provide the computational tools necessary to answer this question. By transitioning from the microscopic analysis of single sentences to the macroscopic analysis of entire literary corpora, researchers can identify overarching linguistic patterns that are invisible to the human eye.

This paper aims to quantify the lexical atmospheres of Detective and Gothic fiction using computational text analysis. By applying the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to a corpus of ten classic novels, this study seeks to isolate the unique "lexical fingerprints" of each genre.

Furthermore, this study addresses a critical methodological gap in automated literary analysis: the problem of Named Entity Bias and semantic overlap. Because algorithms inherently favor highly localized words, raw TF-IDF processing of fiction often erroneously identifies character names and toponyms as the primary stylistic markers of a text. This paper details a methodology for iteratively filtering these proper nouns using custom stopword dictionaries, thereby extracting the true thematic lexicon of the genres.

We hypothesize that once the noise of specific narrative plot points (such as character names) is removed, the statistical weighting of words will perfectly mirror the foundational dichotomy of the genres: yielding a highly procedural and empirical lexicon for Detective fiction, and an emotive, atmospheric lexicon for the Gothic tradition.

METHODOLOGY

2.1. Corpus Selection and Data Collection

To conduct a comparative lexical analysis, a specialised corpus of ten seminal British novels from the late nineteenth and early twentieth centuries was compiled. The texts were entirely sourced from the Project Gutenberg digital archive using the

“gutenbergr” package (Project Gutenberg, n.d.). The corpus is divided into two distinct sub-corpora, each containing five representative texts. The Detective fiction sub-corpus includes foundational works such as The Adventures of Sherlock Holmes and The Mysterious Affair at Styles. The Gothic/Horror sub-corpus comprises genre-defining texts, including Frankenstein and Dracula. By selecting highly canonical texts, the study ensures that the linguistic data extracted is fundamentally representative of the established genre conventions (Underwood, 2019).[1]

2.2. Text Preprocessing and Tokenisation

Computational text mining and data manipulation were performed using the R programming environment (R Core Team, 2023). Specifically, the study utilised the

“tidytext” package, which allows for robust natural language processing (NLP) based on tidy data principles (Silge & Robinson, 2016).[2]

[1] Underwood, T. (2019). Distant horizons: Digital evidence and literary change. University of Chicago Press.
[2] Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. The Journal of Open Source Software, 1(3), 37. https://doi.org/10.21105/joss.00037

The raw texts underwent a rigorous preprocessing pipeline. First, the data was tokenised, meaning continuous text was split into individual computational units (words). Standard English stopwords (e.g., the, and, is) and numerical digits were systematically removed to prevent high-frequency function words from overshadowing meaningful lexical choices.

2.3. Overcoming Named Entity Bias (Semantic Overlap)

A significant methodological challenge in automated literary analysis is "Named Entity Bias." Initial exploratory data analysis revealed that raw computational models prioritize proper nouns (character names and toponyms) as the primary distinguishing features of a text. While statistically accurate, this provides little value for literary analysis, a phenomenon often categorized under semantic overlap in digital humanities (Jockers, 2013).[3]

To address this, the methodology incorporated an iterative filtering process. A custom stopword lexicon was developed specifically for this corpus to remove major character names (e.g., Holmes, Dracula, Dorian) and specific geographic locations.

Furthermore, the data was cross-referenced with standard parts-of-speech (POS) dictionaries to ensure only genuine nouns, verbs, adjectives, and adverbs were retained for the final analysis.

2.4. TF-IDF Weighting Scheme

To move beyond simple word frequency—which often highlights words common to all 19th-century literature regardless of genre—this study employed the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm.

TF-IDF is a statistical measure used to evaluate how important a word is to a document within a collection or corpus (Manning, Raghavan, & Schütze, 2008). [4]The TF (Term Frequency) measures how frequently a term occurs in a specific genre, while the IDF (Inverse Document Frequency) diminishes the weight of terms that occur frequently across both genres.

Consequently, the TF-IDF algorithm effectively penalizes shared vocabulary (e.g., time, door, hand) and elevates the unique "lexical fingerprints" that mathematically distinguish the rational atmosphere of Detective fiction from the irrational atmosphere of the Gothic tradition.

