

Genre 2.0? Embedding-Based Cluster Analysis as a Tool for Text Classification in Medieval Hebrew Literature

Annabelle Fuchs <fuchs_dot_annabelle_at_outlook_dot_com>, University of Haifa  <https://orcid.org/0009-0009-1712-0790>

DOI: <https://doi.org/10.63744/84hnsyf2s5tq>

Abstract

The classification of medieval Hebrew literature has long relied on historically inherited genre labels, often leading to misassignments and blurred textual boundaries. This study applies transformer-based cluster analysis to a corpus of 60 texts, using BEREL embeddings and hierarchical clustering to evaluate whether selected computational methods can provide empirically grounded, complementary insights to traditional genre classifications. The analysis identifies several stable clusters, including a distinct “Narrative” cluster, reinforcing prior research that questions the categorization of certain texts as “Aggadic Midrash.” While some clusters align with established classifications, others highlight ambiguities that challenge conventional taxonomies. The study demonstrates that computational clustering can systematically capture textual affinities, revealing relationships that may remain obscured in traditional approaches. These findings establish a methodological framework for reassessing genre structures in Hebrew literature, laying the groundwork for future research based on expanded datasets and manuscript evidence.

1. Introduction

Medieval Hebrew narrative literature has remained largely underrepresented in the study of Hebrew literary history. In contrast, fields such as Hebrew poetry and religious texts have traditionally occupied a more central position in Jewish Studies (cf. [Rosen and Yassif 2004]; [Yassif 2013]. One consequence of this imbalance is the continued reliance on foundational scholarship from the *Wissenschaft des Judentums*, the 19th- and early 20th-century academic movement that shaped the categorization of Hebrew texts (cf. [Fuchs 2024]). While these early studies were instrumental in establishing the field, they also introduced classificatory frameworks that remain in use despite their limitations. In particular, medieval Hebrew literary genres lack precise definitions, as their categorization continues to reflect historical terminologies rather than clear formal or stylistic distinctions. As [Fuchs 2024] has demonstrated, the term “Midrash” has often been applied indiscriminately to medieval texts that do not align with the genre’s original definition. This overextension of the “Midrash” label has obscured the distinctiveness of Hebrew narrative texts, relegating them to the margins of literary scholarship and limiting their systematic study.

Rather than advancing a renewed genre theory, this study takes an empirical approach, asking: Do Hebrew narrative texts, rather than being subsumed under the broad category of “Midrash,” form a distinct and coherent cluster when analyzed through quantitative methods? If so, this would indicate that computational approaches can detect textual affinities and distinctions that have been overlooked by traditional classifications. To explore this, the study applies embedding-based cluster analysis to a dataset of 60 medieval Hebrew texts, grouping them according to patterns captured by a transformer model rather than relying on predefined genre labels. The aim is to test whether computational methods can reveal textual affinities that align with or challenge existing scholarly categories. The findings lay the groundwork for further research that will refine the dataset and methodology, particularly through the inclusion of manuscript-based texts.

1

2

Transformer models like BEREL 2.0 are deep learning architectures trained on large corpora to recognize contextual patterns in language. Instead of relying on manually defined features, they generate embeddings — dense numerical representations that capture semantic, syntactic, and structural properties of a text. Each text in this study is represented by a vector summarizing its internal linguistic behavior. BEREL 2.0, trained specifically on historical Hebrew, enables the comparison of these vectors, allowing texts to be grouped not by external genre labels but by patterns in how they use language.

The remainder of this paper proceeds as follows: The next section provides an overview of previous computational approaches to Hebrew text analysis, tracing the development of stylometric methods, clustering techniques, and the more recent use of transformer-based embeddings in Jewish literary studies. This is followed by a description of the dataset and methodology, including text selection, preprocessing, and clustering procedures. The results section presents the key findings, highlighting the emerging clusters and their implications for existing classifications. Finally, the discussion considers both the potential and limitations of computational methods, and the conclusion outlines directions for future research.

2. Related Work

This chapter summarizes key computational approaches in Jewish literary studies, focusing on important predecessors in the fields of stylometry, cluster analysis, and embedding-based methods. [Dershowitz et al. 2015] introduced a computational approach to source criticism in the Hebrew Bible, using stylistic markers to distinguish between different authorial layers. The foundation for this work was laid already in 2011 and 2013 by Akiva and Koppel et al., who developed unsupervised methods for detecting multiple authorship, demonstrating the effectiveness of computational stylometry for Hebrew texts [Koppel et al. 2011] [Akiva and Koppel 2013]. [Bar-Asher Siegal and Shmidman 2018] applied similar techniques to reconstruct lost portions of *Mekhilta Deuteronomy*, combining textual pattern recognition with machine learning. While these studies primarily focused on authorship attribution and textual transmission, they demonstrated the viability of computational methods for Hebrew literary analysis, providing a foundation for subsequent work in text classification.

[Van Hecke 2018] demonstrated the effectiveness of hierarchical clustering for analyzing stylistic relationships in the Dead Sea Scrolls, later refining the methodology with [Van Heck and Joode 2021]. This approach was extended in 2020 by [Van der Schans et al. 2020], who applied recurrent neural networks (RNNs) to Biblical Hebrew, using clustering techniques to test diachronic models of linguistic variation. Both studies share a conceptual framework with the present research: They employ clustering algorithms to group texts based on stylistic or linguistic features, visualizing these relationships using dendrograms and dimensionality reduction techniques. The key distinction is that these studies relied on manually selected linguistic features or relatively shallow text representations, while the present study advances this methodology by leveraging BERT-based embeddings, automating feature extraction, and enabling deeper stylistic modeling.

Research in the field over the last years has predominantly applied computational stylometry and cluster analysis to antique texts, such as the Hebrew Bible and the Dead Sea Scrolls. More recent efforts, like those at the Open University of Israel, have shifted focus to modern Hebrew literature (see the project “Computational Stylistic Profiles for the Analysis of Modern Hebrew Prose”). The only known study examining a medieval Hebrew corpus is that of [Tannor et al. 2023], which classified rabbinic texts to identify lost *Midrash Tanhuma* material. For the broader history of computational textual analysis of the Hebrew Bible, see [van Peursen 2023].

A major development in computational Hebrew text analysis has been the introduction of transformer-based language models tailored to historical Hebrew texts. While other models, such as HeBERT [Chriqui and Yahav 2022] were trained on modern Hebrew, their application to pre-modern texts is limited. The introduction of BEREL [Shmidman et al. 2022], followed by its improved version, BEREL 2.0, marked the first model specifically designed for classical Hebrew (cf. BEREL 2.0 on Hugging Face). [Tannor et al. 2023] applied BEREL to classify rabbinic and medieval texts, demonstrating that it outperformed alternative models such as AlephBERT [Seker et al. 2022] in identifying complex language patterns and textual structures. Their findings confirmed that deep-learning-based text representations provide

a more robust foundation for Hebrew text classification than earlier feature-based approaches. This study extends their work by applying BEREL embeddings to medieval Hebrew narrative literature, a domain where computational clustering has not yet been explored.

The present study builds on the methodological foundations established by previous research, adapting them to a new textual corpus and a different literary-historical context. By combining BERT-based embeddings with hierarchical clustering, it explores whether this approach can offer a data-driven complement to traditional genre classification.

9

3. Data and Methodology

3.1. Corpus Selection

The dataset for this study consists of 60 Hebrew texts from antique, rabbinic, and medieval literary traditions, including a few modern Hebrew translations of earlier works originally preserved in other languages. The majority of these texts were obtained through the Git repository of the Sefaria website (Link to GitHub). Sefaria, a widely used online library of Jewish texts, categorizes its collection into the following groups: “Tanakh,” “Mishnah,” “Talmud,” “Kabbalah,” “Liturgy,” “Jewish Thought,” “Tosefta,” “Chasidut,” “Musal,” “Responsa,” and “Second Temple.” These categories reflect a conventional classification of Jewish texts and serve as a point of reference in this study, allowing for a systematic comparison between traditional categorizations and the data-driven clusters that emerge through computational analysis.

10

To ensure that the cluster analysis would produce distinct and meaningful clusters, five texts were selected from each of the following categories: “Tanakh” (= the Hebrew Bible), “Talmud,” “Aggadic Midrash,” “Halakhic Midrash,” “Halakha,” “Liturgy,” “Jewish Thought,” “Musal,” and “Second Temple.” The selection was made largely at random, with the only curating criterion being technical constraints, such as a minimum text length of 20,000 characters (though the shortest included text contains approximately 28,000 characters). This process resulted in a core dataset of 45 texts.

