Integrating Semantic and Statistical Features for Authorial Clustering of Qumran Scrolls

Yonatan Lourie[1] **Jonathan Ben-Dov**[2] **Roded Sharan**[3]
[1]Department of Statistics and Operations Research, Tel Aviv University
[2]Department of Bible, The Lester and Sally Entin Faculty of Humanities, Tel Aviv University
[3]Blavatnik School of Computer Science and AI, Tel Aviv University

Abstract

We present a novel framework for authorial classification and clustering of the Qumran Dead Sea Scrolls (DSS). Our approach combines modern Hebrew BERT embeddings with traditional natural language processing features in a graph neural network (GNN) architecture. Our results outperform baseline methods on both the Dead Sea Scrolls and a validation dataset of the Hebrew Bible. In particular, we leverage our model to provide significant insights into long-standing debates, including the classification of sectarian and non-sectarian texts and the division of the Hodayot collection of hymns.

1 Introduction

The discovery of the Dead Sea Scrolls in the mid-20th century represented a turning point in biblical studies and Jewish history, providing a new view into the religious and cultural world of early Judaism and into the theological background of Christianity (VanderKam and Flint, 2005). We discuss the scrolls found in the caves from Qumran, on the shore of the Dead Sea. This is a large collection of approximately 900 scrolls representing a large variety of compositions (i.e. literary entities, books or treatises), each of them featuring its own development history. A large part of the scrolls represents the writings of a community (Collins, 2010) while many other scrolls potentially originate with wider circles of contemporary Judaism. These distinctions, in particular questions of authorship, classification, and origins remain unresolved, fueling scholarly debates for decades. The present study focuses on two such questions.

The first question is the inner division and composition of the collection of Hodayot hymns, particularly the well-preserved copy $1QH^a$ from Qumran Cave 1. While earlier research distinguished a class of "Teacher Hymns" from the "Community Hymns" in the rest of the collection (Douglas, 1999), this division is now contested or modified (Newsom, 2021; Johnson, 2022). The second question is the distinction between sectarian and non-sectarian scrolls; While this distinction was once considered consensus (Dimant, 2014), it is now debated (Martínez, 2010).

Compounding these challenges is the fragmentary nature of the scrolls. The title "scrolls" may be misleading, since most of the corpus is preserved in fragments except for a handful of more complete scrolls. Many texts are incomplete, with reconstructed or uncertain words, making traditional manual analysis both laborious and subjective.

Recent advances in Natural Language Processing offer new possibilities for analyzing ancient texts like the DSS. However, applying computational techniques to ancient Hebrew presents unique difficulties. Hebrew is a highly inflected and morphologically rich language with ambiguous word boundaries, inconsistent orthography, and the absence of vowels in many texts. Even modern Hebrew NLP tools face accuracy issues (Tsarfaty et al., 2019). These challenges increase when dealing with ancient forms of the language. This complexity, combined with the fragmentary and noisy nature of the DSS, necessitates robust and innovative computational approaches.

To our knowledge, computational approaches to the DSS are limited. Traditional stylometry methods have been applied, as demonstrated by Starr (Starr, 2019), and classifiers for biblical texts have been explored in the Dicta-Tiberias project[1]. Additionally, Yoffee et al. (Bühler et al., 2024) investigated text partitioning in the Bible, highlighting the potential for structural analysis using computational techniques. Van Hecke's (Van Hecke, 2018) work represents one of the only applications of NLP methods to the DSS, using basic computational linguistics techniques like tri-grams. While

[1] https://tiberias.dicta.org.il/

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 12–21
May 4, 2025.

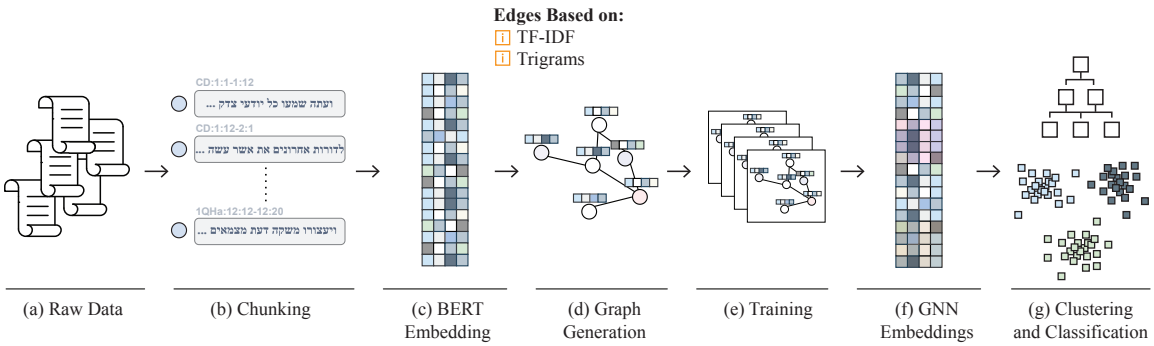

Figure 1: A sketch of the research outline. (**a** and **b**) Data collection and chunking. (**c**) Converting the text to numerical embeddings with BERT semantic model. (**d**) Graph generation based on statistical features. (**e+f**) Graph neural network training and extraction of integrated embeddings. (**g**) Application to clustering and classification questions.

