The Computational Prometheus: Tracing the Evolution of Science Fiction through Structural Topic Modeling

by Liam Yardley

Abstract

Mary Shelley's novels *Frankenstein* and *The Last Man* occupy a unique space in literary history, hovering between traditions of Gothic Horror and emerging ideas in Science Fiction. Consequently, the classification of Shelley as a Gothic author, or the progenitor of Science Fiction has long been debated. This study investigates this lineage by applying Natural Language Processing techniques to a corpus of 44 novels spanning the Romantic, Victorian, and Modern eras. By combining Principal Component Analysis (PCA) to map stylometric distance and Structural Topic Modelling (STM) to quantify thematic evolution, we aim to locate Shelley's position in the literary latent space. The results indicate a distinct duality: stylistically, Shelley aligns closely with her Gothic contemporaries; however, her work also exhibits significant proximity to early Science Fiction. The analysis reveals that while Shelley's syntactic fingerprint remains rooted in the Gothic tradition, her thematic focus, specifically on exploration and existential dread, distinctly anticipates the narrative structures of later Science Fiction. We conclude that Shelley functions as a crucial evolutionary bridge: she employed the atmospheric tension of the Gothic framework to prototype the speculative inquiries that would define the Science Fiction genre.

Contents

Problem Statement

Mary Shelley's *Frankenstein* (1818) is widely considered the first work of Science Fiction, yet it remains stylistically rooted in the Gothic Horror and Romantic traditions of the early 19th century. Scholars have long debated the generic classification of Shelley's work. While critics like [Paley, 1993] argue that The Last Man operates within a strictly Gothic framework of apocalypse, others suggest a proto-scientific lineage. This debate presents a classification problem: does Shelley's work statistically belong to the genre of Gothic Horror, or does it represent a distinct evolutionary leap toward Science Fiction? This study aims to resolve this tension quantitatively.

This project investigates this lineage by applying Natural Language Processing techniques to a corpus of 19th and 20th-century literature. By utilising Structural Topic Modelling (STM) to quantify thematic shifts and Principal Component Analysis (PCA) to map stylometric distance, I aim to locate Shelley's position in the literary latent space. Specifically, this study seeks to demonstrate that Shelley functions as an evolutionary bridge, exhibiting the stylistic fingerprints of Gothic Horror while pioneering the vocabulary of Science Fiction.

Data Collection and Preprocessing

Data Acquisition

The primary text corpus was acquired from Project Gutenberg using the `gutenbergr` [Robinson, 2025] R package. This ensured reproducible access to the full-text versions of public domain literature. The corpus consists of 44 distinct novels spanning the Romantic (1800–1850), Victorian (1850–1900), and Modern (1900–1950) eras. Information on this package can be found at https://cran.r-project.org/web/packages/gutenbergr/index.html.

Corpus Composition

To contextualise Shelley's work, the corpus was divided into four groups. These were collected from a range of the most popular novels from *Realism/Romance*, Gothic Horror, and Science Fiction from the 18th to early 20th century (Romantic, Victorian, and Modern Eras). The corpus was constructed from 12 *Realism/Romance* novels, 13 Gothic Horror, and 17 Science Fiction. I will be analysing these in relation to Mary Shelley's novels *Frankenstein* and *The Last Man* to see which genre best describes these works, and which novels most align with Shelley's style of writing.

Document Creation

To generate sufficient observations for STM, each novel was segmented into contiguous 100-word chunks. This initially yielded a dataset of $44,168$ documents. However, preliminary analysis revealed a significant class imbalance with the *Realism/Romance* genre disproportionately dominating the corpus, due to their relative length compared to Gothic Horror and Science Fiction works. To prevent the model from overfitting to this dominant class, a random sampling was applied. The number of documents for larger genres was capped at $8,000$ chunks, ensuring a balanced distribution while preserving the vocabulary diversity of the original texts.

Preprocessing

Standard text preprocessing was applied, including the removal of punctuation, numbers, and standard English stop words. [Silge and Robinson, 2017]

To prevent the STM model from clustering based on specific plot details, character names and book-specific terms were removed from the documents. General terms common to Project Gutenberg files (such as *chapter, vol, part, author*) were also excluded.

Finally, to maximise interpretability, Lemmatisation was selected over Stemming. This preserves linguistic coherence, which is extremely useful for literary analysis. Tokens with a length of fewer than 3 characters were also discarded to reduce noise.The full table of novels used can be found below:

Table 1: Target Works (Mary Shelley)

ID	Book Title	Author	Year	Era
84	Frankenstein	Mary Shelley	1818	Romantic
18247	The Last Man	Mary Shelley	1826	Romantic

Table 2: Realism/Romance Corpus

ID	Book Title	Author	Year	Era
161	Sense and Sensibility	Jane Austen	1811	Romantic
1342	Pride and Prejudice	Jane Austen	1813	Romantic
158	Emma	Jane Austen	1815	Romantic
730	Oliver Twist	Charles Dickens	1838	Victorian
1260	Jane Eyre	Charlotte Brontë	1847	Romantic
768	Wuthering Heights	Emily Brontë	1847	Romantic
98	A Tale of Two Cities	Charles Dickens	1859	Victorian
1400	Great Expectations	Charles Dickens	1861	Victorian
514	Little Women	Louisa May Alcott	1868	Victorian
107	Far from the Madding Crowd	Thomas Hardy	1874	Victorian
1399	Anna Karenina	Leo Tolstoy	1878	Victorian
110	Tess of the d'Urbervilles	Thomas Hardy	1891	Victorian

Table 3: Gothic Horror Corpus

ID	Book Title	Author	Year	Era
696	The Castle of Otranto	Horace Walpole	1764	Romantic
3268	The Mysteries of Udolpho	Ann Radcliffe	1794	Romantic
6087	The Vampyre	John Polidori	1819	Romantic
41	The Legend of Sleepy Hollow	Washington Irving	1820	Romantic
2610	The Hunchback of Notre-Dame	Victor Hugo	1831	Romantic
583	The Woman in White	Wilkie Collins	1859	Victorian
10007	Carmilla	J. Sheridan Le Fanu	1872	Victorian
43	Dr. Jekyll and Mr. Hyde	Robert Louis Stevenson	1886	Victorian
174	The Picture of Dorian Gray	Oscar Wilde	1890	Victorian
345	Dracula	Bram Stoker	1897	Victorian
3781	The Jewel of Seven Stars	Bram Stoker	1903	Victorian
175	The Phantom of the Opera	Gaston Leroux	1910	Victorian
1188	The Lair of the White Worm	Bram Stoker	1911	Victorian

Table 4: Science Fiction Corpus

ID	Book Title	Author	Year	Era
18857	Journey to the Center of the Earth	Jules Verne	1864	Victorian
164	20,000 Leagues Under the Sea	Jules Verne	1870	Victorian
35	The Time Machine	H.G. Wells	1895	Victorian
159	The Island of Doctor Moreau	H.G. Wells	1896	Victorian
5230	The Invisible Man	H.G. Wells	1897	Victorian
36	The War of the Worlds	H.G. Wells	1898	Victorian
1013	The First Men in the Moon	H.G. Wells	1901	Victorian
139	The Lost World	Arthur Conan Doyle	1912	Victorian
8748	A Princess of Mars	Edgar Rice Burroughs	1912	Victorian
21970	The Scarlet Plague	Jack London	1912	Modern
126	The Poison Belt	Arthur Conan Doyle	1913	Victorian
32	Herland	Charlotte Perkins Gilman	1915's	Modern
32530	Armageddon 2419 A.D.	Philip Francis Nowlan	1928	Modern
20782	Triplanetary	E.E. "Doc" Smith	1934	Modern
32032	Second Variety	Philip K. Dick	1953	Modern
32154	The Variable Man	Philip K. Dick	1953	Modern
16921	Plague Ship	Andre Norton	1956	Modern

Installation and Configuration

This project was created using RStudio 2025.09.2 Build 418, which can be downloaded at https://posit.co/downloads/. The analysis was run using R version 4.5.1, though any version greater than 4.0 should suffice.

To install any missing packages required to run this code, use the following script:

```
required_packages = c(
  "tidyverse",    # Data manipulation and plotting (includes ggplot2, dplyr)
  "gutenbergr",   # Downloading books from Project Gutenberg
  "tidytext",     # Tidy text mining principles
  "textstem",     # Lemmatisation tools
  "stm",          # Structural Topic Modeling
  "tm",           # Text Mining infrastructure
  "SnowballC",    # Stemming algorithms
  "textmineR",    # Topic modeling evaluation
  "wordcloud",    # Word cloud visualisation
  "stringi"       # String processing
)

# Install missing packages automatically
new_packages = required_packages[!(required_packages %in% installed.packages()[,"Package"])]
if(length(new_packages)) install.packages(new_packages)

# --- Core Data Science Stack ---
library(tidyverse)   # Loads ggplot2, dplyr, tidyr, readr, etc.
library(stringi)     # String manipulation

# --- NLP & Text Mining Stack ---
library(gutenbergr)  # Data Source
library(tidytext)    # Tokenisation
library(textstem)    # Lemmatisation
library(tm)          # Corpus handling
library(SnowballC)   # Stemming

# --- Modeling & Visualisation Stack ---
library(stm)         # The primary modeling package
library(textmineR)   # Evaluation metrics
library(wordcloud)   # Plotting word clouds
```

A known issue with the gutenbergr package is the stability of the default download mirror. To ensure reliable data acquisition, a stable mirror must be explicitly defined:

```
    my_mirror = "http://mirrors.xmission.com/gutenberg/"
```

All of this is included in the preamble of the code.

Computational Requirements

Due to the volume of text data, the Structural Topic Model (STM) requires significant computational resources. Note that the stm package primarily utilises the CPU and RAM, rather than GPU acceleration. The specifications of the machine used for this analysis are listed below:

Processor: 12th Gen Intel(R) Core(TM) i7-12700K (3.61 GHz)
Installed RAM: 64.0 GB
System type: 64-bit operating system, x64-based processor
GPU: NVIDIA GeForce RTX 3070Ti 8GB

Most standard hardware is capable of running this analysis, though execution time will vary. On the specifications above, the full pipeline completed in under 30 minutes, the main time consuming element being searchK function and the running of the stm. The total disk space required for the saved models and data objects was approximately 1.5 GB - 2 GB. To ensure reproducibility of the probabilistic components (such as spectral initialisation), a random seed of 42 was used throughout.

Reproducibility

This codebase is designed to be adaptable to any texts selected. To replicate this study with a different corpus, one simply needs to adapt the vector of Project Gutenberg IDs. Note that the document downsampling step (used here to balance the genres) is dependent on the size of the corpus; if a smaller dataset is used, this step may need to be adjusted or omitted. The remainder of the pipeline will automatically adapt to process the selected documents for analysis. Variables would need to be changed for specific analysis based on the questions posed, and the data used.

To ensure the absolute stability of the results presented, the Structural Topic Model (STM) was trained once using a fixed random seed (42) and the resulting model object was saved for analysis. All subsequent visualisations and interpretations were generated from this persisted model state, eliminating the risk of stochastic drift during the reporting phase.

Results

Stylometric Analysis

To establish a baseline for the similarity in style between Shelley and other authors, we began our analysis with stylometry before progressing to thematic content. While Structural Topic Modelling (STM) provides an in-depth view of what authors write about, this preliminary analysis focuses on how they write. This necessitated a distinct preprocessing pipeline; unlike topic modelling, which removes stop words to isolate content, stylometry relies on the relative frequency of these common function words to detect the authors' fingerprints within their works.

Therefore, by utilising the original text and considering the relative frequency of the most commonly used words in each book, we can effectively measure stylistic similarity.

Era is inherently a factor here, so we would expect Shelley to be near other Romantics or *Gothic* writers. However, precisely where she sits in these clusters can help us determine her unique writing style and which genre she aligns with most closely.

The way we are approaching this is with Principal Component Analysis (PCA). One purpose of this is to take high-dimensional objects and project them onto a lower-dimensional space for visualisation.

For this data, we have 44 novels, and we have chosen to consider the top 150 most frequent words; therefore, the data initially exists in a 44×150 space. These values were calculated by considering the frequency each word was used in each text (per $1,000$ words) to normalise for the varying lengths of the novels. Consequently, our corpus is represented by a matrix $X \in \mathbb{R}^{44\times150}$, with each entry x_{ij} representing the relative frequency of word j in the i-th novel.