11

In addition, 12 further texts from Sefaria’s “Aggadic Midrash” category were included, each displaying varying degrees of narrative characteristics. These texts have been previously discussed in scholarly debates concerning their classification, making them particularly relevant for assessing how computational methods distinguish between Midrashic and narrative literature (e.g., [Börner-Klein 2004]; [Börner-Klein 2010]; [Börner-Klein 2016]; [Nikolsky and Atzman 2021]; [Fuchs 2024]). Their inclusion allows for a closer examination of whether narrative tendencies influence their clustering and to what extent they align with or diverge from traditionally defined Midrashim. Furthermore, three additional texts were incorporated into the dataset based on previous research indicating their narrative character. Since they are not available on Sefaria — and thus do not fall within its categorization — these texts were obtained from freely accessible online sources such as Wikitext.

12

Two texts from the previously mentioned category “Aggadic Midrash”, *Petirat Moshe* and *Petirat Aharon*, were transcribed directly from manuscript witnesses rather than drawn from later printed sources. Most other texts in the dataset are based on later textual traditions, including printed editions, later recensions, and editorially standardized versions. The planned move toward manuscript-based sources is discussed further in section “5.2. Limitations & Future Work.”

13

The 60 texts included in this study were preprocessed before analysis to ensure consistency and eliminate non-textual clues that could interfere with clustering. Special characters, annotations, and *nikud* (vowel markings), when present, were systematically removed; in addition, embedded biblical quotations were excluded as part of the preprocessing. Following this, texts exceeding 100,000 characters were truncated to that length, with the selected segment taken from the middle of the text to avoid biases introduced by introductory or concluding sections.

14

3.2. Text Representation

To generate numerical representations of the texts, this study employed BEREL 2.0, a transformer model specifically trained on rabbinical Hebrew texts [Shmidman et al. 2022]. BEREL is based on the BERT (Bidirectional Encoder

15

Representations from Transformers) architecture, which enables it to capture deep contextual and syntactic relationships within texts [Devlin et al. 2018]. BEREL 2.0 was selected as the sole embedding model for this experiment, as it has been found to be the most suitable for processing medieval Hebrew literature (cf. [Tannor et al. 2023]). While future research may include comparative evaluations of other transformer models, BEREL 2.0 proved to be effective for the present task.

Each text was tokenized into 512-token windows with a stride of 256, ensuring overlapping segments that preserved contextual continuity. The CLS token representation was extracted from each window, as this embedding encodes a summary of the segment's linguistic characteristics. To obtain a single representation per text, the CLS embeddings from all windows were averaged, creating a 768-dimensional vector that captured the overall semantic, syntactic, and structural properties of the text. This approach ensured that longer texts were not disproportionately influenced by their opening or closing sections, instead providing a balanced representation of the full document.

Since BERT-based embeddings are inherently high-dimensional, a two-step dimensionality reduction process was applied to facilitate clustering and visualization. The original 768-dimensional embeddings contain both meaningful linguistic patterns and noise, making direct clustering computationally inefficient. To retain the most relevant features, Principal Component Analysis (PCA) first reduced the embeddings to 50 dimensions, filtering out less informative variance while preserving key structural patterns. To enable clustering, Uniform Manifold Approximation and Projection (UMAP) was applied, reducing the data to two dimensions. UMAP was chosen over alternative methods such as t-SNE due to its ability to maintain both local and global relationships between texts.

This study clusters directly in UMAP 2D for computational feasibility and interpretability. Although this necessarily discards some variance, UMAP is designed to preserve meaningful structural relationships even in low dimensions. To mitigate potential instability, UMAP was run 50 times, and the final projection was averaged across runs, preventing random variations from distorting the clusters.

Given the effectiveness of this approach, as shown in the “4. Results” section, there was no immediate need to explore clustering in higher-dimensional spaces for this proof of concept. However, future studies may benefit from testing clustering in higher-dimensional UMAP spaces (e.g., 10D or 20D) to evaluate whether additional variance further refines classification results. For a discussion on the potential limitations of dimensionality reduction, see “5.2. Limitations & Future Work.”

3.3. Clustering Approach

To identify meaningful textual groupings, this study applied hierarchical clustering to the UMAP-reduced embeddings. Clustering, in this context, refers to the process of grouping texts based on numerical representations that capture internal patterns of language use and structure. The more similar two texts are, the closer their corresponding vectors are in the multidimensional space, which results in them being merged earlier in the hierarchical clustering process. Unlike methods such as K-means, hierarchical clustering does not require specifying the number of clusters in advance, making it well-suited for an exploratory study of textual relationships. Ward's method was chosen, as it minimizes intra-cluster variance, ensuring that texts within a cluster are as similar to each other as possible. Instead of setting a fixed number of clusters, the threshold for merging clusters was dynamically adjusted to produce a meaningful number of clusters, ranging between eight and ten.

To illustrate the clustering process, Figure 1 presents a UMAP projection based on a single clustering iteration. This visualization shows how texts are distributed in the two-dimensional space generated by UMAP, providing a simplified view of their relative positions in the high-dimensional embedding space. The projection offers an intuitive representation of the textual relationships captured in one iteration of the pipeline.