tri-grams can effectively capture local orthographic and morphological patterns (Kulmizev et al., 2017), they are limited in their ability to encode deeper semantic relationships. In our study, we build upon this foundational work by incorporating tf-idf (see definition below) and trigram features, which provide a more nuanced representation of word importance across the corpus. We additionally apply modern semantic embeddings from BEREL (Shmidman et al., 2022), a pre-trained BERT model trained on rabbinic Hebrew literature. This hybrid approach allows us to integrate statistical and semantic features, addressing both the fragmentary nature of the DSS and the inherent challenges of processing ancient Hebrew.

Graph Neural Networks (GNNs) have emerged as powerful tools for representing textual data, particularly for tasks involving relationships between textual entities. Models such as TextGCN (Yao et al., 2018) and BertGCN (Lin et al., 2021) have demonstrated success in text classification by leveraging graph structures, where nodes represent documents or words, and edges capture co-occurrence or semantic relationships. Other works (Yang et al., 2021; Huang et al., 2019), explore alternative GNN architectures, showcasing the versatility of graph-based approaches in text-related tasks.

Unsupervised clustering for text data, particularly ancient and fragmentary texts like our corpus, presents substantial difficulties.

(Kipf and Welling, 2016) have shown good results in unsupervised learning by leveraging graph neural networks to generate latent representations that can be used for clustering. However, their application to textual datasets, particularly ancient and Hebrew corpora, has been limited (some related research exists, such as clustering in the Akkadian language (Stekel et al., 2021)). Our work integrates semantic and statistical features of the DSS within a graph neural network architecture to address the challenges posed by the unique characteristics of this corpus. The resulting embeddings of text chunks are used for clustering and classification of the scrolls, providing significant insights into their structure and content.

2 Methods

We developed a novel model for representing the DSS corpus. Below, we describe the data collection and preprocessing, the representation model, hyperparameter tuning and performance evaluation.

2.1 Data Collection

We used transcriptions of the DSS based on the data files prepared by Martin Abegg[2]. This data powers popular biblical software like Accordance and DSS Electronic Library. We used the Text-Fabric (Roorda, 2019) package, enabling the extraction of both textual content and morphological features for each word.

The corpus was filtered using several criteria. Paratextual elements such as document name, fragment, column, and line numbers were removed, along with all reconstructed text. Letters marked as probable or possible by the editors were retained, while textual gaps were excluded to prevent their analysis as inherent characteristics of the documents. Additionally, doubt marks, non-Hebrew characters, and redundant spaces were eliminated.

[2] https://github.com/ETCBC/dss

We focused exclusively on Hebrew scrolls, excluding Aramaic and Greek. Biblical scrolls were not included in the analysis but rewritten biblical texts like 4Q364 were studied since those texts are comparable to other Qumranic material. Finally, the analysis was restricted to Hebrew scrolls containing a minimum of 300 words.

The texts of each composition were divided into fixed-size chunks; after evaluating various chunk sizes and overlapping ratios (details in Appendix A), a chunk size of 100 words with an overlap of 15 words was chosen. This configuration ensures sufficient representation of smaller scrolls, many of which contain fewer than 500 words, while maintaining granularity for within-scroll analyses. This configuration yielded a dataset with 978 text chunks.

In addition, we curated a set of labels for validation purposes: sectarian / non-sectarian classification and composition name labels, e.g. War Scroll, Instruction, etc (Appendix B).

2.2 Document Representation

We represented each text chunk using both *semantic* and *statistical* features. For the semantic component, we rely on two Hebrew BERT-based models: **BEREL**, which is trained on rabbinic texts (closer to DSS Hebrew than modern Hebrew), and **AlephBERT** (Seker et al., 2022), which is more general for Hebrew tasks. Each text chunk is encoded as a 768-dimensional vector by extracting the final hidden representation of the `[CLS]` token from the model's last layer.

For the statistical component, we use *term frequency–inverse document frequency* (tf-idf). Given a term t in a document d, tf-idf is defined as:

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \cdot \log\Big(\frac{N}{\text{df}(t)}\Big),$$

where $\text{tf}(t, d)$ is the frequency of t in d, N is the total number of documents, and $\text{df}(t)$ is the number of documents containing t. Additionally, we include *tri-grams*: sequences of three consecutive characters, to capture local orthographic and morphological features.