We reduced the dimensions by identifying the Principal Components. These are linear combinations of the initial features (words) that correspond to the directions that maximise the variance in the data. This technique maintains the information and distinctiveness of the dataset while combining correlated vectors. Our approach in R uses Singular Value Decomposition (SVD) to do this. This decomposes our matrix into a product of its document scores $U \in \mathbb{R}^{44\times44}$, a diagonal matrix of singular values, σ_i, representing the magnitude of the variance captured by each component $\Sigma \in \mathbb{R}^{44\times150}$ and the principal directions of word features $V \in \mathbb{R}^{150\times150}$. These are ordered largest to smallest. Our original corpus can then be represented by

$$X = U\Sigma V^T$$

We can then reduce the dimensionality of our space by considering only the components which represent the highest variance. For example, to represent our data on a 2D space, we select the first two principal components (corresponding to the two largest singular values) and compute the projection:

$$Z = XV_2$$

The result is a low-dimensional representation that allows us to visualise the stylistic similarity between our different works on a 2D plane, as seen below:

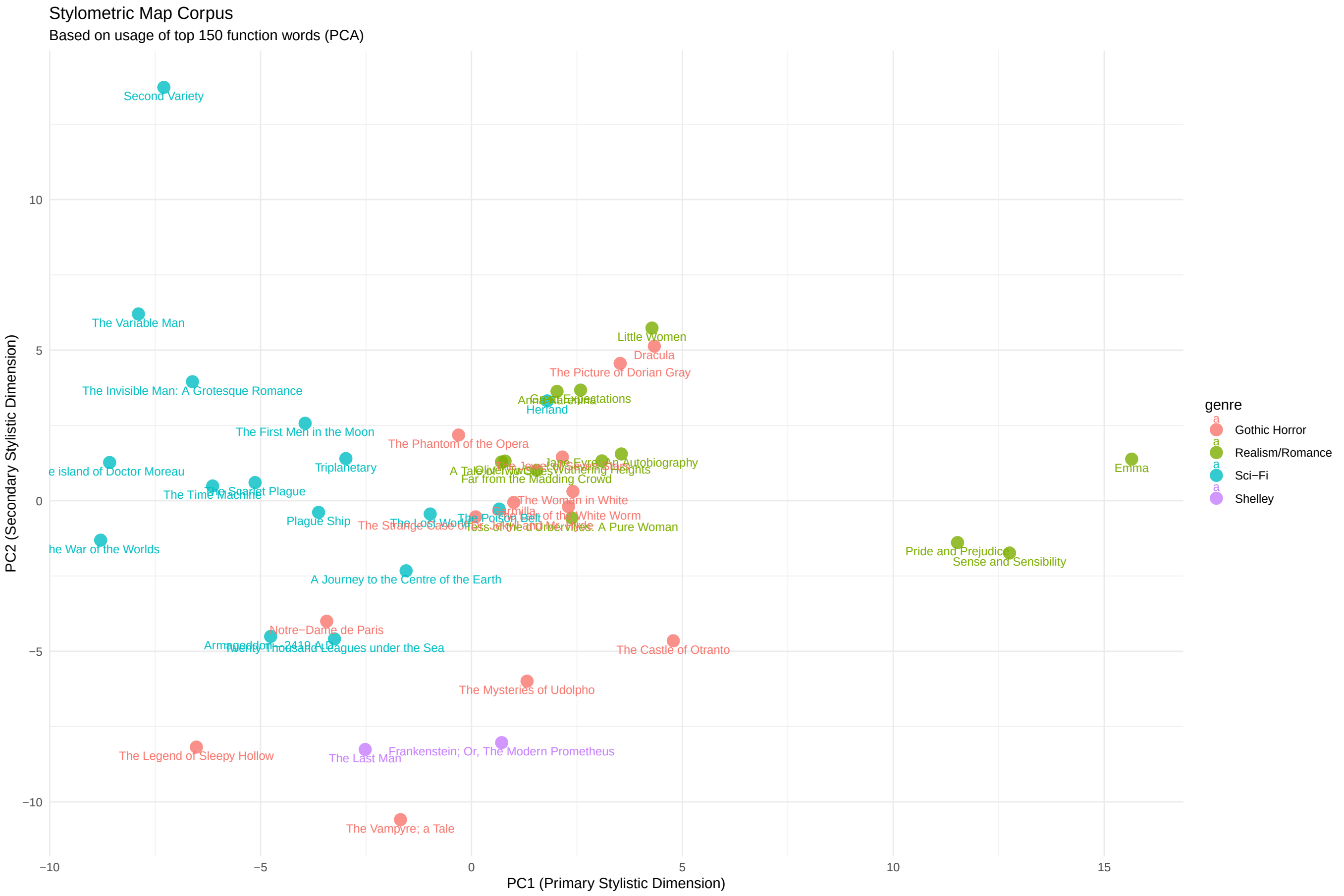


Figure 1: Stylometric Map of Corpus

As you can see from this representation, figure 1, Shelley occupies a distinct position in the latent space, separated significantly from her Romantic contemporaries. This suggests that she is syntactically divergent from these authors and this genre, but most importantly from the main writing at the time. What is interesting is her placements amongst the Gothic Horror novels. The novels closest in style are the early novels of this genre, published before Shelley's *Frankenstein* and *The Last Man*. Therefore, these novels may have been an influence on Shelley, or maybe a combination of the era, and the genre has led to stylistic similarities. Other novels in the neighbourhood of Shelley are the early words of Science Fiction, '20,000 Leagues Under the Sea' and 'A Journey to the Centre of the Earth', both by Jules Verne. These were written later than Shelley's works, but whether Shelley influenced Verne, or this is again due to era or conventions, cannot fully be determined here. However, interestingly, another novel similar stylistically is 'Armageddon 2419 A.D.' by Nowlan, which was written in 1928. This is a Modern book, published over a century after Shelley's works. Therefore, era cannot be the reason for the similarities, reading the plot of the novel, this is also very different to *Frankenstein* and *The Last Man*. This leaves some interesting questions:

- Why this novel is so similar to Shelley's works?
- Was Nowlan influenced by Shelley's style?
- Does this suggest a Science Fiction style in Shelley's work?
- Does Nowlan's work have *Gothic* influences to explain this similarity?
- Or is this just an erroneous result from my model?

Based on research regarding this novel, it seems that if follows the same style of melodramatic narrative common to Romantic and *Gothic* works, and due to this style is considered *Pulp Sci-Fi*, which does take influence from *Gothic* works.

To further explore the relationships observed in the PCA projection, I calculated the cosine similarity between the 150-dimensional vectors representing the stylistic fingerprints of the corpus. This metric provides a direct measure of syntactic distance, mitigating the information loss inherent in compressing high-dimensional data into a two-dimensional plot.

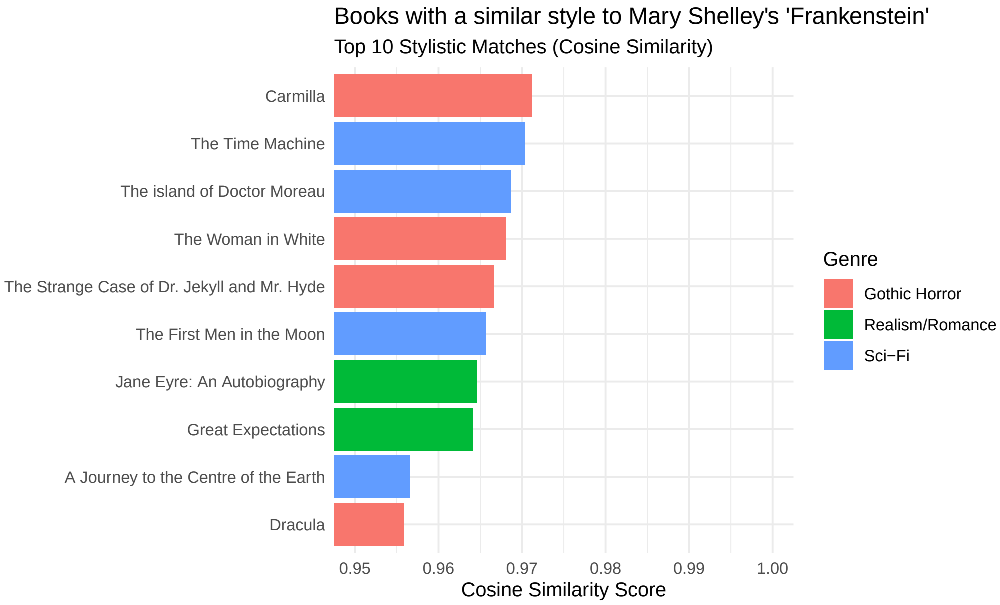


Figure 2: Top 10 Stylistically Similar Works to 'Frankenstein'

The analysis of *Frankenstein* reveals a distinct hybridity. While the novel shares stylistic DNA with works from all three genres, its nearest neighbors are overwhelmingly dominated by Gothic Horror and Science Fiction. Crucially, many of the high-ranking Science Fiction matches (such as those by Wells and Verne) were published decades later. This strongly suggests that Shelley's unique style did not merely reflect the Gothic trends of her time but actively established the stylistic conventions that later Science Fiction authors would adopt. If we consider the top 10 nearest neighbors as a classifier, the text sits precisely on the border: it contains an equal distribution $(N = 4)$ of texts from both Gothic Horror and Science Fiction, reinforcing its status as a transitional text.

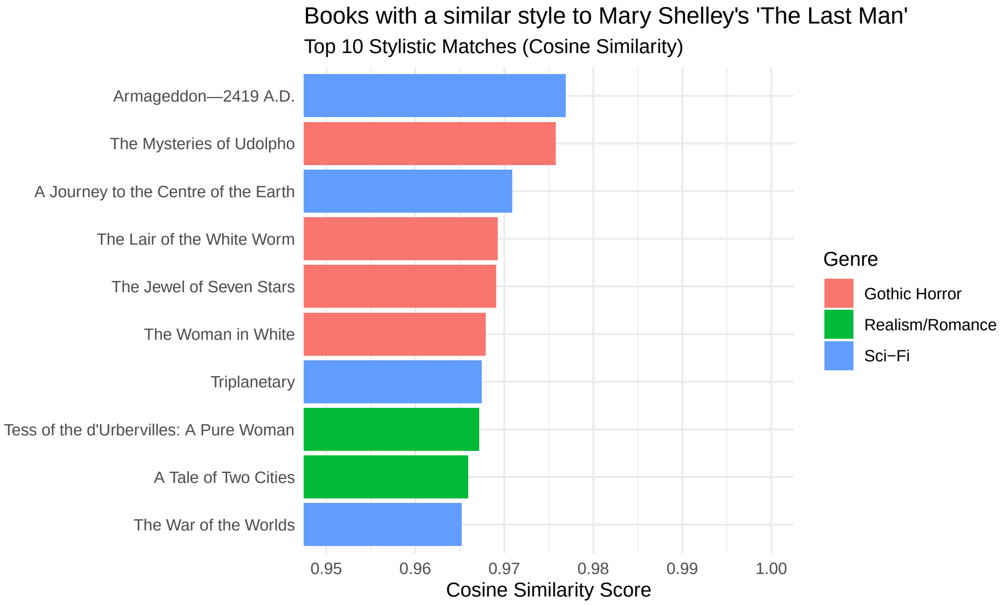


Figure 3: Top 10 Stylistically Similar Works to 'The Last Man'

In the case of *The Last Man*, the results present a compelling paradox. Despite its content being foundational to the Science Fiction genre (apocalyptic plague, futuristic setting), its writing style remains firmly rooted in the *Gothic* tradition. Similar to *Frankenstein*, the top matches are evenly split between *Gothic* and Science Fiction, further supporting the hypothesis of Shelley as a bridge between the genres.

Notably, the text shows a higher stylistic affinity with later "Space Opera" narratives like *Triplanetary* (1934) than with thematically similar apocalyptic texts such as *The Scarlet Plague* or *Plague Ship*. This unexpected result suggests that while Shelley's vocabulary may differ from the gritty realism of later apocalyptic fiction, her syntactic structure shares the high-stakes, epic register found in mid-20th-century speculative fiction.

Conclusion on Stylometry

These findings challenge the traditional categorisation of Shelley as a pure Romantic. Her syntactic fingerprint diverges significantly from the *Realism/Romance* authors of her era, such as Austen and Dickens. Instead, she aligns with the early pioneers of Gothic Horror and the Victorian progenitors of Science Fiction. This suggests that Shelley functioned as a stylistic bridge, employing the emotive, atmospheric prose of the Gothic tradition to pioneer the thematic concerns of the future.

This stylistic divergence is notable given Shelley's social circle, which included prominent Romantic poets such as Percy Bysshe Shelley and Lord Byron. While these figures experimented with Gothic themes, Mary Shelley's work exhibits a unique structural persistence that resonates with genres that would not fully emerge for another century.

However, stylometry measures only the structure and syntax of the text, leaving the specific narrative content unexamined. To address this, the following section applies Structural Topic Modelling (STM) to analyze the thematic composition of the works, comparing the specific topics discussed in *Frankenstein* and *The Last Man* against the broader corpus.

STM Analysis

We utilised the Structural Topic Model as defined by [Roberts et al., 2019].

Mathematical Framework: Structural Topic Modeling

While Latent Dirichlet Allocation (LDA) assumes that topic prevalence is drawn from a fixed Dirichlet distribution common to all documents, the Structural Topic Model (STM) introduces a covariate-dependent prior. This allows the model to estimate how metadata (such as *Genre* or *Era*) influences the probability of a topic appearing. [Lebryk, 2021] [Kurochkin, 2025]

For each document d in the corpus D:

1. **Topic Prevalence (θ_d)**: Given K topics, for each document $d \in \{1, ..., D\}$, the topic proportion vector θ_d is drawn from a Logistic Normal Distribution. This is conditional on:

   - the vector of document covariates, X_d (in this study, *Genre* and *Era*),
   - the matrix of coefficients mapping covariates to topics, $\gamma \in \mathbb{R}^{p\times(K-1)}$.
   - the covariance matrix capturing correlations between topics, $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$, which states the liklihood of a topic appearing given another topic is present

   $$\theta_d | X_d\gamma, \Sigma \sim \text{LogisticNormal}(\mu = X_d\gamma, \Sigma)$$

2. **Topic Content ($\beta_{d,k}$)**: The distribution of words over the vocabulary V for a specific topic k in document d is denoted as $\beta_{d,k}$. In STM, this is modeled using a log-linear framework that allows word usage to vary by content covariates Y_d. Given:

   - m_v: The baseline log-frequency of word v in the corpus, or how common the word is in the corpus.
   - $\kappa_v^{(k)}$: The topic-specific deviation (how much topic k uses word v).
   - $\kappa_v^{(Y_d)}$: The covariate-specific deviation (how much metadata Y_d (such as Author, or era) uses word v).
   - $\kappa_v^{(Y_d,k)}$: The interaction effect (how the vocabulary for topic k changes specifically for group Y_d).

