Written in Time: Unsupervised Discovery of Emotional and Stylistic Structure in the Project Gutenberg Corpus

Tomáš Haufer*

CERGE-EI†

May 26, 2026

Abstract

This project explores whether machine learning and quantitative text analysis can identify emotional, stylistic, and thematic patterns in literature across time. Using the Project Gutenberg (n.d.) corpus, I construct a feature-level dataset of books based on emotion scores, stylometric measures, and semantic fields. The analysis first examines descriptive genre-level differences and then residualises the extracted features on decade, country, and genre fixed effects. Subsequently, I apply clustering and network analysis to map similarities between books and authors in the constructed feature space. Finally, I study temporal changes around selected historical shocks, with World War I serving as the headline case. The results suggest that literary texts contain measurable patterns connected to genre, period, and historical context. In particular, the WWI shock is associated with increases in expressed fear, anger, and conflict-related vocabulary, as well as shifts towards more conflict- and past-oriented clusters. These findings should not be interpreted as strictly causal, as the identification strategy remains limited. Rather, the project serves as an exploratory machine learning exercise showing how large literary corpora can be transformed into structured data and used to study literary change at scale.

*E-mail: tomas.haufer@cerge-ei.cz

†CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, Politickych veznu 7, 111 21 Prague, Czech Republic.

Declaration of AI Use: I used Claude (Sonnet and Opus, Anthropic) as a coding and research assistant for pipeline design and implementation, debugging, and methodological discussion. All project design, analytical decisions, and interpretations were conceived and made independently by me. I take full responsibility for the content, analysis, and conclusions presented in this work.

This paper builds on and partially reuses material from an earlier "Idea Paper" submitted by the author for the *Machine Learning for Social Scientists* course at CERGE-EI.

1 Introduction

Advances in technology, science, governance, and the arts have not occurred in isolation. Progress in one domain has consistently shaped developments in others. Innovation has remained closely tied to the living standards of the people and to their emotions, moods, and opinions.

Furthermore, authors of artistic works do not produce in a vacuum. They are shaped by their education, environment, era, and the ideologies and social groups around them. Knowingly or subconsciously, these influences may flow into their works and thus offer a different lens on their everyday realities.

This project uses the filtered Project Gutenberg (n.d.) corpus and treats the texts as a multi-dimensional data source. For each text, additional information such as the publication year and the author's origin is obtained from public databases. These data allow me to identify patterns in the literature and its authors both within a given time period and across centuries. For each text, emotion scores, stylometric variables, and semantic fields are collected to allow later analysis.[1]

This paper should be read primarily as an exploratory machine-learning project rather than as a publication-ready economics article. Its aim is to test how unsupervised learning and large-scale text analysis can be applied to questions relevant for economics, cultural history, and social change. In this sense, the project is both substantive and methodological: it uses literary texts to study whether historical shocks are reflected in language, while also developing an AI-assisted coding pipeline for corpus construction, metadata enrichment, feature extraction, clustering, and event-study analysis.

The main empirical question is whether literary content is associated with historical shocks after stripping out year × country × genre composition. The results suggest that genre explains most composition-driven variation, but some major events are still associated with measurable changes in expressed emotions and semantic fields. World War I is the clearest case: after 1914, books become more fear-oriented and conflict-oriented relative to their baseline, with expressed fear increasing by approximately 0.16 Standard Deviation (SD). These results should be interpreted as a first step; a future version could refine the metadata, add external economic and historical variables, and revisit the project as a more formal paper intended for publication.

This approach builds on similar large-data-based approaches to textual works and their analysis, popularised by Michel et al. (2011). In a related contribution, Underwood (2019) argues that literature evolves through long, gradual patterns, such as increasing concreteness and differentiation between genres, that traditional, period-based approaches have overlooked.

[1] These data could later be linked to economic variables to study how socioeconomic and political context affected literary production.

These methods have been adopted by economists, who link textual data to actual economic variables. An example of this approach is the study by Baker, Bloom, and Davis (2016), which creates a new index of economic policy uncertainty based on more than 12,000 newspaper articles.[2]

Finally, a methodologically related study by Gorin, Heblich, and Zylberberg (2025)[3] uses 630,000 paintings to recover the socioeconomic context effect and the temporal progression of emotion distributions across European countries.

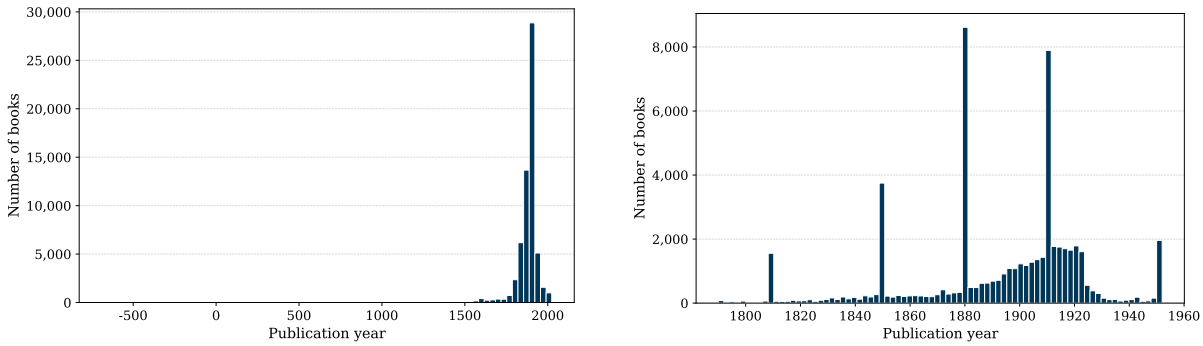
2 Data Manipulation

The core dataset is constructed using the `gutenbergpy` package in Python, covering all English-language books available in the corpus. Initially, metadata are retrieved in order to identify the relevant texts. Subsequently, the corresponding `.txt` files are downloaded and stored. In addition, basic metadata[4] are collected for further analysis.

The publication year is then determined through a hierarchical procedure. First, the publication year is retrieved from Wikidata where available. If this information cannot be obtained, the year is extracted or approximated based on textual summaries. Finally, in cases where neither source provides usable information, the publication date is proxied by adding 35 years to the author's birth year, reflecting the approximate midpoint of an author's productive writing period.

Figure 1 captures the year distribution. At first glance, it is clear that the distribution is heavily biased by the approximation of publication year. In the later parts, this bias is partly addressed by using precision estimation.

Figure 1: Distribution of publication years.



(a) Full range.

(b) Core range, 5th–95th percentile.

Notes: Panel a shows the full range of recovered years; panel b restricts to the central 90% of the distribution. Sharp clustering on round-decade values (1810, 1850, 1880...) reflects the regex extraction of approximate phrases such as "early 19th century" from book summaries.

[2]The authors use large textual data but do not employ any machine learning methods.

[3]This paper was also a major inspiration for this project.

[4]Gutenberg ID, author's birth and death dates, links to the author's and book's Wikipedia pages, a short summary, and the Reading Ease score provided by Project Gutenberg.

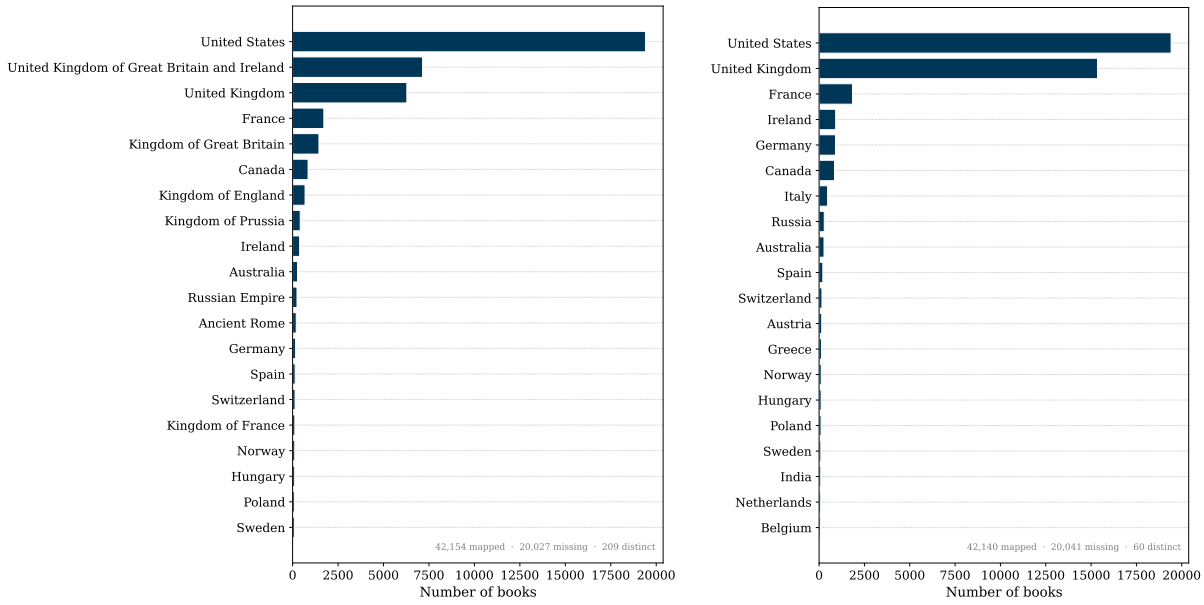
A similar approach is applied to assign the country of origin of each author. In this case, the information is determined according to the following priority cascade:

1. Author's country as identified in Wikidata via the work–author linkage,
2. Author's country obtained from the associated Wikipedia URL,
3. Author's country inferred via name-based search (high confidence[5]),
4. Author's country inferred via name-based search (medium confidence),
5. Country of origin of the work,
6. Author's country inferred via name-based search (low confidence),
7. Country corresponding to the place of publication.

The resulting variable should be interpreted as a proxy for the author's cultural or national origin rather than a strictly defined legal nationality.

The resulting dataset is heavily centred towards the U.S. and UK authors. Figure 2 illustrates the distribution among countries. Figure 3 then shows how the country composition of the corpus evolves over time.

Figure 2: Distribution of author origins, top 20 by frequency.



(a) Raw Wikidata labels. (b) Consolidated country groups.

Notes: Panel a shows the raw historical-state taxonomy returned by Wikidata's P27 (country of citizenship); panel b shows the consolidation into modern country families via the mapping in `utils/country_grouping.py`, with Irish authors reassigned by birthplace and the Austria–Hungary bucket split using original-language signal.

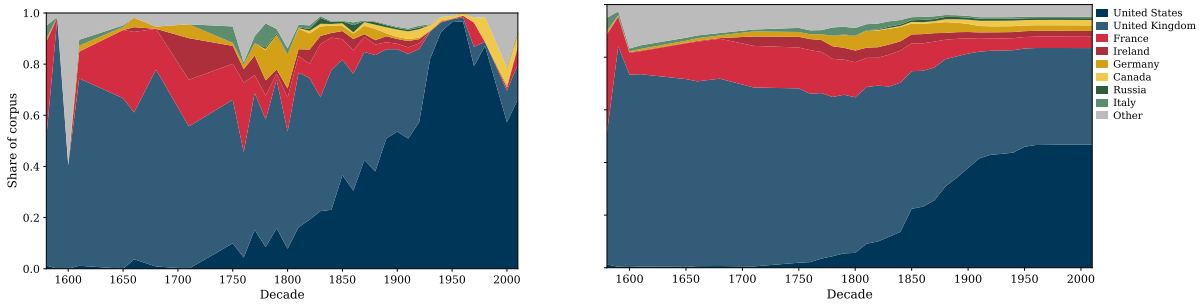
These procedures prioritise accuracy where high-quality sources are available while ensuring broad coverage through fallback mechanisms, at the cost of introducing measurement noise

[5]Confidence levels reflect the estimated reliability of the match between the author and the corresponding Wikidata entry, based on the available identifying information (e.g., name and birth year).

in the latter cases. For each observation, both the final assigned value and its source are recorded, allowing for transparency and potential robustness checks.

Throughout the process, the sample size is gradually lowered as the date of publication and author origin is determinable only for some parts of the corpus. The dataset reduction is captured in Figure 4.

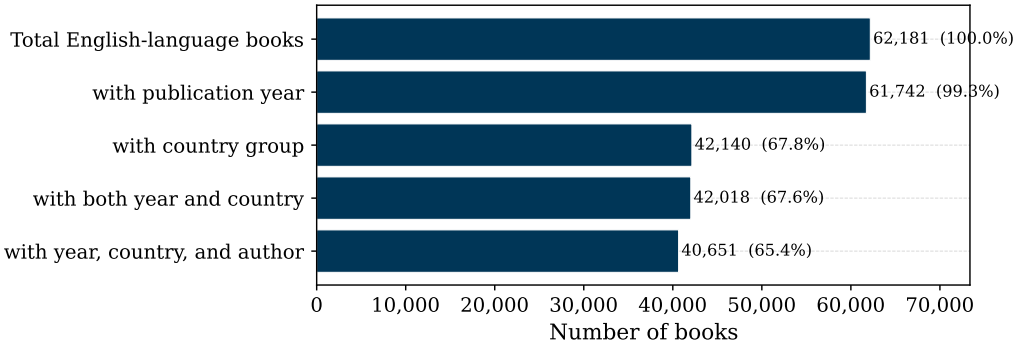
Figure 3: Author-origin composition of the corpus over time.



(a) Per-decade flow. (b) Cumulative.

Notes: Panel a shows the share of books published *within* each decade by country group; panel b shows the share of the corpus published *up to and including* each decade. Decades with fewer than 50 books are excluded. Color order is held consistent across panels.

Figure 4: Sample funnel after metadata enrichment.



Notes: Cascading sample size after the metadata-enrichment pipeline. Starting from the full English-language corpus, each row successively requires an additional metadata field to be non-missing.

The genre of each work is recovered by re-parsing the Resource Description Framework (RDF) for two fields: the Library of Congress Subject Headings (LCSH) (`dcterms:subject`, multi-valued) and Gutenberg's hand-curated bookshelf (`pgterms:bookshelf`, available for ∼30% of books). Each book is then assigned to one of 13 buckets[6] via a priority cascade: the curated bookshelf where present, otherwise an ordered keyword match against the subject list, with narrower categories (poetry, drama, science fiction) checked before broader ones.

For the later temporal analysis, I use a filtered sample from 1700 to 1950. This shorter sample ensures enough data points in each analysed time period, which would not be

[6]Poetry, Drama, six fiction subgenres, History/Biography, Religion/Philosophy, Law, Science/Medicine, and two catch-all categories.

possible outside of the sample. A measure of the precision of the publication year estimate is stored alongside the actual publication year.

3 Feature Extraction

Each book in the working corpus is represented by a fixed-length feature vector encoding three distinct dimensions of its textual content: emotional register, writing style, and semantic field composition. To capture within-book variation, features are first computed at the passage level and subsequently aggregated to the book level, following an approach analogous to Reagan et al. (2016). After full-corpus extraction, the feature matrix consists of 59,969 books × 79 features (76 numeric and 3 metadata).

Each book i is partitioned into P_i non-overlapping passages of fixed length $L = 512$ tokens, yielding $P_i = \lceil |W_i|/L \rceil$ passages per book. The chunk size is chosen to approximate a coherent narrative unit while remaining long enough for stable lexical frequency estimates.

Emotional content is quantified using the NRC Word-Emotion Association Lexicon (NRC) (Mohammad & Turney, 2013), which maps individual words to binary associations across ten dimensions: eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and two polarity classes (positive, negative). Each passage receives a score on each dimension equal to the share of its tokens present in the corresponding lexicon category. The lexicon is applied to lowercased surface forms without lemmatisation. Since it was constructed from contemporary English, some mappings may be anachronistic when applied to nineteenth-century prose; this is acknowledged as a limitation.

Similarly, writing style is captured through a set of surface-level and syntactic features computed using the spaCy natural language processing library (Montani et al., 2023), with the `en_core_web_lg` large English pipeline run on Central Processing Unit (CPU).[7] The following features are extracted per passage:

$$\text{pos}_{ip}^{\tau} = \frac{N_{ip}^{\tau}}{|T_{ip}|}, \qquad \tau \in \{\text{NOUN, VERB, ADJ, ADV}\}, \tag{1}$$

where N_{ip}^{τ} is the count of tokens with Part of Speech (POS) tag τ and T_{ip} is the set of alphabetic tokens in passage p of book i. In addition to these ratios, the feature set includes mean and variance of sentence length, Type-Token Ratio (TTR) as a measure of lexical diversity, negation rate, modal verb frequency as a proxy for expressed uncertainty, punctuation density, and past and present tense ratios as proxies for narrative orientation.

Furthermore, semantic fields capture what a text is *about* in dimensions of economic and historical relevance, complementing the affective signal of the emotion scores. This approach follows the tradition of text-based cultural measurement (Michel et al., 2011; Underwood,

[7] Because of computational and time requirements, I use CERGE-EI's HeavyHorse 4 server with an AMD EPYC 7642 48-core processor.

2019).

Eleven fields are defined: commerce & trade, religion, death & disease, nature, the body, collective framing, individual framing, technology, law & justice, family & kinship, and education. Each field is initialised from a seed word list and expanded via WordNet synsets (Miller, 1995) to cover inflected and synonymous forms. The full seed lists are reported in Appendix A. Scoring is performed on lemmatised passage tokens; each passage receives one normalised frequency score per field.

Finally, the characteristics are aggregated to the book level from individual passages. Let $\mathbf{x}_{ip} \in \mathbb{R}^d$ denote the full feature vector for passage p of book i, concatenating all three feature families. Book-level features are the mean and standard deviation across passages:

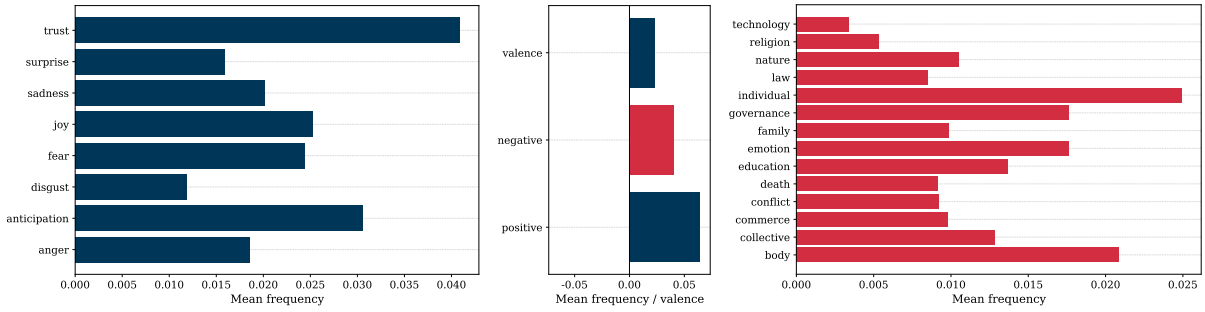
$$\bar{\mathbf{x}}_i = \frac{1}{P_i} \sum_{p=1}^{P_i} \mathbf{x}_{ip}, \qquad \boldsymbol{\sigma}_i = \sqrt{\frac{1}{P_i} \sum_{p=1}^{P_i} \left(\mathbf{x}_{ip} - \bar{\mathbf{x}}_i\right)^2}, \tag{2}$$

leading to the final book-level vector $\mathbf{z}_i = (\bar{\mathbf{x}}_i, \boldsymbol{\sigma}_i) \in \mathbb{R}^{2d}$. Computation of the standard deviation alongside the mean is motivated by Gorin, Heblich, and Zylberberg (2025), who show that the dispersion of emotion scores carries independent information: a book with uniformly high fear reads differently from one that spikes only in select passages, even if both share the same passage-mean.

4 Results of Feature Extraction

The feature extraction has already provided many meaningful outputs. Figure 5 presents the results from emotion scoring and the distribution among the semantic fields.

Figure 5: Mean book-level feature values, full corpus.



(a) NRC emotions. (b) Polarity. (c) Semantic fields.

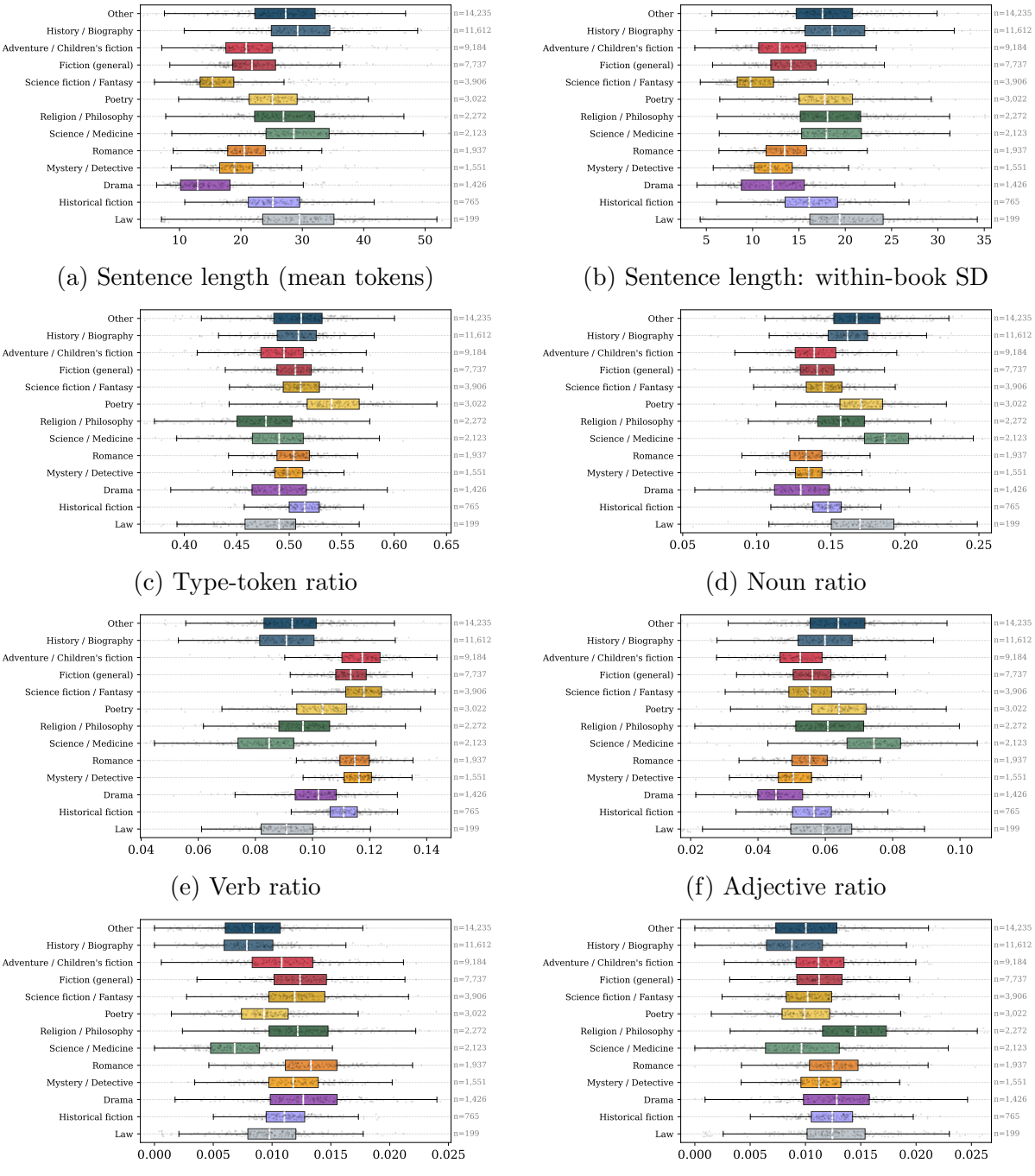
Notes: Sample size: $N = 59{,}969$ books after feature extraction. The sample uses all books regardless of the availability of publication year and author's country group.

