Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Style Classification of Rabbinic Literature for Detection of Lost Midrash Tanḥuma Material

Shlomo Tannor[1], **Nachum Dershowitz**[1], **Moshe Lavee**[2]

[1]Tel Aviv University, Israel
[2]Haifa University, Israel

Corresponding author: Shlomo Tannor , shlomotannor@mail.tau.ac.il

Abstract

Midrash collections are complex rabbinic works that consist of text in multiple languages, which evolved through long processes of unstable oral and written transmission. Determining the origin of a given passage in such a compilation is not always straightforward and is often a matter of dispute among scholars, yet it is essential for scholars' understanding of the passage and its relationship to other texts in the rabbinic corpus.

To help solve this problem, we propose a system for classification of rabbinic literature based on its style, leveraging recent advances in natural language processing for Hebrew texts. Additionally, we demonstrate how this method can be applied to uncover lost material from a specific midrash genre, Tanḥuma-Yelammedenu, that has been preserved in later anthologies.

Keywords

Style classification; text reuse; Jewish studies

I INTRODUCTION

Midrash, an integral genre within Jewish literature, encompasses a range of interpretative and narrative texts that seek to explore and expound upon the meanings of biblical scriptures. These texts incorporate a rich mix of legal, ethical, and philosophical discussions, allegories, parables, and homilies, offering deeper insights into the religious passages they explore.

Midrash anthologies are compilations of these interpretive works. They're inherently complex, containing text in multiple languages, written by different authors over several generations and geographic locations. The anthologists would often merge, quote, or paraphrase earlier sources. This process creates a composite that can make it difficult for scholars to discern the individual components. The identification of sections from particular sources within these anthologies can illuminate various scholarly debates, thereby enriching our understanding of the historical evolution of the rabbinic corpus.

The prospect of automated analysis and classification of rabbinic texts presents immense opportunities. Identifying old manuscripts, revealing lost materials quoted in later works (such as parts of the Tanḥuma literature and *Mekhilta Deuteronomy*), and determining the authorship or dating of a text are all potential applications. Driven by this potential, we turn to state-of-the-art natural language processing (NLP) techniques to explore how we can leverage these tools for midrashic study.

We propose a system for the classification of rabbinic literature. This system detects unique stylistic patterns in the language of the text and can help uncover lost midrashic material quoted in later works. As a test case, we use our method to detect lost sections of the *Midrash Tanḥuma* that are quoted in the *Yalkut Shimoni*.[1]

II RELATED WORK

In recent years, advancements in natural language processing (NLP) and machine learning (ML) have greatly expanded the toolkit available for tasks such as authorship attribution, plagiarism detection, and style classification. These tools have been successfully employed in a variety of contexts, from the analysis of contemporary texts to the examination of historical documents.

In the broad context of textual analysis, Juola [2008] provides a thorough review of authorship attribution, offering a comprehensive understanding of the state of the art in computational methods for authorship attribution and style classification and their applicability to different types of texts.

The application of these techniques to biblical texts has seen particular innovation in recent years. Dershowitz et al. [2015] introduced a method for automatic biblical source criticism, examining preferences among synonyms and other stylistic attributes; technical details may be found in Koppel et al. [2011]. This approach laid a foundation for using stylistic analysis in the context of classical Hebrew texts. Building on that work, Akiva and Koppel [2013] developed an unsupervised algorithm for decomposing multi-author documents, further reinforcing the applicability of NLP and ML models in the field of authorship attribution.

Siegal and Shmidman [2018] applied computational tools to reconstruct *Mekhilta Deuteronomy*, a lost midrash halakha from the school of Rabbi Akiva. Although their research shares a common goal with ours, their approach begins with a list of candidate texts and primarily focuses on eliminating quotes or near-quotes of existing material from other sources. In contrast, our work addresses the problem of generating an initial candidate list for a specific genre.

From a methodological perspective, it is worth noting that our research also bridges the typically separate approaches of text reuse and stylometry or style detection. While text reuse predominantly focuses on content words and semantics, stylometry often concentrates on function words and other habitual linguistic choices that an author may use subconsciously. By combining these two methods, our study offers a unique perspective on text analysis. Forstall and Scheirer [2019] provide an insightful discussion on this topic, while surveying the various computation tools used for the study of intertextuality.