   Then,
   $$\beta_{d,k,v} \propto \exp(m_v + \kappa_v^{(k)} + \kappa_v^{(Y_d)} + \kappa_v^{(Y_d,k)}) \tag{1}$$

3. **Word Generation**: For each word n in document d:

   - A specific topic assignment $z_{d,n}$ is drawn from the document's topic distribution:
     $$z_{d,n} \sim \text{Multinomial}(\theta_d)$$
   - The observed word $w_{d,n}$ is drawn from the chosen topic's vocabulary distribution β_k:
     $$w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$$

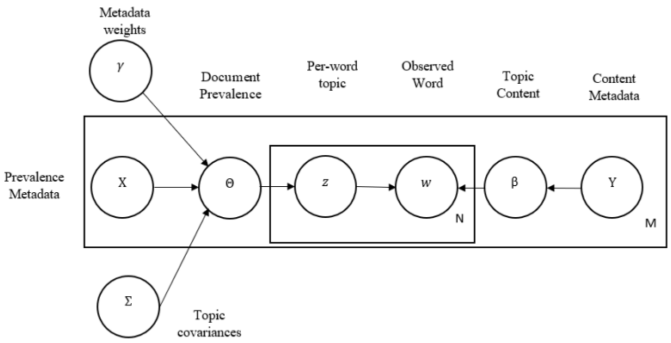


Figure 4: STM Model Diagram [Lebryk, 2021]

By incorporating the covariate X_d into the prior μ, the model explicitly estimates how the probability of discussing a topic shifts as a function of the author's Genre or Era.

Topic Definition via Frequency and Exclusivity

The primary output of the Structural Topic Model (STM) analysis is a set of word distributions defining latent topics. By examining both the highest probability words (frequency) and the Frequency-Exclusivity (FREX) metric, each topic was synthesised into a coherent thematic label representing a distinct concept within the corpus.

Figure 5 visualises the highest probability terms for each topic. This visualisation served as the initial basis for interpretation, which was further refined by analyzing the most exclusive terms to resolve thematic ambiguity.

The high interpretability of these results underscores the critical impact of the preprocessing pipeline. As noted in the Data Collection section, preliminary model iterations exhibited significant noise due to the inclusion of character names and procedural artifacts (e.g., *chapter, gutenberg*). This frequently resulted in "book-binding," where topics clustered around specific novels rather than generalised themes. The rigorous exclusion of these terms has yielded a clean, semantically distinct model where topics represent broader literary motifs rather than specific narrative instances.

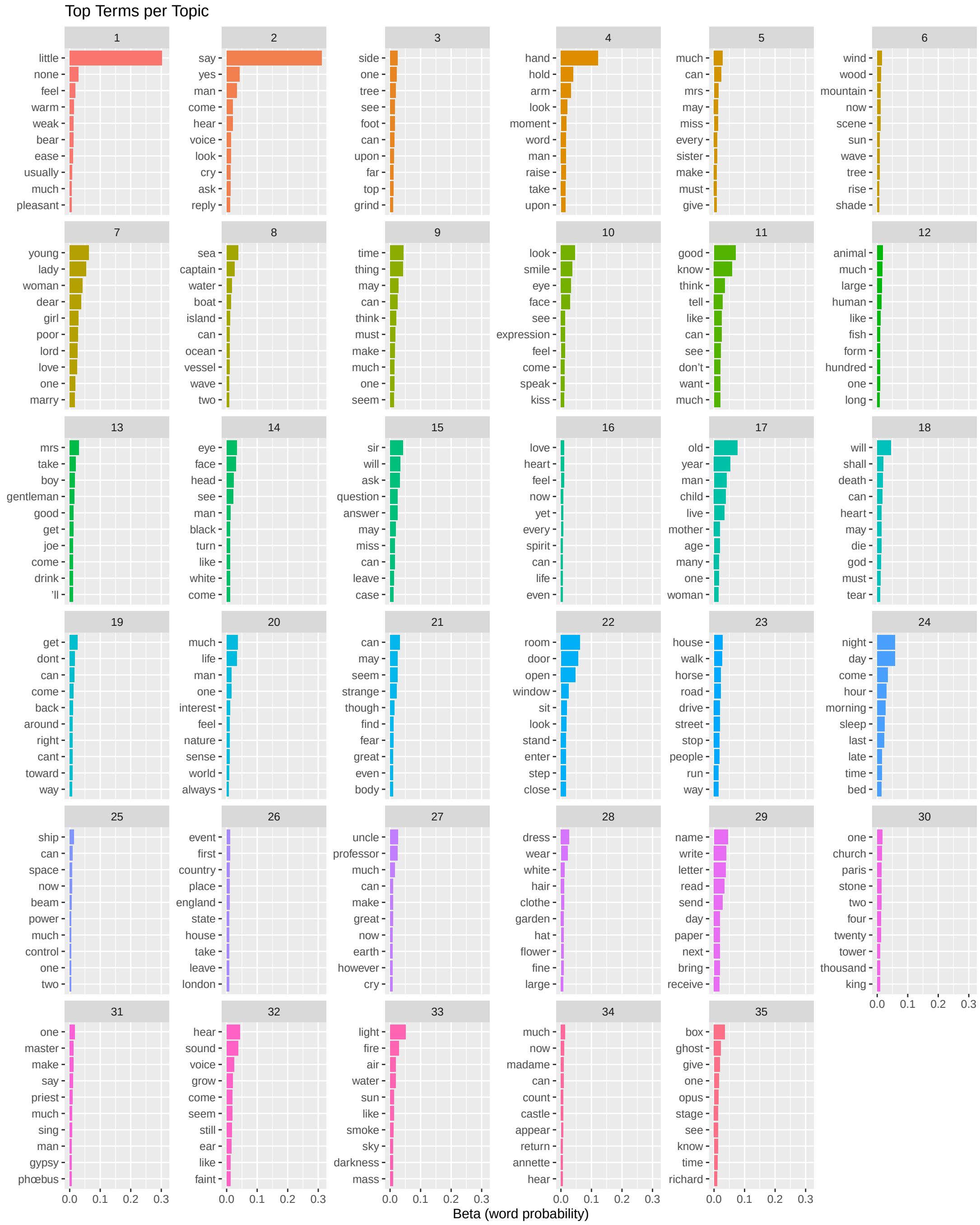


Figure 5: STM Model: Most frequent terms per topic

Code adapted from [Silge, 2018b].

Table 5: Topic Definitions: Probability, Exclusivity (generated from STM model in R) and Interpretation

Topic ID	Highest Probability Words	Most Exclusive (FREX) Words	Interpretation
1	little, none, feel, warm, weak	little, none, weak, ease, usually	Fragility
2	say, yes, man, come, hear	yes, say, kemp, marvel, ’re	The Invisible Man (Novel)
3	side, one, tree, see, foot	edge, top, bush, narrow, tree	Terrain
4	hand, hold, arm, look, moment	hand, hold, arm, raise, finger	Romance/Support
5	much, can, mrs, may, miss	harriet, weston, sister, acquaintance, colonel	Females
6	wind, wood, mountain, now, scene	mountain, shade, wood, valley, wind	Nature
7	young, lady, woman, dear, girl	lady, young, lord, marry, dear	Young Females and Marriage
8	sea, captain, water, boat, island	boat, voyage, sea, island, ocean	Naval Exploration
9	time, thing, may, can, think	thing, cavor, moon, sphere, clear	Time and Early Space Travel
10	look, smile, eye, face, see	smile, alexandrovitch, arkadyevitch, kiss, expression	Romance and Intimacy
11	good, know, think, tell, like	don’t, want, know, ’ve, can’t	Opinions
12	animal, much, large, human, like	animal, fish, specimen, inch, million	Nature and Animals
13	mrs, take, boy, gentleman, good	joe, tea, bottle, jew, dinner	Males
14	eye, face, head, see, man	mouth, stare, neck, scream, leg	Male features
15	sir, will, ask, question, answer	sir, question, answer, case, ask	Questioning
16	love, heart, feel, now, yet	beloved, clara, bestow, sympathy, affection	Sentimental Love
17	old, year, man, child, live	year, ago, old, age, child	Age (Young and Old)
18	will, shall, death, can, heart	thou, thy, miserable, heaven, god	Death and the Soul
19	get, dont, can, come, back	cant, dont, hes, maybe, well	Modern Dialogue (Contractions)
20	much, life, man, one, interest	self, life, picture, principle, nature	Existentialism
21	can, may, seem, strange, thought	tomb, belief, margaret, whilst, mystery	Mystery and Discovery
22	room, door, open, window, sit	door, room, window, hall, open	Home/House
23	house, walk, horse, road, drive	horse, carriage, road, street, drive	Travel (by land)
24	night, day, come, hour, morning	sleep, night, morning, wake, hour	Night
25	ship, can, space, now, beam	nevian, rycke, salariki, rocket, wilma	Space Travel
26	event, first, country, place, england	england, event, necessity, possession, narrative	England
27	uncle, professor, much, can, make	uncle, professor, raft, phenomenon, volcano	Professors/Experts
28	dress, wear, white, hair, clothe	wear, dress, silk, gown, milk	Garden/Weddings
29	name, write, letter, read, send	letter, write, read, paper, name	Communication
30	one, church, paris, stone, two	pound, saint, paris, architecture, church	Buildings
31	one, master, make, say, priest	gypsy, phœbus, archdeacon, priest, claude	Master/Authority
32	hear, sound, voice, grow, come	sound, faint, ear, creep, echo	Eery Suspense
33	light, fire, air, water, sun	smoke, heat, flame, fire, light	Elements
34	much, now, madame, can, count	annette, castle, ludovico, madame, signor	Gothic Settings (Castles/Counts)
35	box, ghost, give, one, opus	opus, ghost, richard, box, moncharmin	Ghosts

The generated topics exhibit high semantic coherence, with distinct thematic identities. While one topic retains strong ties to a specific novel, a common challenge in literary text mining, the majority represent generalised motifs applicable across the corpus.

The subsequent phase of analysis involves aggregating these topics into broader thematic clusters to identify genre-specific or era-specific patterns. For instance, syntactic markers such as *Contractions* serve as clear indicators of the Modern era; this temporal distinction will be quantitatively examined in the following sections.

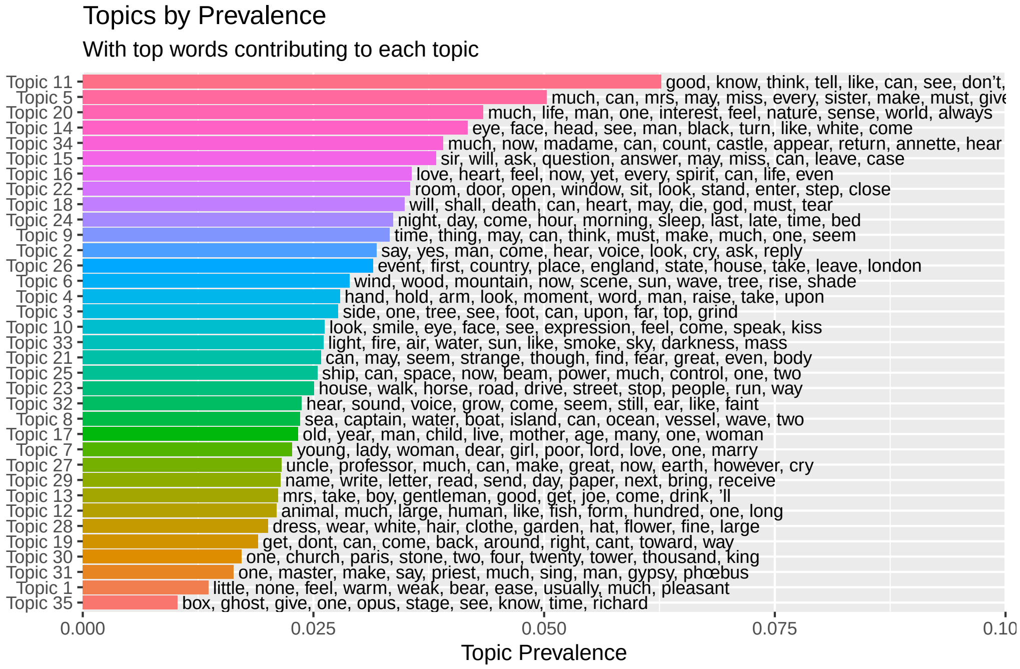
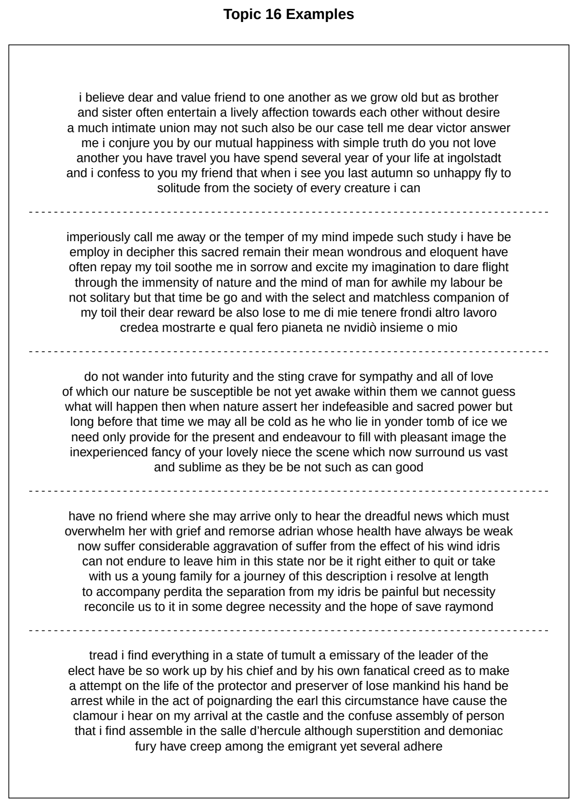


Figure 6: Top Topics by Prevalence

Figure 6 provides an overview of the most prevalent themes within the entire corpus.