Positive emotions and fields dominate the corpus, which in turn yields a positive average valence. *Disgust* and *surprise* are the least frequently expressed emotions, while *trust* and *anticipation* are the most frequently expressed.

Individual and *body* semantic fields dominated the corpus. This may be due to many

instances of epic prose, fiction, romances, etc., which often focus on individuals. Furthermore, the *body* semantic field partly captures physical description, intimacy, gesture, and embodied emotional expression, all of which tend to be explicitly represented in character-driven prose narratives. On the contrary, *religion* and *technology* scored the lowest, as scientific manuals and religious texts are less common than epic prose, lyrical poems, or drama.

Figure 6: Stylometric features across literary genres.



(a) Sentence length (mean tokens)

(b) Sentence length: within-book SD

(c) Type-token ratio

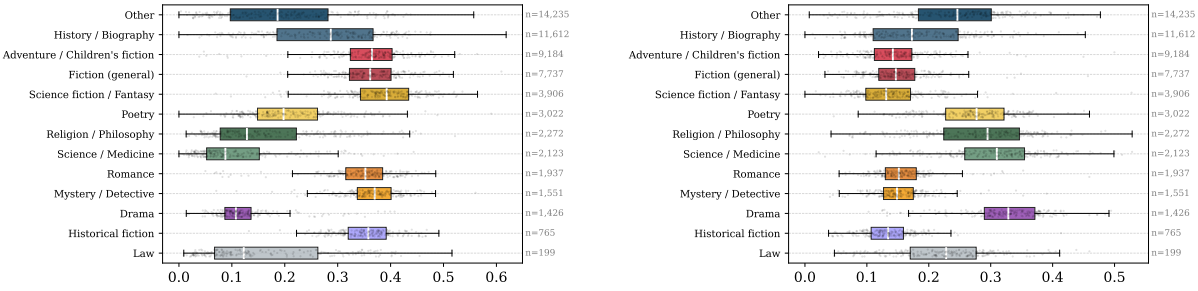
(d) Noun ratio

(e) Verb ratio

(f) Adjective ratio

(g) Negation rate

(h) Modal verb frequency



(i) Past tense ratio

(j) Present tense ratio

Notes: Box plots show the distribution of features for genres with $n \geq 30$. White vertical lines represent the median. Outliers are hidden to preserve scale; instead, a representative random sample of up to 200 books is shown as jittered points for each genre to illustrate density. Color-coding is consistent across all panels.

Figure 6 then shows selected stylometric features across literary genres. Unsurprisingly, *drama* uses the shortest sentences on average, while historical, biographical, and scientific or medical texts use the longest. The *type-token ratio* is relatively similar across genres, with *poetry* standing out for slightly higher lexical diversity. *Scientific* texts contain the highest noun share, whereas epic prose (*fiction, drama, romance, ...*) relies more heavily on verbs. Adjectives are also more common in *scientific* texts, while *drama*, characterised by direct speech, uses them less frequently.

Religious texts feature a higher share of modal verbs, likely reflecting the conditional, prescriptive, and interpretive language characteristic of many religious writings. A feature shared by both *drama* and *scientific* texts is the relatively low use of past tense. Somewhat surprisingly, the same pattern appears in *religious* texts, where one might expect at least an average past-tense ratio given the focus on biblical and historical narratives. These genres consequently use fewer present-tense constructions, whereas epic prose tends to behave in the opposite way.

5 Fixed-Effects Partialling

The book-level feature matrix from Section 3 cannot be directly used for temporal analysis because the composition of the corpus changes over time and across countries. A raw temporal pattern in some feature may therefore reflect changes in the types of books being published rather than changes in the actual content of the books themselves. Following Gorin, Heblich, and Zylberberg (2025), who deal with a similar issue in the context of paintings, I residualise all numeric features on year, country, and genre fixed effects before any further analysis. The resulting residualised matrix is then used in Section 7.6.

To illustrate the intuition, consider a book published in 1885 in the United States, classified as adventure fiction, with a relatively high *fear* score. There are several possible explanations for this. The book may come from a period in which literature was generally more fear-oriented, it may originate from a country where authors used more fear-related language on average, or it may simply belong to a genre that tends to be emotionally darker.

The goal of the fixed-effects approach is to remove these systematic composition effects and isolate the remaining variation that is specific to the individual book.

The econometric intuition follows the Frisch–Waugh–Lovell theorem. Regressing a variable on a set of controls and then analysing the residual variation is numerically equivalent to estimating the full specification directly. In practice, this means that removing the fixed-effect structure beforehand does not bias later analyses that condition on the same variables. Instead, it makes the remaining variation easier to interpret.

In the event-study framework used later in Section 7.6, a residualised feature can therefore be interpreted as a deviation from its expected decade–country–genre baseline.

5.1 Specification

Let y_i denote one of the 76 numeric features for book i, including emotion measures, polarity scores, stylometric variables, and semantic-field frequencies, each measured both as means and within-book standard deviations. The main specification is

$$y_i \;=\; \alpha_{\text{decade}(i)} + \beta_{\text{country}(i)} + \gamma_{\text{genre}(i)} + \varepsilon_i, \tag{3}$$

estimated using Weighted Least Squares (WLS) with weights

$$w_i \;=\; \frac{1}{\left[\text{precision}_i\right]^2}, \tag{4}$$

where precision_i corresponds to the uncertainty measure generated during the publication-year cascade described in Section 2. This ranges from ± 2 years for Wikidata-based dates to roughly ± 25 years for broad summary-based approximations. Weights are normalised to have a sample mean equal to one. As a result, books with more precise publication years receive greater importance in the Fixed Effects (FE) estimation, while books with uncertain dates mainly contribute additional noise.

Country and genre enter as full sets of dummy variables. Books with missing country information are assigned to a separate `Unknown` category so they are not dropped from the sample, which would otherwise remove approximately 32% of observations. Decade is defined as the floor of `Publication Year / 10`, while the alternative coarse-time specification uses 25-year bins. The final sample consists of roughly 54.8K books satisfying the year restrictions ($1700 \le y \le 1950$, precision ≤ 25), consistent with the sample used in Section 6.

For each feature, the stored residualised value is

$$\tilde{y}_i \;=\; \widehat{\varepsilon}_i \;+\; \bar{y}^{\,w}, \qquad \bar{y}^{\,w} \;=\; \frac{\sum_i w_i\, y_i}{\sum_i w_i},$$

that is, the estimated residual shifted back by the weighted grand mean. This keeps the transformed features on approximately the same scale as the original variables and allows

the residualised parquet files to be used in downstream code without additional adjustments. For ratio variables, the shift also helps preserve the approximate $[0, 1]$ range expected by some of the cluster-labelling heuristics, although extreme observations may still slightly exceed these bounds.

Table 1: The four FE specifications.

Tag	**Bin width**	**FE columns**	**Role**
`dec_yc`	decade	year, country	robustness (no genre)
`dec_ycg`	decade	year, country, genre	**headline**
`25y_yc`	25 year	year, country	robustness (coarse time)
`25y_ycg`	25 year	year, country, genre	robustness (coarse time)

Notes: Each produces its own residualised feature parquet. The headline `dec_ycg` specification is used in the later analysis; the remaining specifications are reported in the appendix.

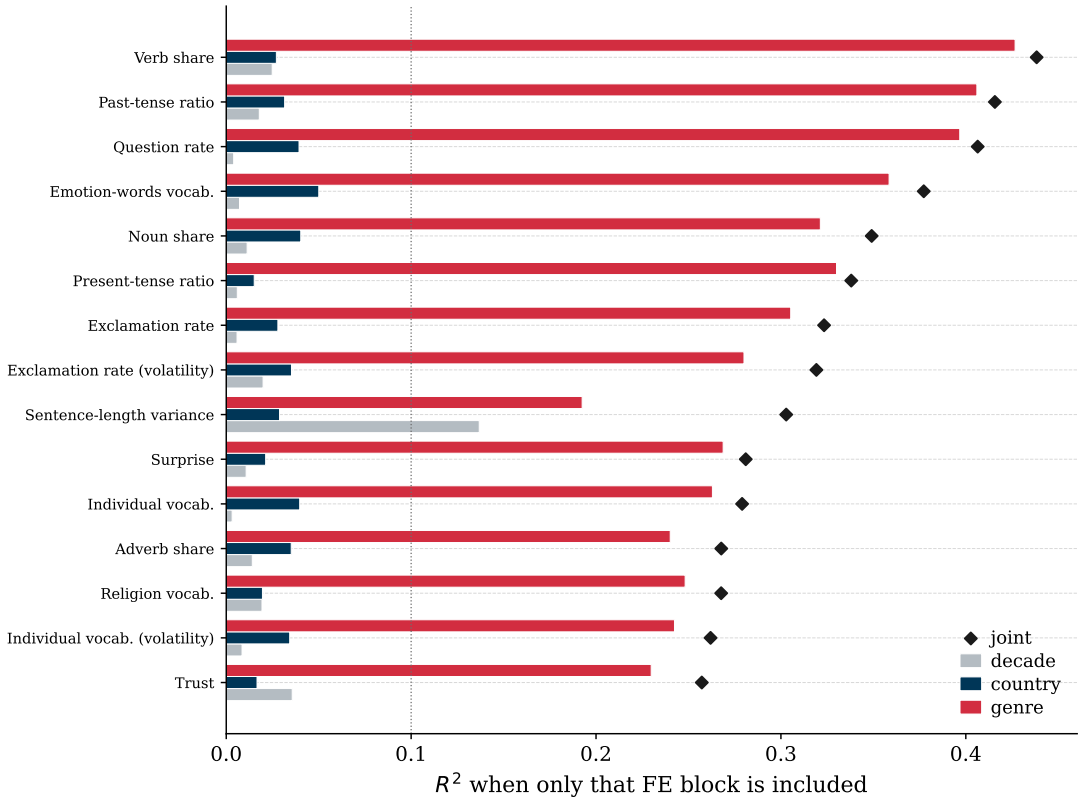
Four related specifications are estimated simultaneously (Table 1). The main specification, `dec_ycg`, includes decade, country, and genre fixed effects. The remaining specifications serve as robustness checks, either using coarser time bins or excluding genre controls.

5.2 Results: Composition R^2 and Block Decomposition

Most of the explainable composition variance comes from genre, while decade fixed effects contribute relatively little. Intuitively, this is not surprising. A romance novel and a history book published in the same year differ substantially more in emotional content and vocabulary than two history books published several decades apart.

This becomes even clearer in the per-block R^2 decomposition shown in Figure 7, where Equation (3) is re-estimated separately for each FE block. For the fifteen most composition-driven features, the joint R^2 values (black diamonds) are already very close to the genre-only estimates (red bars), while decade-only and country-only specifications explain relatively little variation. Religion-related vocabulary provides the clearest example: $R^2_{\text{joint}} \approx 0.27$, of which roughly 0.25 is explained by genre alone, leaving only a small remaining share attributable to decade and country effects.

Figure 7: Per-FE-block R^2 decomposition.



Notes: Top-15 features ranked by the joint R^2 from Equation (3). The black diamond represents the total composition share, while the coloured bars show the explanatory power of individual FE blocks estimated separately (decade = gray, country = navy, genre = red). Because the FE blocks are not orthogonal, the bars do not sum exactly to the joint R^2. The dotted reference line at $R^2 = 0.10$ serves as a rough threshold for features that are meaningfully composition-driven. Genre clearly dominates across all reported features, while decade and country effects remain comparatively small ($\lesssim 0.02$ each).

These results have two main implications. First, from a methodological perspective, residualising the data is essential because the event-study analysis in Section 7.6 would otherwise be driven primarily by changing genre composition. Second, the results also show that a simpler specification using only year and author fixed effects, as originally envisaged in the project proposal, would have absorbed only a small fraction of the relevant variation. The full year × country × genre specification is therefore what makes the later temporal comparisons meaningfully content-comparable.

6 Clustering and Network Construction

The feature matrix is used for two related clustering exercises. First, I cluster books separately within the three main feature spaces: emotion, style, and theme. Second, I construct a literary-similarity network and use Leiden community detection to identify groups of similar authors and books.

These two approaches answer slightly different questions. The per-space k-means clusters describe the main types of feature profiles in the corpus. The network approach, in contrast,

asks which books or authors are close to each other when style and theme are considered jointly. Both analyses are run on the year-filtered sample of 54,769 books that remain after imposing the $[1700, 1950]$ window and the precision threshold ≤ 25, joined with the author, country, and genre information described in Section 2.

6.1 Per-Space Clustering

The 76 numeric book-level features fall into three natural groups: 22 NRC emotion and polarity features (`nrc_*`), 26 stylometric features (`style_*`), and 28 semantic-field features (`sem_*`). Each group contains both passage-level means and within-book standard deviations.

I keep these three feature spaces separate rather than running one combined Principal Component Analysis (PCA). A single combined PCA would mix emotional, stylistic, and thematic information, which would make the resulting clusters harder to interpret. In an earlier prototype, the combined PCA mainly produced stylometric clusters, while the emotion and semantic features appeared only in smaller components. Treating the three spaces separately therefore gives more transparent results.

The pipeline is the same for each feature space $b \in \{\text{emotion}, \text{style}, \text{theme}\}$. First, the features are standardised to mean zero and unit variance. Second, I run PCA and retain 95% of the variance. Third, I estimate k-means for each candidate $k \in \{4, 5, \ldots, 20\}$ with `n_init=10`. Finally, I choose k using several diagnostics: rank consensus over silhouette, Calinski–Harabasz, and Davies–Bouldin scores; Gaussian Mixture Model (GMM)-Bayesian Information Criterion (BIC); and GMM-Integrated Completed Likelihood (ICL). The final choice is then checked manually against the diagnostic output.

The headline run, labelled `first`, uses $k = 4$ for emotion, $k = 8$ for style, and $k = 6$ for theme. Table 2 summarises the main diagnostics. Figures 8– 10 show the resulting partitions in two-dimensional Uniform Manifold Approximation and Projection (UMAP) projections of the corresponding PCA spaces. The UMAP projections are used only for visualisation; the clustering itself is performed on the full retained PCA space.

Table 2: Per-space clustering diagnostics for the headline `first` run.

Space	n_{feat}	PCA dims	Var. ret.	k	10-seed ARI	Bootstrap J
Emotion	22	12	0.955	4	0.985	0.918
Style	26	18	0.958	8	0.971	0.901
Theme	28	19	0.953	6	0.996	0.943

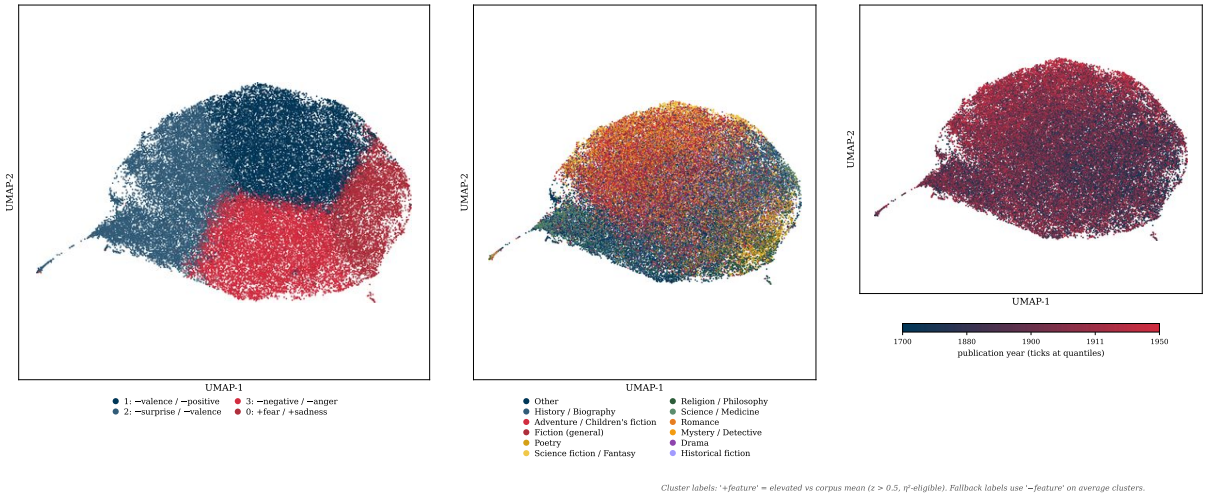
Notes: Stability is reported in two ways: mean pairwise Adjusted Rand Index (ARI) across 10 random-seed re-runs of k-means, and mean per-cluster bootstrap Jaccard (J) across 100 resamples (Hennig, 2007). All clusters exceed the $J \geq 0.85$ stability threshold except one style cluster, which has $J = 0.71$ and is therefore flagged in the diagnostics.

From the UMAP projections, several patterns are visible. In Figure 9, the cluster labels

printed in the legend record deviation directions, while the colours themselves merely indicate cluster identity. The more negation-heavy and interrogative style clusters appear to be associated mainly with epic prose. The third panel also suggests that these books are more common toward the later years of the sample.

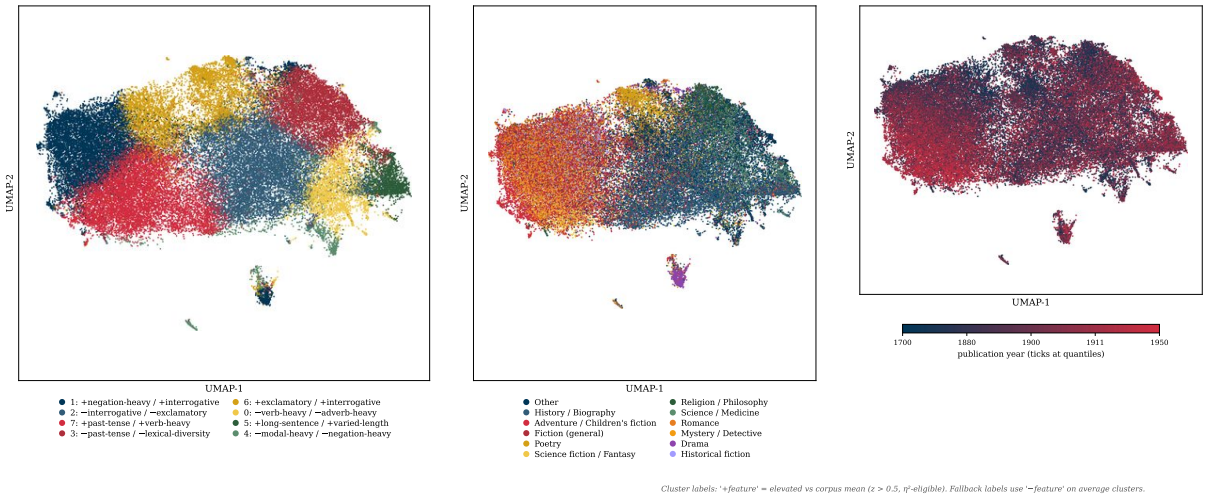
By contrast, the older part of the sample appears to contain more exclamatory and interrogative texts, many of which are classified as poetry. Figure 10 also suggests that poetry is associated with a stronger individual thematic dimension.

Figure 8: UMAP embedding of the emotion-space k-means partition in the headline `first` run ($k = 4$).



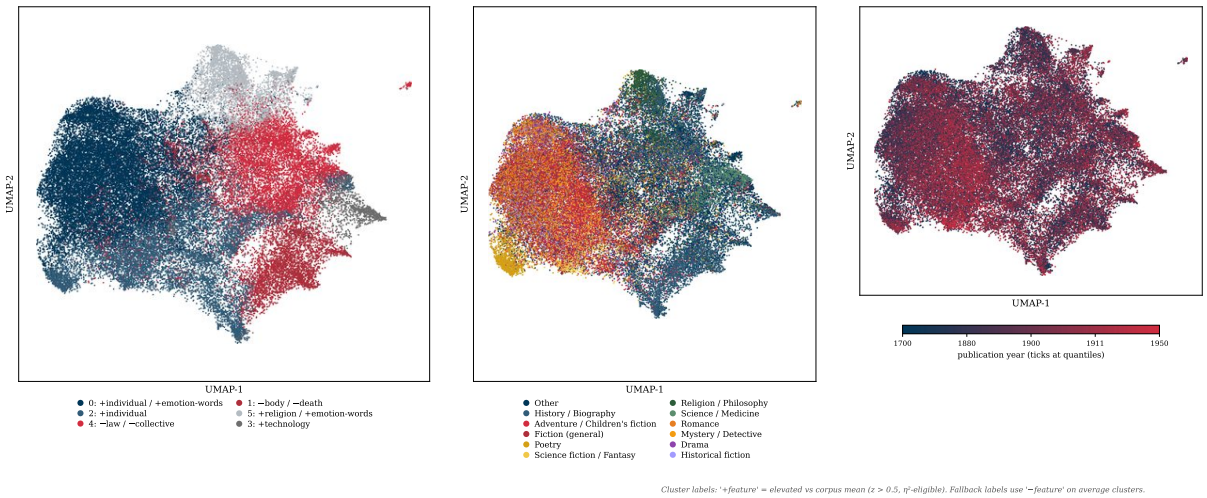
Notes: Each point is one book. Colours encode k-means cluster identity.