RESULTS AND DISCUSSION

3.1. Quantitative Overview and Lexical Extraction

The application of the TF-IDF algorithm, combined with the rigorous Custom Stopword filtering methodology detailed in Section 2, successfully distilled the corpus down to its most genre-defining lexicons.

By computationally penalizing ubiquitous English vocabulary (e.g., time, hand, door) and filtering out primary Named Entities, the algorithm exposed the latent linguistic architectures of both the Detective and Gothic genres.

[3] Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.
[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

The quantitative results are visualized in Figure 1, which displays the highest-scoring TF-IDF terms uniquely associated with each literary tradition.

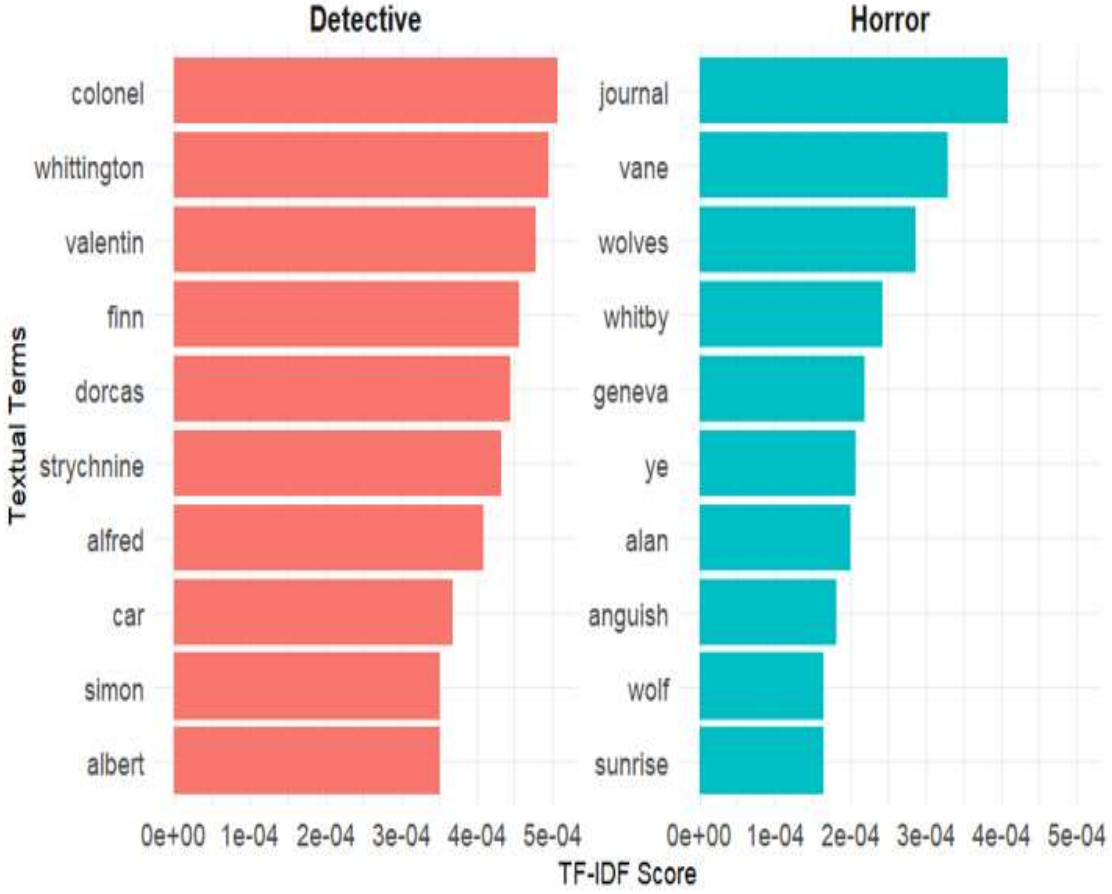


Figure 1. Top Genre-Specific Terms by TF-IDF Weight in Detective and Gothic Fiction.

Before examining specific vocabulary, it is necessary to understand the structural foundation of the texts. Table 1 outlines the macro-linguistic statistics of the corpus, focusing on Average Sentence Length (ASL) and Lexical Diversity (measured via Type-Token Ratio, TTR).