2.3 Graph Construction

We use a graph neural network (GNN) framework where adjacency matrices are derived from cosine similarity between text chunk embeddings (*tf-idf* and *trigram*). For each embedding type, we first create a matrix $\mathbf{A}^\star \in \mathbb{R}^{n_{\text{chunk}} \times n_{\text{chunk}}}$ where $\mathbf{A}^\star_{ij} = \text{cosine\_similarity}(\text{chunk}_i, \text{chunk}_j)$ if $\text{cosine\_similarity}(\text{chunk}_i, \text{chunk}_j) > t$, and $\mathbf{A}^\star_{ij} = 0$ otherwise.

Let $\mathbf{D}^\star$ be the degree matrix of $\mathbf{A}^\star$. We then apply symmetric normalization to obtain the matrix $\tilde{\mathbf{A}}^\star \in \mathbb{R}^{n_{\text{chunk}} \times n_{\text{chunk}}}$:

$$\tilde{\mathbf{A}}^\star = (\mathbf{D}^\star)^{-\frac{1}{2}} \, \mathbf{A}^\star \, (\mathbf{D}^\star)^{-\frac{1}{2}}.$$

We perform this procedure separately for both embedding types (i.e., *tf-idf* and *trigram*), resulting in $\tilde{\mathbf{A}}^\star_{\text{tfidf}}$ and $\tilde{\mathbf{A}}^\star_{\text{trigram}}$. Next, we combine these normalized matrices via element-wise addition:

$$\mathbf{M}_{ij} = \tilde{\mathbf{A}}^\star_{\text{tfidf},ij} + \tilde{\mathbf{A}}^\star_{\text{trigram},ij}.$$

We then apply a threshold h to $\mathbf{M}$ to form our adjacency matrix $\mathbf{A}$, and normalize it using the same symmetric normalization procedure, yielding our final adjacency matrix $\tilde{\mathbf{A}}$. This adjacency matrix represents edges between text chunks whose combined similarity (over tf-idf and trigram) exceeds the threshold h.

2.4 Model

For generating refined text embeddings, we use a Graph Auto Encoder model (Kipf and Welling, 2017). Our model uses a two-layer GNN to encode graph structure and node features into a low-dimensional latent space.

The graph is represented by a normalized adjacency matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{N \times N}$. Node features are represented as a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$, where each node is initialized with a BERT-based embedding vector. The latent space representation is given by $\mathbf{Z} \in \mathbb{R}^{N \times F}$. The GNN layers propagate information through the graph, defined as:

$$\text{GNN}(\mathbf{X}, \mathbf{A}) = \tilde{\mathbf{A}}\text{ReLU}(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}_0)\mathbf{W}_1, \quad (1)$$

where $\mathbf{W}_0$ and $\mathbf{W}_1$ are trainable weight matrices.

The decoder reconstructs the adjacency matrix $\mathbf{A}$ using the inner product of the latent representations. For a given edge $\mathbf{A}_{ij}$, the decoder predicts the edge probability as:

$$\hat{\mathbf{A}}_{ij} = \sigma(\mathbf{Z}_i \cdot \mathbf{Z}_j^T), \quad (2)$$

where $\sigma(\cdot)$ is the sigmoid function, and $\mathbf{Z}_i$ and $\mathbf{Z}_j$ are the latent representations of nodes i and j. The reconstruction loss is computed as:

$$\begin{aligned} \mathcal{L} = &-\frac{1}{|\mathcal{E}+|} \sum_{(i,j)\in\mathcal{E}+} \log \hat{\mathbf{A}}ij \\ &-\frac{1}{|\mathcal{E}-|} \sum_{(i,j)\in\mathcal{E}-} \log(1 - \hat{\mathbf{A}}ij), \end{aligned} \quad (3)$$