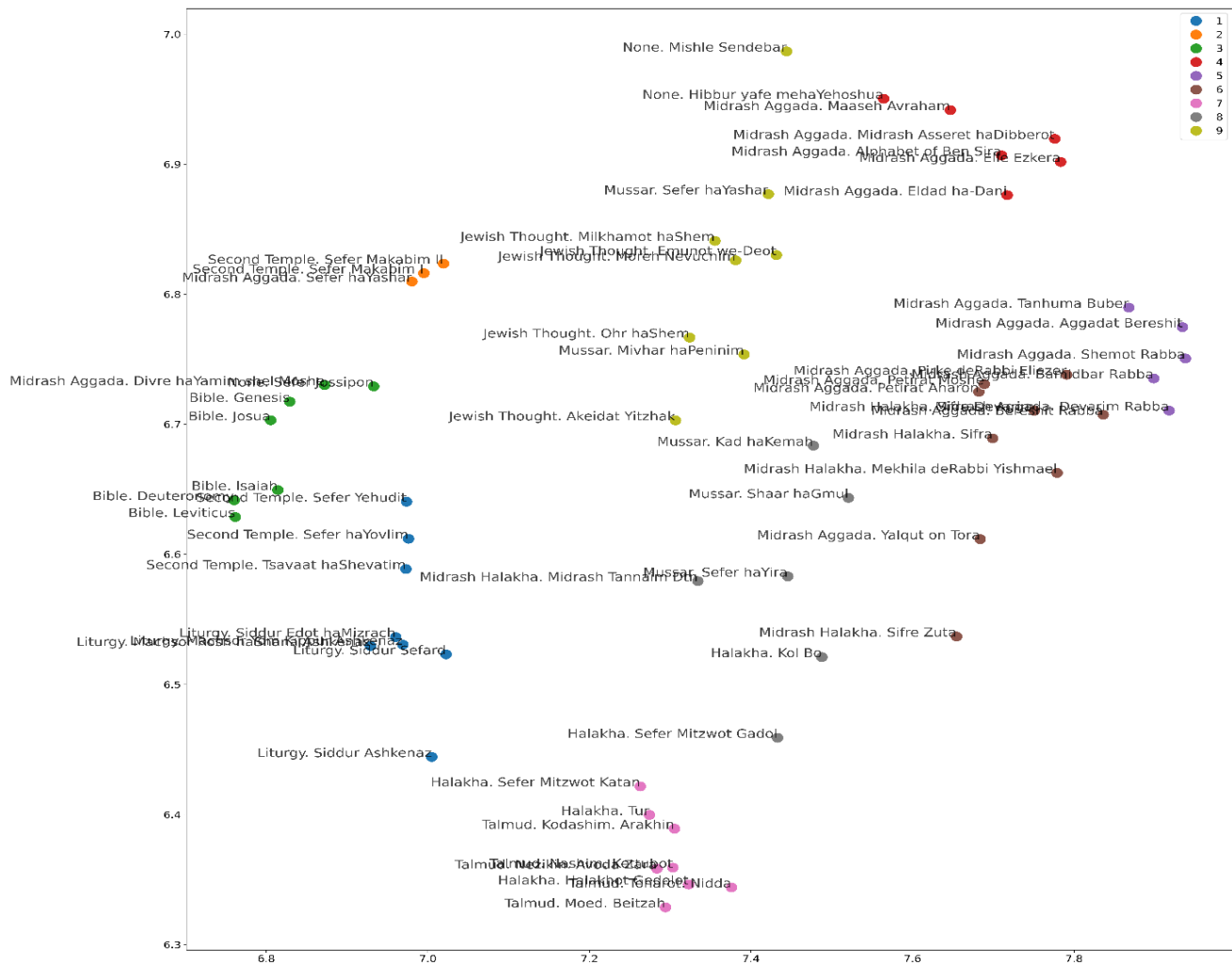


Figure 1. UMAP projection based on one clustering iteration. Each point represents a single text, and colors indicate cluster membership as determined by hierarchical clustering of the UMAP-reduced embeddings in that run. While the projection is simplified to two dimensions, it preserves local relationships between texts and offers an intuitive visualization of the grouping process. In this particular iteration, nine clusters were identified. For clarity and interpretive alignment, the descriptive cluster labels used in Table 1 are applied here, even though the exact groupings differ slightly from the final meta-clusters: 1 Blue: “Liturgy,” 2 Orange: “Pseudo-Bible,” 3 Green: “Bible/Pseudo-Bible,” 4 Red: “Narrative,” 5 Purple: “Midrash Aggada,” 6 Brown: “Midrash–Narrative Mix,” 7 Pink: “Talmud/Halakha,” 8 Grey: “Halakha–Philosophy Mix,” 9 Olive: “Philosophy.”

To ensure the robustness of the clustering results, the entire pipeline, including embedding generation, was repeated 100 times. Since transformer-based models like BEREL 2.0 produce slightly different embeddings in each run, individual cluster assignments showed minor variations across iterations. Rather than relying on a single clustering outcome, a co-occurrence matrix was constructed, tracking how often each text appeared in the same cluster across all 100 runs.

A second round of hierarchical clustering was then applied to this matrix, identifying stable groupings that consistently emerged across runs. These stable clusters — referred to in this article as meta-clusters — reflect recurrent textual affinities independent of run-to-run variation. Their distribution is visualized in a co-occurrence heatmap (Figure 2), which highlights the strength of associations between texts across the 100 iterations.

In addition to storing cluster memberships, the pipeline also recorded the pairwise distance between each pair of texts in every run. These distance values, derived from the position of each text in the reduced embedding space, represent the degree of dissimilarity between texts: the lower the value, the more similar the texts. By averaging these distances across all runs, a new distance matrix was created that reflects the overall proximity of texts throughout the experiment.

A second clustering solution was then derived by applying Ward’s hierarchical clustering directly to the averaged distance matrix. Unlike the meta-clusters, which reflect co-clustering frequency, this second approach sorts texts based

on cumulative pairwise dissimilarity. The result is a dendrogram showing hierarchical relationships among all texts, with cluster boundaries generated by an automated distance threshold (Figure 3).

Both the meta-clusters derived from co-occurrence patterns and the distance-based clusters obtained from averaged pairwise dissimilarities resulted in ten groupings. While the two structures broadly align, subtle differences highlight texts that occupy boundary positions between clusters — cases that are addressed in section “4. Results.”

3.4. Evaluation & Validation

The co-occurrence matrix confirms that clusters formed consistently, indicating that the results are not random but reflect meaningful textual similarities. The emerging clusters align closely with expected classifications, reinforcing confidence in the method. The heatmap in Figure 2 demonstrates that texts with known affinities are repeatedly grouped together across iterations. A complementary clustering structure, derived from the averaged distance measurements between texts, yielded a similar grouping of the corpus, further supporting the robustness of the approach. While factors such as text length and embedding variations were considered potential biases, their influence appears minimal. A more detailed evaluation of methodological limitations follows in section “5. Discussion.”

4. Results

4.1. General Overview

Both clustering approaches applied in this study — meta-clustering based on co-occurrence patterns and hierarchical clustering of averaged distance measurements — converged on a shared structure of ten distinct clusters. To preserve interpretability and analytical coherence, the number of clusters was constrained to a range between eight and ten; within this range, both methods independently arrived at a division into ten groups. This clustering structure is echoed in both visualizations (Figure 2 and Figure 3), highlighting how texts consistently formed meaningful groupings across runs.

Notably, these clusters do not always align with Sefaria’s predefined categories. Rather than relying on external (genre) classifications, they are derived from internal textual patterns — semantic, syntactic, and structural — offering a data-driven reflection of textual relationships. These results reinforce the limitations of traditional classification systems, such as those used by the Sefaria website, and confirm the potential of computational methods to produce more reliable textual groupings. The following table presents the ten clusters identified through both meta-clustering and distance-based clustering, allowing for direct comparison between the two methods.

Text	Distance-Based Cluster (Dendrogram)	Meta-Cluster (Heatmap)
Midrash Aggada. Shemot Rabba	1 (“Midrash Aggada”)	1 (“Midrash Aggada”)
Midrash Aggada. Devarim Rabba	1 (“Midrash Aggada”)	1 (“Midrash Aggada”)
Midrash Aggada. Bamidbar Rabba	1 (“Midrash Aggada”)	1 (“Midrash Aggada”)
Midrash Aggada. Bereshit Rabba	1 (“Midrash Aggada”)	1 (“Midrash Aggada”)
Midrash Aggada. Tanhuma Buber	1 (“Midrash Aggada”)	1 (“Midrash Aggada”)
Midrash Aggada. Aggadat Bereshit	1 (“Midrash Aggada”)	1 (“Midrash Aggada”)
Midrash Halakha. Mekhila	2 (“Midrash/Narrative Mix”)	1 (“Midrash Aggada”)

deRabbi Yishmael		
Midrash Aggada. Pirke deRabbi Eliezer	2 ("Midrash/Narrative Mix")	1 ("Midrash Aggada")
Midrash Aggada. Petirat Moshe	2 ("Midrash/Narrative Mix")	2 ("Midrash/Narrative Mix")
Midrash Aggada. Petirat Aharon	2 ("Midrash/Narrative Mix")	2 ("Midrash/Narrative Mix")
Midrash Halakha. Sifre Zuta	2 ("Midrash/Narrative Mix")	2 ("Midrash/Narrative Mix")
Midrash Aggada. Yalqut on Tora	2 ("Midrash/Narrative Mix")	2 ("Midrash/Narrative Mix")
Midrash Aggada. Eldad ha-Dani	3 ("Narrative")	3 ("Narrative")
Midrash Aggada. Alphabet of Ben Sira	3 ("Narrative")	3 ("Narrative")
Midrash Aggada. Midrash Asseret haDibberot	3 ("Narrative")	3 ("Narrative")
Midrash Aggada. Elle Ezkera	3 ("Narrative")	3 ("Narrative")
Midrash Halakha. Sifre Devarim	3 ("Narrative")	3 ("Narrative")
Midrash Halakha. Sifra	3 ("Narrative")	3 ("Narrative")
None. Hibbur yafe mehaYehoshua	4 ("Narrative II")	4 ("Narrative II")
Midrash Aggada. Maaseh Avraham	4 ("Narrative II")	4 ("Narrative II")
None. Mishle Sendebat	4 ("Narrative II")	4 ("Narrative II")
Mussar. Shaar haGmul	5 ("Philosophy")	5 ("Philosophy")
Mussar. Kad haKemah	5 ("Philosophy")	5 ("Philosophy")
Jewish Thought. Ohr haShem	5 ("Philosophy")	5 ("Philosophy")
Jewish Thought. Akeidat Yitzhak	5 ("Philosophy")	5 ("Philosophy")
Mussar. Mivhar haPeninim	5 ("Philosophy")	5 ("Philosophy")
Mussar. Sefer haYashar	5 ("Philosophy")	5 ("Philosophy")
Jewish Thought. Emunot we-Deot	5 ("Philosophy")	5 ("Philosophy")
Jewish Thought. Moreh Nevuchim	5 ("Philosophy")	5 ("Philosophy")
Jewish Thought. Milkhamot haShem	5 ("Philosophy")	5 ("Philosophy")
Halakha. Sefer Mitzvot	6 ("Halakha/Philosophy")	6