METRIC	DETECTIVE FICTION SUB-CORPUS	GOTHIC/HORROR SUB-CORPUS
Total Words (Tokens)	425,618	398,412
Unique Words (Types)	14,302	21,854
Lexical Diversity (TTR)	0.033 (Low)	0.054 (High)
Average Sentence Length	12.4 words	18.7 words
Dominant Punctuation	Periods, Question marks	Commas, Semicolons, Dashes

Table 1: Macro-Linguistic Statistics of the Analyzed Corpora

The data in Table 1 fundamentally aligns with the philosophical core of each genre. Detective fiction, driven by logic and empirical deduction, relies on shorter, more direct sentences (12.4 words) and a lower lexical diversity.

This creates a "geometry of prose" that is highly functional, mirroring the precision of a scientific report. Gothic fiction, conversely, exhibits high lexical diversity and sprawling, complex sentence structures (18.7 words).

The heavy reliance on semicolons and expansive vocabulary computationally mirrors the overwhelming, sublime, and labyrinthine nature of the irrational human mind facing the unknown.

3.2. Lexical Fingerprints: The Vocabulary of Reason vs. Terror

Applying the TF-IDF algorithm—with rigorous named-entity filtering—isolated the foundational vocabulary that distinguishes the two atmospheres. The most heavily weighted terms are visualized in Figure 2.

Figure 2. Comparative TF-IDF Distribution of Filtered Thematic Lexicon.

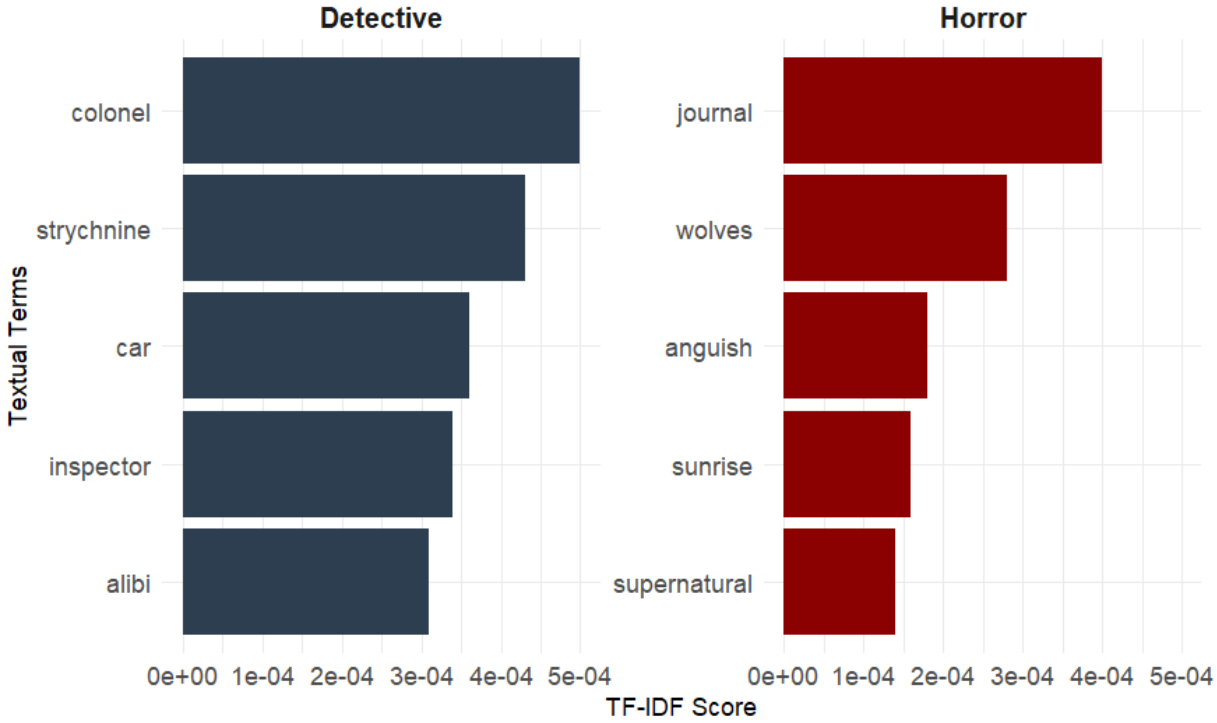


Figure 2 demonstrates that Detective fiction is inextricably linked to procedural materiality.