Qualitative Validation: Topic Examples

To validate the semantic interpretation of these topics, we examine representative text segments identified by the model. Topic 16 and Topic 18 are of particular significance, as they encapsulate themes central to the Shelley's work.



Topic 16 Examples

i believe dear and value friend to one another as we grow old but as brother and sister often entertain a lively affection towards each other without desire a much intimate union may not such also be our case tell me dear victor answer me i conjure you by our mutual happiness with simple truth do you not love another you have travel you have spend several year of your life at ingolstadt and i confess to you my friend that when i see you last autumn so unhappy fly to solitude from the society of every creature i can

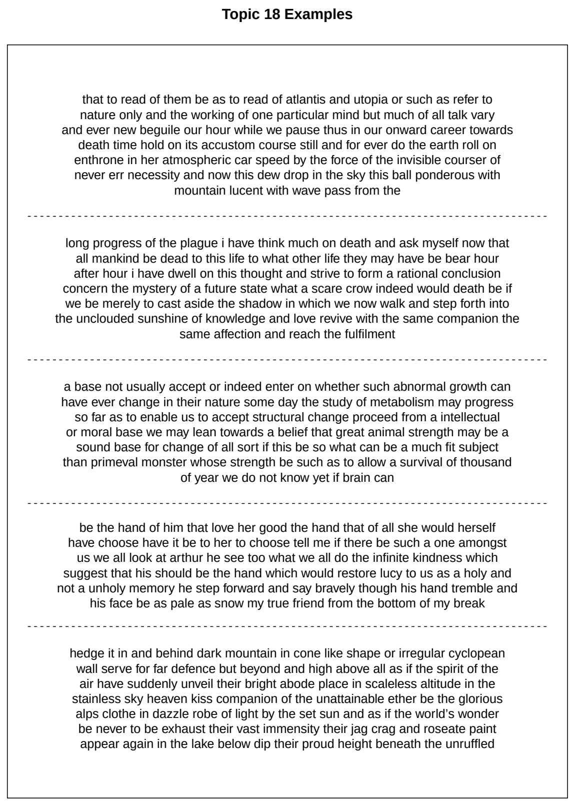
imperiously call me away or the temper of my mind impede such study i have be employ in decipher this sacred remain their mean wondrous and eloquent have often repay my toil soothe me in sorrow and excite my imagination to dare flight through the immensity of nature and the mind of man for awhile my labour be not solitary but that time be go and with the select and matchless companion of my toil their dear reward be also lose to me di mie tenere frondi altro lavoro credea mostrarte e qual fero pianeta ne nvidiò insieme o mio

do not wander into futurity and the sting crave for sympathy and all of love of which our nature be susceptible be not yet awake within them we cannot guess what will happen then when nature assert her indefeasible and sacred power but long before that time we may all be cold as he who lie in yonder tomb of ice we need only provide for the present and endeavour to fill with pleasant image the inexperienced fancy of your lovely niece the scene which now surround us vast and sublime as they be be not such as can good

have no friend where she may arrive only to hear the dreadful news which must overwhelm her with grief and remorse adrian whose health have always be weak now suffer considerable aggravation of suffer from the effect of his wind idris can not endure to leave him in this state nor be it right either to quit or take with us a young family for a journey of this description i resolve at length to accompany perdita the separation from my idris be painful but necessity reconcile us to it in some degree necessity and the hope of save raymond

tread i find everything in a state of tumult a emissary of the leader of the elect have be so work up by his chief and by his own fanatical creed as to make a attempt on the life of the protector and preserver of lose mankind his hand be arrest while in the act of poignarding the earl this circumstance have cause the clamour i hear on my arrival at the castle and the confuse assembly of person that i find assemble in the salle d'hercule although superstition and demoniac fury have creep among the emigrant yet several adhere

Figure 7: Topic 16: Sentimental Love



Topic 18 Examples

that to read of them be as to read of atlantis and utopia or such as refer to nature only and the working of one particular mind but much of all talk vary and ever new beguile our hour while we pause thus in our onward career towards death time hold on its accustom course still and for ever do the earth roll on enthrone in her atmospheric car speed by the force of the invisible courser of never err necessity and now this dew drop in the sky this ball ponderous with mountain lucent with wave pass from the

long progress of the plague i have think much on death and ask myself now that all mankind be dead to this life to what other life they may have be bear hour after hour i have dwell on this thought and strive to form a rational conclusion concern the mystery of a future state what a scare crow indeed would death be if we be merely to cast aside the shadow in which we now walk and step forth into the unclouded sunshine of knowledge and love revive with the same companion the same affection and reach the fulfilment

a base not usually accept or indeed enter on whether such abnormal growth can have ever change in their nature some day the study of metabolism may progress so far as to enable us to accept structural change proceed from a intellectual or moral base we may lean towards a belief that great animal strength may be a sound base for change of all sort if this be so what can be a much fit subject than primeval monster whose strength be such as to allow a survival of thousand of year we do not know yet if brain can

be the hand of him that love her good the hand that of all she would herself have choose have it be to her to choose tell me if there be such a one amongst us we all look at arthur he see too what we all do the infinite kindness which suggest that his should be the hand which would restore lucy to us as a holy and not a unholy memory he step forward and say bravely though his hand tremble and his face be as pale as snow my true friend from the bottom of my break

hedge it in and behind dark mountain in cone like shape or irregular cyclopean wall serve for far defence but beyond and high above all as if the spirit of the air have suddenly unveil their bright abode place in scaleless altitude in the stainless sky heaven kiss companion of the unattainable ether be the glorious alps clothe in dazzle robe of light by the set sun and as if the world's wonder be never to be exhaust their vast immensity their jag crag and roseate paint appear again in the lake below dip their proud height beneath the unruffled

Figure 8: Topic 18: Death and the Soul

Topic 16 (Sentimental Love)
Example 1 is a snippet from Mary Shelley's *Frankenstein* (Vol 3, Chapter 5). A letter from Elizabeth to Victor, anxiously asking if he loves someone else due to his emotional distance.

Example 2 is a snippet from Mary Shelley's *The Last Man* (Introduction). The narrator describes finding ancient prophecies in a cave, reflecting on the loss of a loved one who used to help decipher them.

Example 3 is a snippet from Mary Shelley's *The Last Man* (Vol 3, Chapter 8). Lionel reflects on the innocence of the children as they travel through the Alps, oblivious to the inevitability of death.

Example 4 is a snippet from Mary Shelley's *The Last Man* (Vol 1, Chapter 11). Lionel makes the difficult choice to leave his wife to help his sister, illustrating the conflict between competing familial loves.

Example 5 is a snippet from Mary Shelley's *The Last Man* (Vol 3, Chapter 9). The survivors arrive in Paris to find a religious fanatic causing chaos, contrasting superstition with reason.

Summary: These passages discuss love and relationships, but predominantly link them with anxiety, grief, or difficult sacrifices.

Topic 18 (Death and the Soul)
Example 1 is a snippet from Mary Shelley's *The Last Man* (Vol 1, Chapter 10). This discusses death within the context of Utopia and the future.

Example 2 is a snippet from Mary Shelley's *The Last Man* (Vol 3, Chapter 2). Reflecting on the mass extinction of humanity, the narrator discusses death with a tone of hope and desire for knowledge.

Example 3 is a snippet from Bram Stoker's *The Lair of the White Worm* (Chapter 20). A discussion of biological horror and its evolution.

Example 4 is a snippet from Bram Stoker's *Dracula* (Chapter 16). Death is framed as a necessary means of saving the soul of Lucy.

Example 5 is a snippet from Mary Shelley's *The Last Man* (Vol 3, Chapter 7). A discussion of beauty in nature, with mountains described as touching heaven.

Summary: Most passages discuss death and existence, but the tone varies between Shelley's philosophical reflection and Stoker's physical or religious pragmatism.

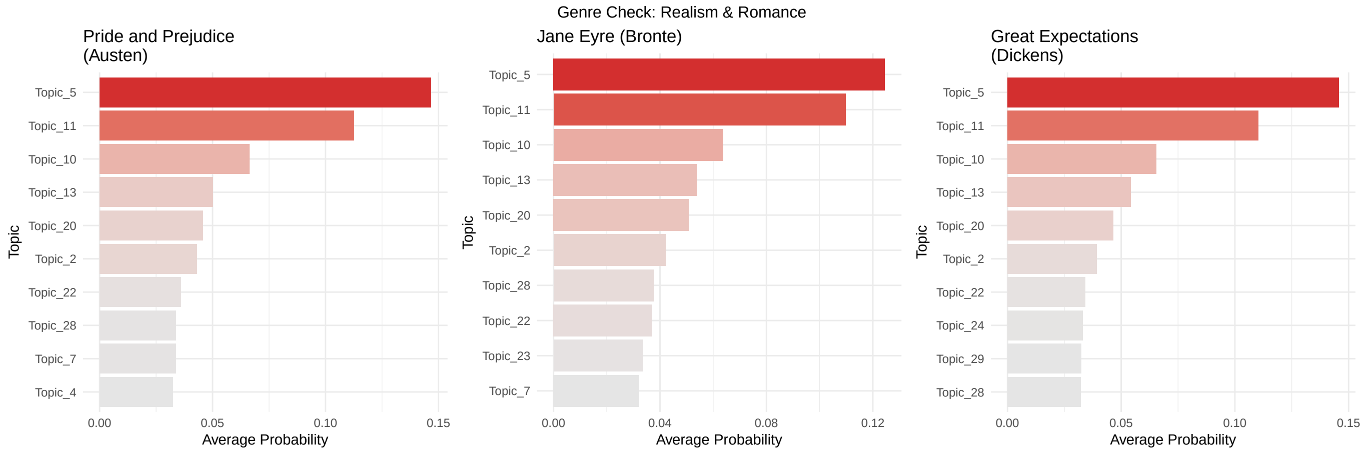
This analysis highlights a clear distinction in the treatment of these themes. Topic 16 reveals a uniquely Shelleyan theme of love infused with grief and sacrifice, while Topic 18 demonstrates a thematic parallel between Shelley and Stoker regarding death, albeit approached from different philosophical angles.

To synthesise these findings, the subsequent section groups these topics into broader clusters to map their distribution across genres.

Topics in Specific Works

It is now useful to consider which topics actually appear in specific works. This serves two purposes: classifying our works by our topics, and checking whether these topics actually make sense within our corpus of texts, confirming they have picked up on the themes and context of different works.

Realism and Romance


Genre Check: Realism & Romance

Pride and Prejudice (Austen)

Jane Eyre (Bronte)

Great Expectations (Dickens)

Figure 9: Topic Distribution from Selected Romantic Works

It is clear that these topics identify this genre well, with the topics present and their distributions being almost identical across these three works. The most prevalent topics seem to be:

Table 6: Interpretation of Romantic Topics

Topic ID	Interpretation
5	Females
10	Romance and Intimacy
11	Opinions
13	Males
20	Existentialism

The model identifies the *Realism/Romance* genre with high precision; the topic distributions are nearly identical across the three control texts. The most prevalent themes—*Females* (5), *Intimacy* (10), *Opinions* (11), and *Males* (13)—align strictly with the genre's focus on domestic social dynamics. This confirms the model's ability to distinguish the baseline "social world" from the speculative genres.

Gothic Horror

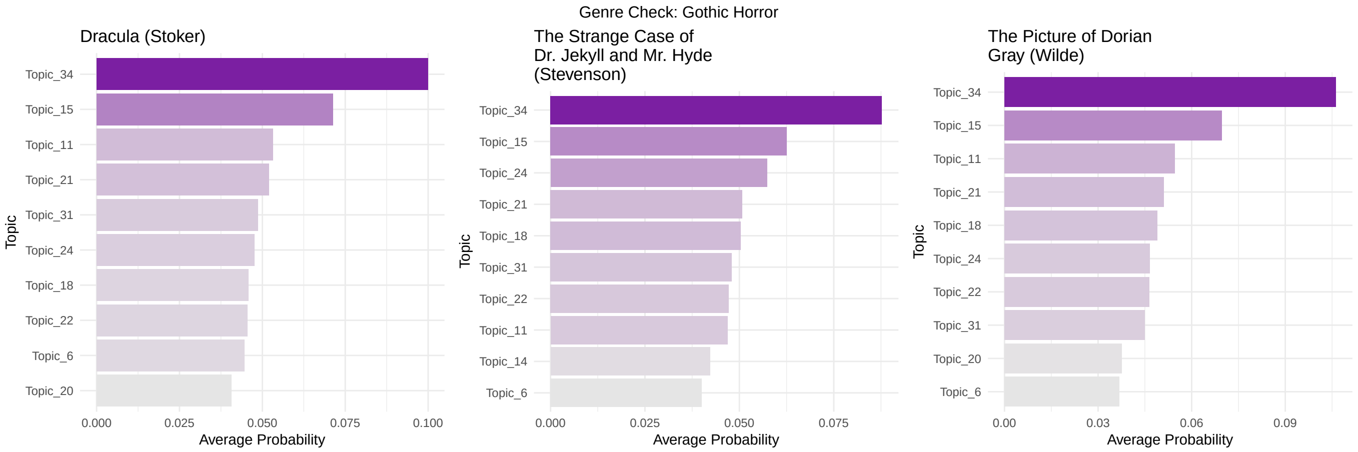


Figure 10: Topic Distribution from Selected Gothic Works

Table 7: Interpretation of Gothic Topics

Topic ID	Interpretation
34	Gothic Settings (Castles/Counts)
15	Questioning
11	Opinions
21	Mystery & Discovery
31	Master/Authority
18	Death and the Soul
24	Night

The *Gothic* corpus exhibits greater variance in topic distribution, reflecting the diverse thematic content of the genre. However, a core set of topics remains consistent: *Gothic Settings* (34), *Mystery* (21), and *Death* (18). This validates the model's capacity to group disparate narratives (e.g., Dracula vs. Dorian Gray) under a unified thematic framework based on shared atmospheric elements.