Gadol	Mix")	("Halakha/Philosophy Mix")
Halakha. Kol Bo	6 ("Halakha/Philosophy Mix")	6 ("Halakha/Philosophy Mix")
Mussar. Sefer haYira	6 ("Halakha/Philosophy Mix")	6 ("Halakha/Philosophy Mix")
Bible. Josua	7 ("Bible/Pseudo-Bible")	7 ("Bible/Pseudo-Bible")
Bible. Genesis	7 ("Bible/Pseudo-Bible")	7 ("Bible/Pseudo-Bible")
None. Sefer Jossipon	7 ("Bible/Pseudo-Bible")	7 ("Bible/Pseudo-Bible")
Midrash Aggada. Divre haYamim shel Moshe	7 ("Bible/Pseudo-Bible")	7 ("Bible/Pseudo-Bible")
Bible. Leviticus	7 ("Bible/Pseudo-Bible")	8 ("Bible")
Bible. Deuteronomy	7 ("Bible/Pseudo-Bible")	8 ("Bible")
Bible. Isaiah	7 ("Bible/Pseudo-Bible")	8 ("Bible")
Second Temple. Sefer Makabim I	8 ("Pseudo-Bible")	7 ("Bible/Pseudo-Bible")
Midrash Aggada. Sefer haYashar	8 ("Pseudo-Bible")	7 ("Bible/Pseudo-Bible")
Second Temple. Sefer Makabim II	8 ("Pseudo-Bible")	7 ("Bible/Pseudo-Bible")
Liturgy. Siddur Sefard	9 ("Liturgy")	9 ("Liturgy")
Liturgy. Siddur Ashkenaz	9 ("Liturgy")	9 ("Liturgy")
Liturgy. Machsor Yom Kippur Ashkenaz	9 ("Liturgy")	9 ("Liturgy")
Liturgy. Machsor Rosh haShana Ashkenaz	9 ("Liturgy")	9 ("Liturgy")
Liturgy. Siddur Edot haMizrach	9 ("Liturgy")	9 ("Liturgy")
Second Temple. Sefer haYovlim	9 ("Liturgy")	9 ("Liturgy")
Second Temple. Sefer Yehudit	9 ("Liturgy")	9 ("Liturgy")
Second Temple. Tsavaat haShevatim	9 ("Liturgy")	9 ("Liturgy")
Halakha. Tur	10 ("Talmud/Halakha")	10 ("Talmud/Halakha")
Halakha. Sefer Mitzwot	10 ("Talmud/Halakha")	10

Katan		("Talmud/Halakha")
Midrash Halakha. Midrash Tannaim Dtn	10 ("Talmud/Halakha")	10 ("Talmud/Halakha")
Talmud. Nezikin. Avoda Zara	10 ("Talmud/Halakha")	10 ("Talmud/Halakha")
Talmud. Moed. Beitzah	10 ("Talmud/Halakha")	10 ("Talmud/Halakha")
Talmud. Kodashim. Arakhin	10 ("Talmud/Halakha")	10 ("Talmud/Halakha")
Talmud. Nashim. Kettubot	10 ("Talmud/Halakha")	10 ("Talmud/Halakha")
Talmud. Toharot. Nidda	10 ("Talmud/Halakha")	10 ("Talmud/Halakha")
Halakha. Halakhot Gedolot	10 ("Talmud/Halakha")	10 ("Talmud/Halakha")

Table 1. The color scheme used in Table 1 mirrors that of Figure 2 and Figure 3, providing a consistent visual reference for comparing cluster assignments.

To maintain clarity and transparency throughout this research paper, each text is named according to a consistent convention. The name begins with its traditional classification as listed in Sefaria, followed by a period, and then the individual text title (e.g., “Midrash Aggada. Shemot Rabba”). This structure ensures that the original genre assignment remains visible, allowing for direct comparison between traditional categorizations and the clusters identified in this study.

The cluster names used in this study were assigned post hoc based on the composition of each group, aiming to provide a neutral and descriptive representation of the texts contained within them. Whenever possible, the labels were aligned with traditional literary classifications, ensuring that they reflect recognizable textual traditions rather than imposing new genre definitions.

This is particularly relevant when considering the assignment of the term “Narrative” to one of the clusters, which includes several texts traditionally classified as “Aggadic Midrash.” In previous research [Dan 1974] [Yassif 2009] [Fuchs 2024], it has been observed that certain texts labeled as “Midrash” exhibit a distinctly narrative structure while lacking the formal characteristics of midrashic literature, raising questions about their classification. However, defining clear criteria for distinguishing narrative texts from Midrashim has remained a challenge.

The cluster analysis presented here offers an empirical perspective on this issue, as one of the clusters emerged with a distinctive stylistic profile aligning with features typically associated with narrative texts. This grouping was not imposed in advance and only recognized retrospectively based on its clear separation from other clusters and its shared textual characteristics. By relying on measurable textual similarities, rather than preconceived genre definitions, this approach provides a data-driven means of reassessing textual classifications. The naming conventions used throughout this research paper serve as interpretative tools, describing the observed clustering patterns without enforcing rigid genre distinctions.

4.2. Visualization of Results

Two visualizations offer insights into the structure and stability of the detected textual groupings: a co-occurrence heatmap (Figure 2) and a dendrogram based on averaged distance measurements (Figure 3). Both visualizations employ Ward’s method for clustering, though they rely on different types of data extracted from the clustering process.

The co-occurrence heatmap (Figure 2) presents the frequency with which each pair of texts appeared in the same cluster across 100 full iterations of the pipeline. Each cell in the matrix represents a pairwise relationship, with warmer colors indicating higher co-clustering consistency and cooler colors reflecting weaker associations. This visualization highlights the most stable clusters in the corpus; while, for example, the groups labeled “Talmud/Halakha,” “Liturgy,” and “Philosophy” show strong internal cohesion, other clusters, such as the transitional group “Midrash/Narrative Mix,” display more fluid boundaries, suggesting stylistic overlap or ambiguity in textual affinities.

The heatmap provides an interpretable summary of long-run clustering behavior. Unlike single-run visualizations, it emphasizes robust relationships by averaging over many iterations. Cluster assignments in this visualization form the basis for the meta-cluster structure used in the descriptive table above.



Figure 2. Heatmap of co-occurrence frequencies across 100 clustering iterations. Each cell shows how often a pair of texts was assigned to the same cluster. Warmer colors indicate stronger co-clustering tendencies. Clusters were generated via hierarchical clustering of this co-occurrence matrix using Ward’s method. The color scheme matches the cluster labels in Table 1: Red: “Narrative II,” Blue: “Midrash Aggada,” Green: “Midrash–Narrative Mix,” Purple: “Midrash Aggada,” Orange: “Halakha–Philosophy Mix,” Brown: “Philosophy,” Pink: “Talmud/Halakha,” Teal: “Liturgy,” Grey: “Bible,” Black: “Pseudo-Bible/Bible.”

In contrast to the heatmap, which is based on co-occurrence patterns, the dendrogram (Figure 3) visualizes the average pairwise distances between texts across the 100 clustering runs. These distances reflect how similar or dissimilar each text is to every other text, based on their relative positions in the embedding space. The dendrogram arranges the texts hierarchically, grouping the most similar pairs first, and successively merging them into larger clusters.

The horizontal axis in the dendrogram represents distance: the further to the right the merging point, the more similar the texts. Cluster divisions were produced by applying a threshold of 0.26 to the tree structure, resulting in a division into ten groups. However, the dendrogram also allows for a more nuanced reading: even when texts fall into different

clusters, their proximity in the tree reveals how closely related they remain.