Figure 9: UMAP embedding of the style-space k-means partition in the headline `first` run ($k = 8$).



Notes: As in Figure 8, for the stylometric feature block.

Figure 10: UMAP embedding of the theme-space k-means partition in the headline `first` run ($k = 6$).



Notes: As in Figure 8, for the semantic-field feature block.

6.2 Leiden Network Communities

The network analysis complements the per-space clustering. While k-means groups books within a specific feature family, the network approach asks which books or authors are similar when style and theme are considered jointly. This follows the logic of Gorin, Heblich, and Zylberberg (2025), who use network structure to study similarity among paintings.

The author-level network is the main specification. I first aggregate books to authors using the precision-weighted mean of their features and drop authors represented by only one book (`min_books = 2`). This leaves 5,617 authors. The stylometric and semantic-field blocks are then reduced separately by PCA, each to approximately 16 components retaining 95% of the variance. Reducing the two blocks separately prevents the larger block from dominating the embedding. I exclude the emotion block because emotion is treated as an outcome of community structure rather than as an input defining that structure.

The resulting 32-dimensional author embedding is L_2-normalised, so inner products correspond to cosine similarities. I then build an exact k-nearest-neighbour graph with $k_{\text{NN}} = 105$ using Facebook AI Similarity Search (FAISS). Edges are union-symmetrised and weighted by cosine similarity. Communities are recovered using Leiden (Traag, Waltman, & Eck, 2019) with the modularity quality function, $\gamma = 1$, and edge weights given by the similarity scores. The book-level network is constructed in the same way, except that the author-aggregation step is omitted.

Table 3 reports the basic network statistics. The author-level network has 7 communities and modularity 0.573, while the book-level network has 14 communities and modularity 0.664. The cross-seed Leiden ARI is 0.855 at the author level and 0.795 at the book level, suggesting that the recovered community structure is not mainly driven by random initialisation.

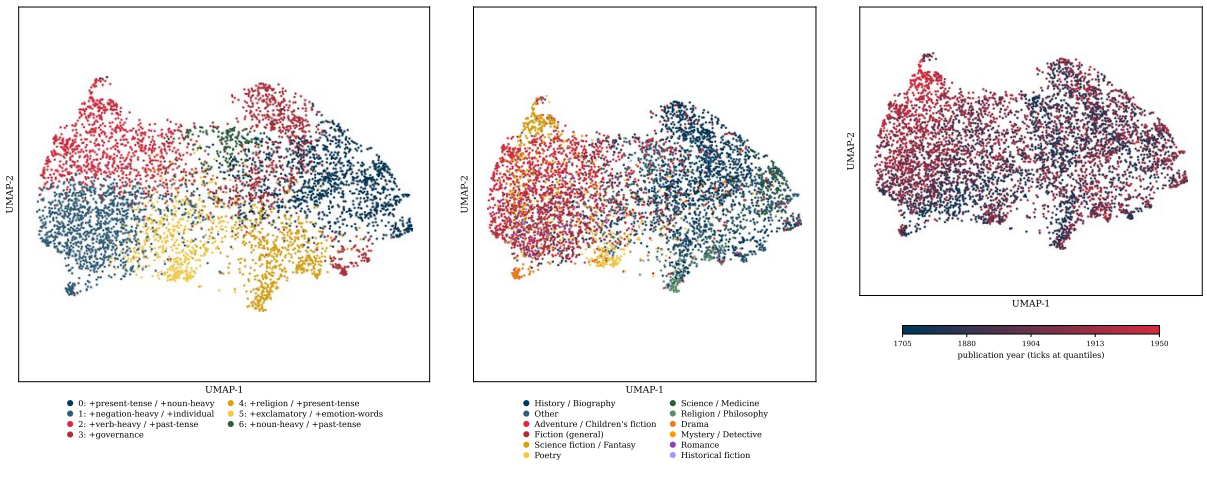
Table 3: Network statistics.

Quantity	Author level	Book level
Entities	5,617	54,769
Edges	410,475	4,301,496
Mean degree	146.2	157.1
Modularity	0.573	0.664
# communities	7	14
Leiden 10-seed ARI	0.855	0.795

Notes: Edges are union-symmetrised cosine k-nearest-neighbour edges from the `first` run. The book-level network is denser because no author aggregation is applied. The author-level network is treated as the main specification because authors are the more natural unit for studying literary communities.

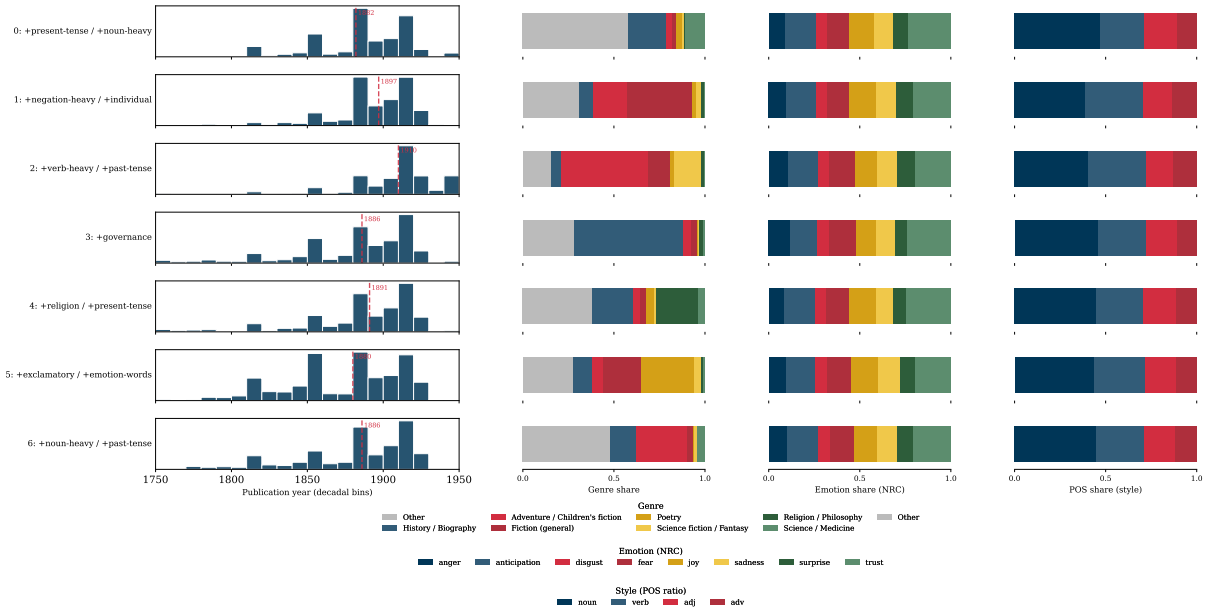
Several patterns emerge from Figure 11. First, authors with higher verb share and past-tense ratio appear to be concentrated together. These authors are also mostly associated with *science fiction/fantasy* and other fiction genres, which are among the newest genres in the corpus. Similarly, authors using stronger *governance* vocabulary seem to cluster close to the *history/biography* category. Finally, authors employing more exclamatory language and emotion words again appear close to poetry. Figure 12 presents a more concise summary of each cluster's characteristics.

Figure 11: Author-level Leiden communities in UMAP space (`first` run).



Notes: Each point represents one author with at least two books in the corpus. The UMAP embedding is computed from the 32-dimensional block-PCA author features, and colours indicate Leiden community identity. Seven communities are recovered with modularity 0.573.

Figure 12: Author-level community dashboard (`first`).



6.3 Robustness, Residualised Re-Clustering, and Cluster Labels

A central concern in clustering exercises is that results may depend too much on the chosen number of clusters or on random initialisation. I therefore run the full clustering procedure under several alternative configurations. These runs vary the number of clusters in the emotion, style, and theme spaces and, in one case, the network neighbourhood size. The details of the four robustness runs are reported in Appendix C.1.

I evaluate stability in three ways. First, within each run, k-means is re-estimated with 10 different random seeds. Second, I use a 100-iteration bootstrap Jaccard procedure to check whether clusters remain stable under resampling.[8] Third, I compare partitions across the different robustness runs using pairwise ARI. The headline run is highly stable in all three feature spaces: the 10-seed ARI is above 0.97 and the mean bootstrap Jaccard exceeds 0.85 for all spaces, although one small style cluster falls below the cluster-level threshold.

As an additional check, I repeat the clustering pipeline on the Stage 4 residualised feature parquet `book_features_resid_dec_ycg.parquet`. This produces a fifth run, `resid_dec_ycg`, using the same cluster counts as the headline `first` run. This comparison asks whether the cluster structure survives after removing year × country × genre composition. If the ARI between `first` and `resid_dec_ycg` is high, the clusters mainly capture within-cell content differences. If it is low, then the raw clusters were likely driven by genre, country, or time composition rather than by the textual patterns of interest. The cross-run comparisons are reported in Appendix C.1.

Finally, I assign interpretable labels to the recovered clusters using a simple η^2-based

[8]The reported robustness runs were selected after checking that the resulting partitions were stable under the diagnostics described here. I therefore use them to illustrate sensitivity to alternative choices of k and k_{NN}, rather than to search for an unstable specification.

heuristic. I first identify the features that separate the clusters most clearly, based on a one-way ANOVA of each feature on cluster identity. For each cluster, I then keep the most relevant of these features and use the direction of their deviation from the average cluster centroid to form the label. This produces labels such as "+fear / +sadness", "−valence / −positive", or "−negative / −anger". Additional details on the labelling procedure and the corresponding cluster-discrimination tables are reported in Appendix C.3.

7 Temporal Event Studies

With the residualised features from Section 5 and the cluster structure from Section 6, the next question is whether the textual features change around major historical shocks after composition has been partialled out. To test this, I use a simple shock-window pre/post regression with clustering-robust inference across 12 historical events covered by the corpus.

7.1 Estimator and Shocks

For each historical shock $t^\star$ and each residualised feature $\tilde{y}$, I restrict the sample to books published within the window $[t^\star - 5,\, t^\star + 5]$ and estimate

$$\tilde{y}_i \;=\; \alpha + \beta \cdot \mathbb{1}\big[\text{year}_i \geq t^\star\big] + \varepsilon_i, \qquad w_i \;=\; \tfrac{1}{[\text{precision}_i]^2}, \tag{5}$$

by precision-weighted least squares. The coefficient of interest is β, which captures the average change in the residualised feature after the onset of the shock. Since $\tilde{y}$ has already been purged of year × country × genre composition, β can be read as the shock-specific deviation from the smooth decade–country–genre baseline.

A simple analytical WLS standard error is likely to be too optimistic. Books published in the same year may share unobserved conditions, such as censorship, paper supply, publication delays, or a general wartime atmosphere. I therefore report a block-bootstrap-by-year standard error. In each of 200 replications, publication years within the event window are resampled with replacement and Equation (5) is re-estimated on the year-resampled book sample. Across the 912 event studies (76 features × 12 shocks), the bootstrap standard error is on average about 2.4× larger than the analytical WLS standard error, so the correction is empirically important. The reported 95% confidence intervals are Wald-style, $\hat{\beta} \pm 1.96 \cdot \text{SE}_{\text{boot}}$, which keeps the forest plots easy to interpret. The percentile-bootstrap interval $[Q_{0.025}, Q_{0.975}]$ is also stored for transparency.

Table 4: The 12 historical shocks used in Section 7.6.

Year	Short	Full name	Type	Span
1756	7YW	Seven Years' War	war	1756–1763
1775	Am.Rev.	American Revolution	war	1775–1783
1789	Fr.Rev.	French Revolution	revolution	1789–1799
1803	Napoleon	Napoleonic Wars	war	1803–1815
1848	1848	1848 Revolutions	revolution	1848–1849
1861	Civil War	US Civil War	war	1861–1865
1873	1873	Long Depression	economic	1873–1879
1893	1893	Panic of 1893	economic	1893–1897
1914	WWI	World War I (*headline*)	war	1914–1918
1918	Flu	Spanish Flu	pandemic	1918–1920
1929	1929	Great Depression	economic	1929–1939
1939	WWII	World War II	war	1939–1945

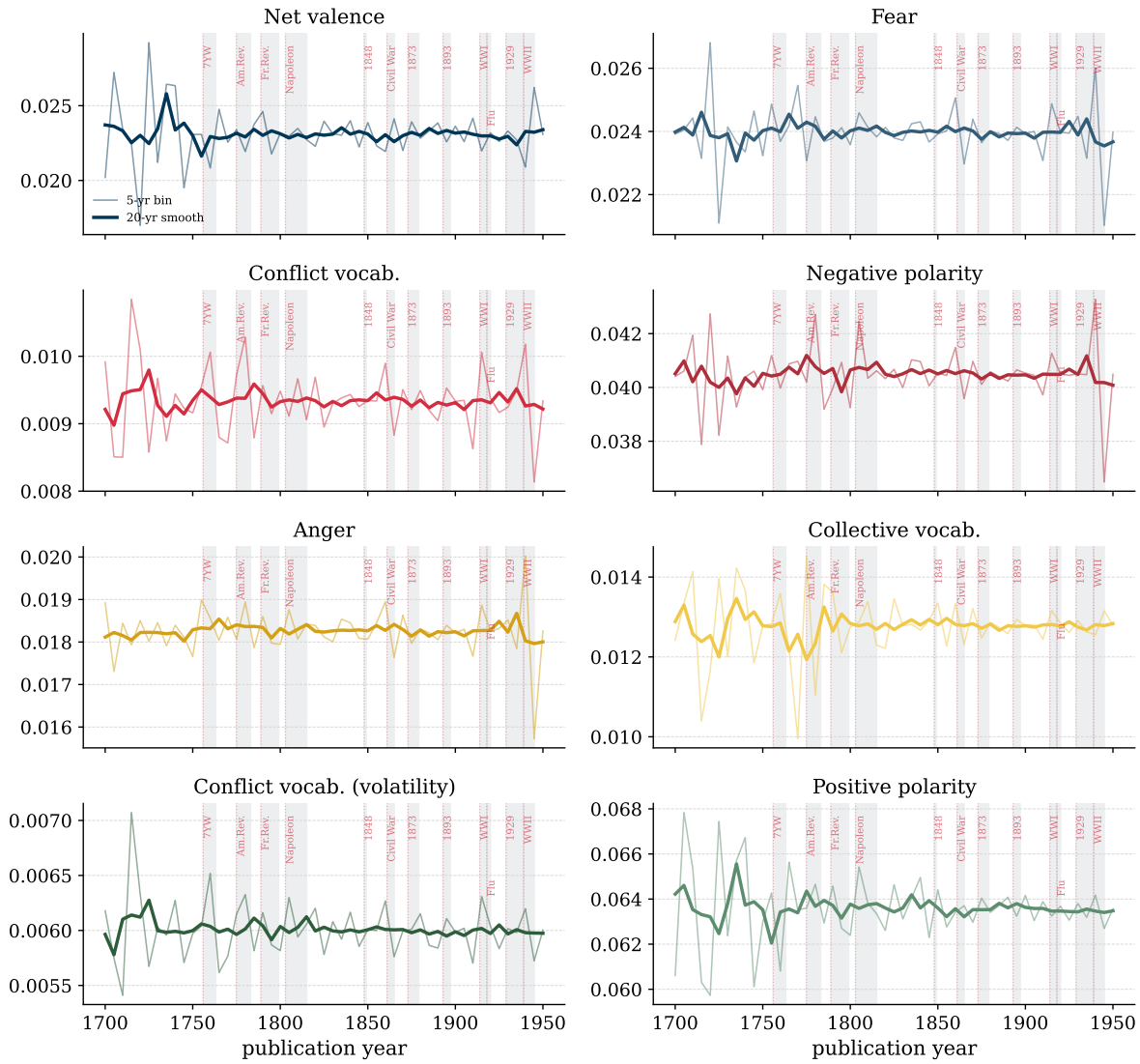
Notes: The *Year* column is the event onset $t^\star$ used in Equation (5); the *Span* column is shown in the figures as a grey shaded interval but does not enter the regression.

Two caveats are important. First, the WWI (1914) and Spanish Flu (1918) windows overlap by four years. Their coefficients are therefore correlated and should be interpreted jointly. Second, the pre-1820 shocks (Seven Years' War, American Revolution, French Revolution, and Napoleonic Wars) have relatively thin samples, around $\sim$ 100–400 books per ± 5 year window. These estimates are reported for completeness, but their confidence intervals are wide and they should not be interpreted too strongly.

7.2 Descriptive Temporal Trends

Before the formal regressions, Figures 13 and 14 show 5-year-bin weighted means of the top twelve residualised features by $|\hat{\beta}|$ at WWI. A 20-year centred rolling smoother is added to show the longer trend. Grey shaded areas mark the duration of shocks and small rotated labels mark shock onset.

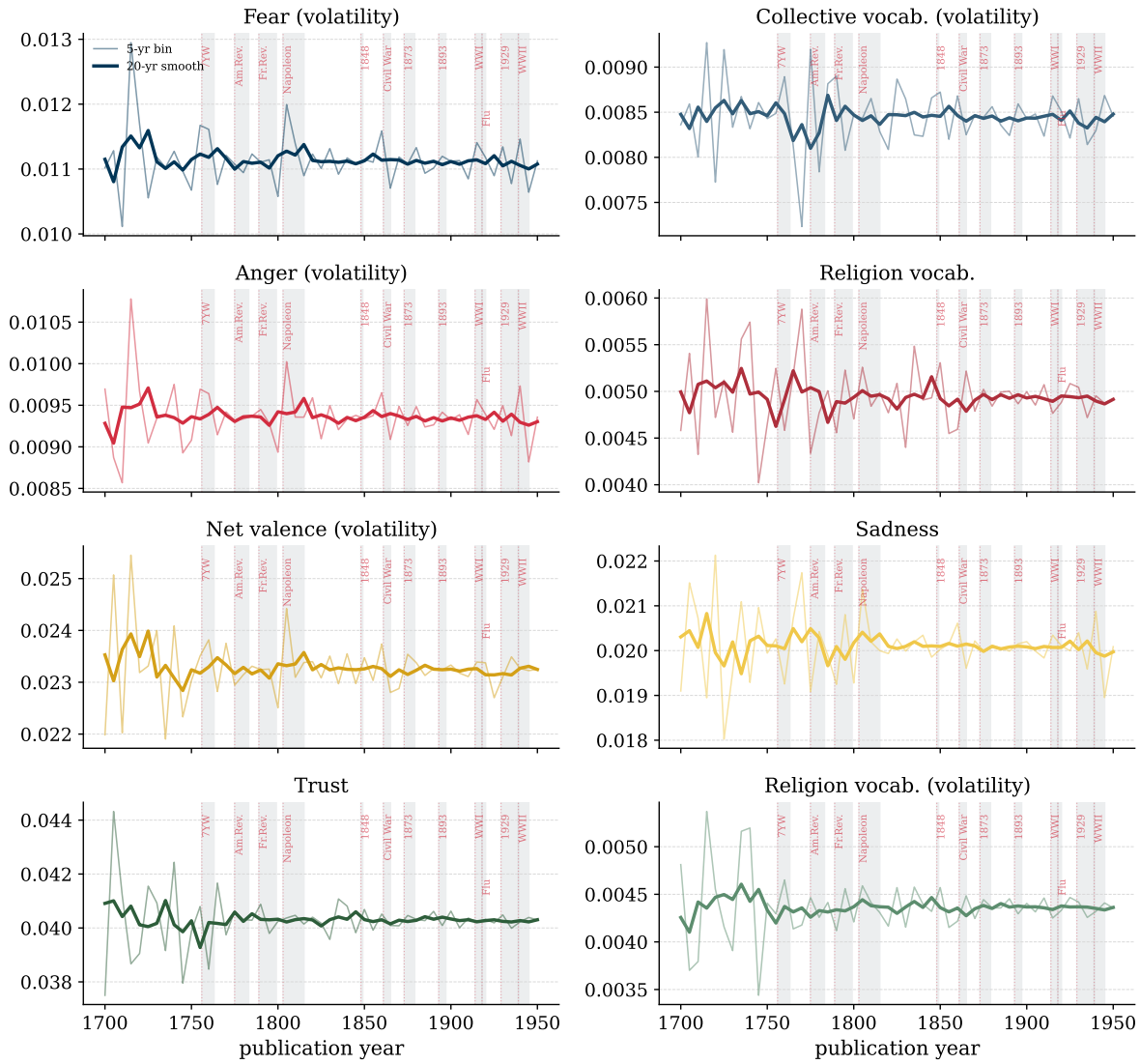
Figure 13: Temporal trends of the top residualised features by $|\hat{\beta}|$ at WWI 1914: features 1–8.



Notes: For each feature, the light line is the 5-year-bin precision-weighted mean of the residualised series; the heavy line is a 20-year centred rolling mean. Grey shaded areas mark shock duration; red dotted lines mark shock onset.

These figures give the first visual suggestion of the main result: fear, conflict vocabulary, and negative polarity increase around 1914–1918, while positive polarity and net valence decline. By contrast, semantic fields such as trust and religion vocabulary show weaker temporal structure.

Figure 14: Temporal trends of the top residualised features by $|\hat{\beta}|$ at WWI 1914: features 9–12.



Notes: As in Figure 13.