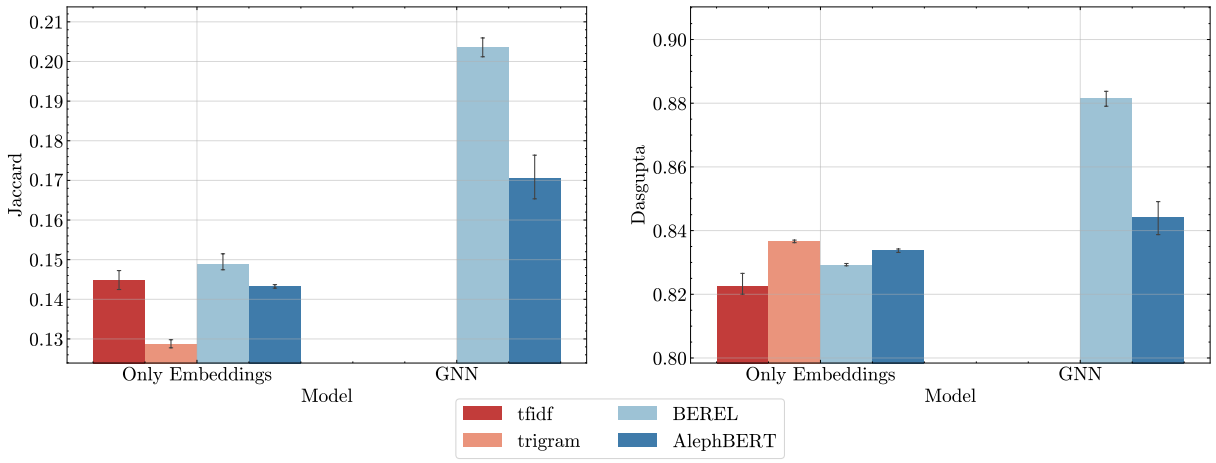

Figure 2: Unsupervised scroll clustering results using different feature extraction methods. Red bars correspond to classical NLP features, blue bars to Hebrew BERT embeddings).

where $\mathcal{E}+$ and $\mathcal{E}-$ are the sets of positive and negative edges, respectively. Negative edges are sampled using negative sampling (Veličković et al., 2018), which is a technique that randomly selects non-existing edges to serve as negative examples in the training. We specifically ensure that the number of negative edges matches the number of positive edges.

During training, we apply a dropout rate of 0.2 along with batch normalization to prevent overfitting.

2.5 Clustering and evaluation

We used two clustering algorithms: agglomerative clustering with Ward's linkage (Jr., 1963) for hierarchical clustering, and k-means clustering for flat clustering. Both Ward's linkage and k-means clustering optimize the same objective: minimizing within-cluster variance, expressed as squared Euclidean distances. The number of clusters is set to the number of compositions in the corpus, reflecting the basic distribution of our dataset.

To assess the clustering performance, we used the Jaccard measure (Rousseeuw, 1987) for external evaluation, where the labels correspond to the **compositions**. We also used the Dasgupta objective (Dasgupta, 2015), which is a custom cost function for evaluating hierarchical clustering models. This method calculates the cost function over a hierarchy of points, given pairwise similarities between those points. In our approach, these similarities are determined by the adjacency of text chunks: for any two consecutive chunks, we assign a similarity score of 1.

2.6 Baselines

We compared our method against several baseline embedding models, including character trigrams, tf-idf and BERT. For each baseline, we applied the same clustering procedure as in our proposed method, using k-means on the text chunk embeddings, with k set to the number of scrolls or compositions.

2.7 Parameter Tuning

To determine the optimal hyper-parameter configuration, we performed a grid search over a range of values for the number of edges (derived from t and h), graph construction methods, hidden dimensions, and learning rates. The optimization criterion was based on minimizing the training loss, with early stopping applied when the loss improvement was less than epsilon. For each set of hyper-parameters, we used 10-fold nested cross-validation to evaluate the embedding performance, measured as the average over all folds. The models were trained using the Adam optimizer (Kingma and Ba, 2017) with a weight decay regularization term of 5e-4.

2.8 Code and data availability

The code developed for this paper has been made publicly available [3], and the resulting dataset has been uploaded to the Hugging Face Hub [4] to facilitate future research efforts. All of our algorithms were implemented in Python 3.10 and executed on

[3] https://github.com/yonatanlou/QumranNLP

[4] https://huggingface.co/datasets/yonatanlou/QumranDataset

a personal MacBook Pro (M2, 2022, 16 GB RAM, 512 GB SSD).

3 Results

We developed a new method that applies a GNN architecture to integrate semantic and linguistic features for clustering of the Qumran Scrolls. First, we identified the optimal parameters by tuning the unsupervised model on the entire Qumran corpus, demonstrating that our algorithm outperforms baseline methods. We then validated the algorithm on the Hebrew Bible dataset, achieving similar performance and confirming its robustness. Finally, we used the trained model to extract improved text embeddings, which were applied to address various well-known research questions. We used the BEREL model which yields the best results across our experiments.

3.1 Model evaluation

We evaluated our model based on its performance on the entire Qumran corpus, yielding a GNN with 978 nodes and 9,391 edges. This evaluation explored different initial BERT embeddings and compared them to our baseline methods (see Figure 2). Interestingly, classical methods like tri-grams and tf-idf demonstrated competitive results compared to the BERT embeddings, likely due to the unique characteristics of our corpus. The GNN-based methods in these experiments were based on tf-idf and trigram based similarities, yielding the best overall performance.

3.2 Text clustering

The present section will address two research questions about the homogeneity and coherency of the Qumran corpus, both on a large and a smaller scale. It was apparent earlier on that some scrolls reflect the vocabulary, style and theology of a separate community with its ideas and institutions. The core texts of that group were found as well preserved scrolls in Qumran Cave 1 and were published in the 1950s. It was also clear that some compositions did not belong to the Yahad community but were rather a heritage of wider circles. The dividing line between the categories seemed clear at first (Newsom, 1990; Dimant, 2014), but in recent decades it has been debated. Many new fragmentary compositions were found in Cave 4 whose social identity is uncertain, and in addition the definition of a "sect" and the outright connection between it and its literary product was problematized. We therefore venture to test whether these categories could be achieved using advanced clustering techniques. A first research question pertains to the composition and inner-variety of the collection of Hodayot hymns.