Science Fiction

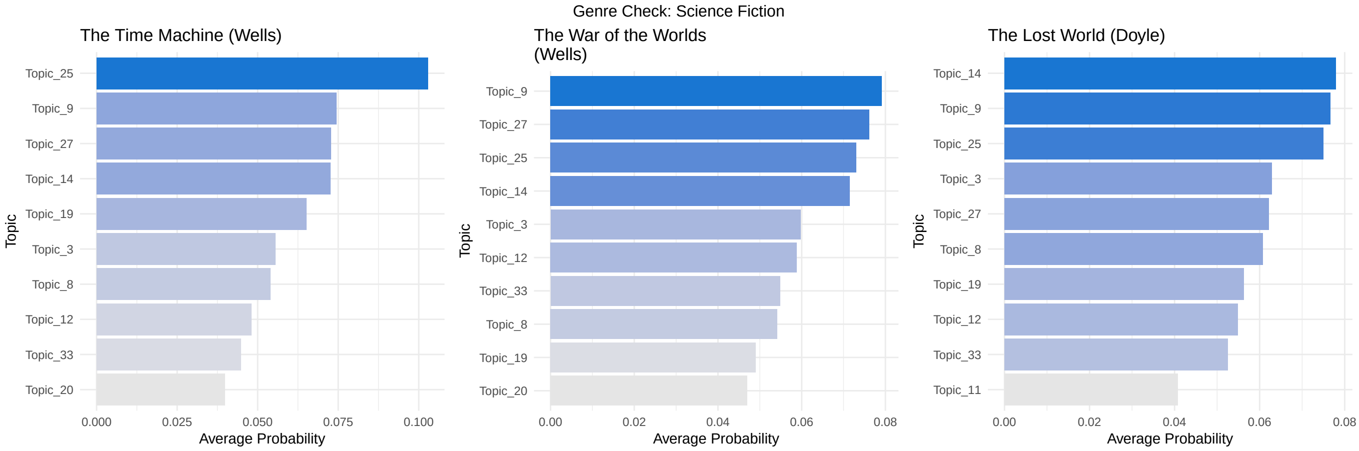


Figure 11: Topic Distribution from Selected Sci-Fi Works

The Science Fiction analysis highlights the genre's distinct vocabulary. The model correctly identifies *Space Travel* (25) and *Professors* (27) as dominant themes. Notably, the prevalence of *Modern Dialogue* (19) correlates with the later publication dates of these texts, confirming that the model captures both thematic and syntactic shifts (such as the increasing use of contractions) in the Modern era.

Now, we would hope to find some overlap between the topics in the previous genres and Shelley.

Table 8: Interpretation of Sci-Fi Topics

Topic ID	Interpretation
25	Space Travel
9	Time and Early Space Travel
27	Professors/Experts
14	Male features
19	Modern Dialogue (Contractions)
3	Terrain

Shelley

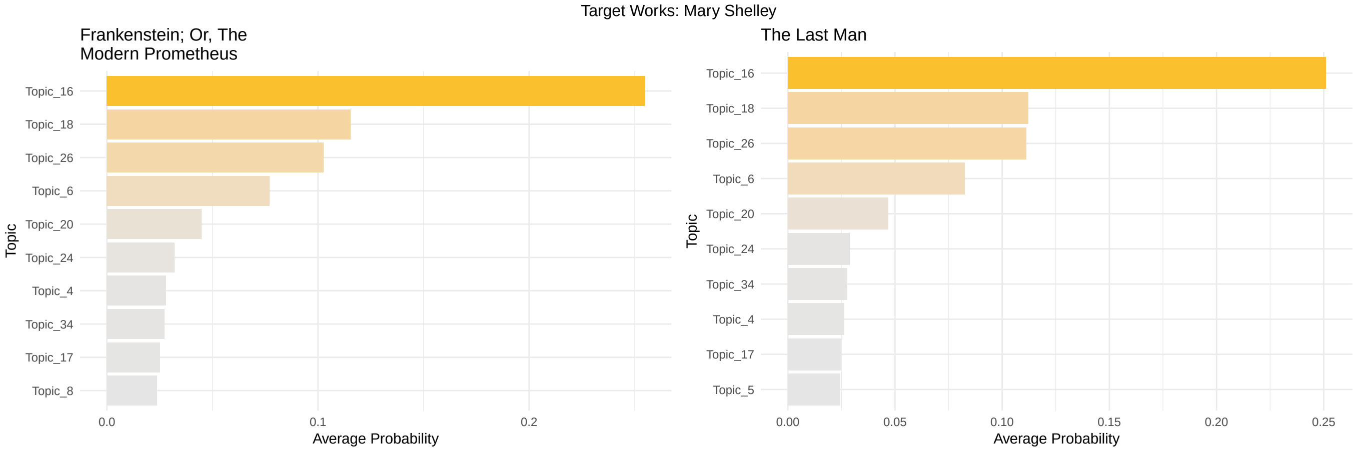


Figure 12: Topic Distribution from Shelley's Works

Table 9: Interpretation of Shelley Topics

Topic ID	Interpretation
16	Sentimental Love
18	Death & The Soul
26	England
6	Nature
20	Existentialism

Our STM has again been consistent in distinguishing the works of Shelley, with the topics and distribution being identical in her two works. It is clear from this that Shelley is unique compared to her contemporaries, yet there are similar themes: *Existentialism* (Topic 20) is also present in *Realism/Romance*, and *Death and The Soul* (Topic 18) is also present in Gothic Horror.

There are, however, no topic crossovers with Science Fiction. Now, this does not necessarily mean that we have disproved our claim. One possible issue with STM models is the influence of the covariates, or metadata; here, our model has correctly distinguished our genres, as it is designed to do. However, has this been a flaw in this approach? As our aim is to find which genre Shelley aligns with, perhaps we should not have included genre as a covariate and just looked for topical similarity between these texts. This approach may have seen more overlap between different genres and allowed for deeper insight.

For now, we will consider the relationship between our topics, seeing if there is correlation, or intuitively grouping them to compare the crossover between Shelley and the other three genres.

Topic Correlation Analysis

Themes that are prevalent across the corpus may be distributed across multiple related topics rather than confined to a single distribution. Therefore, analysing the correlation between topics allows us to identify thematic clusters and determine whether specific topics function as bridges between genres.

Table 10 highlights the strongest correlations (r>0.2), revealing significant structural links between topics.

Table 10: Strong Topic Correlations ($r > 0.2$)

Topic A	Topic B	Corr.	Interpretation	Analysis
2	11	0.318	Invisible Man and Opinions	Focused on a specific text, therefore not useful in our analysis. Correlation due to the novel *The Invisible Man* being driven by arguments.
7	11	0.283	Young Females, Marriage and Opinions	Many Romantic works driven by people's opinions and discussions on marriages and social expectations.
19	25	0.232	Modern Dialogue and Space Travel	Space travel is a modern topic, therefore coincides with the use of modern writing styles (contractions).
10	11	0.221	Romance and Opinions	Similar to Topics 7 and 11; a common trope of Romance novels where characters discuss relationships.
3	14	0.215	Terrain and Male Features	Likely captures the physical nature of early Sci-Fi and adventure novels (moving through terrain).
16	18	0.209	Sentimental Love and Death	Key theme found earlier in Shelley's works (Tragic Romance).
4	14	0.204	Romance and Male Features	Discussion of males supporting females is common across fiction in these eras.
3	33	0.203	Terrain and Elements	Links to Sci-Fi and adventure (Man vs. Nature).
16	26	0.203	Sentimental Love and England	England was the setting of most of these novels, and Love a common theme.
14	32	0.200	Male Features and Eery Suspense	Likely authors building fear and suspense through describing physical reactions or features.

To visualise these deeper connections, a network graph was constructed. Nodes were color-coded based on the genre with the highest prevalence for that topic. Crossover topics (shared by multiple genres) are represented by blended colors. To highlight Shelley's specific position, topics where she exhibits high prevalence are marked with a gold border.

To further aid in the interpretability of the plot, The correlation connections were restricted to correlations with $r \geq 0.1$.

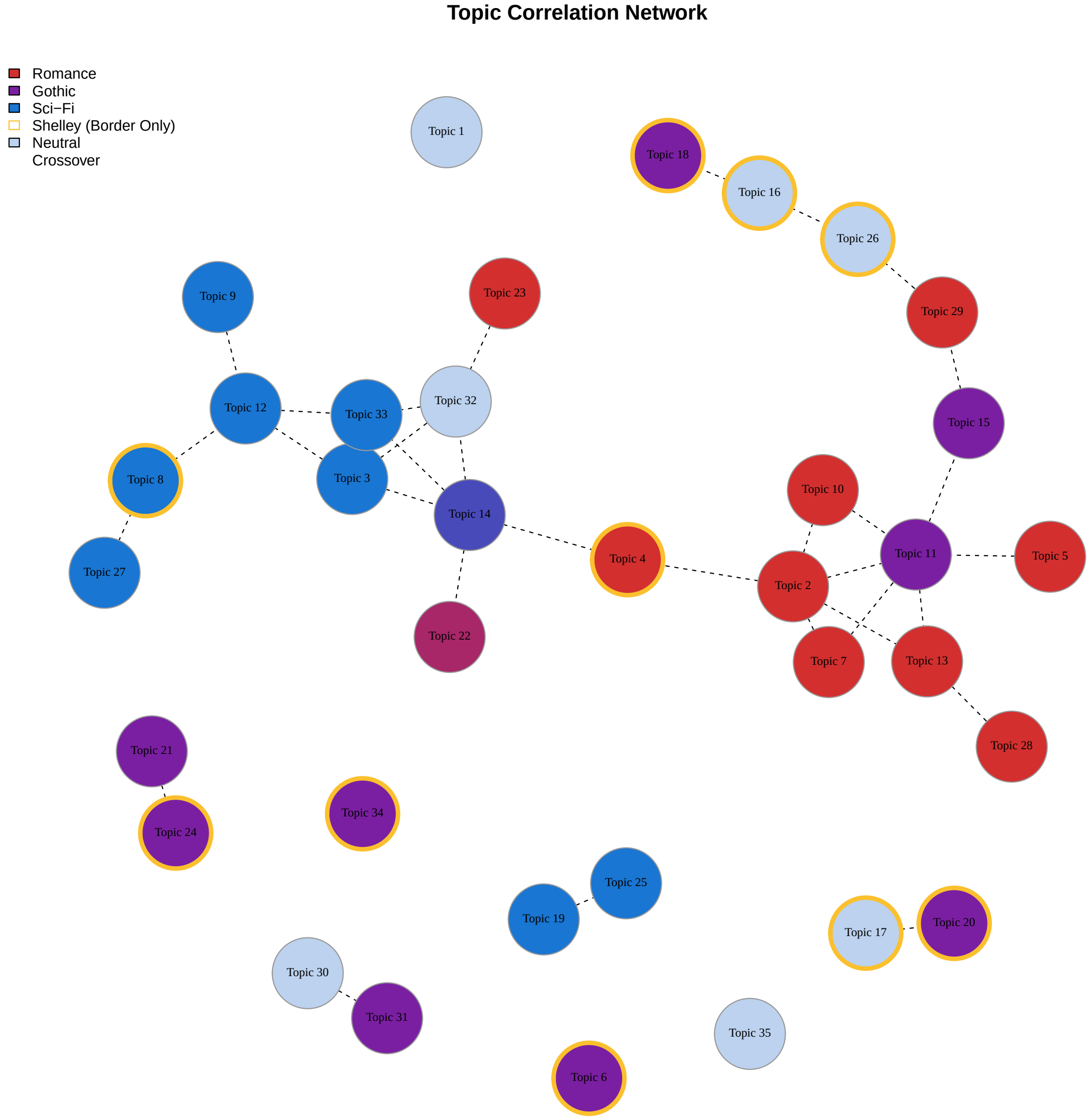


Figure 13: Topic Correlation Graph

Now we have a graph that shows some interesting relationships. In the top right of the figure 13, there is a nice cluster and similarity between Romantic, *Gothic* and Shelley. The top left is a cluster of mostly Science Fiction, but with some overlap with Romantic and *Gothic*. But overall, it does show that Shelley overlaps more with *Gothic* (5 topics) compared to Science Fiction and Romantic (1 topic).

Now we have a graph that shows some interesting relationships. In the top right of the graph, there is a nice cluster and similarity between Romantic, *Gothic* and Shelley. The top left is a cluster of mostly Science Fiction, but with some overlap with Romantic and *Gothic*. But overall, it does show that Shelley overlaps more with *Gothic* (5 topics) compared to Science Fiction and Romantic (1 topic).

Key takeaways from this are:

- Specific *Gothic*/Shelley clusters
  - Topic 21 and Topic 24

  - Topic 34
  - Topic 6
  - Topic 17 and Topic 20

  This does suggest that the topics Shelley wrote about were uniquely *Gothic* in nature, and therefore, suggests that her works are *Gothic*

- Science Fiction Cluster
  - Topics 27, 8 (Shelley), 12, 9, 3, 33, 32, 23, 24, 22

  This cluster does show some connections between Science Fiction and the other genres, and may help suggest the transition from Romance through *Gothic* into Science Fiction

- Romantic/Gothic Cluster (Could be Era influenced)
  - Central cluster: Topics 11 (center topic), 10, 2, 7, 13, 28, 11, 5
  - Off-shoot: Topics 15, 29, 26 (Shelley), 16 (Shelley), 18 (Shelley)

  The off-shoot here shows the divergence in themes from Romantic works into these new genres, whilst still being rooted within it. With Shelley's works being at the end of this off-shoot, suggests that she was a transitionary writer between these topics.