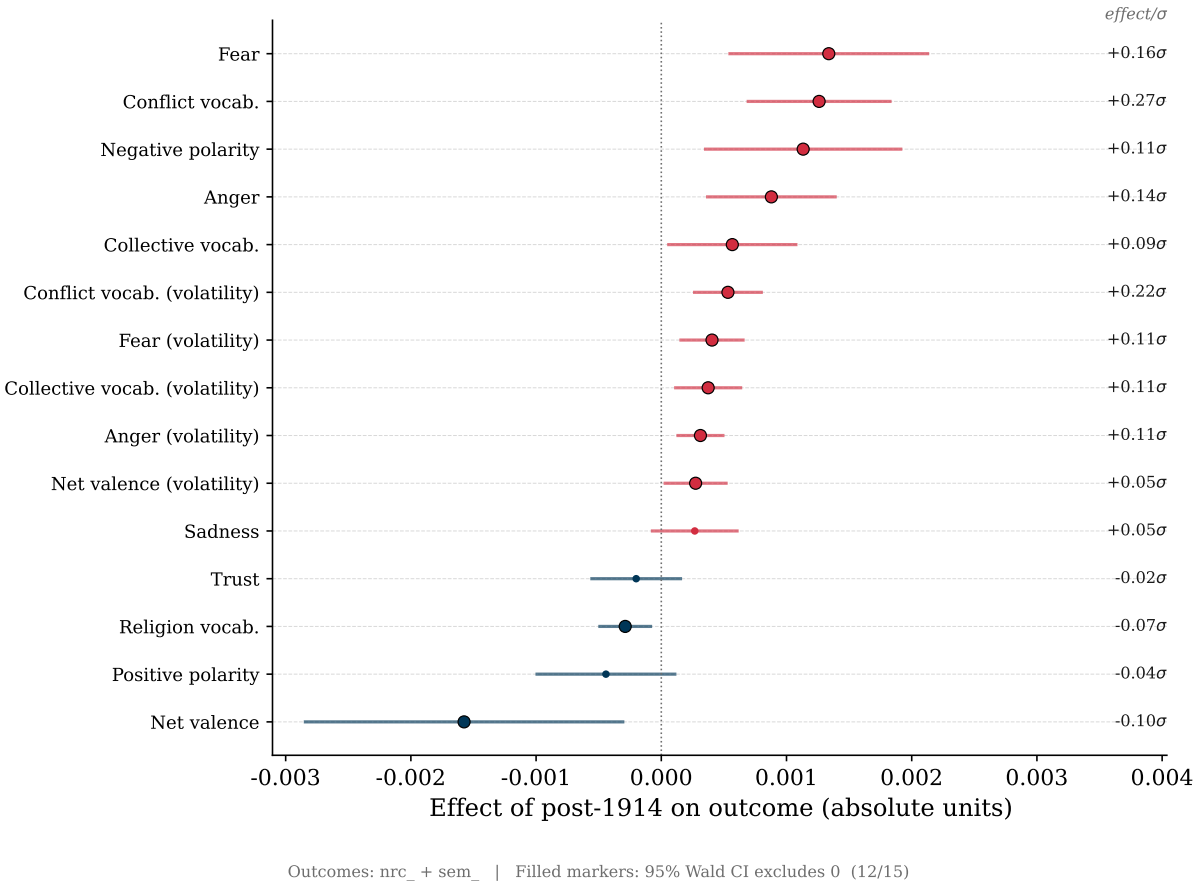
7.3 Headline Event Study: World War I

Figure 15 presents the headline event-study result: the top-20 features by $|\hat{\beta}|$ around 1914, restricted to emotion and semantic-field outcomes. Stylometric features are reported separately in Section 7.5. Coefficients are shown in absolute units on the main x-axis. The right-hand side of the plot also reports each coefficient in units of the within-sample SD of the residualised outcome, allowing absolute effects and effects in standard-deviation units to be compared.

The results show a clear post-1914 shift. Twelve of the top fifteen features have 95% Wald confidence intervals excluding zero. Expressed fear rises by about +0.16 SD, conflict

vocabulary by about +0.27 SD, expressed anger by about +0.14 SD, and negative polarity by about +0.11 SD. Net valence falls by roughly −0.10 SD. Taken together, these results suggest that post-1914 books are not only more likely to mention war-related themes, but are also more fearful and more conflict-oriented than their decade–country–genre baseline would predict. This is the central empirical result of the paper. It is also consistent with the four-run robustness sweep in Section 6.3 and the residualised re-clustering check.

Figure 15: WWI event study: top-20 residualised features by $|\hat{\beta}|$.



Notes: Rows are ordered by $|\hat{\beta}|$. Horizontal segments are 95% Wald confidence intervals from block-bootstrap-by-year standard errors (200 replications). Filled markers indicate intervals excluding zero. The right margin shows the effect in units of the outcome's within-sample residualised SD (the figure labels this axis σ). Outcomes are restricted to NRC emotion/polarity and semantic- field features; style features are reported in Section 7.5.

For each of the other eleven shocks, I also produce an individual forest plot so that each event can be discussed separately if needed. The remaining figures are included in the Appendix G for completeness.

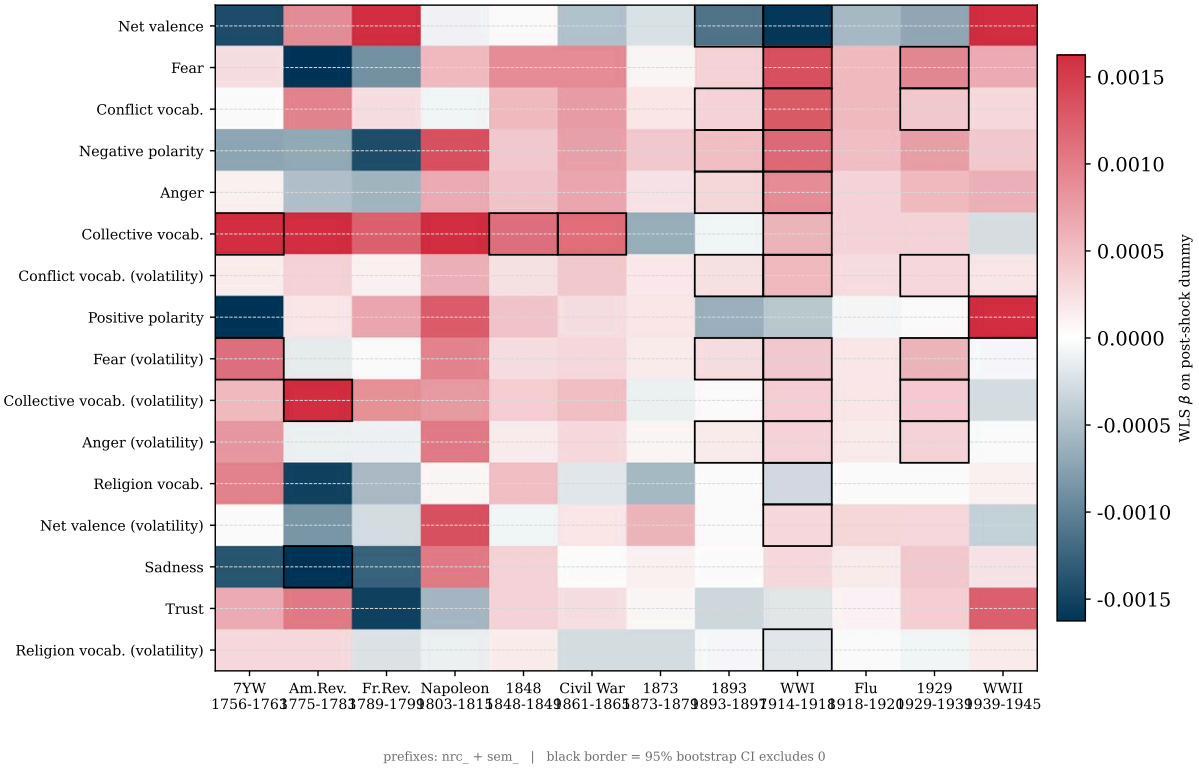
7.4 Cross-Shock Structure

The cross-shock heatmap in Figure 16 extends the WWI result to all twelve shocks. Rows show the top features by overall $|\hat{\beta}|$ across all events, while columns show shocks in chronological order. Cell colour captures the sign and size of $\hat{\beta}$, with red indicating positive coefficients and navy indicating negative coefficients. Cells with a black outline have a 95%

Wald Confidence Interval (CI) excluding zero.

Two patterns stand out. First, WWI is the strongest event in the feature-level analysis, with several significant positive shifts in fear, conflict, and negative language. Second, the pre-1820 shocks are especially noisy. Their coefficients vary in sign and only a few cells are statistically significant, which is consistent with the thin-sample caveat discussed in Section 7.1. The Great Depression and WWII show weaker but still negative-valence patterns. By contrast, the 1873 Long Depression and the 1893 Panic do not produce a clear emotional shift, suggesting that medium-size financial shocks are not strongly reflected in the emotional register of contemporary literature within a ±5-year window.

Figure 16: Event-study coefficients across all 12 shocks.



Notes: Rows are residualised features selected by overall $|\hat{\beta}|$ across shocks. Columns are the 12 shocks in Table 4. Colour encodes $\hat{\beta}$, with red indicating positive values and navy indicating negative values, on a common scale. Black-outlined cells have 95% Wald CI excluding zero. Style features are excluded; see Section 7.5 for their separate panel.

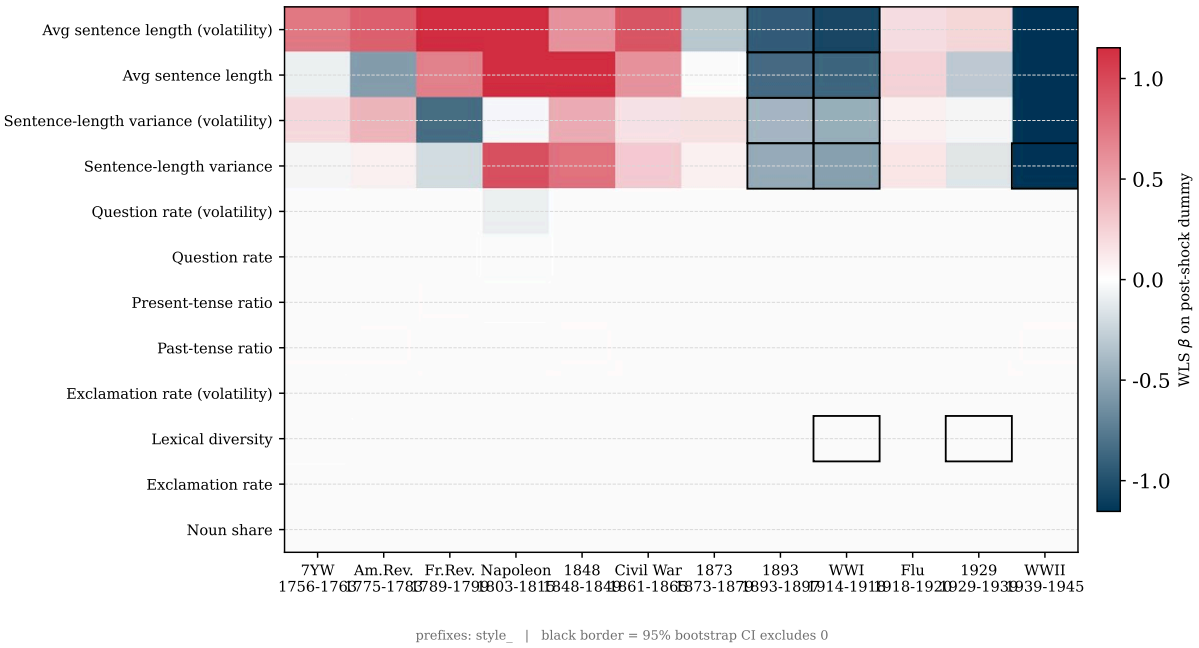
7.5 Style Features: A Placebo

Figure 17 repeats the same cross-shock analysis, but only for stylometric features. These features are empirically close to zero. Average sentence length, type–token ratio, noun/verb ratios, tense ratios, and modal/negation rates do not systematically move around discrete historical shocks within a ±5-year window.

This is reassuring for two reasons. First, it suggests that long-run literary modernisation, such as prose becoming shorter, denser, or more direct over the 19th and 20th centuries, is mainly a slow trend rather than an immediate response to shocks. This supports the use of

year FE in Section 5. Second, it justifies keeping the headline figures focused on emotion and semantic-field outcomes. Adding style features would mostly dilute the signal and increase the multiple-testing burden without adding much information.

Figure 17: Robustness: style features only.



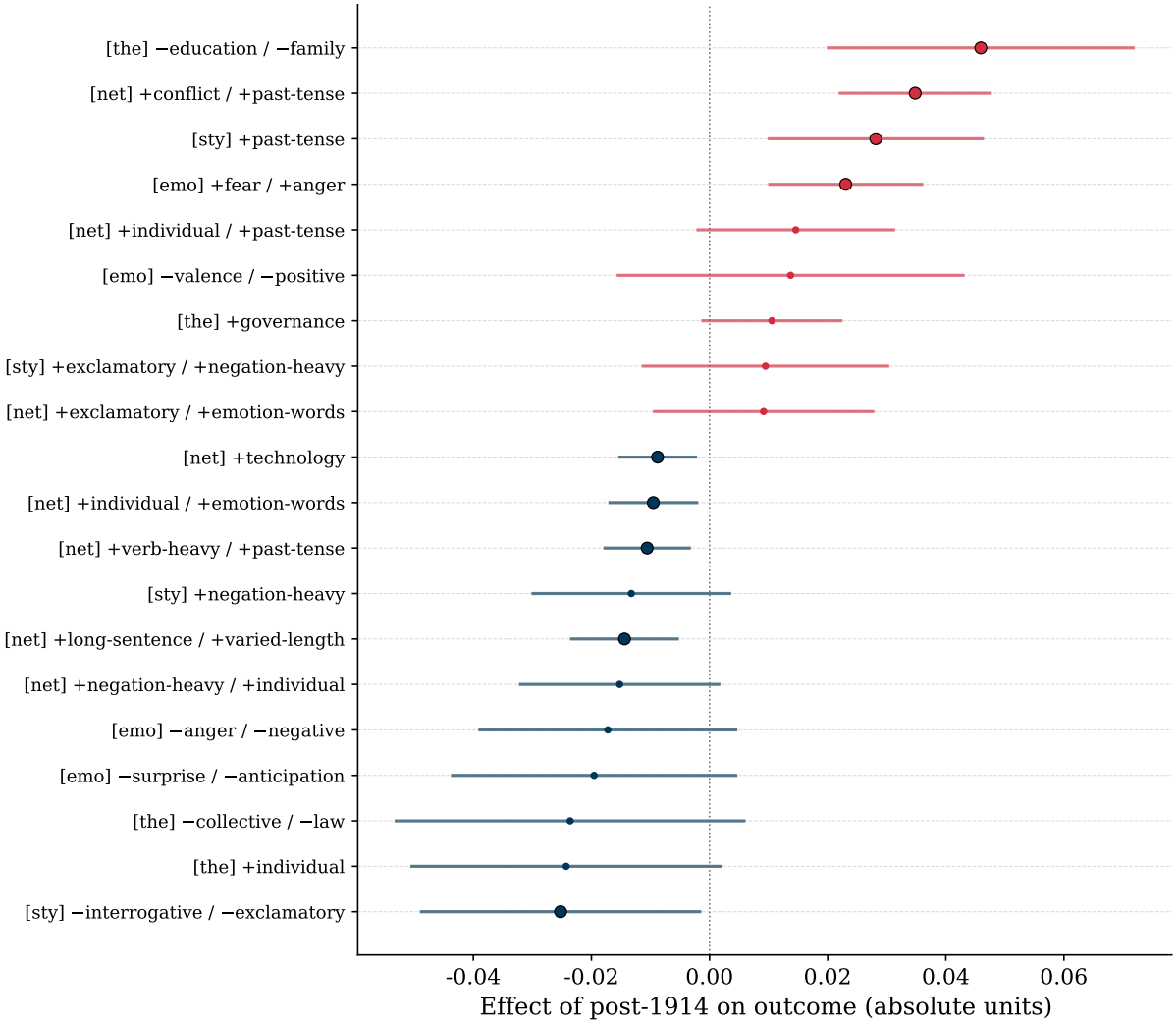
Notes: Analogous to Figure 16, restricted to stylometric features (`style_*`). Coefficients are concentrated near zero and only a few cells are statistically significant.

7.6 Cluster-Level Event Studies

Sections 7.3–7.5 focus on the within-feature question: does a given residualised feature $\tilde{y}$ shift around shock $t^\star$? A related question is whether the composition of books shifts toward particular clusters after a shock. For example, are post-WWI books more often classified into high-fear or war-related clusters than books in the pre-WWI baseline?

To answer this, I estimate the linear-probability version of Equation (5) for each cluster c across the four spaces and two network levels from Section 6. The outcome is the indicator $\mathbb{1}[\text{book } i \in \text{cluster } c]$, while the right-hand side and the bootstrap procedure remain the same. Figure 18 reports the WWI result, and Figure 19 reports the cross-shock grid. Cluster labels on the y-axis are the η^2-gated labels from Section 6.3.

Figure 18: Cluster-membership event study at WWI 1914.

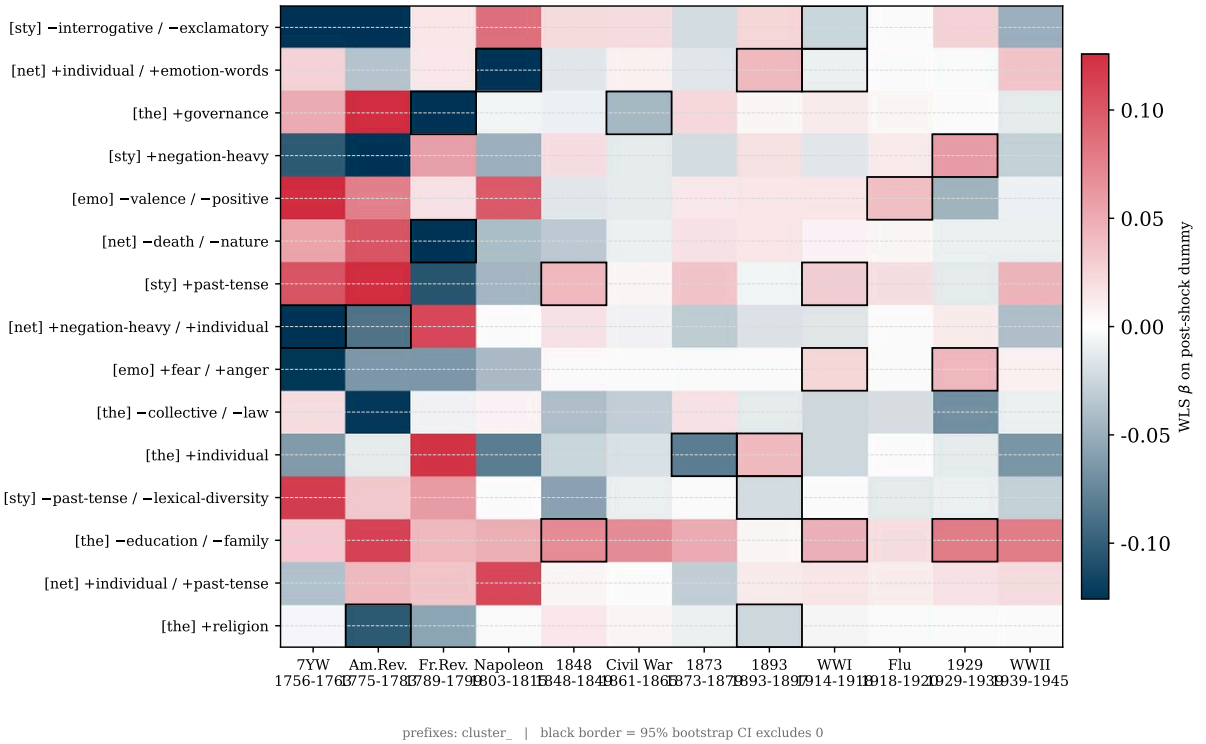


Outcomes: cluster_ | Filled markers: 95% Wald CI excludes 0 (9/20)

Notes: Outcomes are 0/1 indicators for cluster membership in each of the per-space and network partitions from Section 6. Cluster labels are the η^2-gated labels from `data/results/clusters/resid_dec_ycg/cluster_names.json`. Inference details follow Figure 15.

The cluster-level results support the feature-level findings. After 1914, books are more likely to belong to the "-education/-family" theme cluster and the "+conflict/+past-tense" network cluster. This pattern may signal growing scepticism toward humanist values and the emergence of the so-called "Lost Generation" - authors deeply affected by the war experience, whose disillusionment and trauma were reflected in their literary work.

Figure 19: Cluster-membership event studies across all 12 shocks.



Notes: Analogous to Figure 16, with cluster-membership outcomes replacing residualised feature outcomes. WWI again produces the clearest pattern, while pre-1820 shocks remain noisy.

8 Conclusion

Literature has never been isolated from its contemporary world. Literary theorists have long identified schools, movements, and periods, but this project approached literary change from a different angle. Instead of classifying texts primarily through close reading or established literary categories, I used quantitative and machine learning methods to explore whether large-scale patterns in perceived emotions, style, and themes can be identified across time.

Using the Project Gutenberg (n.d.) corpus and a methodology inspired by Gorin, Heblich, and Zylberberg (2025), the project transformed literary texts into a structured dataset of emotional, stylistic, and thematic features. Even the descriptive results already suggested meaningful variation across genres. Epic prose appeared more verb-heavy, scientific texts used more adjectives, and poetry stood out through higher lexical diversity measured by the type-token ratio. These findings indicate that the extracted features capture at least some interpretable dimensions of literary form and content.

The following stages showed that much of this variation is connected to genre. After residualising the features on decade, country, and genre fixed effects, genre explained the largest share of feature-level variability, while decade and country fixed effects contributed less, with the partial exception of sentence length. This suggests that literary form is strongly structured by genre, but also that some temporal and contextual variation remains

after accounting for these broad categories.

The clustering and network analysis then provided a broader view of how books and authors relate to each other in the constructed feature space. The identified clusters were not intended to reproduce historical literary movements directly. Still, many of them were interpretable in terms of genre, period, or thematic orientation. For example, science-fiction and fantasy-related clusters were more closely aligned with the later period of the corpus. This points to the potential of unsupervised methods as an exploratory tool for mapping literary structure, even if such results require careful interpretation and stronger grounding in literary theory. In future work, a more focused clustering strategy could aim at identifying smaller and more historically specific clusters, which may better approximate literary schools, movements, or generations than the broader clusters used in this project.