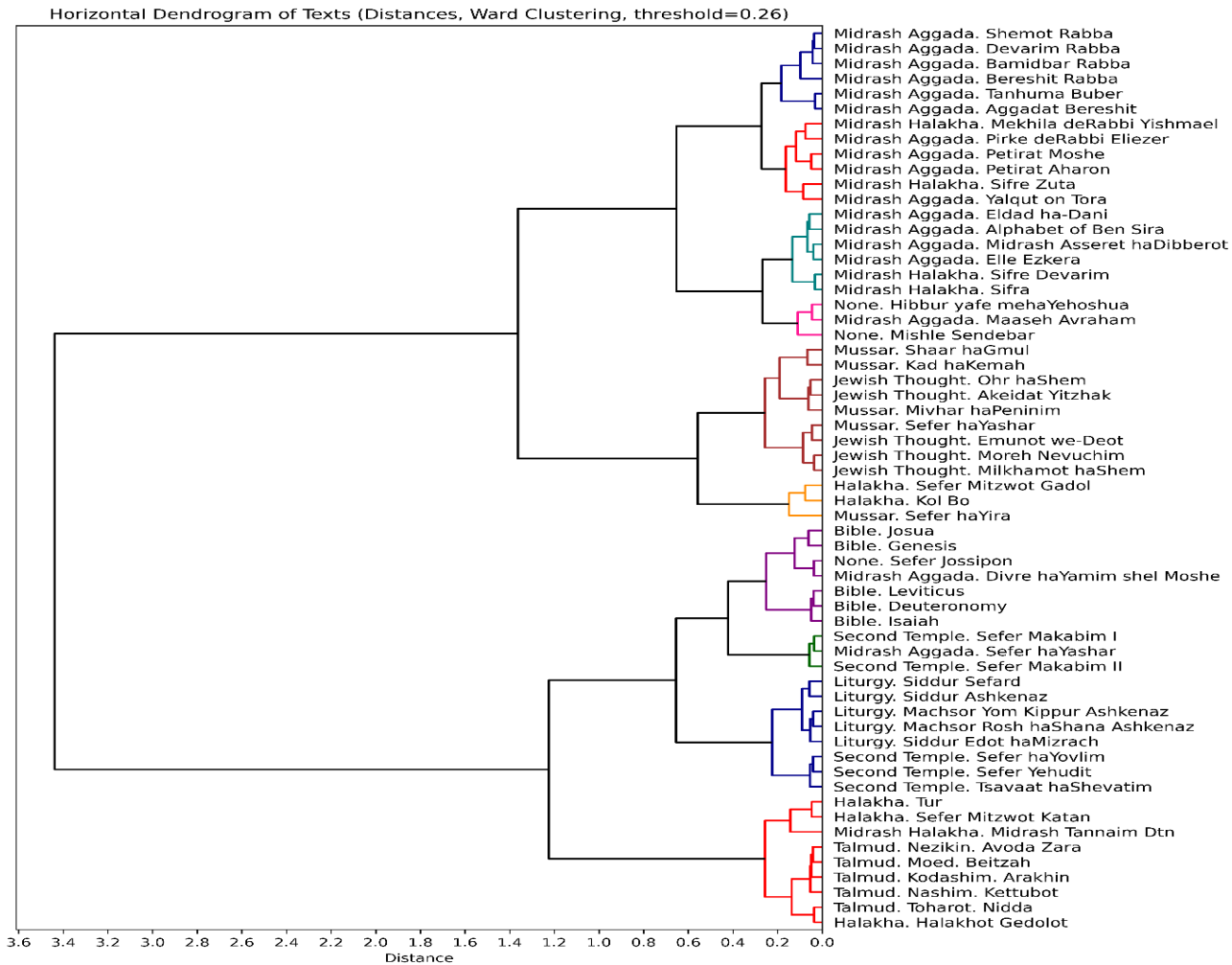


Figure 3. Dendrogram showing hierarchical clustering of averaged pairwise distances between texts (across 100 runs). Ward's method was used for clustering. A threshold of 0.26 was applied to determine ten clusters. The horizontal axis represents dissimilarity; shorter branches indicate greater similarity. The color scheme corresponds to Table 1: Blue: "Midrash Aggada," Red: "Midrash/Narrative Mix," Teal: "Narrative," Pink: "Narrative II," Brown: "Philosophy," Orange: "Halakha-Philosophy Mix," Purple: "Pseudo-Bible/Bible," Green: "Pseudo-Bible," Blue II: "Liturgy," Red II: "Talmud/Halakha."

4.3. Key Findings

The clustering results reveal stable and meaningful groupings, though these do not always align with Sefaria's predefined categories. Rather than replicating traditional genre classifications, the clusters reflect patterns in textual structure, language use, and composition, allowing texts to be grouped based on internal features rather than external labels. This highlights how embedding-based similarities can transcend historical or thematic classifications, resulting in clusters that are broader and more flexible than conventional genre categories.

The following discussion of cluster contents proceeds in the order of appearance in Table 1, which lists both the meta-cluster and distance-based cluster assignments.

The first cluster may be described using the label "Midrash Aggada." In the distance-based clustering, this group is relatively cohesive, comprising six core texts: *Bereshit Rabba*, *Shemot Rabba*, *Bamidbar Rabba*, *Devarim Rabba*, *Tanhuma Buber*, and *Aggadat Bereshit*. The meta-clustering approach includes two additional works: *Pirke deRabbi Eliezer* and *Mekhilta deRabbi Yishmael*. The latter, however, already belongs to a more mixed grouping of midrashic and narrative works in the distance-based clustering.

Despite these minor differences in cluster boundaries, both approaches confirm the internal stylistic coherence of the core aggadic midrashim. Their textual similarities outweigh distinctions in historical context or thematic content, supporting the conclusion that their classification as “Aggadic Midrashim” is not merely conventional but grounded in measurable linguistic and structural features. Notably, the included texts span from approximately the 4th to the 12th centuries, yet they are consistently grouped together by the model. This suggests that the features captured by BEREL, such as lexical choice, syntactic structure, and discursive patterns, reflect a unique stylistic cluster profile that persists across centuries, and therefore includes both earlier and later Midrashim alike.

42

The next group can be described as “Midrash/Narrative Mix.” It appears in both clustering outcomes, albeit with slight differences in text allocation, as previously discussed. In both cases, the works in this cluster are positioned between the “Midrash Aggada” and the “Narrative” groups, suggesting that they share features with both. Most of the included texts are traditionally classified as “Midrash Aggada,” though *Sifre Zuta* stands out as an exception, having been traditionally identified as a halakhic midrash. The cluster analysis indicates that all of these works exhibit at least partial narrative characteristics, positioning them stylistically between the more homogeneous “Midrash Aggada” and “Narrative” clusters.

43

The third group, described as the “Narrative” cluster, is identical in both clustering approaches. It consists of six works and is closely related to a fourth cluster that also contains texts with predominantly narrative character. The separation of these two groups results from the threshold applied during the clustering process: Had the cutoff been set slightly higher, they would have merged into a single, broader cluster.

44

The emergence of this distinct Narrative cluster is a central finding of the study. This grouping includes texts that are traditionally labeled as “Aggadic Midrashim,” not only in Sefaria’s classification but also in broader scholarly discourse. Some even bear the word “Midrash” in their titles, such as *Midrash Aseret ha-Dibberot* (“The Midrash of the Ten Commandments”). Despite these labels, the clustering consistently identified their stronger affinity with other works characterized by narrative structure, suggesting that their stylistic and linguistic features align more closely with narrative compositions than with midrashic works from the “Midrash Aggada” group. The clustering process, grounded in internal textual features rather than inherited genre designations, highlights these affinities and challenges the conventional classification schemes that have long shaped the field.

45

The next cluster, broadly labeled “Philosophy,” appears identically in both clustering outcomes. It brings together texts that Sefaria distinguishes as either “Musal” or “Jewish Thought,” but which the clustering process grouped into a single coherent category. From a computational perspective, this suggests that the boundary between these two designations may be less rigid than traditionally assumed.

46

A related cluster contains just three texts, *Sefer Mitzvot Gadol*, *Kol Bo*, and *Sefer haYira*, and also appears identically in both clustering outcomes. Conventionally, these works are classified as either halakhic (*Sefer Mitzvot Gadol*, *Kol Bo*) or ethical (*Sefer haYira*). In the meta-clustering, they show partial overlap with both the “Talmud/Halakha” and “Philosophy” clusters, while in the distance-based clustering, they are positioned closer to the broader “Philosophy” group. Nonetheless, these three works demonstrate a consistent stylistic affinity with one another that sets them apart from their neighboring clusters. Their internal cohesion was strong enough for both clustering methods to isolate them as a distinct group. For the purposes of this study, this cluster is labeled “Halakha–Philosophy Mix,” reflecting the dual influence of legal and ethical-philosophical traditions within these texts.

47

A group of works broadly categorized as “Bible/Pseudo-Bible” appears in both clustering outcomes, though with minor internal differences. The overall cluster consists of ten texts that are either part of the biblical canon or imitate its style. Both clustering methods separate the group into two smaller subclusters but do so in different ways.

48

In the distance-based clustering, three works, *Sefer Makkabim I*, *Sefer Makkabim II*, and the midrashic *Sefer haYashar*, are set apart as a stylistically coherent subgroup. These texts of different backgrounds imitate the language and structure of the biblical narrative, marking them as “pseudo-biblical” compositions. In contrast, the meta-clustering approach isolates three canonical books, *Leviticus*, *Deuteronomy*, and *Isaiah*, as a separate cluster, while grouping *Genesis* and *Joshua* together with pseudo-biblical texts.