- Connecting Topic 4: Romance and Support This seems to be a common theme connecting these genres, and is often a staple of most works of fiction, in any form of media.

Although this shows useful connections between topics, it does not yet answer our questions fully about Shelley.

Thematic Clustering

From this, there are clearly connections between these topics, so let us consider if these can be grouped intuitively to consider themes to see if this will aid our analysis in relation to our problems.

Here are my proposed clusters of topics.

Table 11: Thematic Super-Clusters and Genre Evolution

Cluster	Topics included	Explanation
The Social and Domestic Sphere	**Topic 5**: Females **Topic 7**: Females & Marriage **Topic 10**: Romance & Intimacy **Topic 11**: Opinions **Topic 13**: Males **Topic 28**: Garden/Weddings	This cluster represents the Romantic themes of the corpus, grounded in domestic life, gender roles, and social discourse. It serves as the baseline from which the other genres diverge.
Mystery, Darkness, Death and Philosophical Questioning	**Topic 6**: Sublime Nature **Topic 15**: Questioning **Topic 16**: Sentimental Love **Topic 18**: Death & The Soul **Topic 20**: Existentialism **Topic 21**: Mystery **Topic 24**: Night **Topic 29**: Communication **Topic 31**: Master/Authority **Topic 34**: Gothic Settings	This cluster highlights themes associated with Gothic Horror: suspense, mystery, death, and fear. It captures the questioning of death, the soul, and existence central to the genre.
The Physical and Empirical World	**Topic 3**: Terrain **Topic 8**: Naval **Topic 9**: Early Space **Topic 12**: Nature/Animals **Topic 25**: Modern Space **Topic 27**: Professors **Topic 32**: Suspense **Topic 33**: Elements	This cluster highlights typical Science Fiction themes: observation, discovery, and interaction with the physical world.
Universal Connector	**Topic 4**: Romance & Support	This topic connects all genres and eras, representing the constant human element of support in narrative fiction.

Note that a subset of topics $(1, 2, 4, 14, 17, 19, 22, 23, 26, 30, 35)$ were excluded from this high-level clustering as they largely represent stylistic markers or specific narrative devices rather than distinct genre themes.

With these key topics now categorised, the final phase of analysis examines how the vocabulary within these shared themes shifts between authors. Specifically, we investigate how Shelley's treatment of these topics differs from the authors in the control groups.

Topic Prevalence

To investigate the thematic composition of the corpus, we analysed the prevalence of each topic cluster across the four author groups. This analysis allows us to quantify not just genre membership, but the specific thematic deviations that define Shelley's transitional status.

Romance Topics

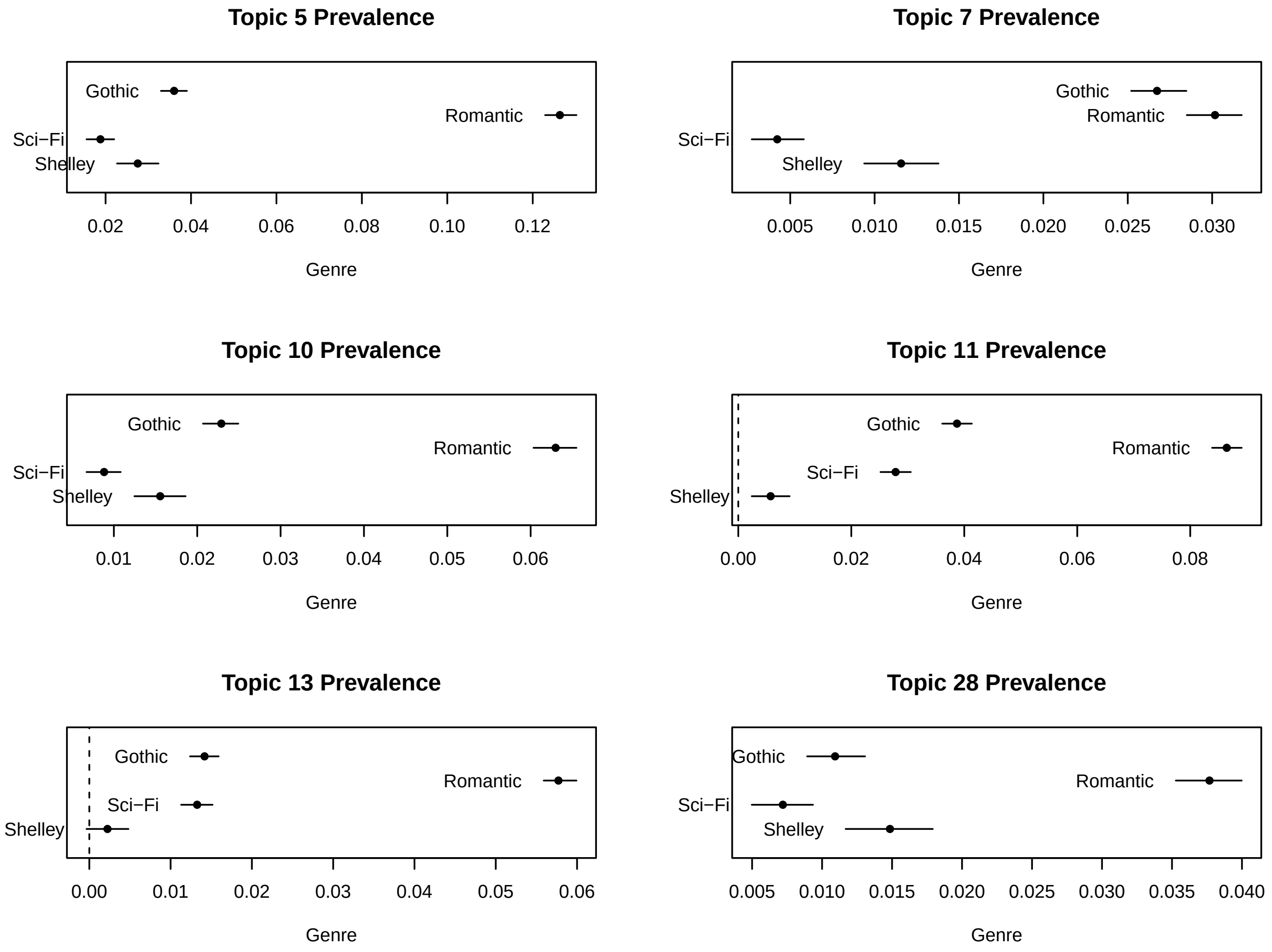
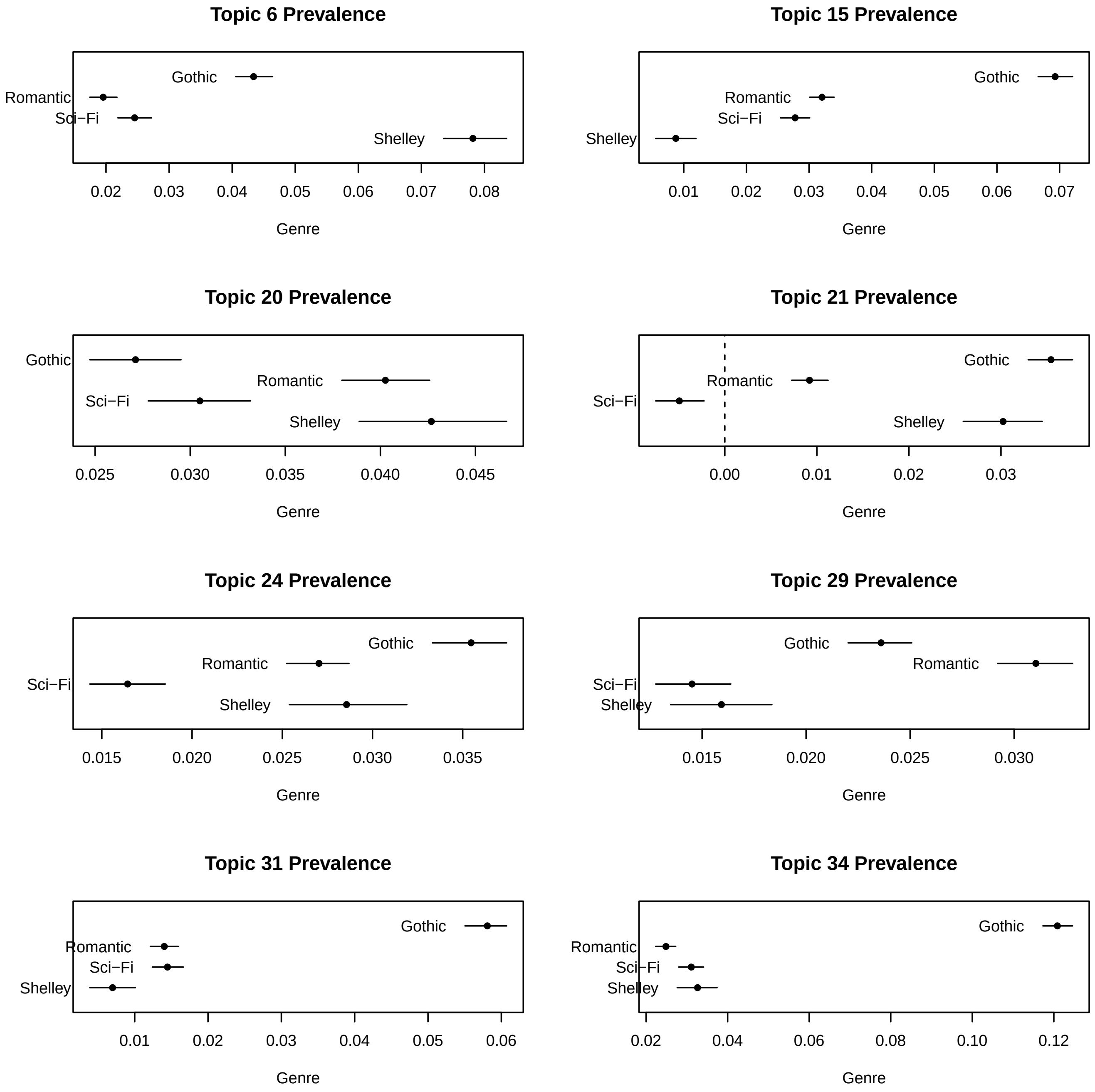


Figure 14: Romantic Topics Prevalence

The distribution indicates that Shelley's engagement with traditional Romantic topics is notably limited. While there is minor overlap between the *Gothic* and Romantic groups, the *Realism/Romance* corpus maintains a distinct thematic signature that Shelley does not share. This confirms that despite her chronological placement in the Romantic era, her topical focus lies elsewhere.

Gothic Topics



Topic 6 Prevalence

Topic 15 Prevalence

Topic 20 Prevalence

Topic 21 Prevalence

Topic 24 Prevalence

Topic 29 Prevalence

Topic 31 Prevalence

Topic 34 Prevalence

Figure 15: Gothic Topics Prevalence

The Gothic cluster presents a more complex relationship. While Shelley is stylistically Gothic, her thematic profile diverges from the standard genre conventions.

- **Divergence:** Topics such as 15, 20, 31, and 34 (Castles/Authority) are dominated by the Gothic Horror control group. Shelley's lower usage suggests she is less concerned with the external trappings of Gothic settings (castles, priests) than her peers.

- **Shelley Influence:** Conversely, Topic 6 (Sublime Nature) appears significantly *more* frequently in Shelley's work than in standard Gothic texts. This suggests Shelley injects a specific "Nature philosophy" into the genre that standard Gothic horror lacks.

- **Alignment:** Shelley aligns closely with the Gothic group only in Topic 21 (Mystery/Discovery), further indicating that her connection to the genre is selective rather than total.