The temporal analysis offered the most direct connection between literary patterns and historical context. Using selected historical events as shock variables, both trend comparisons and event-study-style regressions suggested that some periods were associated with measurable changes in textual features. In the headline case of World War I, the shock was associated with a 0.16 SD increase in expressed fear, a 0.27 SD rise in conflict vocabulary, and a 0.14 SD increase in expressed anger. The cluster-level results supported this feature-level evidence, as post-1914 books became more likely to belong to more conflict-, fear-, and past-oriented clusters.

These results should be interpreted with caution. Although I conducted several robustness checks, the identification strategy remains limited, and the estimated effects should not be read as strictly causal. The project therefore does not claim to prove that historical shocks directly changed literature in a causal sense. Rather, it shows that large literary corpora can be transformed into structured quantitative data and used to detect patterns that are consistent with broader historical and cultural developments.

The main limitation of the project is its breadth. The pipeline combines feature extraction, residualisation, clustering, network construction, and temporal analysis within one project. Each of these components would deserve a more focused treatment. The clustering part could go deeper into smaller and more historically specific clusters, which may better approximate literary schools, movements, or generations. The network analysis could also be developed further, especially with stronger links to literary theory.

The temporal part would require a stronger identification strategy and a more micro-level approach. One possible direction would be to trace the intensity of specific words, names, places, or events in texts and connect them more directly to historical shocks. Another useful extension would be to bring in historical time series, such as macroeconomic indicators, conflict intensity, political instability, or demographic changes, and use them as additional controls or explanatory variables. This could make the analysis less dependent on broad event dummies and more connected to the actual historical and textual mechanisms

through which context enters literature.

Overall, this project is best understood as an exploratory machine learning exercise rather than a finished causal study of literary history. Many aspects would need improvement, but the project shows that literary texts can be transformed into emotional, stylistic, and thematic data, and that these data can be used to study literary change at scale. In this sense, it serves less as a final answer and more as an overview of possible future directions, which I would like to revisit in a more focused form.

References

Baker, **S. R.**, **Bloom**, **N.**, & **Davis**, **S. J.** (2016). Measuring economic policy uncertainty*. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Gorin, **C.**, **Heblich**, **S.**, & **Zylberberg**, **Y.** (2025, June). State of the art: Economic development through the lens of paintings. https://doi.org/10.3386/w33976

Hennig, **C.** (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271. Retrieved May 26, 2026, from https://ideas.repec.org//a/eee/csdana/v52y2007i1p258-271.html

Michel, **J.-B.**, **Shen**, **Y. K.**, **Aiden**, **A. P.**, **Veres**, **A.**, **Gray**, **M. K.**, **The Google Books Team**, **Pickett**, **J. P.**, **Hoiberg**, **D.**, **Clancy**, **D.**, **Norvig**, **P.**, **Orwant**, **J.**, **Pinker**, **S.**, **Nowak**, **M. A.**, & **Aiden**, **E. L.** (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Miller, **G. A.** (1995). WordNet: A lexical database for english. *Commun. ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mohammad, **S. M.**, & **Turney**, **P. D.** (2013). Crowdsourcing a word–emotion association lexicon [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00460.x]. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Montani, **I.**, **Honnibal**, **M.**, **Honnibal**, **M.**, **Boyd**, **A.**, **Landeghem**, **S. V.**, & **Peters**, **H.** (2023, October 16). *Explosion/spaCy: V3.7.2: Fixes for APIs and requirements*. Zenodo. https://doi.org/10.5281/zenodo.10009823

Project Gutenberg. (n.d.). Project Gutenberg. https://www.gutenberg.org/

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

Traag, **V.**, **Waltman**, **L.**, & **Eck**, **N. J. v.** (2019). From louvain to leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Underwood, **T.** (2019, February 14). *Distant horizons: Digital evidence and literary change*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226612973.001.0001

A Semantic Field Seed Lists

Field	Seed words
Commerce	trade, merchant, money, price, market, debt, profit, goods, coin, bank, wage, labour, wealth, commerce, capital, credit, stock, exchange, tariff, manufacture, industry, export, import, shopkeeper
Religion	god, church, prayer, sin, soul, faith, heaven, divine, holy, sacred, priest, worship, scripture, salvation, pious, sermon, temple, bishop, monk, baptism, communion, confession, gospel, psalm
Death	death, die, dead, grave, burial, corpse, murder, kill, perish, mourn, funeral, tomb, coffin, plague, disease, pestilence, famine, casualty, slaughter, massacre, wound, agony
Nature	forest, river, sky, earth, mountain, harvest, season, field, sun, storm, flood, ocean, tree, garden, flower, rain, wind, meadow, valley, animal, bird, fish, horse, wilderness
Body	heart, hand, blood, face, eye, breast, flesh, tears, breath, pain, head, arm, lip, mouth, bone, skin, hair, finger, foot, chest
Collective	we, our, us, together, community, people, society, nation, public, fellow, common, united, mutual, collective, solidarity
Individual	I, my, myself, alone, solitary, private, personal, self, individual, lonely, singular, independent
Technology	engine, machine, telegraph, railway, electricity, invention, steam, factory, laboratory, experiment, mechanism, apparatus, locomotive, motor, device, engineer, mechanical, patent, telephone, wire, iron, steel, furnace, mill
Law	court, judge, trial, jury, verdict, sentence, crime, prison, justice, punishment, guilty, innocent, lawyer, attorney, statute, witness, testimony, convict, magistrate, plea, defendant, prosecution, constable, sheriff
Family	mother, father, son, daughter, wife, husband, child, marriage, family, household, heir, parent, brother, sister, uncle, aunt, cousin, nephew, widow, orphan, bride, wedding, kinship, domestic
Education	school, university, teacher, student, book, learn, study, scholar, wisdom, lesson, pupil, professor, lecture, college, academy, knowledge, instruction, tutor, examination, library, curriculum, diploma

Table 5: Seed word lists for semantic field scoring. Each list is expanded via WordNet synsets before matching.

B Stage-4 Partialling — Detailed Diagnostics

Figure 20: Per-FE-block R^2 decomposition (full).

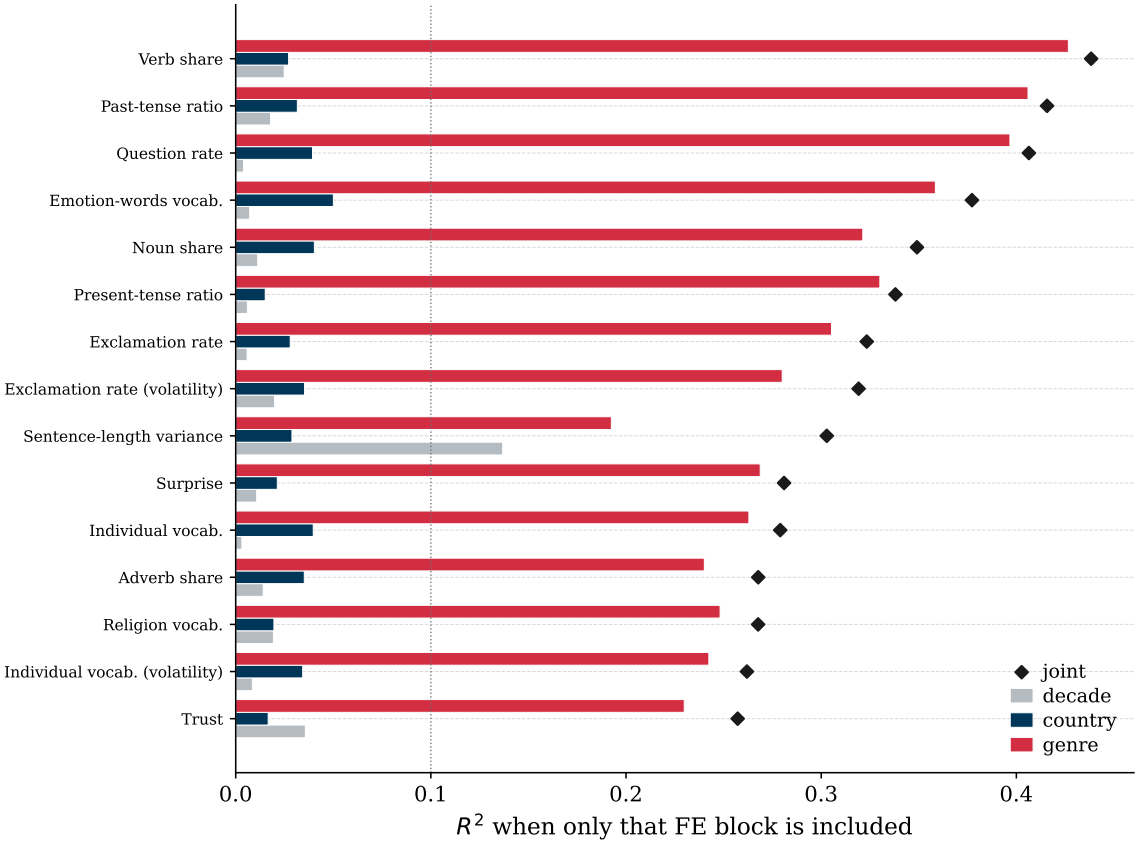


Table 6: Summary of the four FE specifications. Full per-feature R^2 in `data/features/partialling_log.json`.

Spec	Median R^2	Max R^2	n_{features}	$n_{\text{regressors}}$
`dec_yc`	0.019	0.149	76	41
`dec_ycg`	0.110	0.268	76	54
`25y_yc`	0.022	0.151	76	24
`25y_ycg`	0.109	0.270	76	37

C Additional Clustering Diagnostics and Robustness UMAPs Across Run Tags

This appendix reports additional implementation and robustness details for the clustering analysis in Section 6. The full implementation is in `pipeline/s05_clustering.py`. The per-space clustering functions are in `features/clustering_spaces.py`, and the network functions are in `utils/clustering_network.py`.

C.1 Robustness Configurations

The Stage 5 clustering module is run four times under different cluster counts and one alternative network neighbourhood size. These runs are labelled `first`, `second`, `third`, and `fourth`. Each run is saved separately in `data/results/clusters/<tag>/`, so the resulting partitions can be compared directly.

Table 7: Stage 5 robustness configurations.

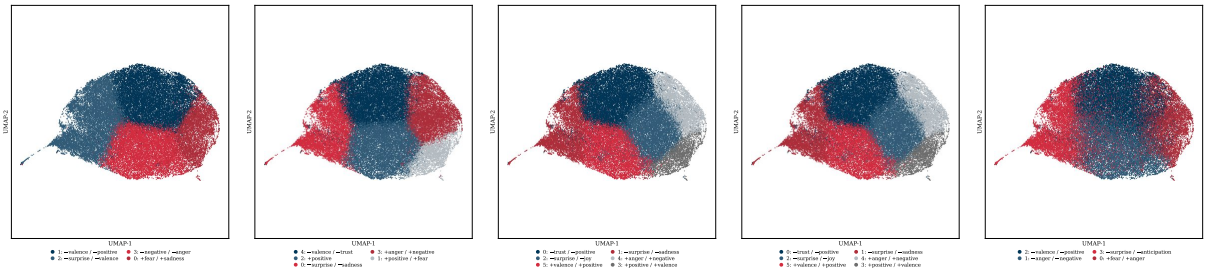
Run tag	k_{emo}	k_{sty}	k_{thm}	k_{NN}	Role
`first`	4	8	6	105	Headline run, informed by GMM-BIC and small-cluster inspection.
`second`	5	4	5	100	Fewer style and theme clusters; tests against over-segmentation.
`third`	6	6	4	40	Sparser network; tests whether Leiden communities depend on neighbourhood size.
`fourth`	6	8	10	105	More theme clusters; tests against under-segmentation.

Stability is evaluated in three ways. First, `cluster_stability_kmeans` re-fits k-means with 10 different random seeds. Values of ARI close to one indicate that the partition is driven by the feature geometry rather than by initialisation. Second, `cluster_stability_bootstrap` uses a 100-iteration bootstrap procedure. Books are resampled with replacement, k-means is re-fitted, and original books are assigned to their nearest bootstrap centroid. The mean per-cluster Jaccard index is then used to assess stability. Following Hennig (2007), I treat mean $J \geq 0.85$ as evidence of a stable cluster. Third, I compare the partitions across run tags using pairwise ARI. This checks whether similar structure appears when k or k_{NN} is changed.

C.2 Residualised Re-Clustering

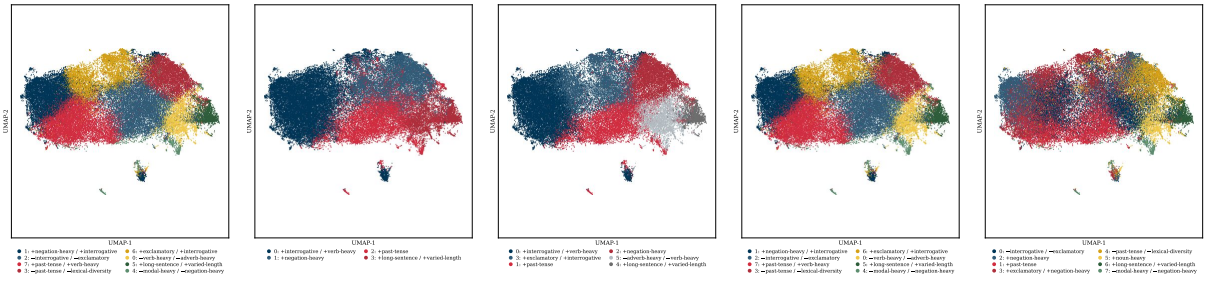
The clustering pipeline is also repeated on the residualised feature matrix `book_features_resid_dec_ycg.parquet`. This run is labelled `resid_dec_ycg` and uses the same cluster counts as the headline `first` run. Comparing `first` with `resid_dec_ycg` shows how much of the original clustering structure survives after year × country × genre composition is removed.

Figure 21: Cross-run ARI between the robustness runs and the residualised re-clustering — emotion space.



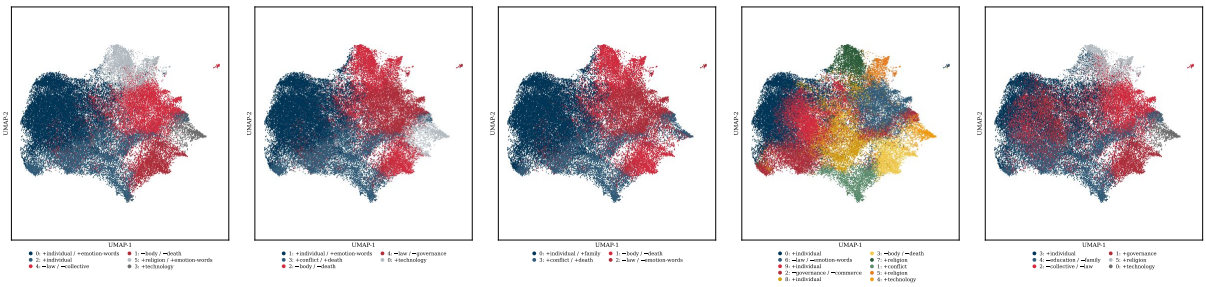
Notes: Each cell reports the pairwise ARI between two run tags' emotion partitions. The `first–resid_dec_ycg` comparison shows how much of the emotion-space cluster structure survives FE partialling.

Figure 22: Cross-run ARI between the robustness runs and the residualised re-clustering — style space.



Notes: As in Figure 21, for the style space.

Figure 23: Cross-run ARI between the robustness runs and the residualised re-clustering — theme space.



Notes: As in Figure 21, for the theme space.

C.3 Cluster Labels and Discrimination Tables

Cluster labels are produced through a two-step η^2-gated procedure that is summarised in Section 6.3. The full procedure is reported here.

Step 1: Eligibility filter. For each feature f I compute the between-cluster effect size from a one-way ANOVA of the feature on cluster identity:

$$\eta^2_f = \frac{SS_{\text{between}}}{SS_{\text{total}}} = \frac{\sum_c n_c \left(\bar{y}_{f,c} - \bar{y}_f\right)^2}{\sum_i \left(y_{f,i} - \bar{y}_f\right)^2},$$

where $\bar{y}_f$ is the corpus grand mean of feature f and $\bar{y}_{f,c}$ is its cluster-c mean. Only features whose η^2 rank places them in the top quartile of the partition are kept as eligible candidates for labelling. This restricts the label vocabulary to features that genuinely separate the clusters.

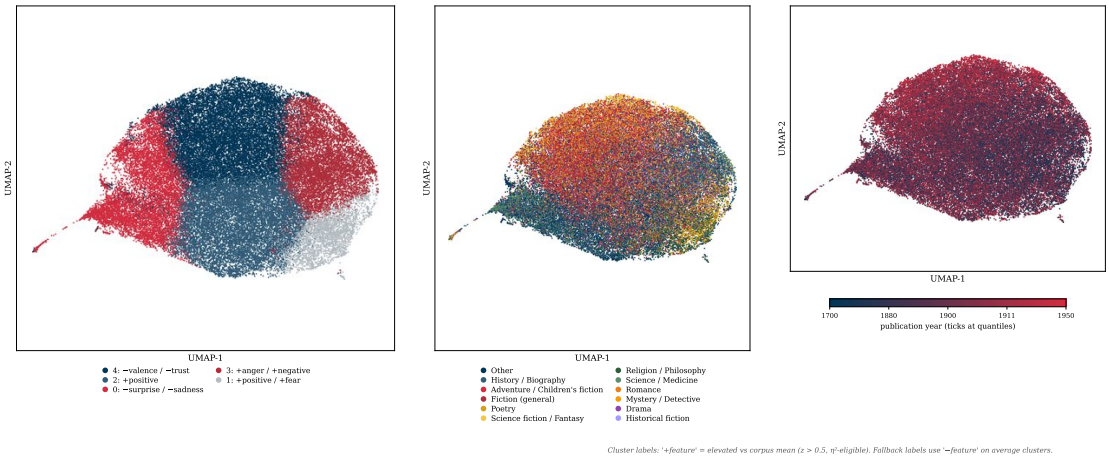
Step 2: Between-cluster z-score of the cluster centroid. Among the eligible features, I compute, for each cluster c and feature f, the standardised deviation of the cluster centroid from the mean of all cluster centroids:

$$z_{c,f} \;=\; \frac{\bar{y}_{f,c} - \mu_f}{\sigma_f}, \qquad \mu_f = \tfrac{1}{k}\sum_c \bar{y}_{f,c}, \quad \sigma_f = \mathrm{sd}\big(\bar{y}_{f,1}, \ldots, \bar{y}_{f,k}\big).$$

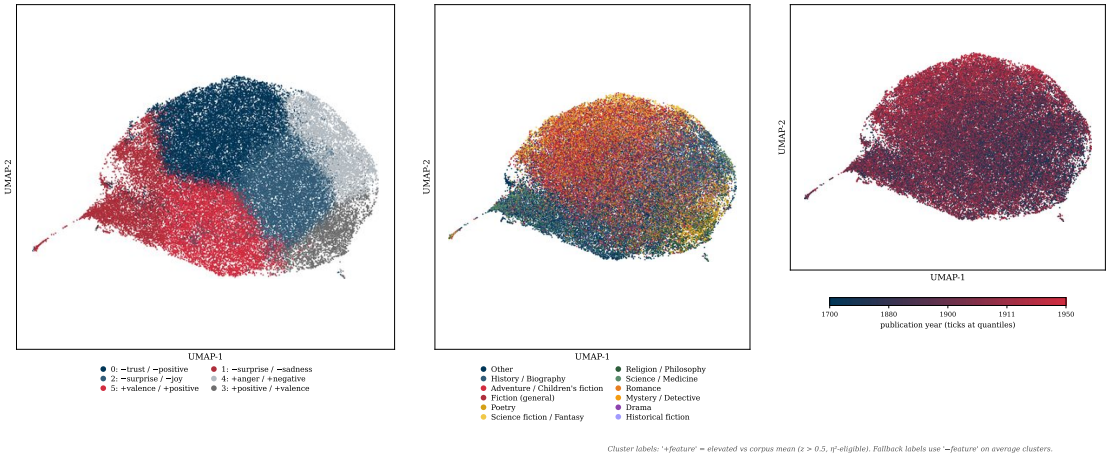
The top two features by $|z_{c,f}|$ above the threshold $|z| \geq 0.5$ define the label, with their sign indicating direction. Example labels in the emotion space include "+fear / +sadness", "−valence / −positive", and "−negative / −anger".

The complete label files are stored as `data/results/clusters/<tag>/cluster_names.json`. For the paper figures, the automatically generated labels are manually refined for readability. Below are the UMAPs for alternative specifications.

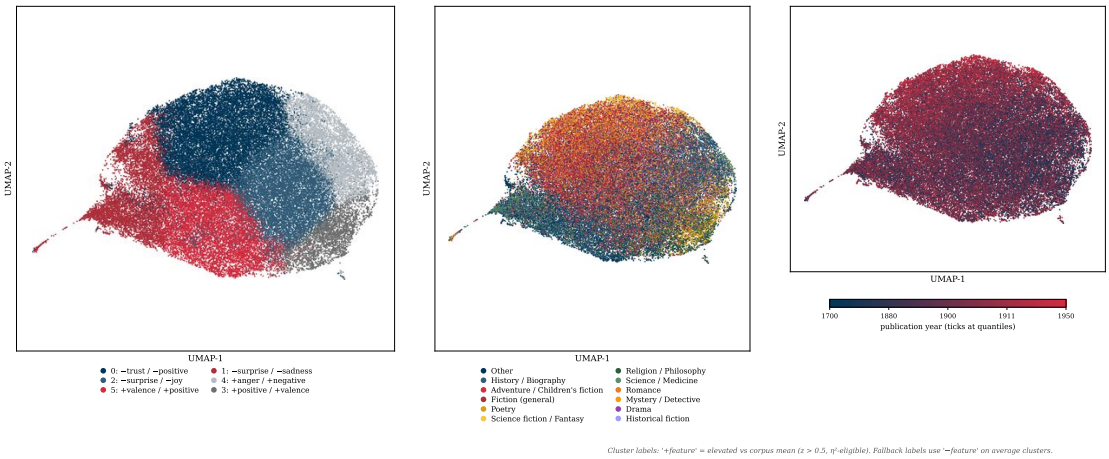
Figure 24: Emotion space across run tags.