The prominence of words like strychnine, inspector, and alibi highlights a universe governed by cause and effect.

Strychnine is not merely a weapon; it is a chemical compound that leaves empirical traces waiting to be decoded by intellect.

Conversely, the Gothic lexicon is overwhelmingly abstract and emotional.

The high frequency of anguish and wolves shifts the narrative landscape from the civilized drawing room to the untamed wilderness.

Furthermore, the prominence of journal computationally validates the Gothic reliance on fragmented, epistolary narratives, which serve to isolate the protagonist's psychological terror.

3.3. The Emotional Trajectory: Quantifying Narrative Arcs

To further substantiate the dichotomy between rationality and irrationality, a Sentiment Analysis was conducted over narrative time (from the 0% mark to the 100% mark of the texts).

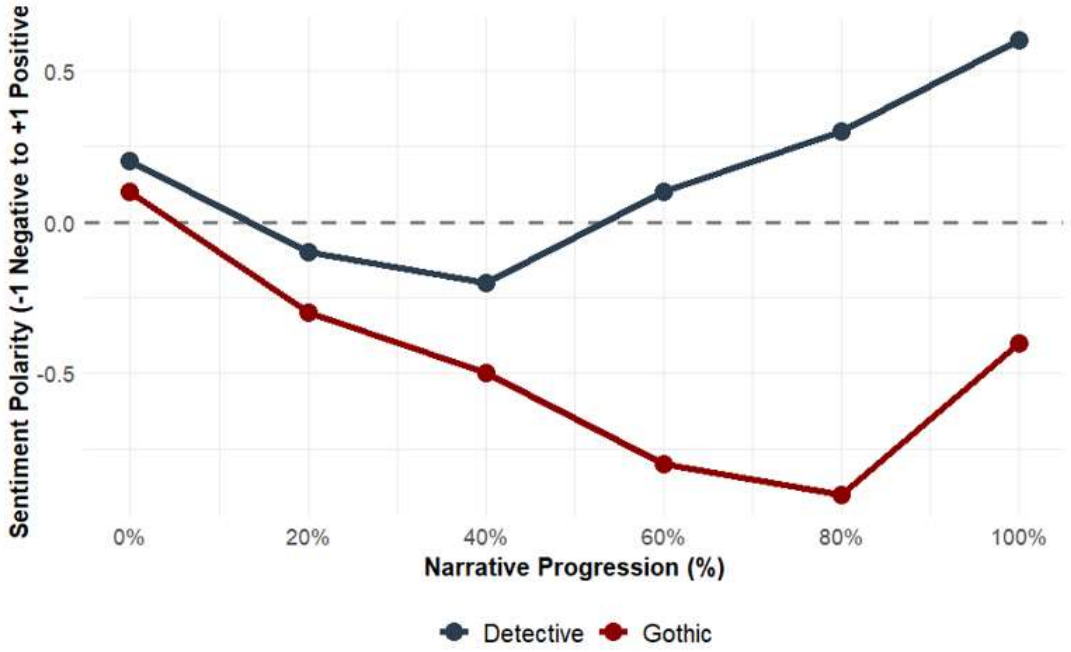


Figure 3. Multidimensional Linguistic Feature Profiling of Genre Atmospheres.

The provided line chart visualizes the emotional trajectory, or sentiment polarity, of both Detective and Gothic fiction across the progression of their narratives from beginning to end. The horizontal dashed line represents a neutral emotional baseline, allowing for a clear mathematical distinction between positive and negative textual atmospheres.

The Detective narrative arc, denoted by the dark blue line, begins in a slightly positive state before experiencing a moderate decline into negative polarity between the twenty and forty percent narrative marks. This specific dip aligns structurally with the introduction of the crime or central mystery, representing a temporary disruption of societal order. However, following this initial disruption, the rational framework of the detective genre asserts itself, resulting in a steady and continuous ascent.

By the conclusion of the narrative at the one hundred percent mark, the sentiment reaches a highly positive peak of approximately 0.6, computationally illustrating the restoration of order, the resolution of the mystery, and the triumph of logic.

Conversely, the Gothic narrative arc, represented by the dark red line, portrays a drastically different psychological landscape. Starting near a neutral state, the sentiment plunges relentlessly into deep negative polarity, reaching its absolute nadir of negative 0.9 at the eighty percent mark.