Finally, the synergy between technology and humanities research is increasingly appreciated. A significant example is the Ithaca project [Assael et al., 2022], a joint venture between DeepMind and the University of Oxford, which focuses on the restoration and classification of ancient Greek epigraphs. This project demonstrates the potential of such interdisciplinary collaboration and offers a model for similar initiatives. This collaboration model resonated with our approach and influenced how we conducted our research.

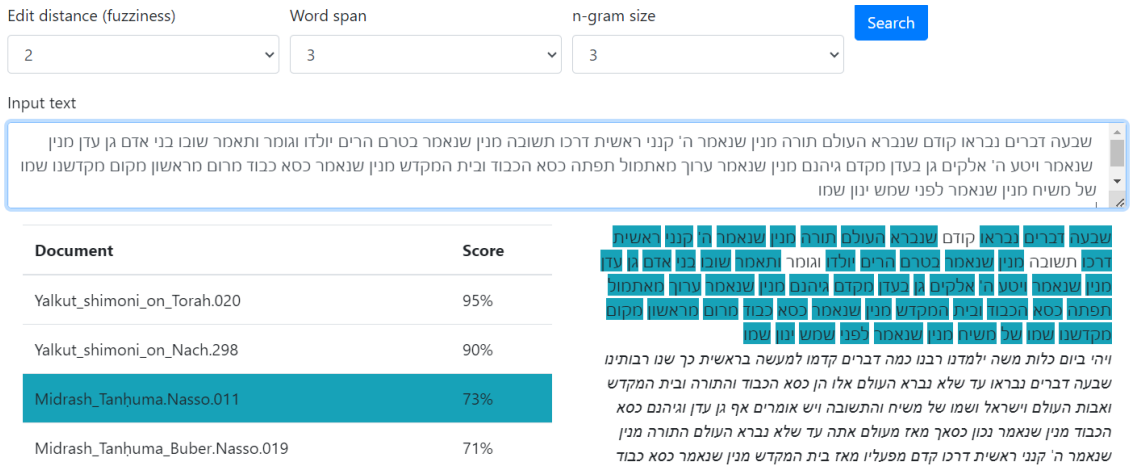

Figure 1: The text-reuse engine, RWFS, shows how a medieval midrash paragraph is reusing early material from various sources including *Midrash Tanḥuma*.

III METHOD

3.1 Dataset

Our training dataset was extracted from Sefaria's resources.[2] We use the raw text files and divide them into the following categories of rabbinic texts:

Mishnah. In this category we include all tractates of the Mishnah and the Tosefta. Both collections are generally dated to the second century CE and consist of rabbinic rulings and debates, organized by topic.

Midrash Halakhah. These collections are dated to around the same time of the Mishnah, but they are organized according to the Pentateuch and focus more on the exegesis of biblical verses. In this class we include: *Mekhilta d'Rabbi Yishmael*, *Mekhilta d'Rashbi*, *Sifra*, *Sifre Numbers*, and *Sifre Deuteronomy*.

Jerusalem Talmud. We include all tractates of the Jerusalem Talmud, omitting the Mishnah passages that provide the basis for discussion. These texts for the most part are written in a mixture of Hebrew and Palestinian Aramaic and are roughly dated to the 4th c. CE.

Babylonian Talmud. We include all tractates of the Babylonian Talmud, omitting the Mishnah passages that provide the basis for discussion. These texts for the most part are written in a mixture of Hebrew and Babylonian Aramaic and are roughly dated to the 5th c.

Midrash Aggadah. In this category we include early midrash works assumed to have been composed during the amoraic period (up to the 5th c.) or slightly later. The works included in training are: *Genesis Rabbah*, *Leviticus Rabbah*, and *Pesikta de-Rav Kahanna*. Like midrash halakhah these works follow the order of verses in the Bible, but in contrast they focus less on deriving rulings (halakhah) and more on expounding on the biblical narrative. Other works that we did not use during training but which we partially associate with this category include: *Ruth Rabbah*, *Lamentations Rabbah*, and *Canticles Rabbah*.

Midrash Tanḥuma. In this category we include later midrashic works that make up what is referred to as Tanḥuma-Yelammedenu Literature. The works included in training are: *Midrash*

[1] A medieval midrash anthology, attributed to Simeon of Frankfort, 13th century CE, by the publisher of the second printed edition (Venice, 1566).

[2] https://github.com/Sefaria/Sefaria-Export.