(a) `second` ($k = 5$).

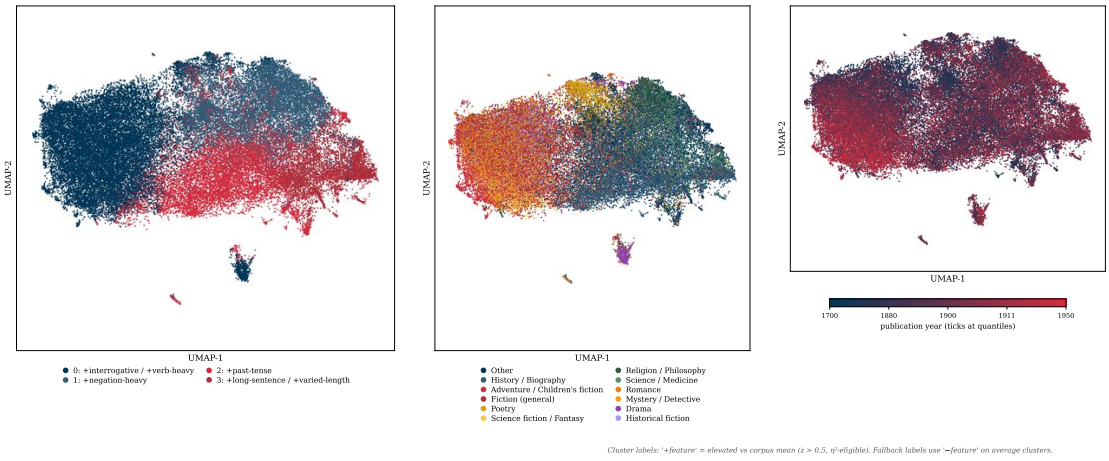


(b) `third` ($k = 6$).

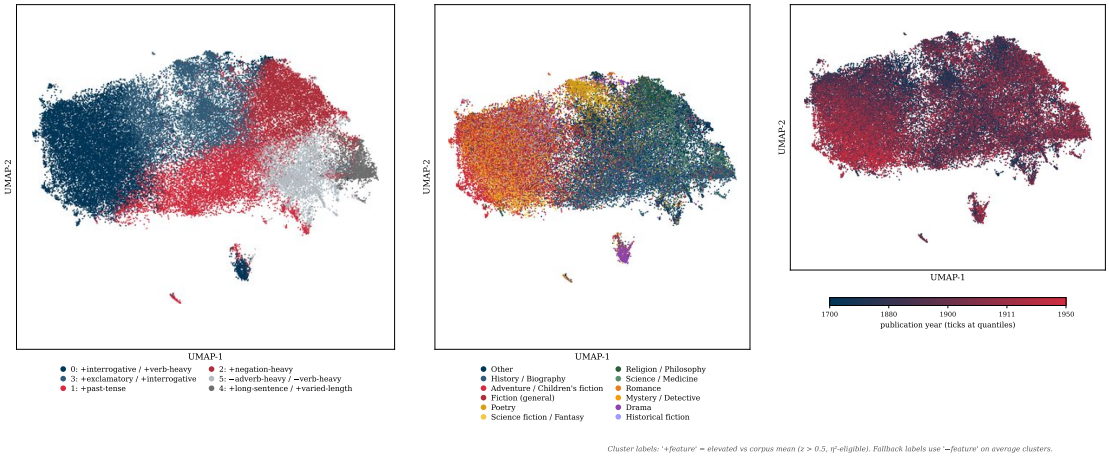


(c) `fourth` ($k = 6$).

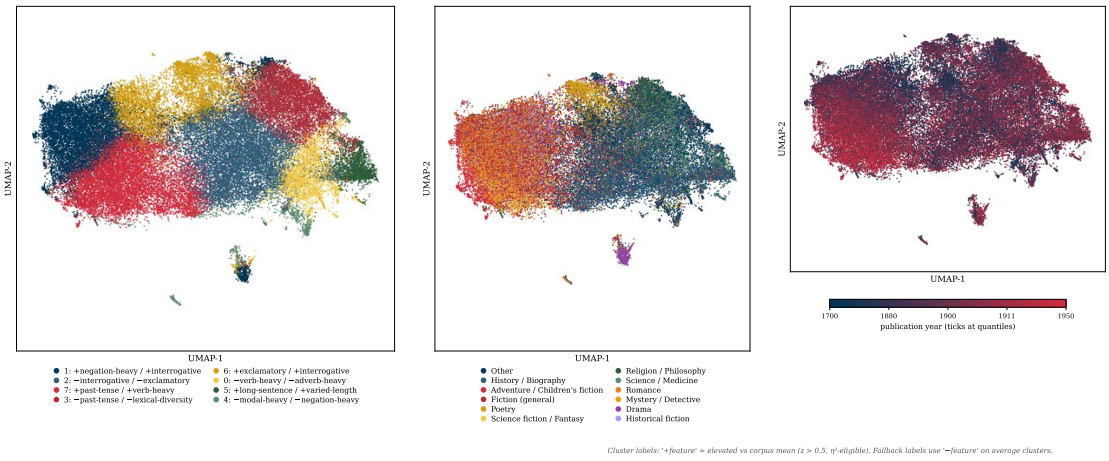
Figure 25: Style space across run tags.



(a) `second` ($k = 4$).

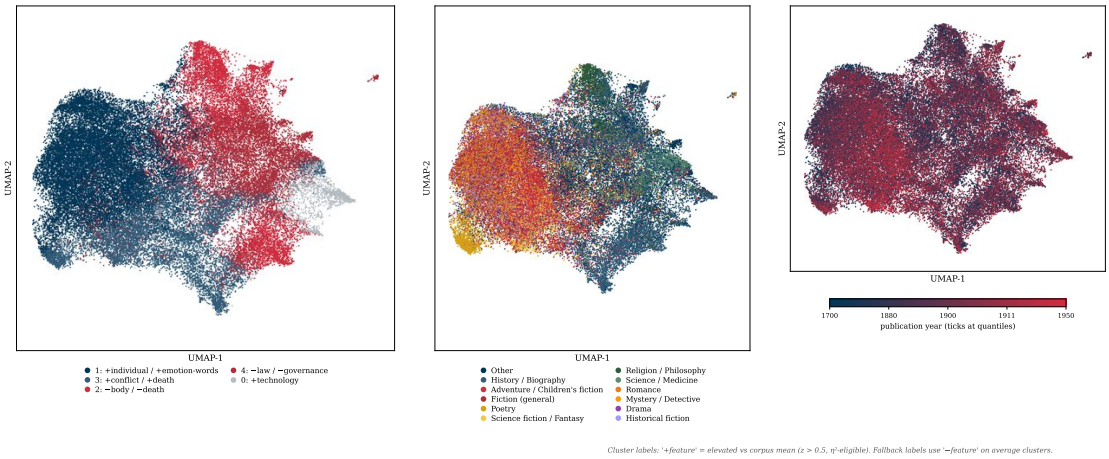


(b) `third` ($k = 6$).

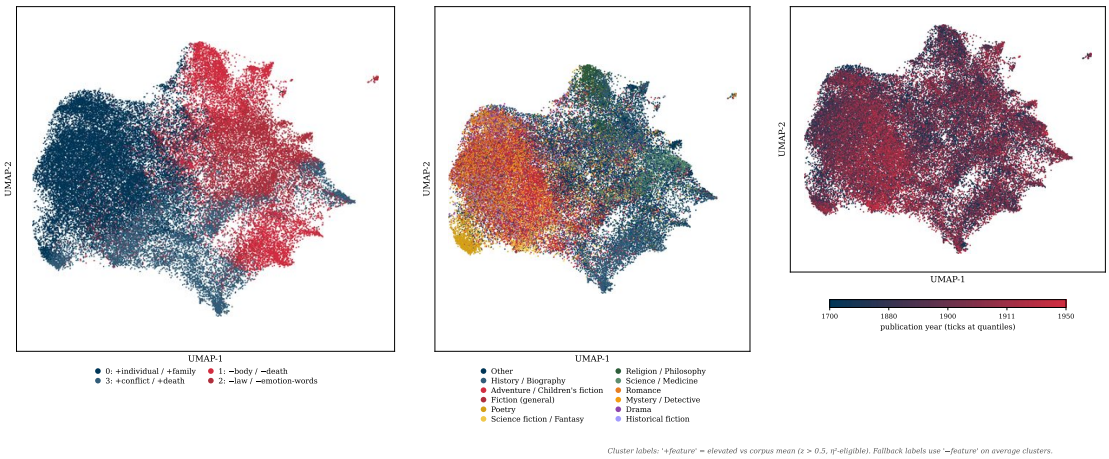


(c) `fourth` ($k = 8$).

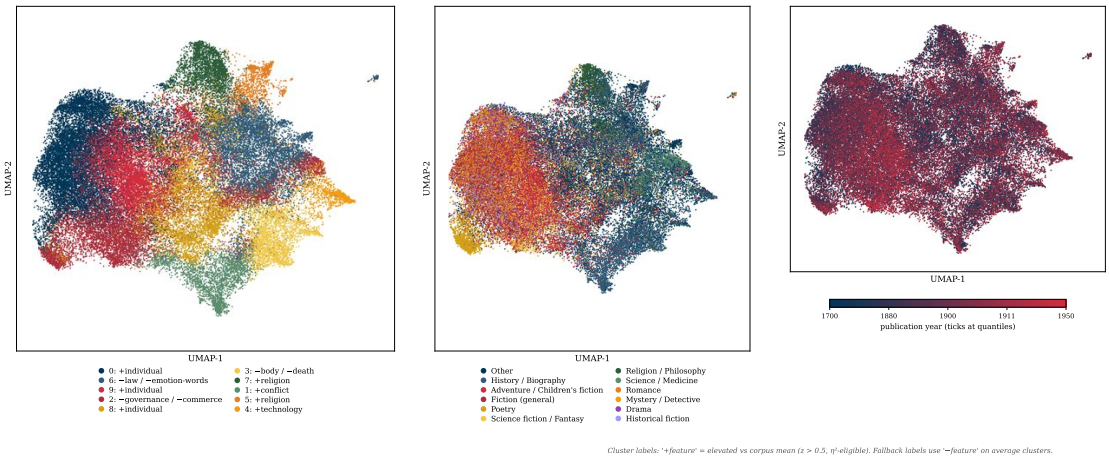
Figure 26: Theme space across run tags.



(a) `second` ($k = 5$).

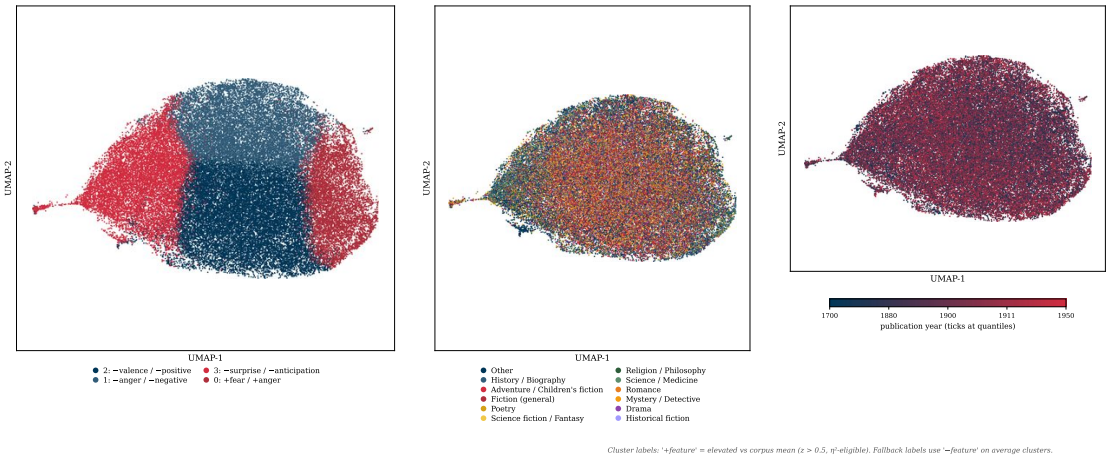


(b) `third` ($k = 4$).

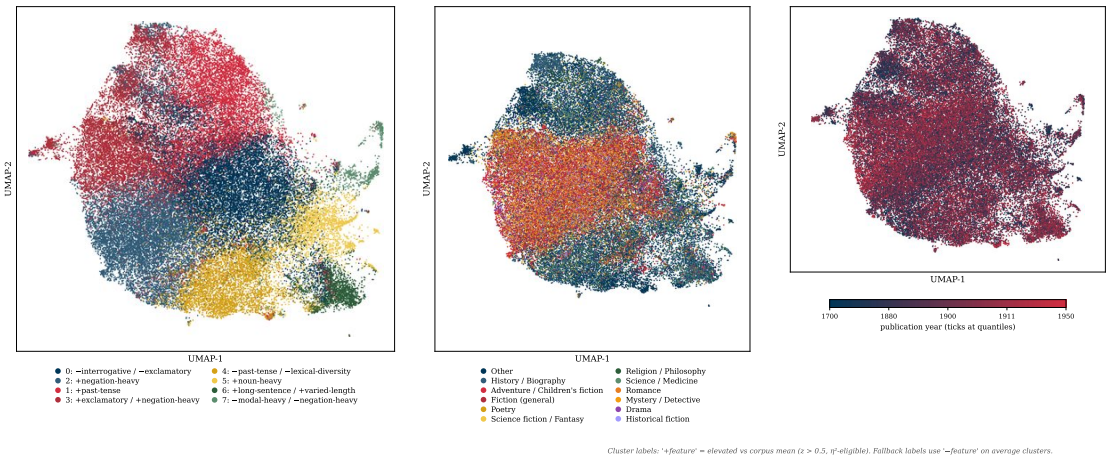


(c) `fourth` ($k = 10$).

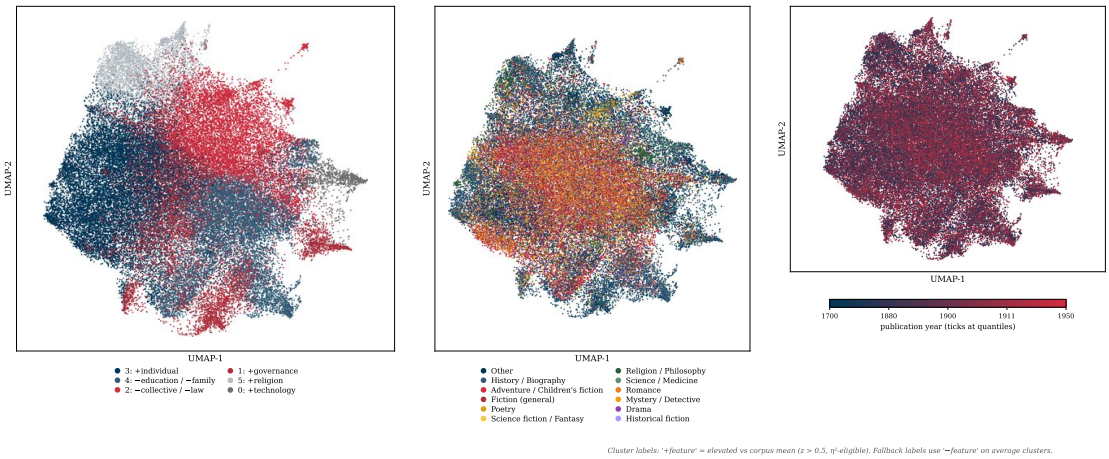
Figure 27: Residualised re-clustering (`resid_dec_ycg`).



(a) Emotion.



(b) Style.



(c) Theme.

D Network: Book-Level and Community Dashboards

Figure 28: Book-level Leiden communities (`first`).

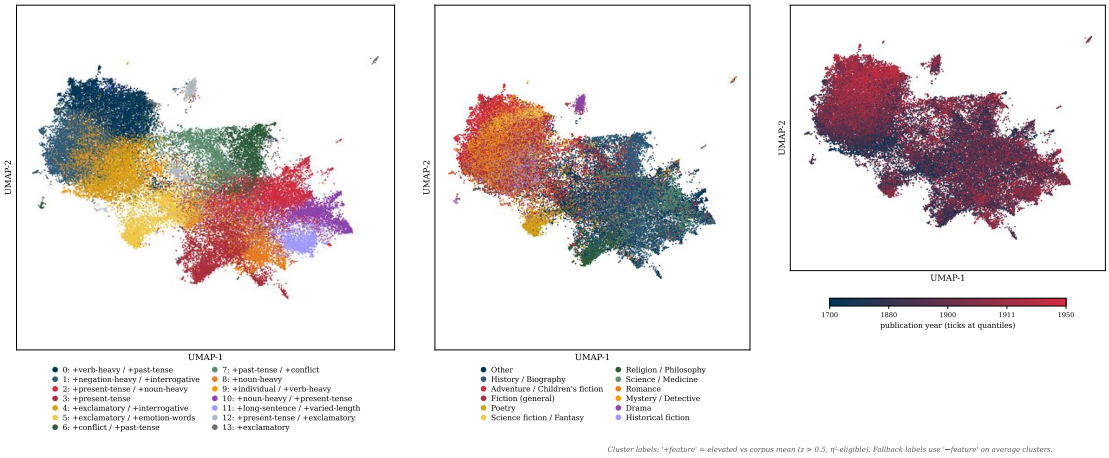


Figure 29: Author-level community dashboard (`first`).

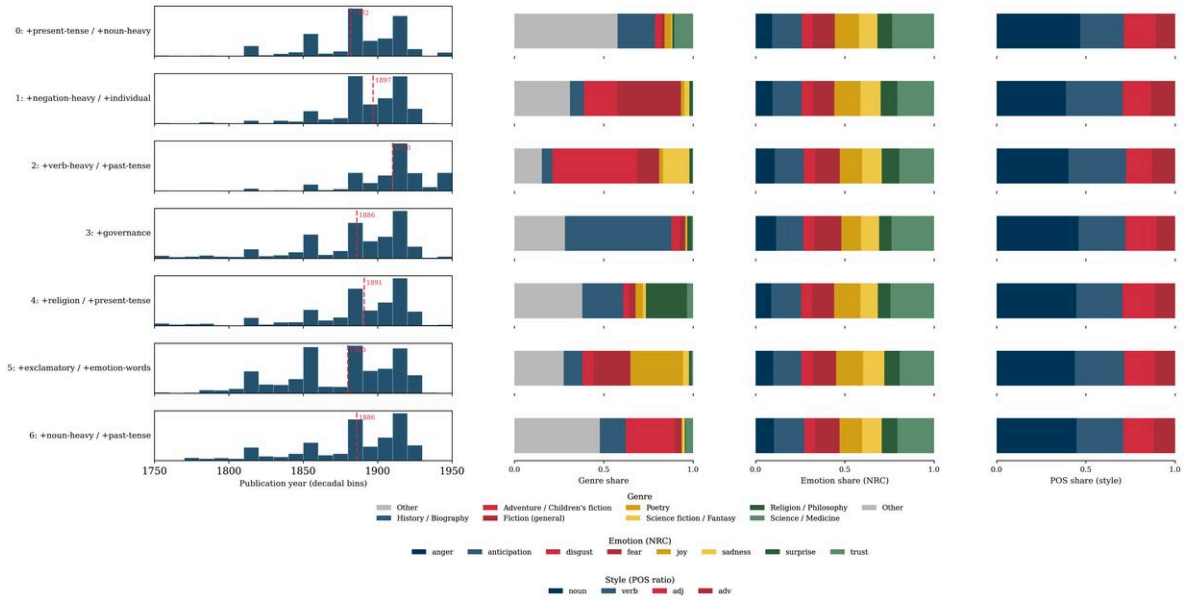


Figure 30: Book-level community dashboard (`first`).