Clustering within the Hodayot. We examine one of the large scrolls from Cave 1, the Hodayot scroll 1QH^a containing religious hymns on the life of the community. These hymns have parallels in fragmentary copies from Cave 4, but they will not be examined here as the problem is clearest on the largest copy. A prevalent theory distinguished two types of hymns: Teacher Hymns and Community Hymns, as defined in (Douglas, 1999; Johnson, 2022) based on earlier studies. While the hymns are quite similar, experts detected in them different vocabulary and themes. The exact extent of the Teacher Hymns within 1QH^a is debated. Douglas saw them as a block of material concentrated in columns 10-17, and the Community Hymns in columns 2-8, 18-24. Douglas considers Columns 9 and 18-20 as transition material enveloping the Teacher Hymns, that can therefore belong to each of the groups. Other studies saw the Teacher Hymns as a dispersed group through the entire scroll. The existence and extent of Teacher Hymns are examined here. We applied our GNN model to the Hodayot composition. Using these embeddings, we applied hierarchical clustering to perform unsupervised clustering. Our analysis (Figure 3) identified the following clusters:

- Cluster 1 (Purple) and cluster 2 (Red) contain Teacher Hymns from columns 10-16, with two chunks from the enveloping material (columns 18-19).
- Cluster 3 (orange) and cluster 4 (green) contain primarily Community Hymns from columns 1-9, 18-23, with several transitional chunks from column 9 and two hymns from columns 14 and 15. Three chunks from columns 11, 12, 15 stand at the edge of the cluster.

The results overwhelmingly confirm the existence of a distinct category of Teacher Hymns. Moreover, the results confirm that Teacher Hymns are clustered at the center of 1QH^a, with only a few outliers. These outliers will be discussed in a dedicated article intended for the Qumran research

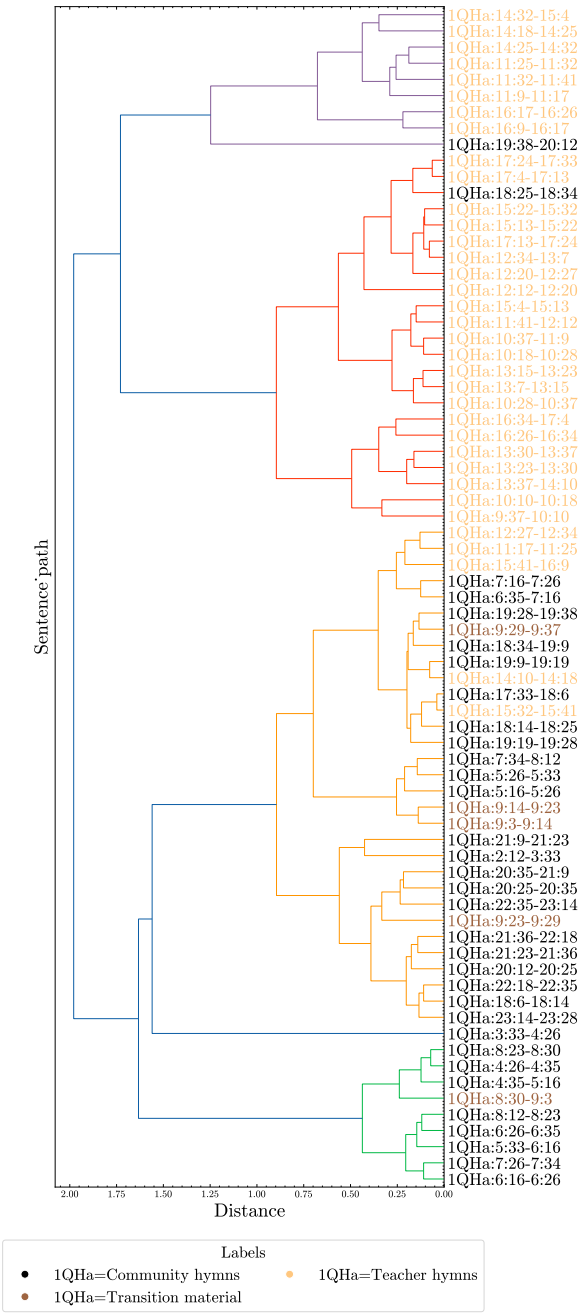

Figure 3: Dendrogram of the Hodayot composition clustering using the best-performing GNN model.

community. Notable is the identity of 3:33–4:26, which received a cluster of its own, as well as the presence of 14:10–18 and 15:32–41 within the cluster of Community Hymns. The blurred identity of the transition passages, as well as some outliers, may be attributed to a unifying "Maskil" redaction (Johnson, 2022).