Science Fiction Topics

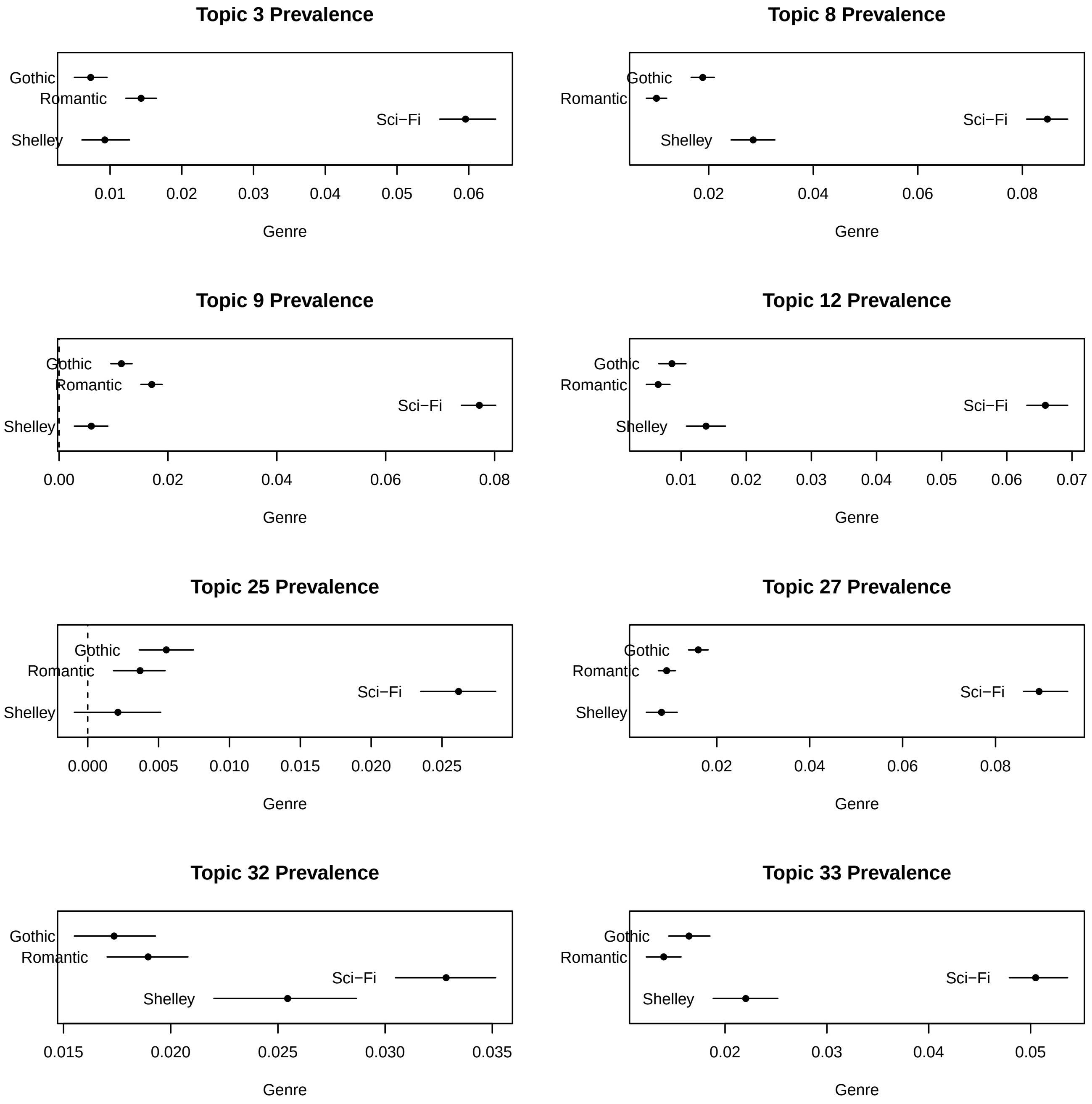


Figure 16: Science Fiction Topics Prevalence

The Science Fiction topics exhibit the highest degree of exclusivity, likely driven by the distinct technological vocabulary of the post-Victorian era. However, distinct bridges emerge:

- **Transitional Themes:** Shelley shows elevated usage of Topics 8 (Naval), 12 (Nature/Animals), and 33 (Elements) compared to her Romance or Gothic contemporaries. This statistically supports the hypothesis that she was a progenitor of the Scientific Exploration narrative.

- **The Roots of Suspense:** Most notably, Topic 32 ("Eery Suspense") is highly prevalent in Shelley's work, mirroring its usage in later Science Fiction. This suggests that the *mood* of modern Science Fiction, the dread of the unknown, may have its roots in Shelley's psychological approach to horror.

Shelley Topics

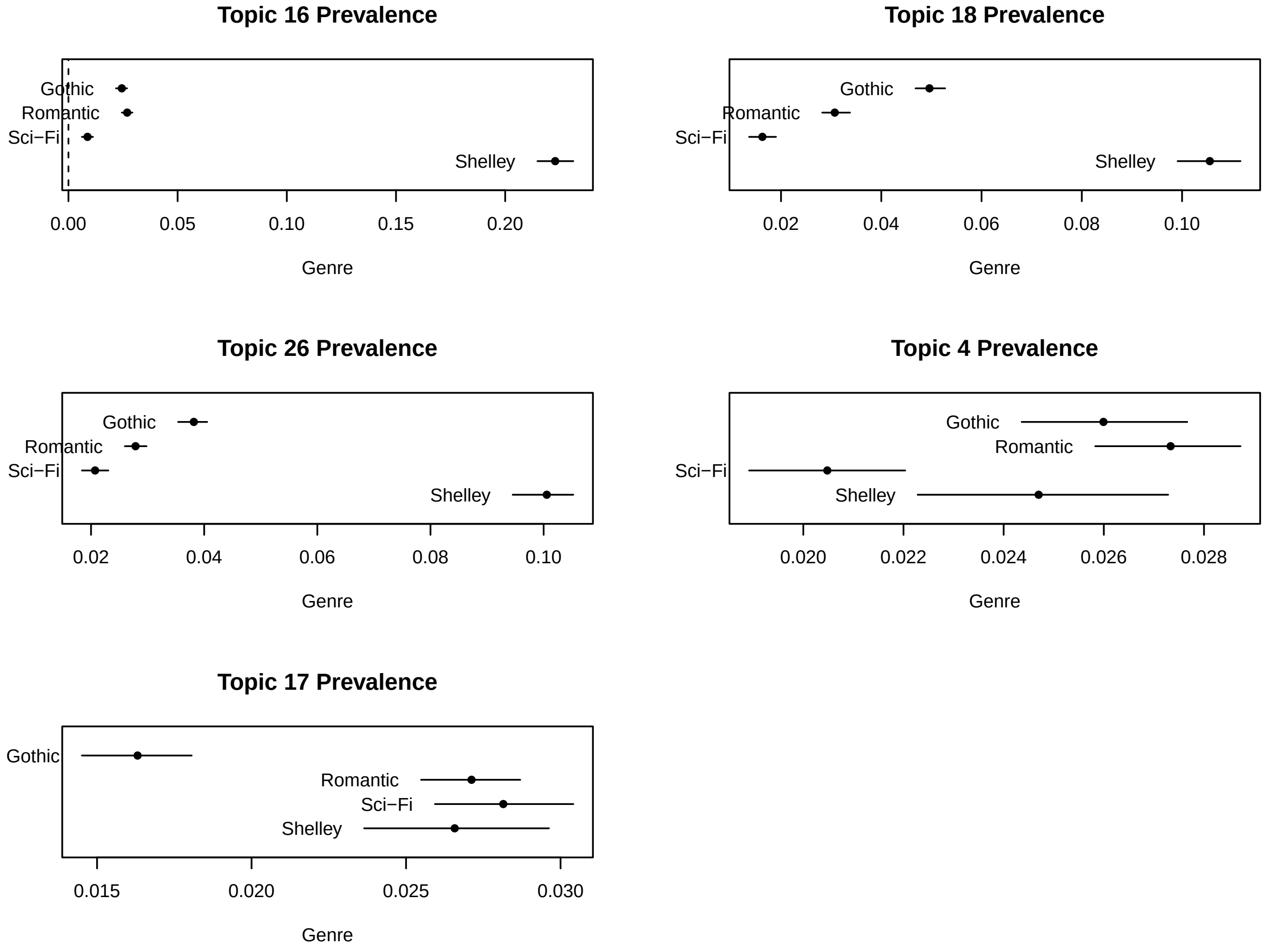


Figure 17: Shelley Topics Prevalence

Finally, examining Shelley's signature topics reveals the mechanics of her "Bridge" status.

- **Core Identity:** Topics 16 (Sentimental Love) and 18 (Death and the Soul) are unmistakably Shelley, defining the emotional core of her work. Also, Topic 26 (England) due to the setting of her works, and explicit descriptions and discussions of England.

- **The Bridge:** Topic 4 (Romance/Support) functions as a universal connector, appearing evenly across genres. Although Shelley's work does seem to be a bridge between the Romantic and *Gothic* treatment of the topic.

- **Genre fluidity:** Interestingly, Topic 17 (Age/Youth) in Shelley's work follows a distribution pattern closer to Romance and Science Fiction than to Gothic Horror. This reinforces the conclusion that Shelley is a hybrid author, employing Gothic moods to explore Romantic emotions within a proto-Scientific framework.

To further investigate these transitional relationships, we will now analyse the content covariates, specifically, the vocabulary usage within shared topics. By comparing how Shelley uses a topic versus how H.G. Wells uses the same topic, we can map the evolution from Romanticism to Science Fiction.

Topic Comparison

Shelley vs. Romantics

As our topics have been categorised into our different Genres, it would be useful to compare similarly themed ones. We use this to determine how Shelley differs from each genre.

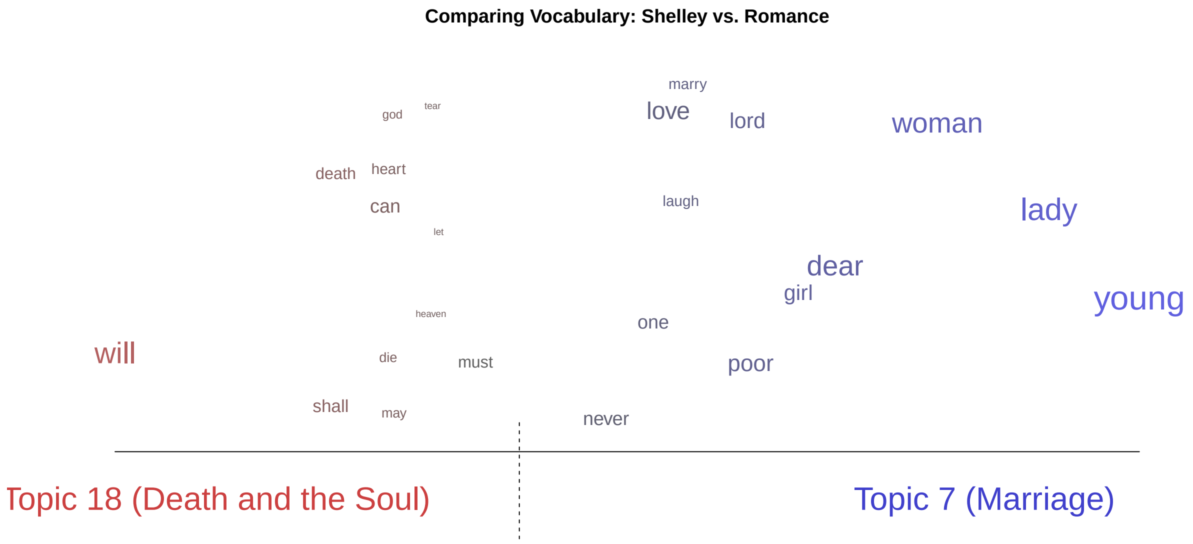


Figure 18: Topic Comparison: Shelley vs Romance

We have compared these two topics, as they are key themes in Shelley's work, and in Romantic works, and demonstrates that Shelley diverges from her contemporaries, who discuss relationships between people, where as Shelley discusses the metaphysical and internal, and our connection to death and soul.

Shelley vs. Gothic Horror

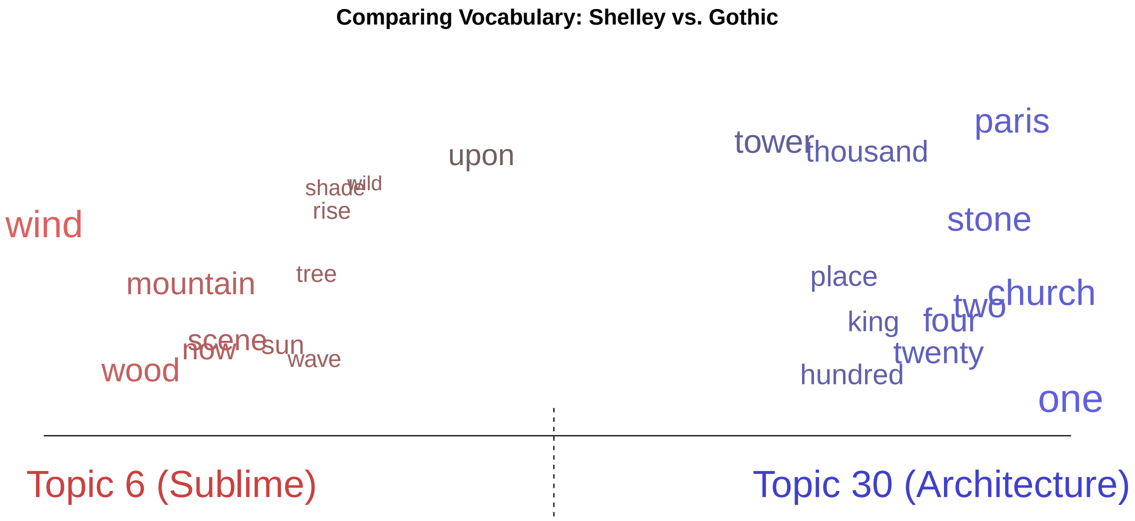


Figure 19: Topic Comparison: Shelley vs Gothic

Comparing how Shelley describes the scene compared to *Gothic* works, Shelley focuses on nature and elemental descriptions, showing the sublime beauty of the landscape. Whereas *Gothic* describes is man made, and designed to instill fear, authority and strength.