49

Both clustering approaches yield results that are interpretable: Either the canonical books (with the exception of *Genesis* and *Joshua*) form a separate group, or the pseudo-biblical works of later origin cluster together. The placement of *Genesis* and *Joshua* alongside biblical imitations reflects their influence as stylistic models, while the grouping of later pseudo-biblical compositions underscores their shared narrative form. Although it is difficult to determine exactly which features led to the separation in each case, the resulting groupings align well with known patterns and raise no unexpected contradictions. 50

The “Liturgy” cluster is likewise identical in both clustering outcomes and forms a cohesive group overall. Within the cluster, however, the visualizations reveal two internally distinct subgroups. One consists of the prayer books — texts labeled as *Siddur* or *Machzor* — while the other includes works traditionally classified as “Second Temple.” Although these two subgroups originate from different historical contexts and were composed for different purposes, their proximity in the embedding space suggests a level of formal or discursive similarity. The transformer-based model recognizes consistent textual features across both types — perhaps in terms of formulaic structure, elevated diction, or liturgical framing — which brings them into alignment despite their thematic and functional differences. The clustering thus reflects both their shared stylistic core and their internal differentiation. 51

The final cluster, labeled “Halakha,” is the most internally unified group in the dataset and appears identically in both clustering outcomes. The distinctive linguistic and structural features characteristic of talmudic-halakhic discourse set them apart from all other works included in the study. Their strong affiliation, detected across both the co-occurrence and distance-based analyses, reflects and reinforces traditional genre attributions. The halakhic texts form a vector-space signature, further validating the clustering approach used in this study. 52

4.4. Notable Observations and Interpretive Implications

Atypical placements within the clustering process can offer valuable entry points for deeper textual analysis. Whether through unexpected affinities or hybrid textual profiles, such results highlight irregularities that may hold particular relevance for scholars working on the individual texts. More broadly, the analysis demonstrates that computational clustering can uncover patterns of literary behavior that transcend conventional genre categories, inviting both methodological reflection and new avenues of literary interpretation. 53

One such case is the placement of *Midrash Tannaim*. While this work consistently aligns with halakhic texts in the distance-based clustering, the meta-clustering reveals an interesting pattern: across several iterations, *Midrash Tannaim* also clustered with liturgical texts. This positioning suggests the presence of stylistic features that link it not only to halakhic discourse but also to formal or structural traits found in liturgical compositions. Given that parts of this work are later reconstructions, the observed behavior may reflect influences from the sources used in this process, offering valuable clues for scholars investigating the text’s editorial history and compositional layers. 54

Another example of interpretive interest is the work *Mishle Sendebār*. As a Hebrew version of a widely circulated medieval frame-tale narrative, it exists in multiple cultural and linguistic traditions. *Mishle Sendebār* clusters consistently with the “Narrative” group in both clustering outcomes. The co-occurrence heatmap, however, reveals a weaker secondary affinity with the “Philosophy” cluster. This outcome reflects a certain stylistic hybridity, likely connected to the text’s Arabic *Vorlage*. 55

A third case that invites closer attention is the clustering of *Petirat Moshe* and *Petirat Aharon*. The analysis shows that these texts do not clearly align with the “Midrash Aggadah” cluster, as one might expect based on traditional classification. This outcome reflects recent scholarship that places them in a more narrative framework [Fuchs 2024]. Rather than forcing a definitive categorization, the embedding-based analysis identifies stylistic affinities with both narrative and midrashic texts. This dual positioning offers a compelling, data-driven starting point for integrating traditional genre assignments with more recent interpretive developments, demonstrating the method’s potential to mediate between competing classifications. 56

The scattering of texts traditionally grouped under “Midrash Halakha” across several clusters further underscores the model’s sensitivity to internal textual variation. While some of these texts align with the “Halakha” group, others are 57

distributed across the “Narrative” and “Midrash/Narrative Mix” clusters. This dispersion may be partly influenced by technical factors such as text truncation (see “5.2. Limitations & Future Work”), but in many cases, only detailed case studies will uncover the textual or structural features that shaped these results. Their varied placement should be seen not as anomalies, but as productive starting points for future research.

A last pattern to be pointed out is the clustering of texts composed in a pseudo-biblical style from different historical periods. It supports the idea that biblical imitation functions as a stylistic marker independent of historical or canonical context. This reinforces the view that so-called “Rewritten Bible” traditions can be empirically identified through computational analysis, and that “rewriting” is best understood as an ongoing literary method rather than a discrete genre category (cf. [Zsengellér 2014]).

This article does not aim to offer a comprehensive interpretation of each individual observation derived from the presented clustering results. Rather, it serves as a proof of concept for embedding-based cluster analysis in the study of Hebrew literature. The findings presented here demonstrate that the method can reveal meaningful textual relationships and genre-blurring tendencies — details that merit closer examination in future research.

5. Discussion

5.1. Implications

One of the challenges in applying embedding-based methods to literary classification is the opacity of machine-generated clusters. Unlike traditional literary categories, which are explicitly defined based on historical, thematic, or formal criteria, computational clusters emerge from statistical patterns in textual features — patterns that remain largely uninterpretable from a data science perspective. Coherent and meaningful clusters emerge, but the model’s rationale for linking specific texts cannot always be fully explained, as it draws on intricate and largely opaque similarity structures in the embedding space.

From a humanistic perspective, this ambiguity does not diminish the value of the results but instead offers two distinct paths for further research. First, scholars can attempt to interpret the clusters post hoc, identifying the underlying stylistic, rhetorical, or structural features that justify their formation. For example, what linguistic elements unify “Rewritten Bible” texts across different historical periods? Second, these clusters can serve as a starting point for broader literary and cultural inquiries. Scholars might, e.g., compare how specific themes, such as the Death of Moses, manifest in texts within the “Narrative” cluster versus those in the “Midrash Aggada” cluster. By grounding such studies in empirically derived groupings rather than inherited genre labels, we increase the likelihood of identifying meaningful textual relationships.

The results of this study demonstrate that embedding-based cluster analysis is not merely a means of validating existing classifications; it actively challenges and refines them. The clusters identified here do not simply replicate conventional genre labels but instead reveal measurable textual affinities, highlighting both expected relationships and areas where traditional classifications obscure significant textual connections.

One of the most striking findings is the emergence of a distinct cluster that, based on its composition, has been retrospectively labeled “Narrative.” The cluster itself formed automatically; its label was only assigned post-analysis, reinforcing the argument that certain texts traditionally classified as “Aggadic Midrash” are, in fact, better understood as independent narrative compositions.

It should be mentioned that this does not mean narrativity is confined to this cluster. Many texts outside of it, such as those categorized as “Bible/Pseudo-Bible,” “Talmud/Halakha,” and “Midrash Aggada,” contain strong narrative elements yet do not cluster with the “Narrative” group. The texts in the “Narrative” cluster are so stylistically dominated by narrative structure that they form a computationally detectable unit, separate from texts that incorporate narrative elements within a broader literary framework.

By analyzing texts based on measurable intrinsic similarities rather than inherited classifications, this study reveals new

relationships between works that may have been overlooked due to historical misclassifications. Traditional taxonomies have often grouped texts together based on assumptions about their function, theme, or historical context, whereas embedding-based clustering highlights affinities rooted in linguistic form, structural organization, and deeper representational patterns. As a result, certain connections have remained obscured, while other groupings may have been artificially imposed. The findings here demonstrate how computational methods can help correct and refine these historical classifications, allowing us to identify clusters that more accurately reflect textual composition.

Importantly, embedding-based clustering does not replace traditional literary scholarship — rather, it enhances it by providing a systematic and empirical basis for text classification. The results confirm that computational methods can reliably distinguish well-defined textual traditions, while also highlighting ambiguous cases that demand further literary and philological analysis. By extending these methods, future research can explore broader questions of intertextuality, manuscript transmission, and literary influence within Hebrew literature.

66

5.2. Limitations & Future Work

The results of this study provide meaningful insights into the classification of medieval Hebrew texts, although several limitations must be considered when interpreting the findings.

67

First, the textual sources used in this analysis present certain constraints. The corpus consists of online-available versions, which often reflect later textual traditions rather than manuscript-based readings. Manuscript sources are generally preferable, as they preserve earlier textual layers and minimize the influence of editorial reshaping found in printed editions. Future research should aim to work directly with manuscript evidence, which would allow for a more precise understanding of textual features without the influence of later edited text versions. New transcriptions of handwritten Hebrew sources are expected to become available soon through the ERC-funded project MiDRASH (<https://doi.org/10.3030/101071829>). Comparative experience with other medieval corpora may also offer useful points of reference (cf. [Guéville and Wrisley 2024]). The fact that two texts, *Petirat Moshe* and *Petirat Aharon*, were transcribed from manuscript sources rather than from later recensions had no known impact on the presented results. An additional constraint lies in the fact that, although all texts were cleaned and standardized, the corpus as a whole was not manually checked for editorial inconsistencies.