Classification of sectarian scrolls. The clustering of sectarian compositions produced less definitive results although it did point out two main clusters of sectarian compositions. The clustering depends on three categories: 1) core texts of the yahad community based on their vocabulary and content: War Scroll, Community Rule, Rule of the Congregation, Hodayot, CD (Damascus Document), Pesharim and similar documents (Dimant, 2014). 2) texts that do not display sectarian features (such as Apocryphom Jeremiah), 3) Other compositions which, while displaying some similar features, are not fully consistent and their identity remains debated (for example Shir Shabbat - the Songs of the Sabbath Sacrifice). In this section we clustered compositions rather than chunks. To this end, we averaged each composition's chunk embeddings to obtain one vector per composition. Then we performed hierarchical clustering on these composition-level embeddings. The resulting dendrogram is provided in Figure 4. The labeling in this diagram is based on the composition level (e.g., Hodayot), with some labels representing groups of related compositions, such as Pesharim, Rewritten Pentateuch, and Calendrical Texts. The classification as sectarian or non-sectarian is provided in Appendix B. It is important to note that the clustering process was entirely unsupervised and only later compared with predefined labels.

Our expectation was that sectarian and non-sectarian texts would cluster separately. The Calendrical Texts and the Copper Scroll served as test cases, as their distinct linguistic profiles should stand out in an unsupervised clustering scheme. The dendrogram shows two main clusters of sectarian compositions, with text marked in black. Some appear close to texts of uncertain identity, whose sectarian status is now further supported. The yellow cluster includes seven clearly sectarian texts, such as the Pesharim, MMT, the War Scroll, and the Community Rule, making it strongly indicative of sectarian content. The Apocryphal Psalms (11Q11) also appear here, but the composition's small size and the wide dispersion of its embeddings limit the reliability of its clustering, especially after averaging those embeddings. At the edge of the green cluster is CD, a prominent sectarian text. Its proximity to the main sectarian cluster reflects its sectarian character, whereas its literary diversity in terms of genre and content accounts for its location outside the core sectarian cluster.

The gray cluster at the top of the dendrogram contains core sectarian texts such as Hodayot and The Rule of the Blessings, next to the two wisdom texts Instruction and Mysteries. It also includes poetic compositions such as the Collections

of Psalms, Barkhi Nafshi, Songs of Maskil, Shir Shabbat, and Daily Prayers, alongside the prayers in Dibre Hameorot. This grouping highlights a set of poetic and prayer texts whose sectarian identity has been debated, now shown to align closely with core sectarian compositions.

The Calendrical Texts and the Copper Scroll cluster separately as expected, forming a distinct group with high distance between other compositions.

Analysis of the embeddings for each composition revealed that the dendrogram representation might be misleading for compositions with high dispersion across text chunks, such as the Apocryphal Psalms, Para Kings, Festival Prayers, and the Book of Tobit. Their average representation in the dendrogram does not fully reflect their true properties. For instance, the entire red cluster consists of such compositions. Upon analyzing these text chunks, we found that most are highly fragmentary, while the Book of Tobit exhibits significant stylistic differences between its text chunks.

The green cluster at the bottom of the dendrogram highlights the model's ability to capture the stylistic characteristics of rewritten Bible texts. This cluster includes the Temple Scroll, Rewritten Pentateuch texts, Dibre Moshe, Books of Tobit, and the Book of Jubilees. This generic grouping overrides the sectarian/non-sectarian division, generally placing the compositions in this cluster in the non-sectarian group and instead aligns them based on genre. While this clustering does not directly address the sectarian question, it demonstrates the model's ability to identify main genres. In summary, the model successfully confirms the grouping of core organizational texts and aligns Dibre Hameorot and Instruction with them. It accurately identifies clear non-sectarian texts such as The Book of Tobit and The Book of Jubilees. These categorizations align with common labels and highlight some unexpected groupings. However, the model does not produce distinct results for the Rewritten Bible and prayer genres, reflecting the complexity of these categories.