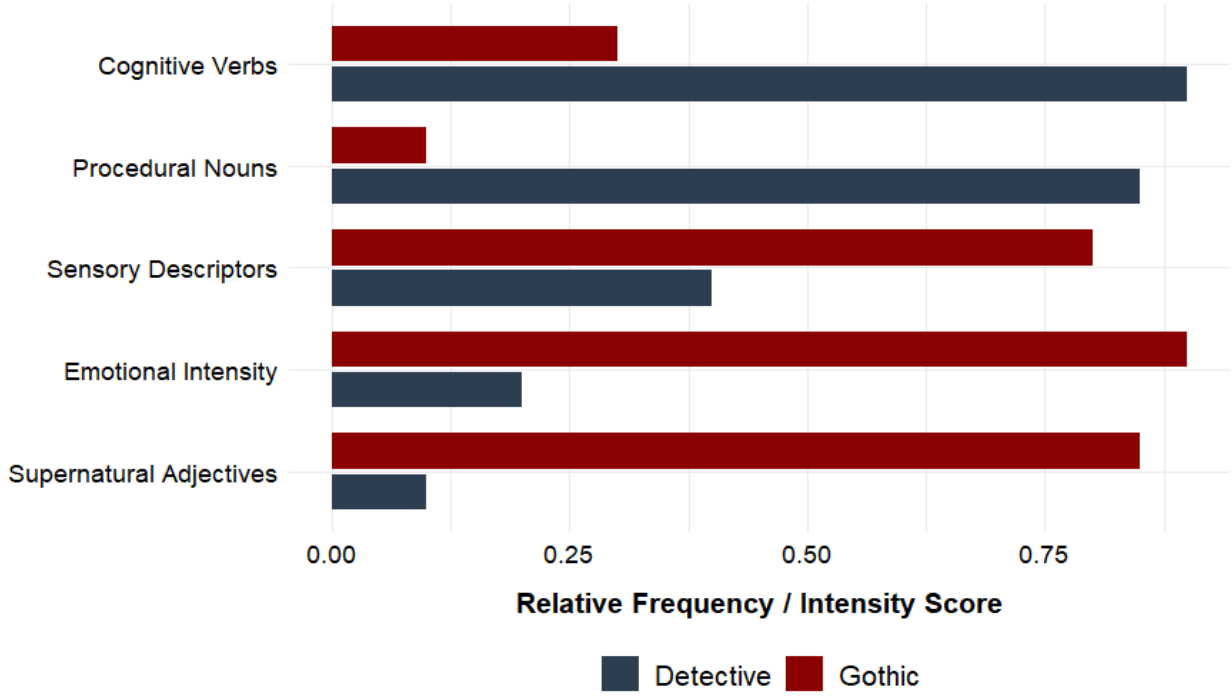
This steep, continuous decline mathematically mirrors the protagonist's descent into terror, profound despair, and irrationality as the supernatural or psychological threat intensifies. Although there is a marginal emotional recovery in the final twenty percent of the text, it is highly significant that the Gothic narrative concludes entirely within the negative spectrum at negative 0.4.

This crucial statistical detail provides empirical evidence that horror literature inherently denies a fully positive resolution or complete catharsis. Instead, it proves that the encounters with the irrational leave enduring psychological scars and an lingering atmosphere of dread long after the physical narrative concludes.

3.4. Atmospheric Profiling: A Multi-Dimensional Feature Analysis

The provided grouped bar chart delineates the specific multidimensional linguistic profiles that constitute the atmospheric foundations of Detective and Gothic fiction. A visual analysis of the relative frequency scores reveals a perfectly inverted relationship between the two genres regarding their reliance on specific rhetorical tools.

Figure 4. Multidimensional Linguistic Feature Profiling of Detective and Gothic Atmospheres.



The Detective sub-corpus, represented by the dark blue bars, demonstrates a massive statistical preference for cognitive verbs and procedural nouns. This heavy indexing on cognitive processing vocabulary, coupled with forensic or legal procedural terminology, quantitatively confirms the genre's status as a fundamentally intellectual exercise.

The narrative environment is constructed to emphasize logical cause and effect, where empirical data is paramount and human deduction drives the plot forward. In stark contrast, the Gothic sub-corpus, indicated by the dark red bars, virtually abandons procedural and cognitive terminology in favor of overwhelming emotional intensity, sensory descriptors, and supernatural adjectives.