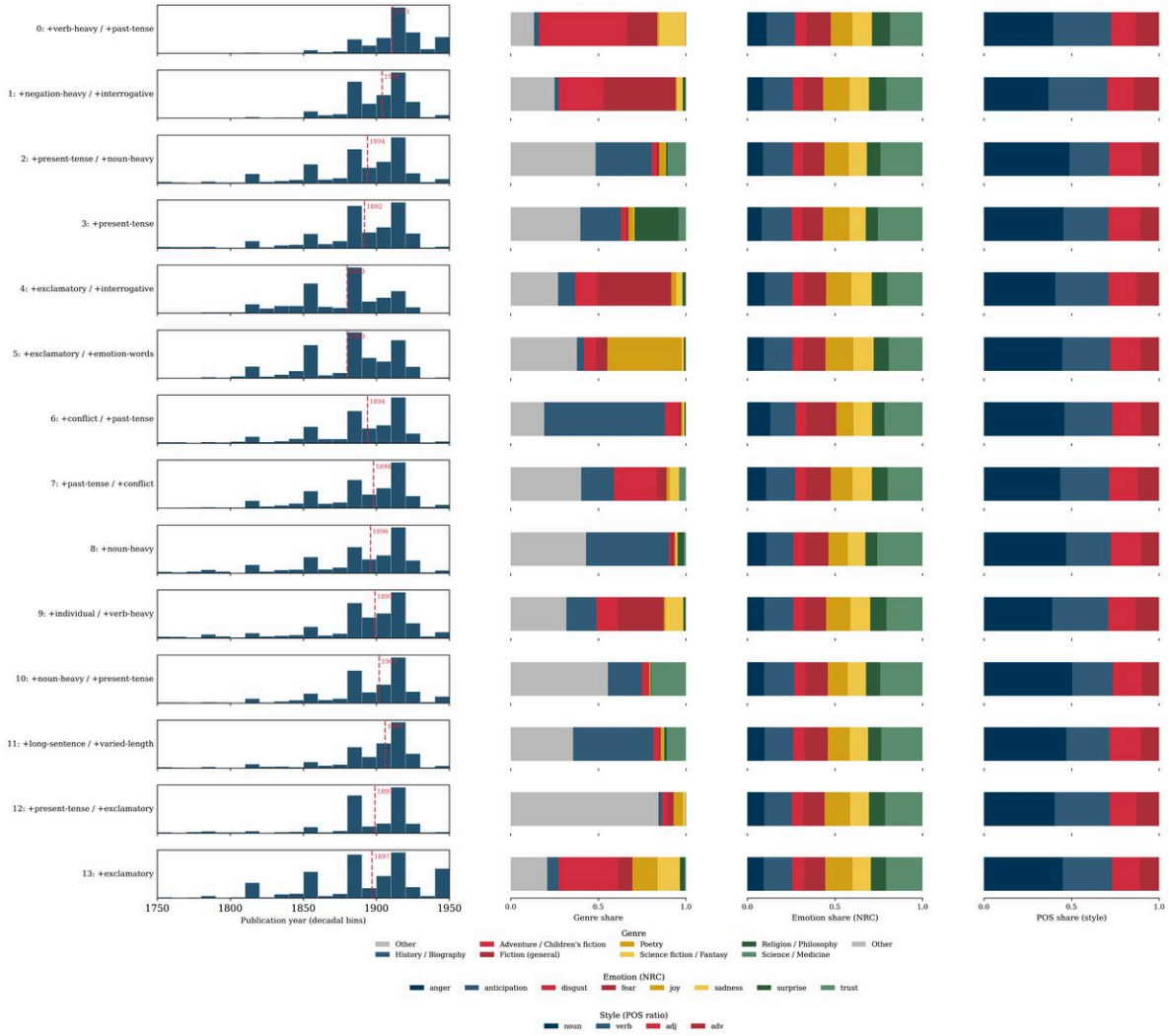
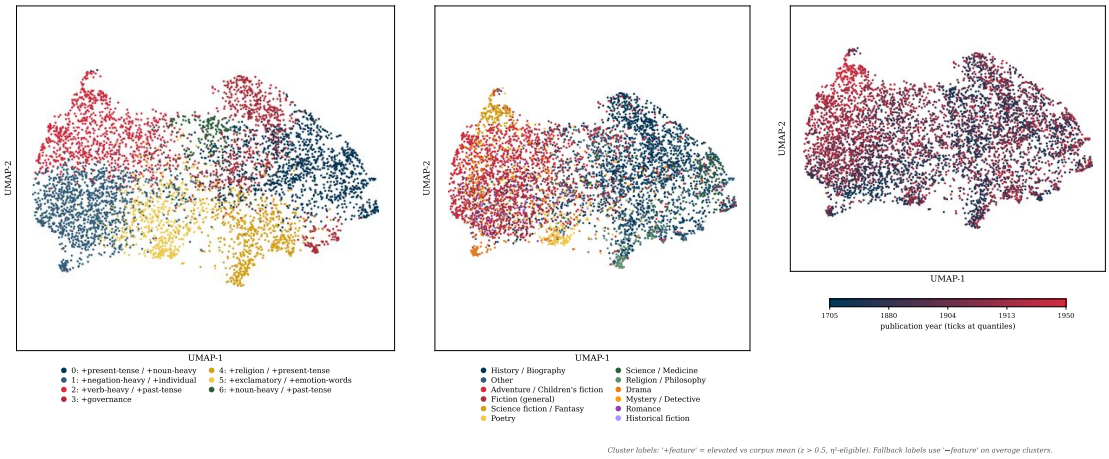
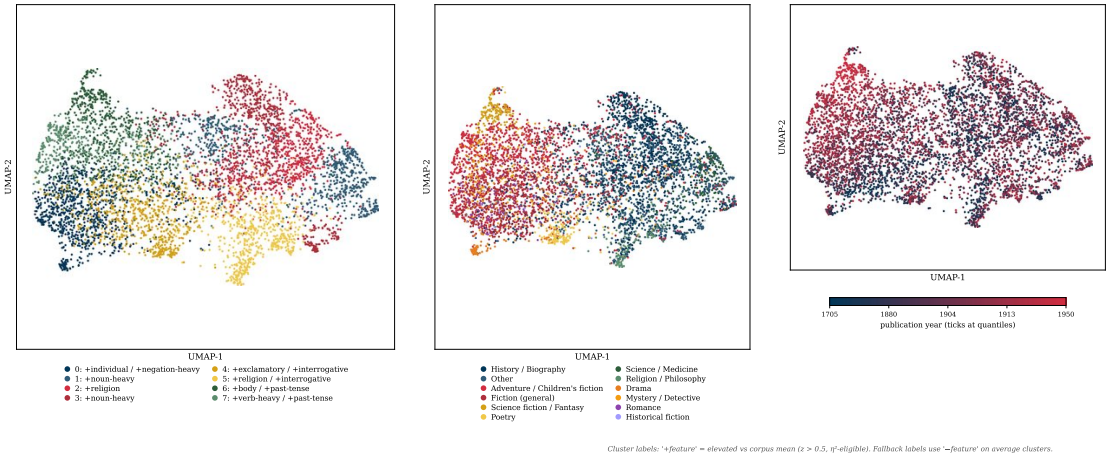


Figure 31: Author-level networks across run tags.



(a) `second`.

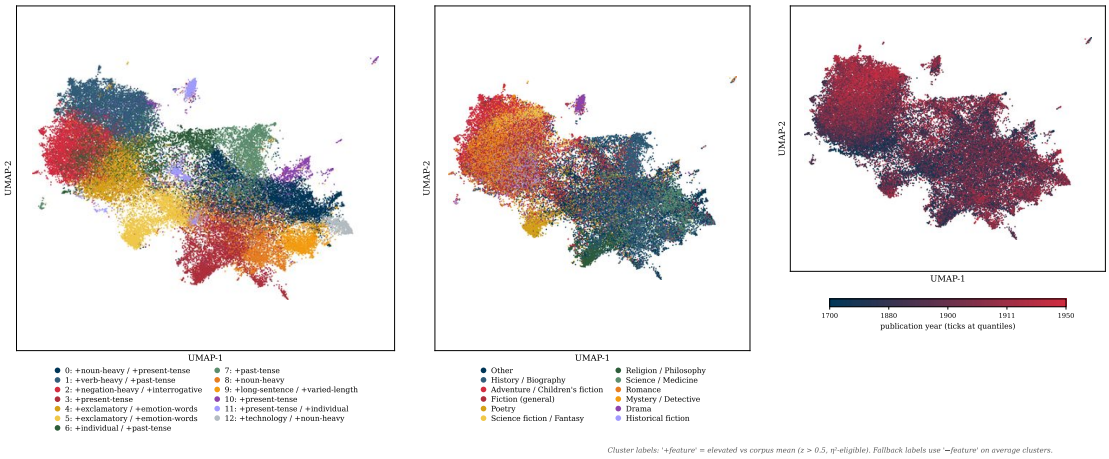


(b) `third` ($k_{\mathrm{NN}} = 40$).

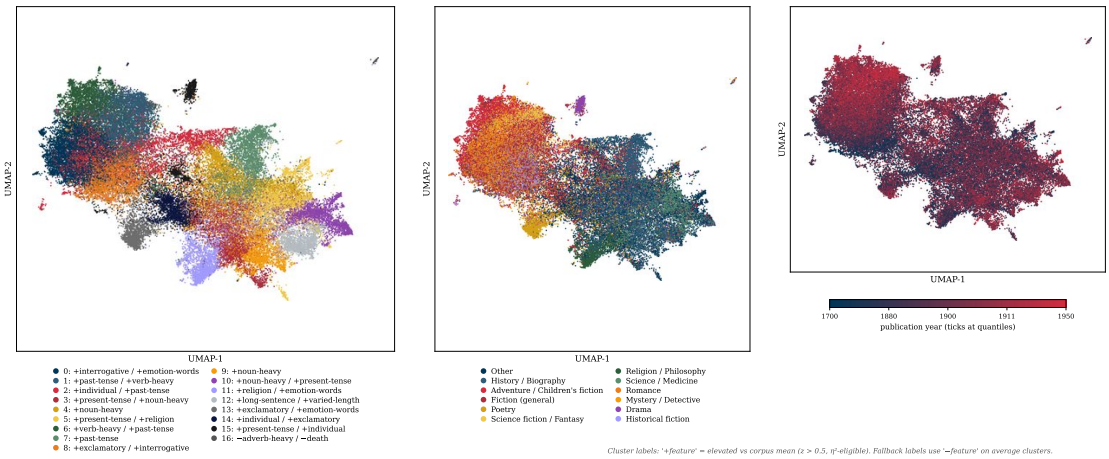


(c) `fourth`.

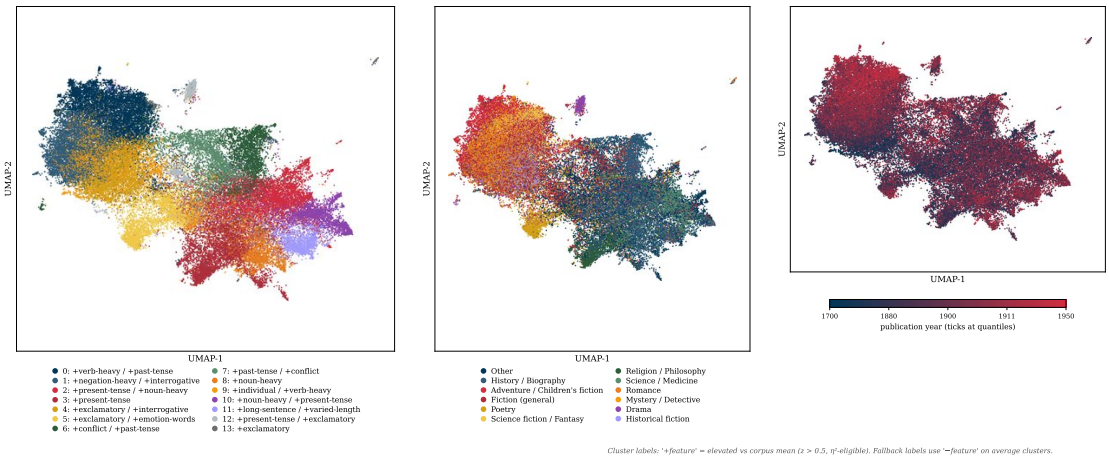
Figure 32: Book-level networks across run tags.



(a) `second`.



(b) `third`.



(c) `fourth`.

E Cross-Run ARI Heatmaps (Network Spaces)

Figure 33: Cross-run ARI on author-level network partitions.

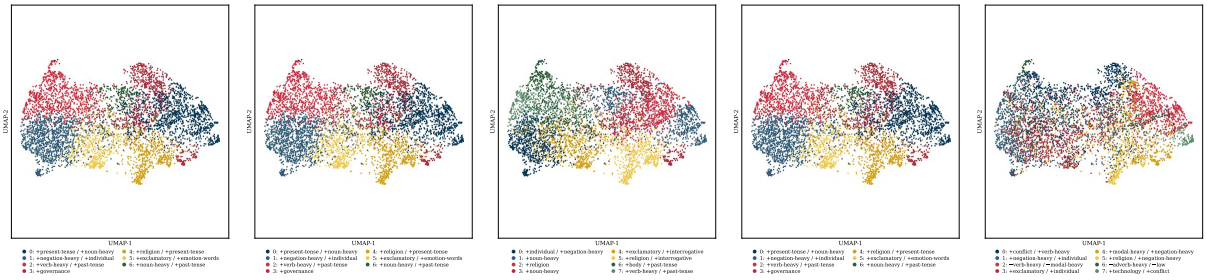
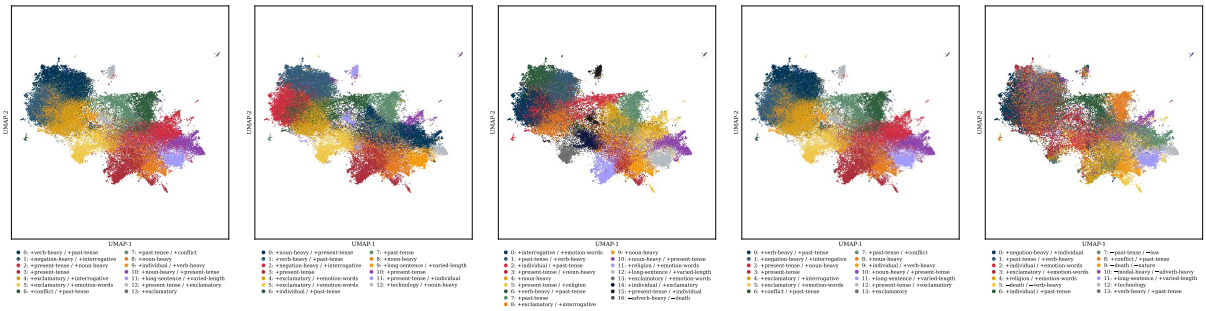


Figure 34: Cross-run ARI on book-level network partitions.



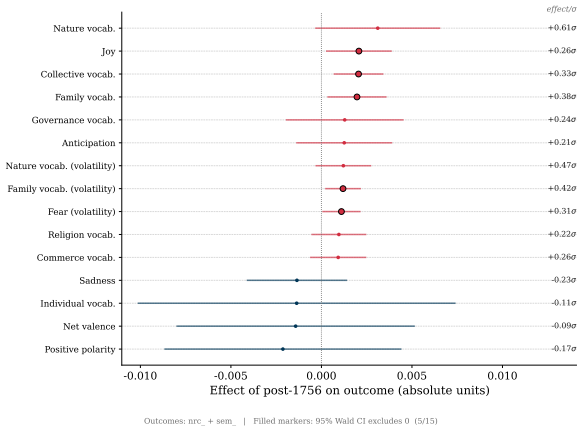
F Stage-6 Headline Summary Table

Table 8: Top features by $|\hat{\beta}|$ at WWI (1914), event-study summary. Bootstrap SE from 200 year-block resamples; Wald CIs.

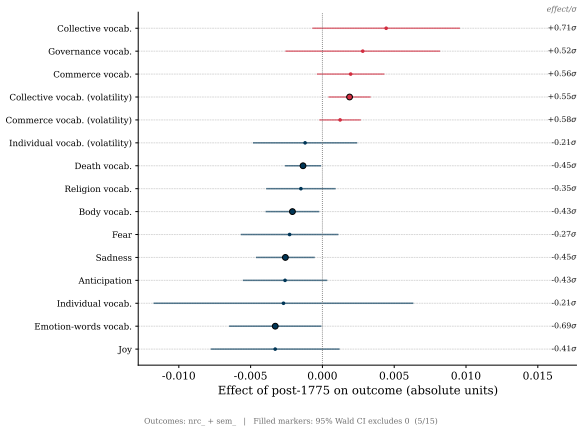
Feature	$\hat{\beta}$	SE (boot)	CI lo	CI hi	n_{pre}	n_{post}
Avg sentence length (volatility)	-1.0369	0.4933	-2.0037	-0.0701	10038	4996
Avg sentence length	-0.8637	0.2279	-1.3103	-0.4171	10038	4996
Sentence-length variance	-0.5397	0.1281	-0.7907	-0.2887	10038	4996
Sentence-length variance (volatility)	-0.4572	0.1519	-0.7550	-0.1594	10038	4996
Present-tense ratio	-0.0032	0.0029	-0.0090	0.0026	10038	4996
Past-tense ratio	0.0026	0.0040	-0.0053	0.0105	10038	4996
Lexical diversity	-0.0022	0.0006	-0.0034	-0.0010	10038	4996
Net valence	-0.0016	0.0007	-0.0029	-0.0003	10038	4996
Fear	0.0013	0.0004	0.0005	0.0021	10038	4996
Exclamation rate	0.0013	0.0013	-0.0012	0.0039	10038	4996
Conflict vocab.	0.0013	0.0003	0.0007	0.0018	10038	4996
Negative polarity	0.0011	0.0004	0.0003	0.0019	10038	4996
Anger	0.0009	0.0003	0.0004	0.0014	10038	4996
Verb share	0.0008	0.0004	-0.0000	0.0017	10038	4996
Question rate	-0.0007	0.0006	-0.0019	0.0005	10038	4996

G Per-Shock Event-Study Forest Plots

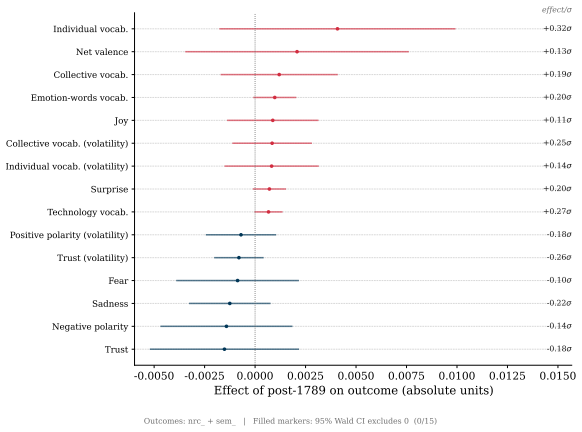
Figure 35: Pre-1820 shocks.



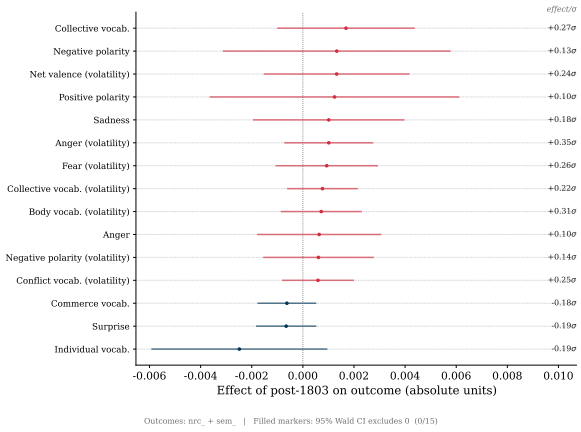
(a) Seven Years' War (1756).



(b) American Revolution (1775).

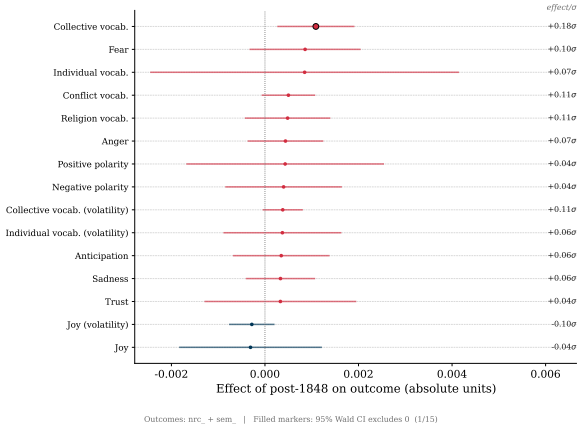


(c) French Revolution (1789).

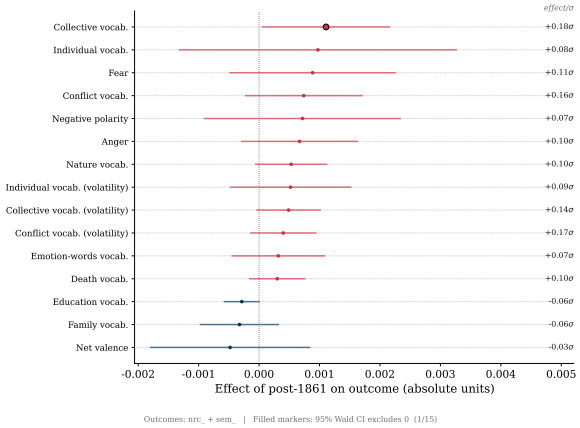


(d) Napoleonic Wars (1803).

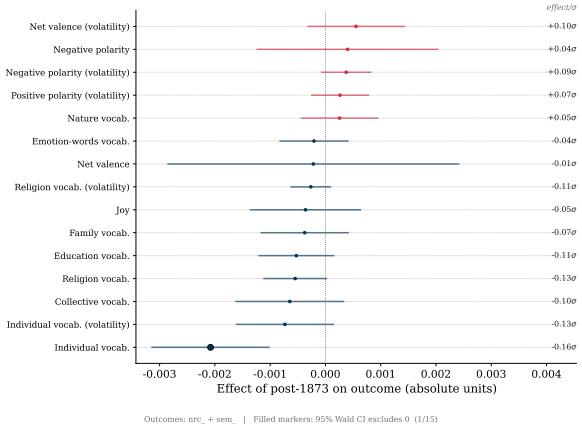
Figure 36: 19th-century shocks.



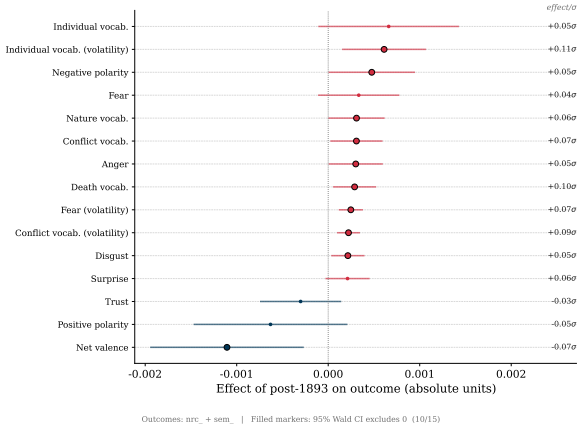
(a) 1848 Revolutions.



(b) US Civil War (1861).

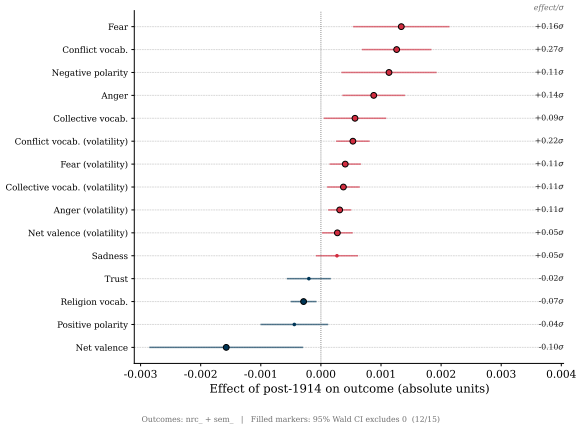


(c) Long Depression (1873).