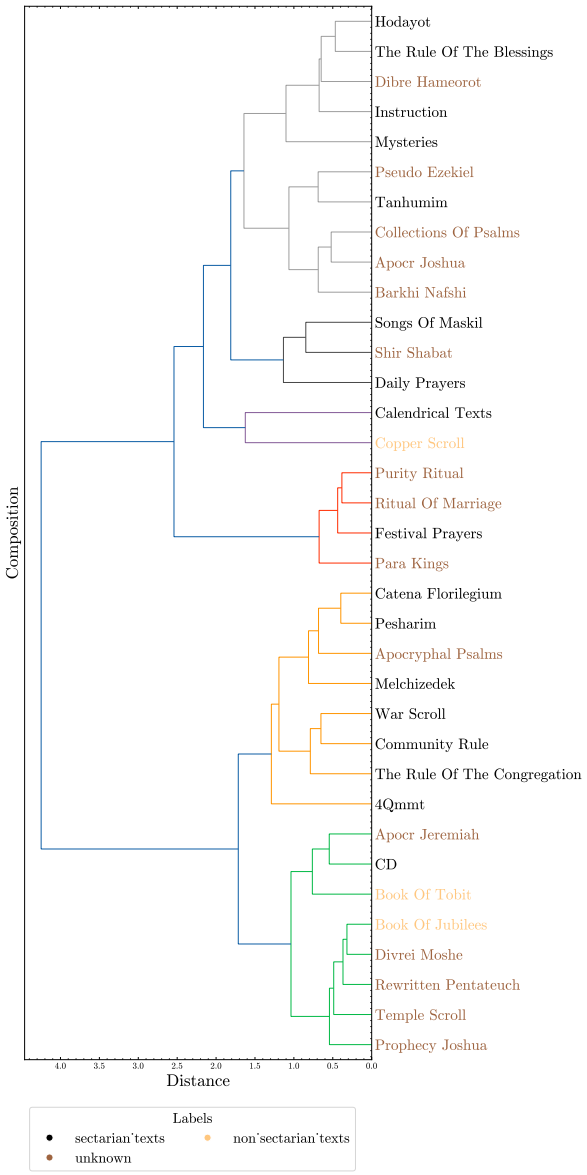

Figure 4: Dendrogram of compositions with sectarian or non sectarian label.

4 Conclusions

Our research introduces a novel GNN-based method that effectively integrates semantic and linguistic features for clustering Qumran texts. By training an unsupervised model on the entire corpus, we identified optimal parameters and demonstrated that our approach outperforms baseline methods.

The model's ability to capture complex relationships between text fragments and represent the text with improved embeddings, allowed us to address significant research questions related to authorship and sectarian classification within the DSS corpus. Our clustering results demonstrate that the clustering aligns closely with traditional divisions established in the literature.

While our model provided promising results, some of its aspects could be improved in future work. Sentence-transformer models are particularly effective for processing chunks of text and offer the potential for greater precision. While there

is currently no pre-trained Hebrew model available, these models could be fine-tuned on non-debatable text chunks to create a robust embedding space specifically tailored to the DSS corpus.

5 Limitations

While our study demonstrates promising results in clustering and classifying the Dead Sea Scrolls, it has several limitations that warrant consideration. Specifically, the fragmentary nature of the DSS corpus poses inherent challenges. The preprocessing steps in this work could be improved, and a comprehensive study dedicated to this topic alone would be beneficial. Moreover, the corpus is continually refined through ongoing manual work, which means that the data used in this study may differ slightly from future versions. The present text of the scrolls is essentially that of the Discoveries in the Judaean Desert (DJD) series, as further refined editions lack a comprehensive electronic repository.

While our clustering results align with traditional scholarly divisions, the evaluation relies on predefined labels that may be subjective. The ground truth for sectarian classification and text authorship is not absolute, thus limiting the objectivity of the performance metrics.

Acknowledgments

This work was supported by a research grant from the Israel Ministry of Innovation, Science and Technology (grant no. 1001577565).

References

Axel Bühler, Gideon Yoffe, Nachum Dershowitz, Eli Piasetzky, Israel Finkelstein, Thomas Römer, and Barak Sober. 2024. Exploring the stylistic uniqueness of the priestly source in genesis and exodus through a statistical/computational lens. *Zeitschrift fur die Alttestamentliche Wissenschaft*, 136(2):165–190. Publisher Copyright: © 2024 Walter de Gruyter GmbH. All rights reserved.

John J Collins. 2010. *Beyond the Qumran Community: The Sectarian Movement of the Dead Sea Scrolls*. Wm. B. Eerdmans Publishing.

Sanjoy Dasgupta. 2015. A cost function for similarity-based hierarchical clustering. *Preprint*, arXiv:1510.05043.

Devorah Dimant. 2014. *History, ideology and Bible interpretation in the Dead Sea Scrolls: Collected studies*, volume 90. Mohr Siebeck.

Michael C Douglas. 1999. The teacher hymn hypothesis revisited: New data for an old crux. *Dead Sea Discoveries*, pages 239–266.

Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3444–3450, Hong Kong, China. Association for Computational Linguistics.

Michael B Johnson. 2022. Look who's talking: Reconsidering the speaker in the 'teacher hymns'(1qha). In *Emerging Sectarianism in the Dead Sea Scrolls*, pages 313–341. Brill.

Joe H. Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Thomas N. Kipf and Max Welling. 2016. Variational graph auto-encoders. *Preprint*, arXiv:1611.07308.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *Preprint*, arXiv:1609.02907.

Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling. 2017. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 382–389, Copenhagen, Denmark. Association for Computational Linguistics.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. Bertgcn: Transductive text classification by combining GCN and BERT. *CoRR*, abs/2105.05727.