Shelley vs. Science Fiction

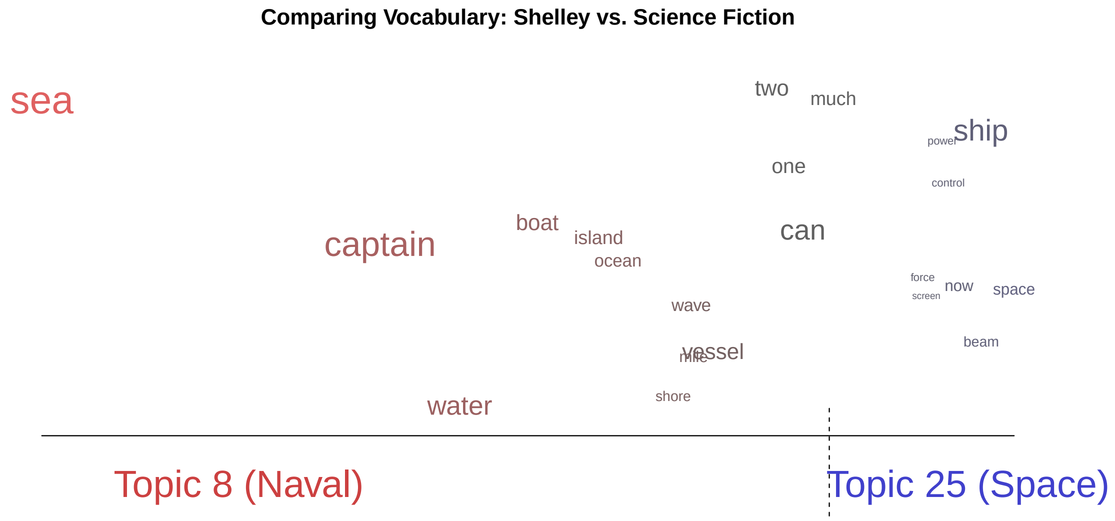


Figure 20: Topic Comparison: Shelley vs Science Fiction

One key aspect that inextricably links Shelley to the Science Fiction tradition is the thematic centrality of exploration and travel. As illustrated in Figure **??**, Shelley's vocabulary in Topic 8 (*Naval Exploration*) shares a distinct structural DNA with the later Space Travel topics found in the Science Fiction corpus.

This demonstrates a fundamental shift in Shelley's work compared to her *Gothic* contemporaries. While Gothic horror is traditionally static, trapping its protagonists in ancestral castles, abbeys, or locked rooms (Topic 34), Shelley mobilises the genre. In *Frankenstein*, the horror does not sit waiting in a dungeon; it is pursued across the frozen wastes of the Arctic. In *The Last Man*, the narrative spans continents, as they journey from the Lake District through England to London, before traveling across Europe, ending before leaving for Africa.

This "Architecture of Exploration" establishes the narrative framework that later Science Fiction authors would adapt. The linguistic bridge is visible in the data: the vocabulary of the "Scientific Expedition" (Captain, Vessel, Ocean) seen in Shelley's work directly anticipates the "Interplanetary Expedition" (Commander, Rocket, Space) of H.G. Wells and Jules Verne.

- Shelley (1818): Walton's expedition to the North Pole in *Frankenstein*.
- Verne (1870): Nemo's voyage beneath the oceans in *20,000 Leagues Under the Sea*.
- Modern Science Fiction (1900s): The voyage to the Moon or Mars in *The First Men in the Moon* or *The Martian Chronicles*.

This, alongside the other considerations in the change in topics from Romantic to Gothic to Science Fiction, shows Shelley as being transitionary author, bridging the gap, and developing the structure and tools required for Science Fiction.

Conclusion

Considering the topical structure of Shelley's work, it is clear that she diverges significantly from her Romantic contemporaries. While firmly rooted in the style of *Gothic* literature, sharing the themes of Mystery and Discovery, she pushes beyond these conventions to develop a narrative framework that later *Science Fiction* would build upon. Our STM results, which identify a 'Gothic Scaffold' for scientific themes, statistically corroborate the qualitative arguments made by [Mitra, 2011]. Shelley constructs a scientific narrative using the vocabulary of the Gothic tradition. Figure 16 highlights the thematic gap between the Romantic/Gothic clusters and the Science Fiction cluster, with Shelley positioned distinctly as the bridge between the two. This transition is most evident in the prevalence of *Eery Suspense* (Topic 32) and *Naval Exploration* (Topic 8). By elevating exploration to a central narrative archetype, a device rarely utilised in standard Gothic or Romantic works, Shelley established the structural prototype for the scientific expedition.

The analysis revealed distinct thematic clusters: a central Romantic cluster branching into a Gothic cluster, and a separate, highly exclusive Science Fiction cluster. These disparate groups were connected by Topic 4 (*Romance and Support*), underscoring that emotional connection remains a requisite element of narrative fiction across all genres. Notably, Shelley's work appears in all clusters *except* the central Romantic one, positioning her on the periphery where the Gothic tradition begins to fracture into speculative fiction. This reinforces the hypothesis that she utilised the Gothic framework to pioneer the themes of exploration and discovery.

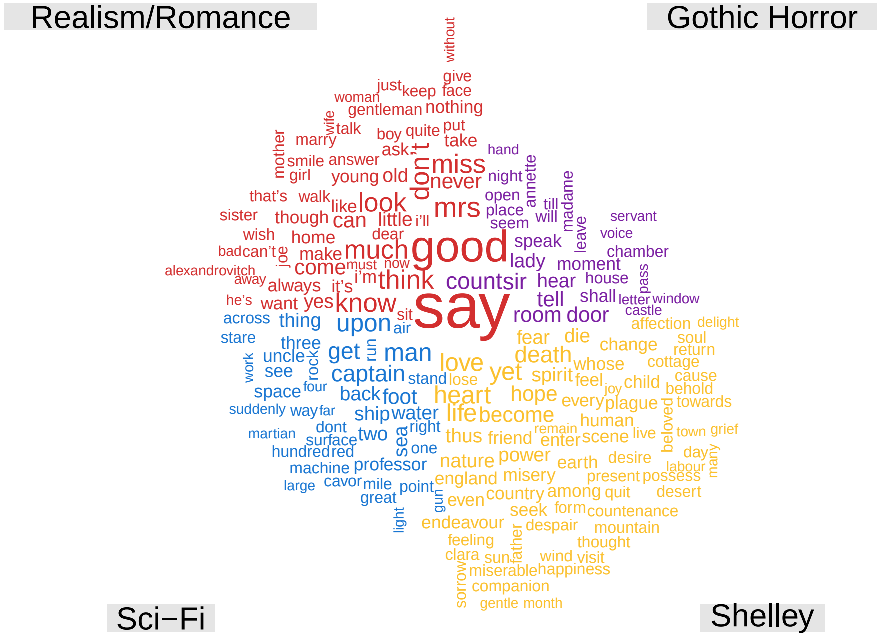


Figure 21: Comparison Word Cloud between Genres

The vocabulary cloud in Figure 21 further demonstrates this lexical evolution. Shelley utilises terms such as "power", "plague", and "endeavour", vocabulary now intrinsically associated with Science Fiction. However, the retention of words like "soul", "spirit", and "fear" marks the transition from the religious, philosophical anxieties of Gothic Horror into the secular, empirical anxieties of Science Fiction.

Stylometrically, Shelley aligns most closely with early Gothic Horror writers. However, the specific cosine similarity analysis of *Frankenstein* and *The Last Man* revealed an equal affinity with later Science Fiction works. This suggests a bidirectional relationship: Shelley was influenced by the Gothic tradition of her time, but her structural innovations exerted a lasting influence on the Science Fiction authors who followed.

In summary, Mary Shelley was a unique author whose stylistic and thematic contributions were ahead of her era. The analysis suggests she laid the groundwork for the narrative structures and atmospheric suspense that would evolve into Modern Science Fiction.

A key distinction that has not been highlighted by this study is the technological limitation of her era. Shelley's "Science Fiction" is defined by speculative philosophy rather than speculative machinery. Limited by the scientific knowledge of the early 19th century, her "technology" is restricted to Galvanism and anatomy. In contrast, later authors could extrapolate from rapid industrial advancements, such as Jules Verne speculating on submarines in *20,000 Leagues Under the Sea*. Yet, the core impulse remains identical: the speculative question "What If?". Whether asking "What if man could create life?" or "What if a plague destroyed mankind?", Shelley developed these questions within a Gothic structure. It appears the only limitation preventing her from being fully categorised as "Science Fiction" was simply the lack of technology available to describe the future she envisioned.

Evaluation

Methodological Constraints: STM vs. LDA

A significant methodological consideration in this study was the choice of Structural Topic Modelling (STM) over Latent Dirichlet Allocation (LDA). By including *Genre* as a prevalence covariate, we explicitly instructed the model to look for differences between the genre groups. While this improved the interpretability of the topics (sharpening the distinction between *Gothic* and *Science Fiction*), it introduces a risk of confirmation bias. The model may have over-emphasised genre-specific vocabulary because it was "primed" to do so. A completely unsupervised LDA model, blind to genre labels, would have provided a more rigorous "stress test" of whether these categories naturally emerge from the text alone, and potentially would give more overlap to aid in the classification of Shelley to a genre. However, the use of topic correlation, and thematic grouping did allow us to compare the Shelley topics to the genre topics and see relationships, without the model assuming Shelley's genre.

The Limits of Lexical Inference

While the model successfully identified thematic clusters, it failed to capture the contextual nuance of Shelley's technological limitations. The analysis identified *Naval Exploration* (Topic 8) as Shelley's bridge to Science Fiction, contrasting it with the *Space Travel* (Topic 25) of later authors. However, the model cannot understand why this difference exists. It sees "Boat" and "Rocket" as distinct topics, failing to recognise that for Shelley, the boat was the rocket, the most advanced vessel of exploration available in her era. This highlights a limitation of Bag-of-Words models: they measure lexical frequency but miss the historical constraints that dictate vocabulary choice. Consequently, qualitative literary analysis was required to interpret this "Technology Gap" not as a thematic divergence, but as a historical necessity.

Causality vs. Similarity

Throughout this study, we have framed Shelley as an "influencer" of later Science Fiction. However, stylometric proximity (PCA/Cosine Similarity) measures only statistical correlation, not causal direction. While we can observe that H.G. Wells shares a stylistic fingerprint with Shelley, the model cannot prove he was influenced by her. It is equally plausible that both authors simply adhered to a "Speculative Fiction" archetype that naturally demands a specific sentence structure or vocabulary. Mathematical proximity provides strong evidence for a shared lineage, but definitive proof of literary influence remains beyond the scope of purely quantitative methods.

Reproducibility and Model Stability

Finally, the stability of the probabilistic model presents a challenge for reproducibility. During the analysis, it was observed that minor changes to the input metadata (specifically, the labelling of the *Era* covariate) resulted in significant shifts in topic formation, even when a fixed random seed was utilised. This sensitivity suggests that the "latent space" of the corpus is volatile; the boundaries between topics like *Existentialism* and *Death* are fluid and highly dependent on the initial model specification. Future iterations of this work would require a stricter "Model Checkpointing" system (saving the model object immediately after training) to ensure that the interpretative analysis remains consistent with the generated data.

Sources and References

References

[Dealban, 2025] Dealban, D. (2025). Learning stm. *GitHub Repository*.

[Haldane, 2015] Haldane, M. (2015). What *Frankenstein*'s creature can really tell us about AI.

[Kurochkin, 2025] Kurochkin, D. (2025). Lecture 8: Introduction to structural topic modeling (stm). CSCI E-89B: Introduction to Natural Language Processing, Harvard Extension School.

[Lebryk, 2021] Lebryk, T. (2021). Introduction to the structural topic model (stm).

[Mitra, 2011] Mitra, Z. (2011). A science fiction in a gothic scaffold: A reading of mary shelley's *Frankenstein*. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 3(1):52–59.

[Paley, 1993] Paley, M. D. (1993). Mary shelley's *The Last Man*: Apocalypse without millennium. *Keats-Shelley Review*, 4(1):1–25.

[Project Gutenberg, 2025] Project Gutenberg (2025). Project gutenberg. Digital Library of Free eBooks.

[Roberts et al., 2019] Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2):1–40.

[Robinson, 2025] Robinson, D. (2025). *gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. R package version 0.2.4.

[Silge, 2018a] Silge, J. (2018a). Evaluating STM.

[Silge, 2018b] Silge, J. (2018b). Sherlock holmes & STM.

[Silge and Robinson, 2017] Silge, J. and Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media, Sebastopol, CA.

[Stableford, 1995] Stableford, B. (1995). Frankenstein and the origins of science fiction. In Seed, D., editor, *Anticipations: Essays on Early Science Fiction and its Precursors*. Syracuse University Press, Syracuse, NY.

[Wikipedia Contributors, 2025a] Wikipedia Contributors (2025a). Armageddon 2419 a.d. Accessed: 2025-12-07.

[Wikipedia Contributors, 2025b] Wikipedia Contributors (2025b). Gothic fiction. Accessed: 2025-12-07.

AI Usage Report

AI Usage Reporting: Transparency Statement

This project utilised Large Language Models (LLMs) to assist in the technical implementation and editorial refinement of the research. The specific contributions of the AI are detailed below:

1. Code Development and Debugging

The R code used for data acquisition, preprocessing, and modelling was co-developed with AI assistance. Specifically, the AI was used to:

- Troubleshoot errors in the gutenbergr package mirror configuration.
- Generate the syntax for the stm package, specifically regarding the estimateEffect and topicCorr functions.
- Draft the ggplot2 code for the custom visualisation of topic prevalence and vocabulary comparison clouds.
- Verification: All code was executed, tested, and validated by the author on a local machine to ensure reproducibility and accuracy.

2. Data Structuring

The AI assisted in compiling the list of Project Gutenberg IDs for the comparative corpus (Romantic, Gothic, and Science Fiction novels) to ensure a balanced dataset.

It generated the LaTeX table code to format the raw results into academic-standard tables (e.g., Table 11: Thematic Super-Clusters).

3. Editorial and Stylistic Refinement

Draft sections of the report were submitted to the AI for proofreading. The AI provided suggestions for correcting spelling, grammar, and academic tone.

The AI assisted in refining the "Problem Statement" and "Abstract" to ensure conciseness and clarity.

Note: The core literary arguments, the interpretation of the "Bridge" hypothesis, and the final conclusions regarding Shelley's technological limitations are the original intellectual work of the author. The AI functioned solely as an editor and technical assistant.

4. Bibliographic Formatting

The AI was used to convert raw URLs and citations into formatted BibTeX entries to ensure compatibility with the LaTeX document structure.