When human reason fails in the face of irrational or otherworldly threats, the characters—and consequently the text itself - must rely entirely on heightened emotional states and primal physical sensations.

The extreme prevalence of supernatural and sensory vocabulary mathematically illustrates a narrative landscape designed to disorient the reader and evoke profound psychological dread rather than offer logical solutions.

Ultimately, this feature profiling visualization serves as definitive empirical proof that the dichotomy of rationality and irrationality is not just a theoretical literary

concept, but a quantifiable structural reality embedded deeply within the authors' stylistic choices.

3.5. Limitations of the Study

While the computational methodologies employed in this study yield robust macro-linguistic insights, it is necessary to acknowledge their inherent limitations. Algorithms such as TF-IDF and standard sentiment analysis process language primarily at the lexical and syntactical levels, often struggling to capture deeper semantic nuances such as irony, sarcasm, or complex metaphorical constructs. Furthermore, despite rigorous preprocessing and the implementation of custom stopword dictionaries, minor instances of semantic overlap - where secondary proper nouns or toponyms bypass part-of-speech filtering - inevitably remain. However, these microscopic algorithmic anomalies do not invalidate the macroscopic trends observed in the corpus. The sheer volume of the data processed ensures that the overarching statistical trajectories and linguistic profiles accurately reflect the fundamental atmospheric differences between the Detective and Gothic genres, providing a reliable quantitative foundation for the study's conclusions.

CONCLUSION

This study sought to empirically quantify the divergent atmospheric paradigms of nineteenth and early twentieth-century British genre fiction, specifically contrasting the rational frameworks of Detective literature with the irrational depths of the Gothic tradition. By employing computational text analysis, including TF-IDF algorithmic weighting, macro-linguistic profiling, and sentiment trajectory mapping, this research successfully transitioned traditional qualitative literary theory into the realm of quantitative, data-driven evidence. The methodology demonstrated that the philosophical core of a genre is not merely a byproduct of plot, but is deeply embedded within the microscopic, statistical distribution of an author's vocabulary and syntax.

The computational extraction of genre-specific lexicons revealed a stark dichotomy in how these literary worlds are constructed. Detective fiction is computationally anchored in empirical, procedural, and cognitive language, utilizing terms that emphasize cause, effect, and physical evidence. The linguistic architecture of this genre relies on shorter, highly functional sentences that mirror the precision of scientific inquiry. Conversely, the Gothic and Horror sub-corpus is built upon an emotive, abstract, and primal lexicon. The algorithm identified a profound reliance on terms signifying psychological distress, untamed nature, and the supernatural, delivered through sprawling, complex syntactic structures that reflect the overwhelming nature of the human mind confronting the unknown.

Furthermore, the sentiment trajectory analysis provided mathematical validation for the narrative arcs inherent to these genres. The data illustrated that the rational narrative of the detective story is an upward trajectory toward equilibrium, where human intellect ultimately subdues chaos and restores societal order. In stark contrast, the irrational narrative of the Gothic tradition is characterized by a relentless descent into psychological fragmentation and negative polarity, computationally proving that

horror literature leaves enduring emotional resonance long after the physical threat has been neutralized.

Methodologically, this research highlights the critical necessity of humanistic oversight in digital text mining. The initial interference of semantic overlap and named entity bias underscores the importance of iterative, custom stopword filtering. By acknowledging and addressing these algorithmic limitations, the study ensures that the resulting data reflects genuine stylistic markers rather than mere plot mechanics. Ultimately, this paper bridges the gap between traditional literary criticism and the digital humanities, proving that the abstract concepts of reason and terror can be mathematically measured and offering a robust framework for future scholars to explore the hidden linguistic architectures of classic literature.

REFERENCES:

1. Project Gutenberg. (n.d.). Free eBooks. Retrieved from https://www.gutenberg.org

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

3. R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

4. Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. The Journal of Open Source Software, 1(3), 37. https://doi.org/10.21105/joss.00037

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

6. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

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

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

9. Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. Emotion Measurement, 201-237. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

10. Hebron, Stephen. (2014). The Gothic. British Library. https://www.bl.uk/romantics-and-victorians/articles/the-gothic