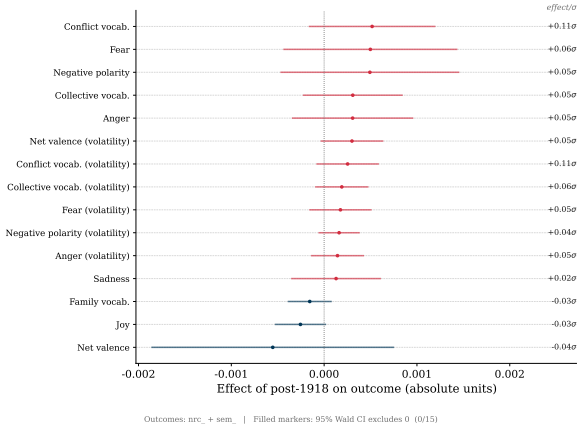


(d) Panic of 1893.

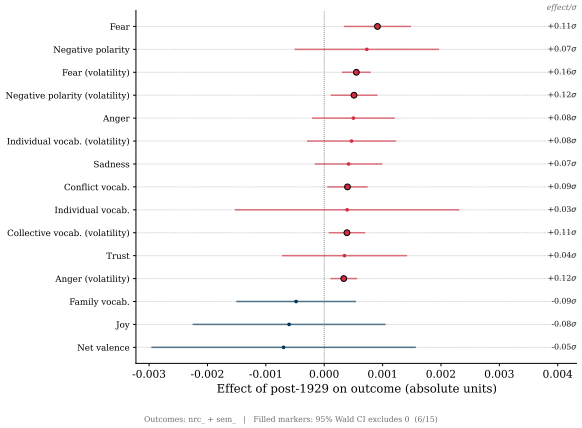
Figure 37: 20th-century shocks.



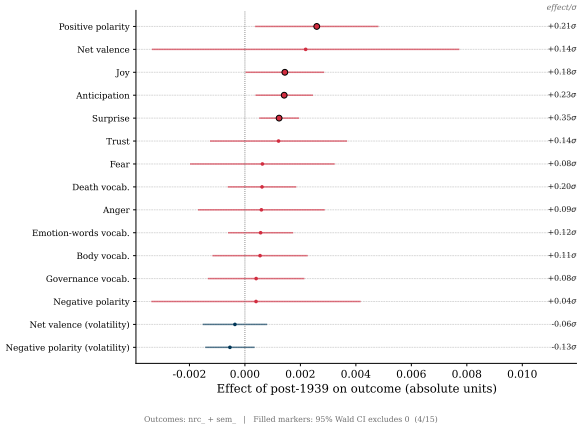
(a) World War I (1914).



(b) Spanish Flu (1918).



(c) Great Depression (1929).



(d) World War II (1939).

H Cluster Discrimination Tables

Emotion

Table 9: Per-cluster diagnostic — emotion space (`first` run). Columns: cluster ID, η^2-gated label, size, share, median publication year, IQR low/high, top genre, top z-scored features, bootstrap mean Jaccard $\bar{J}$ with 5th/95th percentiles, and a verdict (HIGH if $\bar{J} \geq 0.85$, OK if ≥ 0.75, WEAK if ≥ 0.60, else UNSTABLE).

Cluster	Label	n	Share	Median year	IQR low	IQR high	Top genre	Top features	Boot. $\bar{J}$	J_{q05}	J_{q95}	Verdict
0	+fear / +sadness	6980	0.127444	1882	1850	1910	History / Biography (29%)	disgust:+1.37, sadness:+1.36, negative:+1.35	0.894885	0.788952	0.984367	HIGH
1	-valence / -positive	18002	0.328690	1903	1880	1913	Adventure / Children's fiction (20%)		0.912016	0.790075	0.981860	HIGH
2	-surprise / -valence	16353	0.298581	1906	1880	1914	Other (27%)		0.936474	0.854735	0.989676	HIGH
3	-negative / -anger	13434	0.245285	1891	1869	1910	Other (33%)	valence:+1.37, joy:+0.97, positive:+0.91	0.930553	0.861704	0.985474	HIGH

Table 10: Top discriminating features — emotion space (`first` run). Per cluster, the top features by within-cluster z-score together with their corpus-wide η^2 and the η^2 rank among all features.

Cluster	Label	Feature	z	η^2	η^2 rank
0	+fear / +sadness	disgust	1.365786	0.282542	15
0	+fear / +sadness	sadness	1.356410	0.346496	12
0	+fear / +sadness	negative	1.352047	0.382373	8
1	-valence / -positive	surprise	0.434123	0.097561	21
1	-valence / -positive	anger	0.336377	0.515846	3
1	-valence / -positive	negative	0.325349	0.524616	2
2	-surprise / -valence	valence	0.003014	0.379936	9
2	-surprise / -valence	positive	-0.945362	0.250962	18
2	-surprise / -valence	anticipation	-0.953265	0.132364	20
3	-negative / -anger	valence	1.368806	0.379936	9
3	-negative / -anger	joy	0.965911	0.261196	17
3	-negative / -anger	positive	0.913409	0.404462	7

Style

Table 11: Per-cluster diagnostic — style space (`first` run).

Cluster	Label	n	Share	Median year	IQR low	IQR high	Top genre	Top features	Boot. $\bar{J}$	J_{q05}	J_{q95}	Verdict
0	-verb-heavy / -adverb-heavy	3457	0.063120	1900	1880	1910	Other (42%)	noun-heavy:+2.12, lexical-diversity:+1.99, adjective-heavy:+1.74	0.921256	0.800091	0.975463	HIGH
1	+negation-heavy / +interrogative	12016	0.219394	1901	1880	1912	Fiction (general) (29%)	negation-heavy:+1.82, interrogative:+1.67, verb-heavy:+1.29	0.890164	0.543266	0.978539	HIGH
2	-interrogative / -exclamatory	11751	0.214556	1892	1862	1910	Other (41%)	past-tense:+1.32, lexical-diversity:+0.98, present-tense:+0.75	0.895688	0.656785	0.976726	HIGH
3	-past-tense / -lexical-diversity	6713	0.122569	1898	1876	1911	Other (45%)	adjective-heavy:+1.61, modal-heavy:+1.58, present-tense:+1.46	0.953936	0.880102	0.981785	HIGH
4	-modal-heavy / -negation-heavy	1565	0.028575	1897	1870	1911	Other (44%)	present-tense:+0.74	0.856620	0.256941	0.974259	HIGH
5	+long-sentence / +varied-length	1885	0.034417	1908	1880	1913	History / Biography (47%)	long-sentence:+2.46, varied-length:+2.39, verb-heavy:+1.25	0.956522	0.931086	0.982608	HIGH
6	+exclamatory / +interrogative	6537	0.119356	1880	1850	1905	Poetry (24%)	exclamatory:+2.14, lexical-diversity:+1.61, interrogative:+1.51	0.878394	0.682730	0.976082	HIGH
7	+past-tense / +verb-heavy	10845	0.198013	1910	1888	1919	Adventure / Children's fiction (32%)	past-tense:+1.66, verb-heavy:+1.27, adverb-heavy:+0.97	0.852676	0.330032	0.982865	HIGH

Table 12: Top discriminating features — style space (`first` run).

Cluster	Label	Feature	z	η^2	η^2 rank
0	-verb-heavy / -adverb-heavy	noun-heavy	2.117999	0.363683	12
0	-verb-heavy / -adverb-heavy	lexical-diversity	1.993436	0.355150	15
0	-verb-heavy / -adverb-heavy	adjective-heavy	1.735964	0.405258	9
1	+negation-heavy / +interrogative	negation-heavy	1.815217	0.483074	7
1	+negation-heavy / +interrogative	interrogative	1.669292	0.500287	6
1	+negation-heavy / +interrogative	verb-heavy	1.293208	0.575615	3
2	-interrogative / -exclamatory	past-tense	1.316588	0.326776	18
2	-interrogative / -exclamatory	lexical-diversity	0.980530	0.176813	26
2	-interrogative / -exclamatory	present-tense	0.747197	0.333960	16
3	-past-tense / -lexical-diversity	adjective-heavy	1.609579	0.270361	24
3	-past-tense / -lexical-diversity	modal-heavy	1.579119	0.401185	10
3	-past-tense / -lexical-diversity	present-tense	1.458080	0.384674	11
4	-modal-heavy / -negation-heavy	present-tense	0.735872	0.384674	11
4	-modal-heavy / -negation-heavy	lexical-diversity	-0.165465	0.176813	26
4	-modal-heavy / -negation-heavy	exclamatory	-0.242162	0.458901	8
5	+long-sentence / +varied-length	long-sentence	2.459091	0.665788	1
5	+long-sentence / +varied-length	varied-length	2.387564	0.548718	4
5	+long-sentence / +varied-length	verb-heavy	1.249090	0.300954	19
6	+exclamatory / +interrogative	exclamatory	2.141887	0.458901	8
6	+exclamatory / +interrogative	lexical-diversity	1.609878	0.176813	26
6	+exclamatory / +interrogative	interrogative	1.510311	0.195523	25
7	+past-tense / +verb-heavy	past-tense	1.659970	0.582985	2
7	+past-tense / +verb-heavy	verb-heavy	1.274618	0.575615	3
7	+past-tense / +verb-heavy	adverb-heavy	0.974222	0.356273	14

Theme

Table 13: Per-cluster diagnostic — theme space (`first` run).

Cluster	Label	n	Share	Median year	IQR low	IQR high	Top genre	Top features	Boot. $\bar{J}$	J_{q05}	J_{q95}	Verdict
0	+individual / +emotion-words	22318	0.407493	1900	1880	1912	Fiction (general) (28%)	individual:+1.84, family:+1.63, emotion-words:+1.47	0.979893	0.936147	0.991563	HIGH
1	-body / -death	5430	0.099144	1898	1866	1910	History / Biography (53%)	governance:+1.88, law:+1.61, commerce:+1.48	0.922382	0.719212	0.979552	HIGH
2	+individual	11303	0.206376	1904	1880	1915	Adventure / Children's fiction (25%)	nature:+1.83, death:+1.75, conflict:+1.58	0.950634	0.861286	0.981645	HIGH
3	+technology	1361	0.024850	1908	1880	1913	Other (46%)	technology:+2.04, body:+1.77, commerce:+1.08	0.906215	0.680415	0.979552	HIGH
4	-law / -collective	9373	0.171137	1897	1872	1910	Other (42%)		0.954056	0.877299	0.982731	HIGH
5	+religion / +emotion-words	4984	0.091000	1890	1875	1910	Other (35%)	religion:+2.01, education:+1.99, emotion-words:+1.32	0.942590	0.572097	0.988490	HIGH

Table 14: Top discriminating features — theme space (`first` run).

Cluster	Label	Feature	z	η^2	η^2 rank
0	+individual / +emotion-words	individual	1.838915	0.390162	6
0	+individual / +emotion-words	family	1.625936	0.265105	13
0	+individual / +emotion-words	emotion-words	1.470533	0.349009	7
1	-body / -death	governance	1.881658	0.265974	12
1	-body / -death	law	1.610522	0.059382	28
1	-body / -death	commerce	1.479593	0.193095	19
2	+individual	nature	1.833304	0.241008	15
2	+individual	death	1.750956	0.311344	9
2	+individual	conflict	1.579701	0.286399	10
3	+technology	technology	2.037750	0.466516	1
3	+technology	body	1.769151	0.110312	27
3	+technology	commerce	1.076704	0.233021	16
4	-law / -collective	nature	0.233802	0.241008	15
4	-law / -collective	religion	-0.318338	0.434750	3
4	-law / -collective	technology	-0.394515	0.466516	1
5	+religion / +emotion-words	religion	2.006229	0.434750	3
5	+religion / +emotion-words	education	1.989432	0.193148	18
5	+religion / +emotion-words	emotion-words	1.318355	0.187034	20

Network — author level

Table 15: Per-community diagnostic — author-level Leiden communities (`first` run).

Cluster	Label	n	Share	Median year	IQR low	IQR high	Top genre	Top features	Boot.	$\bar{J}$	J_{q05}	J_{q95}	Verdict
0	+present-tense / +noun-heavy	8276	0.151107	1882	1869	1910	Other (56%)	noun-heavy:+2.10, lexical-diversity:+1.96, technology:+1.83	NaN		NaN	NaN	
1	+negation-heavy / +individual	12612	0.230276	1897	1880	1910	Fiction (general) (36%)	family:+1.81, negation-heavy:+1.72, individual:+1.62	NaN		NaN	NaN	
2	+verb-heavy / +past-tense	8693	0.158721	1910	1895	1921	Adventure / Children's fiction (47%)	law:+1.82, verb-heavy:+1.44, past-tense:+1.22	NaN		NaN	NaN	
3	+governance	3333	0.060856	1886	1850	1910	History / Biography (60%)	conflict:+2.07, governance:+2.03, commerce:+1.94	NaN		NaN	NaN	
4	+religion / +present-tense	3363	0.061403	1891	1865	1910	Other (35%)	religion:+2.10, education:+1.99, modal-heavy:+1.75	NaN		NaN	NaN	
5	+exclamatory / +emotion-words	3175	0.057971	1880	1850	1907	Poetry (29%)	exclamatory:+1.89, lexical-diversity:+1.72, emotion-words:+1.32	NaN		NaN	NaN	
6	+noun-heavy / +past-tense	726	0.013256	1886	1850	1910	Other (47%)	collective:+2.01, nature:+1.68, past-tense:+1.59	NaN		NaN	NaN	

Table 16: Top discriminating features — author-level Leiden communities (`first` run).

Cluster	Label	Feature	z	η^2	η^2 rank
0	+present-tense / +noun-heavy	noun-heavy	2.099711	0.083406	40
0	+present-tense / +noun-heavy	lexical-diversity	1.960179	0.153172	26
0	+present-tense / +noun-heavy	technology	1.833643	0.058280	49
1	+negation-heavy / +individual	family	1.810492	0.043207	51
1	+negation-heavy / +individual	negation-heavy	1.720281	0.285259	5
1	+negation-heavy / +individual	individual	1.617883	0.269277	6
2	+verb-heavy / +past-tense	law	1.818891	0.092433	38
2	+verb-heavy / +past-tense	verb-heavy	1.439349	0.405505	1
2	+verb-heavy / +past-tense	past-tense	1.222339	0.319980	4
3	+governance	conflict	2.071381	0.125924	30
3	+governance	governance	2.025116	0.082545	41
3	+governance	commerce	1.944163	0.094555	37
4	+religion / +present-tense	religion	2.095229	0.212343	10
4	+religion / +present-tense	education	1.990502	0.064227	47
4	+religion / +present-tense	modal-heavy	1.753557	0.058725	48
5	+exclamatory / +emotion-words	exclamatory	1.892702	0.198218	13
5	+exclamatory / +emotion-words	lexical-diversity	1.722742	0.074759	43
5	+exclamatory / +emotion-words	emotion-words	1.322774	0.150899	27
6	+noun-heavy / +past-tense	collective	2.007151	0.085751	39
6	+noun-heavy / +past-tense	nature	1.683827	0.078290	42
6	+noun-heavy / +past-tense	past-tense	1.594104	0.068161	46

Network — book level

Table 17: Per-community diagnostic — book-level Leiden communities (`first` run).

Cluster	Label	n	Share	Median year	IQR low	IQR high	Top genre	Top features	Boot.	$\bar{J}$	J_{q05}	J_{q95}	Verdict
0	+verb-heavy / +past-tense	7596	0.138692	1911	1902	1922	Adventure / Children's fiction (50%)	verb-heavy:+1.62, law:+1.58, death:+1.49	NaN		NaN	NaN	
1	+negation-heavy / +interrogative	7220	0.131826	1904	1880	1913	Fiction (general) (41%)	negation-heavy:+1.93, family:+1.67, adverb-heavy:+1.67	NaN		NaN	NaN	
2	+present-tense / +noun-heavy	6483	0.118370	1894	1867	1910	Other (46%)	present-tense:+1.23, noun-heavy:+1.12, lexical-diversity:+1.09	NaN		NaN	NaN	
3	+present-tense	5800	0.105899	1892	1878	1910	Other (38%)	religion:+3.12, education:+3.11, modal-heavy:+1.28	NaN		NaN	NaN	
4	+exclamatory / +interrogative	4154	0.075846	1880	1850	1900	Fiction (general) (42%)	exclamatory:+1.89, interrogative:+1.52, emotion-words:+1.23	NaN		NaN	NaN	
5	+exclamatory / +emotion-words	4024	0.073472	1880	1850	1909	Poetry (42%)	nature:+1.90, exclamatory:+1.83, lexical-diversity:+1.56	NaN		NaN	NaN	
6	+conflict / +past-tense	3382	0.061750	1894	1870	1910	History / Biography (69%)	conflict:+3.15, past-tense:+1.02, death:+0.87	NaN		NaN	NaN	
7	+past-tense / +conflict	3094	0.056492	1898	1870	1910	Other (39%)	collective:+2.48, nature:+1.64, death:+1.07	NaN		NaN	NaN	
8	+noun-heavy	2902	0.052986	1896	1857	1910	History / Biography (47%)	commerce:+3.21, governance:+2.68, law:+2.18	NaN		NaN	NaN	
9	+individual / +verb-heavy	2821	0.051507	1899	1880	1910	Fiction (general) (26%)	individual:+2.43, adverb-heavy:+1.28, verb-heavy:+1.11	NaN		NaN	NaN	
10	+noun-heavy / +present-tense	2798	0.051087	1902	1880	1912	Other (55%)	technology:+3.38, noun-heavy:+2.28, body:+1.85	NaN		NaN	NaN	
11	+long-sentence / +varied-length	2206	0.040278	1906	1880	1912	History / Biography (46%)	long-sentence:+3.43, varied-length:+3.28, verb-heavy:+2.16	NaN		NaN	NaN	
12	+present-tense / +exclamatory	1541	0.028136	1899	1880	1910	Drama (44%)	present-tense:+1.88, exclamatory:+1.43, individual:+1.39	NaN		NaN	NaN	
13	+exclamatory	748	0.013657	1897	1855	1913	Adventure / Children's fiction (34%)	family:+0.97, nature:+0.74, exclamatory:+0.52	NaN		NaN	NaN	

Table 18: Top discriminating features — book-level Leiden communities (`first` run).

Cluster	Label	Feature	z	η^2	η^2 rank
0	+verb-heavy / +past-tense	verb-heavy	1.617189	0.570408	3
0	+verb-heavy / +past-tense	law	1.576747	0.175085	44
0	+verb-heavy / +past-tense	death	1.493660	0.266020	27
1	+negation-heavy / +interrogative	negation-heavy	1.925029	0.432997	11
1	+negation-heavy / +interrogative	family	1.668658	0.248498	32
1	+negation-heavy / +interrogative	adverb-heavy	1.668304	0.311604	22
2	+present-tense / +noun-heavy	present-tense	1.230195	0.260992	28
2	+present-tense / +noun-heavy	noun-heavy	1.123093	0.117854	52
2	+present-tense / +noun-heavy	lexical-diversity	1.094764	0.208302	39
3	+present-tense	religion	3.115331	0.345838	16
3	+present-tense	education	3.109620	0.148982	49
3	+present-tense	modal-heavy	1.279766	0.276980	26
4	+exclamatory / +interrogative	exclamatory	1.887535	0.449438	10
4	+exclamatory / +interrogative	interrogative	1.515740	0.213672	36
4	+exclamatory / +interrogative	emotion-words	1.229115	0.464880	7
5	+exclamatory / +emotion-words	nature	1.903597	0.209661	38
5	+exclamatory / +emotion-words	exclamatory	1.830281	0.449438	10
5	+exclamatory / +emotion-words	lexical-diversity	1.563356	0.171307	47
6	+conflict / +past-tense	conflict	3.152085	0.379054	14
6	+conflict / +past-tense	past-tense	1.024958	0.588426	1
6	+conflict / +past-tense	death	0.874421	0.125847	51
7	+past-tense / +conflict	collective	2.482260	0.360907	15
7	+past-tense / +conflict	nature	1.636004	0.209661	38
7	+past-tense / +conflict	death	1.065235	0.266020	27
8	+noun-heavy	commerce	3.210623	0.212295	37
8	+noun-heavy	governance	2.680490	0.316353	21
8	+noun-heavy	law	2.175753	0.070605	54
9	+individual / +verb-heavy	individual	2.427816	0.571290	2
9	+individual / +verb-heavy	adverb-heavy	1.277910	0.311604	22
9	+individual / +verb-heavy	verb-heavy	1.112137	0.570408	3
10	+noun-heavy / +present-tense	technology	3.383029	0.218192	35
10	+noun-heavy / +present-tense	noun-heavy	2.283103	0.400284	13
10	+noun-heavy / +present-tense	body	1.845059	0.173805	45
11	+long-sentence / +varied-length	long-sentence	3.425608	0.556599	4
11	+long-sentence / +varied-length	varied-length	3.278401	0.495721	5
11	+long-sentence / +varied-length	verb-heavy	2.162116	0.173634	46
12	+present-tense / +exclamatory	present-tense	1.879762	0.458643	8
12	+present-tense / +exclamatory	exclamatory	1.426520	0.411363	12
12	+present-tense / +exclamatory	individual	1.392783	0.571290	2
13	+exclamatory	family	0.965426	0.248498	32
13	+exclamatory	nature	0.736486	0.209661	38
13	+exclamatory	exclamatory	0.515863	0.411363	12

I List of Abbreviations

ARI Adjusted Rand Index

BIC Bayesian Information Criterion

CI Confidence Interval

CPU Central Processing Unit

FAISS Facebook AI Similarity Search

FE Fixed Effects

GMM Gaussian Mixture Model

ICL Integrated Completed Likelihood

LCSH Library of Congress Subject Headings

LPM Linear Probability Model

NRC NRC Word-Emotion Association Lexicon

PCA Principal Component Analysis

POS Part of Speech

RDF Resource Description Framework

SD Standard Deviation

TTR Type-Token Ratio

UMAP Uniform Manifold Approximation and Projection

WLS Weighted Least Squares