68

Second, the size of the corpus imposes certain limitations. While 60 texts provide a preliminary overview, they do not allow for fine-grained distinctions between subcategories. This limitation becomes evident in cases such as *Tanhuma Buber*, which clustered with “Midrash Aggada” due to the absence of other texts from the *Tanhuma-Yelammedenu* genre in the dataset (cf. [Bregman 2003]; [Nikolsky and Atzman 2021]). Likely, a more extensive dataset would enable clearer distinctions between only loosely related texts or cases like the “Halakha/Philosophy Mix” micro-cluster. Furthermore, corpus selection was constrained by availability. Texts were chosen not only based on scholarly relevance but also on accessibility, meaning that some works were omitted simply due to a lack of reliable online versions.

69

In addition, the corpus contains a relative imbalance, as the Midrash category is more strongly represented than others. This reflects the central role of the Midrash–narrative distinction in the present study. The results indicate that this imbalance did not distort the clustering: Smaller but more distinct categories, such as “Talmud,” “Liturgy,” and “Philosophy,” still formed stable groups, while the “Midrashic–narrative” spectrum displayed internal differentiations that align with philological expectations. This suggests that the embeddings captured genuine textual similarities rather than simply reproducing prior group sizes. As a proof of concept, the present study provides a foundation for further research, and its results strongly encourage follow-up studies including broader and more balanced corpora.

70

A further limitation concerns text truncation. To ensure uniformity, all texts were analyzed as 100,000-character samples, always taken from the middle section of each text. On the one hand, this approach provides a consistent basis for comparison; on the other hand, it may have affected the classification of longer texts. For example, the work *Sifra* was assigned to the “Narrative” cluster. It is possible that selecting a different text segment would have resulted in stronger alignment with the “Talmud/Halakha” group. This suggests that truncation may have influenced the clustering results in some cases. Future studies should explore the impact of different segmentation strategies.

71

Additionally, the dataset includes texts of varying lengths, ranging from 28,000 to 100,000 characters. No strong correlation between text length and clustering behavior was observed, but the possibility of length-based biases cannot be entirely ruled out. A more controlled dataset — where text length is systematically balanced — could help determine whether this factor has any impact on clustering results. 72

When interpreting the findings of this study, the potential effect of dimensionality reduction on clustering accuracy must also be taken into account. While the two-step reduction (PCA to 50D, UMAP to 2D) effectively structured the data for insightful preliminary observations, clustering directly in 2D may have discarded some variance. This choice prioritized interpretability and computational feasibility. The results demonstrate that stable, meaningful clusters emerged despite this reduction. Future studies, however, may explore clustering in higher-dimensional spaces (e.g., 10D or 20D) to assess whether additional variance improves classification accuracy. 73

Another consideration is the choice of BEREL embeddings. No alternative models were tested for comparison. The decision to use BEREL 2.0 was based on prior research indicating its effectiveness for handling pre-modern Hebrew texts, and the results of this study confirm that it produced meaningful and expected clusters. Future research could still explore whether other transformer-based models, potentially fine-tuned on medieval Hebrew corpora, yield similar or even improved results. Testing multiple embeddings could help determine whether stylistic clustering outcomes are model-dependent or if they remain stable across different representations of textual features. 74

A further methodological limitation concerns the interpretability of embedding-based clustering. Since transformer models encode texts as dense vectors without an explicit decoder, the rationale for grouping particular texts cannot be directly recovered. In this respect, the method applied here produces clusters whose internal logic must be analyzed post hoc through philological and literary analysis, which provides a humanistic framework for interpretation rather than an explanation of the encoding process itself. Future research may seek to complement this approach with more fine-grained examinations on the digital side, for instance, by combining embedding-based clustering with selected stylometric features, or by experimenting with segment-level representations rather than aggregated text embeddings. Such strategies will not render the embedding process transparent, but they may provide additional interpretive traction and help clarify whether the observed affinities hold consistently across different portions of a text. 75

Despite the outlined limitations, the overall clustering patterns align well with expected textual relationships, confirming the robustness of the presented method. 76

6. Conclusion

This study has demonstrated that embedding-based clustering is an effective tool for identifying textual affinities in Hebrew literature. The results confirm that computational methods can both validate traditional classifications and highlight areas where inherited taxonomies require refinement. The emergence of a distinct “Narrative” cluster supports the argument to reconsider some historically inherited genre labels in medieval Hebrew literature, suggesting that certain texts traditionally categorized as “Aggadic Midrash” may be more accurately described as independent narrative compositions under contemporary classification frameworks. At the same time, clusters such as “Liturgy,” “Talmud/Halakha,” or “Bible/Pseudo-Bible” proved to be highly stable, reinforcing established textual groupings. Other clusters, such as “Midrash/Narrative Mix,” exhibit a more fluid internal structure, which provides valuable insights into texts that occupy intermediate positions between literary traditions. 77

Beyond its immediate findings, this study demonstrates that computational approaches provide a powerful, data-driven means of identifying textual relationships. Categorization remains an essential scholarly tool — allowing for meaningful comparisons, distinctions, and analyses of parallel developments. Without classification, it would be impossible to trace literary dependencies, influences, and thematic connections across texts. However, within the study of medieval Jewish literature, existing classifications have often lacked precision, relying on overly broad terminology that has not proven analytically useful. Instead of simply redefining literary genres, this study argues that a parallel approach is now possible — one that, in addition to traditional reasoning, empirically derives textual relationships from semantic, structural, and rhetorical patterns. This is a methodological enhancement that has only recently become feasible, as advances in computational analysis allow for objective, data-driven classification systems. 78

A valuable outcome of this study is that the clustering methodology and computational pipeline developed here can be easily reapplied to larger and more refined corpora. As more manuscript-based texts become available, the same code and algorithm can be used to test whether the same stable clusters emerge in a more comprehensive dataset. This ensures that the approach remains scalable and adaptable, allowing for continued refinement of textual classifications without requiring significant modifications to the underlying methodology.

Although these findings are preliminary, they provide a solid foundation for further research. Future studies should expand the dataset, incorporate manuscript-based texts, and refine clustering methodologies. Furthermore, the effectiveness of BEREL 2.0 embeddings in this experiment suggests that additional comparative tests with other models could further enhance the reliability of transformer-based classifications. While immediate follow-up research will likely focus on refining the computational approach, the present study has already yielded meaningful results on the level of literary analysis. The observed cluster structures not only validate the method but also point to interpretive insights that merit deeper engagement within Jewish literary studies. A more detailed investigation into medieval Hebrew narrative literature is already planned as part of the upcoming Digital Humanities postdoctoral project “Dynamics in the Polysystem: A Study of Hebrew Narrative Prose and its Changing Status within the Jewish Literary Canon of the Middle Ages,” which will be conducted at the University of Haifa from November 2025 onward, funded as a Walter Benjamin Fellowship by the DFG (German Research Association).