Florentino García Martínez. 2010. Beyond the sectarian divide: The" voice of the teacher" as an authority-conferring strategy in some qumran texts. In *The Dead Sea Scrolls*, pages 227–244. Brill.

Carol A Newsom. 1990. 'sectually explicit' literature from qumran. *The Hebrew Bible and Its Interpreters*, 1:167–87.

Carol A Newsom. 2021. A farewell to the hodayot of the community. *Dead Sea Discoveries*, 28(1):1–19.

Renyi Qu, Ruixuan Tu, and Forrest Bao. 2024. Is semantic chunking worth the computational cost? *Preprint*, arXiv:2410.13070.

Dirk Roorda. 2019. Text-fabric: handling biblical data with ikea logistics. *HIPHIL Novum Journal for Bible and Digital Resources*, 5(2):126–135.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56, Dublin, Ireland. Association for Computational Linguistics.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. 2022. Introducing berel: Bert embeddings for rabbinic-encoded language. *Preprint*, arXiv:2208.01875.

J. Starr. 2019. Classifying the aramaic texts from qumran: A statistical analysis of linguistic features. *Palestine Exploration Quarterly*, 151(2):160–163.

Moshe Stekel, Amos Azaria, and Shai Gordin. 2021. Word sense induction with attentive context clustering. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 144–151, NIT Silchar, India. NLP Association of India (NLPAI).

Reut Tsarfaty, Shoval Sadde, Stav Klein, and Amit Seker. 2019. What's wrong with Hebrew NLP? and how to make it right. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 259–264, Hong Kong, China. Association for Computational Linguistics.

P. Van Hecke. 2018. Computational stylometric approach to the dead sea scrolls: Towards a new research agenda. *Dead Sea Discoveries*, 25(1):57–82.

James VanderKam and Peter Flint. 2005. *The meaning of the Dead Sea scrolls: their significance for understanding the Bible, Judaism, Jesus, and Christianity*. A&C Black.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2018. Deep graph infomax. *Preprint*, arXiv:1809.10341.

Tianchi Yang, Linmei Hu, Chuan Shi, Houye Ji, Xiaoli Li, and Liqiang Nie. 2021. Hgat: Heterogeneous graph attention networks for semi-supervised short text classification. *ACM Trans. Inf. Syst.*, 39(3).

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018. Graph convolutional networks for text classification. *CoRR*, abs/1809.05679.

A Pre-processing parameter evaluation

A.1 Chunk Size Evaluation

When dividing the DSS texts into chunks, a critical factor was selecting an appropriate chunk size to balance representation and analytical granularity. Chunk sizes between 25 and 150 words were evaluated, with overlapping ratios of 5%, 10%, and 15%. Smaller chunk sizes increase the granularity of the analysis but risk fragmenting the text excessively, while larger chunk sizes reduce granularity and limit the number of chunks for shorter scrolls. This limitation is particularly problematic for scrolls containing fewer than 500 words, as larger chunks may result in only one or two chunks per scroll, hindering within-scroll clustering. While more advanced chunking techniques exist (Qu et al., 2024), we chose to use a fixed-size chunking method with overlap due to the highly fragmentary and unordered nature of our corpus, which lack the clear structural organization seen in the Bible.

The evaluation showed that chunk sizes of 100 and 150 words yielded the best performance across intrinsic and extrinsic clustering metrics. While 150-word chunks slightly outperformed in some cases, 100-word chunks were ultimately chosen to allow for better representation of shorter scrolls and greater flexibility in downstream tasks. An overlap of 15 words was selected as it provided a good balance between minimizing information loss and computational efficiency.

B Labeling details

We categorized the compositions into sectarian, non-sectarian, and texts with undetermined identity as follows:

- **Sectarian texts**: Calendrical Texts, Catena and Florilegium, CD, Community Rule, Daily Prayers (4Q503), Festival Prayers (4Q509), Hodayot, Instruction (Musar Lamevin), Melchizedek, Mysteries, Pesharim, Rule of Blessings, Rule of the Congregation, Songs of Maskil, Tanhumim, War Scroll, and 4QMMT.
- **Non-sectarian texts**: Book of Jubilees, Book of Tobit, and Copper Scroll (3Q15).
- **Texts with undetermined identity**: Apocryphal Psalms, Apocryphon Jeremiah, Apocryphon Joshua, Barkhi Nafshi, Collections of Psalms (4Q380-381), Dibre Hameorot,

Dibre Moshe (1Q22), Para Kings (4Q382), Prophecy Joshua, Pseudo-Ezekiel, Rewritten Pentateuch, Ritual of Marriage (4Q512), Shir Olat Hashabbat, Temple Scroll, and Purity Ritual (4Q274).

The full list of labels, including labels for sectarian, composition, and genre, is available online.[5]

[5] https://github.com/yonatanlou/QumranNLP/blob/main/Data/Qumran_labels.csv