Works Cited

- Akiva and Koppel 2013** Akiva, N. and Koppel, M. (2013) “A generic unsupervised method for decomposing multi-author documents”, *Journal of the American Society for Information Science and Technology*, 64(11), pp. 2256-2264. Available at: <https://doi.org/10.1002/asi.22924>.
- Bar-Asher Siegal and Shmidman 2018** Bar-Asher Siegal, M. and Shmidman, A. (2018) “Reconstruction of the Mekhilta Deuteronomy using philological and computational tools”, *Journal of Ancient Judaism*, 9(1), pp. 2-25. Available at: <https://doi.org/10.30965/21967954-00901002>.
- Börner-Klein 2004** Börner-Klein, D. (2004) *Pirke de-Rabbi Elieser: Nach der Edition Venedig 1544 unter Berücksichtigung der Edition Warschau 1852*. Berlin: Walter de Gruyter.
- Börner-Klein 2010** Börner-Klein, D. (2010) “Tell me who I am: reading the Alphabet of Ben Sira”, in H. Liss and M. Oeming (eds.) *Literary Construction of Identity in the Ancient World*. Winona Lake, IN, pp. 135-144. Available at: <https://doi.org/10.5325/j.ctv1bxh3w5.11>.
- Börner-Klein 2016** Börner-Klein, D. (2016) “Yalkut Shimoni Numbers on Proselytes”, in C. Cordoni and G. Langer (eds.) *Let the Wise Listen and Add to Their Learning (Prov 1:5). Festschrift for Günter Stemberger on the Occasion for his 75th Birthday*. Berlin, Boston: De Gruyter, pp. 431-444. Available at: <https://doi.org/10.1515/9783110435283-024>.
- Bregman 2003** Bregman, M. (2003) *The Tanhuma-Yelammedenu Literature: Studies in the Evolution of the Versions*. Piscataway, NJ: Gorgias Press.
- Chriqui and Yahav 2022** Chriqui, A. and Yahav, I. (2022) “HeBERT and HebEMO: a Hebrew BERT model and a tool for polarity analysis and emotion recognition”, *INFORMS Journal on Data Science*, 1(1), pp. 81-95. Available at: <https://doi.org/10.1287/ijds.2022.0016>.
- Dan 1974** Dan, J. (1974) *Hebrew Story in the Middle Ages [Hebrew]*. Jerusalem: Keter.
- Dershowitz et al. 2015** Dershowitz, I., et al. (2015) “Computerized source criticism of biblical texts”, *Journal of Biblical Literature*, 134(2), pp. 253-271. Available at: <https://muse.jhu.edu/article/582808>.
- Devlin et al. 2018** Devlin, J., et al. (2018) “BERT: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint*, arXiv:1810.04805. Available at: <https://arxiv.org/abs/1810.04805>.
- Fuchs 2024** Fuchs, A. (2024) *Neither “Minor” nor “Midrash” – “Medieval Hebrew Bibletales” as an alternative to the concept of the “Minor Medieval Midrashim” [German]*. University Library Heidelberg. Available at: <https://doi.org/10.11588/heidok.00035472>.
- Guéville and Wrisley 2024** Guéville, E. and Wrisley, D.J. (2024) “Transcribing medieval manuscripts for machine learning”, *Journal of Data Mining & Digital Humanities*, Available at: <https://doi.org/10.48550/arXiv.2207.07726>.

- Koppel et al. 2011** Koppel, M., et al. (2011) “Unsupervised decomposition of a document into authorial components”, in D. Lin, Y. Matsumoto and R. Mihalcea (eds.) *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR: Association for Computational Linguistics, pp. 1356-1364. Available at: <https://aclanthology.org/P11-1136/>.
- Nikolsky and Atzman 2021** Nikolsky, R. and Atzman, A. (2021) *Studies in the Tanhuma-Yelammedenu Literature*. Leiden: Brill.
- Rosen and Yassif 2004** Rosen, T. and Yassif, E. (2004) “The study of Hebrew literature of the Middle Ages: major trends and goals”, in M. Goodman (ed.) *The Oxford Handbook of Jewish Studies*. Oxford: Oxford University Press, pp. 241-294. Available at: <https://doi.org/10.1093/oxfordhb/9780199280322.013.0011>.
- Seker et al. 2022** Seker, A., et al. (2022) “AlephBERT: language model pre-training and evaluation from sub-word to sentence level”, in S. Muresan, P. Nakov and A. Villavicencio (eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin: Association for Computational Linguistics, pp. 46-56. Available at: <https://doi.org/10.18653/v1/2022.acl-long.4>.
- Shmidman et al. 2022** Shmidman, A., et al. (2022) “Introducing BEREL: BERT embeddings for rabbinic-encoded language”, *arXiv preprint*, (arXiv:2208.01875). Available at: <https://doi.org/10.48550/arXiv.2208.01875>.
- Tannor et al. 2023** Tannor, S., et al. (2023) “Style classification of rabbinic literature for detection of lost midrash tanhuma material”, *Journal of Data Mining & Digital Humanities*, Available at: <https://doi.org/10.46298/jdmdh.11375>.
- Van der Schans et al. 2020** van der Schans, Y., et al. (2020) “Clustering biblical texts using re-current neural networks”, *Proceedings of the Network Institute Academy Assistants Program 2018/2019*. Available at: <https://doi.org/10.5281/zenodo.4003508>.
- Van Hecke 2018** van Hecke, P. (2018) “Computational stylometric approach to the Dead Sea Scrolls: towards a new research agenda”, *Dead Sea Discoveries*, 25(1), pp. 57-82.
- Van Heck and Joode 2021** Van Hecke, P. and Joode, J.D. (2021) “Promises and challenges in designing stylometric analyses for classical Hebrew”, in S. Fassberg (ed.) *Hebrew Texts and Language of the Second Temple Period: Proceedings of an Eighth Symposium on the Hebrew of the Dead Sea Scrolls and Ben Sira*. Leiden, Boston: Brill, pp. 349-374. Available at: https://doi.org/10.1163/9789004447981_021.
- van Peursen 2023** van Peursen, W. (2023) “Computational Linguistic Analysis of the Biblical Text”, in W. A. Ross and E. Robar (eds.) *Linguistic Theory and the Biblical Text: Semitic Languages and Cultures, Vol. 20*. Cambridge: Open Book Publishers, pp. 223-272. Available at: <https://doi.org/10.11647/obp.0358.05>.
- Yassif 2009** Yassif, E. (2009) *The Hebrew folktale: history, genre, meaning*. Bloomington: Indiana University Press.
- Yassif 2013** Yassif, E. (2013) “The Hebrew story in the Middle Ages: an introduction”, *Jewish Studies Quarterly*, 20(1), pp. 3-8.
- Zsengellér 2014** Zsengellér, J. (2014) *Rewritten Bible after fifty years: texts, terms, or techniques. A last dialogue with Géza Vermès*. Leiden, Boston: Brill.

Recommendations

DHQ is testing out three new article recommendation methods! Please explore the links below to find articles that are related in different ways to the one you just read. We are interested in how these methods work for readers—if you would like to share feedback with us, please complete our short evaluation survey. You can also visit our documentation for these recommendation methods to learn more.

SPECTER Recommendations

Below are article recommendations generated by the SPECTER model:

1. Cluster Analysis in Tracing Textual Dependencies – a Case of Psalm 6 in 16th-century English Devotional Manuals, 2023, Jerzy Wójcik, The John Paul II Catholic University of Lublin
2. Textual Reuse in the Eighteenth Century: Mining Eliza Haywood’s Quotations, 2016, Douglas Ernest Duhaime, University of Notre Dame
3. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, 2017, Christof Schöch, University of Würzburg, Germany
4. The Brain Is Deeper Than the Sea: Sea and Spar Between, Computational Stuplimity, and Fragmentation, 2022, Nathanael Moore, University of Otago
5. Encoding for Endangered Tibetan Texts, 2007, Linda E. Patrik, Department of Philosophy, Union College

DHQ Keyword Recommendations

Below are article recommendations generated by DHQ Keywords:

1. Shakespeare His Contemporaries: collaborative curation and exploration of Early Modern drama in a digital environment, 2014, Martin Mueller, Northwestern University
2. Project Quintessence: Examining Textual Dimensionality with a Dynamic Corpus Explorer, 2023, Samuel Pizelo, UC Davis; Arthur Koehl, UC Davis; Chandni Nagda, University of Illinois at Urbana-Champaign; Carl Stahmer, UC Davis
3. Canonical References in Electronic Texts: Rationale and Best Practices, 2014, Joel Kalvesmaki, Dumbarton Oaks
4. Words, Patterns and Documents: Experiments in Machine Learning and Text Analysis, 2009, Shlomo Argamon, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology; Mark Olsen, ARTFL Project, University of Chicago
5. A Synoptic Primer: Review of Wissensrohstoff Text: Eine Einführung in das Text Mining (2022), 2023, Michael Richter, University of Leipzig

TF-IDF Recommendations

Below are article recommendations generated by the TF-IDF Model:

1. Cluster Analysis in Tracing Textual Dependencies – a Case of Psalm 6 in 16th-century English Devotional Manuals, 2023, Jerzy Wójcik, The John Paul II Catholic University of Lublin
2. How Technology Means: Texts, History, and Their Associated Technologies, 2018, Joshua L. Mann, Expositus
3. Tiresias: A Novel Approach for Mining Book Indices, 2022, Moshe Blidstein, University of Haifa; Daphne Raban, University of Haifa
4. Canonical References in Electronic Texts: Rationale and Best Practices, 2014, Joel Kalvesmaki, Dumbarton Oaks
5. Conjectural Criticism: Computing Past and Future Texts, 2010, Kari Kraus, College of Information Studies and Department of English, University of Maryland



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.