Tanḥuma, *Tanḥuma Buber*, and *Deuteronomy Rabbah*. Other works that we did not use during training but we partially associate with this category include *Exodus Rabbah* starting from Section 15[3] and *Numbers Rabbah* starting from Section 15.[4]

We divide these works into continuous blocks of 50 words. We then clean the text by removing vowel signs, punctuation and metadata. In order to neutralize the effect of orthography differences, we also expand common acronyms and standardize spelling for common words and names.

After cleaning and normalizing the data, we split our dataset into training (80%) and validation (20%) sets. Finally, we downsample all majority classes in the validation set to get a balanced dataset.

3.2 Models

Baseline. For our baseline model we use a logistic regression model over a bag of n-grams encoding. We include unigrams, bigrams, and trigrams. We use the default parameters from scikit-learn [Pedregosa et al., 2011] but set `fit_intercept=False` to reduce the impact of varying text length and set `class_weight="balanced"` to deal with class imbalance in the training data. This type of model is highly interpretable, enabling us to see the features associated with each class. Finally, we choose this model as our baseline as it generally achieves reasonable results without the need to tune hyperparameters.

AlephBERT. The next model we evaluate is AlephBERT [Seker et al., 2022] – a Transformer model trained with the masked-token prediction training objective on modern Hebrew texts. While this model obtains state-of-the-art results for various tasks on modern Hebrew, performance might not be ideal on rabbinic Hebrew, which differs significantly from modern Hebrew. We train the pretrained model on the downstream task using the Huggingface Transformers framework [Wolf et al., 2020] for sequence classification, using the default parameters for three epochs.

BEREL. The third model we evaluate is BEREL [Shmidman et al., 2022] – a Transformer model trained with a similar architecture to that of BERT-base [Devlin et al., 2019] on rabbinic Hebrew texts. In addition to the potential benefit of using a model that was pretrained on similar texts to those of the target domain, BEREL also uses a modified tokenizer that does not split up acronyms that would otherwise be interpreted as multiple tokens with punctuation marks in between. (Acronyms marked by double apostrophes [or the like] are very common in rabbinic Hebrew.) We train the pretrained model on our downstream task in an identical fashion to the training of the AlephBERT model.

Morphological. Finally, we also train a model that focuses only on morphological features in the text, in an attempt to neutralize the impact of content words. We expect this type of model to detect more "pure" stylistic features that help discriminate between the different textual sources. To extract features from the text, we use a morphological engine for rabbinic Hebrew created by DICTA.[5] We then train a logistic regression model over an aggregation of all morphological features that appear in a given paragraph.

[3] See "Exodus Rabbah," *Encyclopaedia Judaica*, for the rationale behind this division.
[4] See "Numbers Rabbah," *Encyclopaedia Judaica*, for the rationale behind this division.
[5] https://morph-analysis.dicta.org.il/.

	Validation Acc
Baseline	0.867
AlephBERT	0.879
BEREL	**0.922**
Morphological	0.560

Table 1: Model accuracy on validation set.

3.3 Text Reuse Detection

To achieve our end goal of detecting lost Tanḥuma material, we combine our style classification model with a filtering algorithm based on text reuse detection.

For reuse detection, we utilize RWFS (Rolling Window Fuzzy Search) by Schor et al. [2021].[6] RWFS uses fuzzy full-text search on windows of n-grams. The system is built on top of a Lucene index,[7] and uses a web-based interface to provide easy querying to technological and non-technological users.

For our corpus of texts for this engine we use all biblical and early rabbinic works using the texts available on Sefaria. We use 3-gram matching and permit a Levenshtein distance of up to 2 for each individual word in the n-gram. The match score for each retrieved document is given by the number of n-gram matches divided by the length of the query and the results are sorted accordingly (Figure 1).

3.4 Detecting Lost Tanḥuma Candidates

Tanḥuma-Yelammedenu Literature is a name given to a genre of late midrash works, some of which are lost and only scarcely preserved in anthologies or Genizah fragments (Bregman, 2003, Nikolsky and Atzmon, 2021). One of the lost works was called "Yelammedenu," and we know about it since it is cited in various medieval rabbinic works such as *Yalkut Shimoni* and the *Arukh*.[8] While lost Tanḥuma material is explicitly cited in some works, it is often quoted without citation in other midrash anthologies.

To find candidates for "lost" Tanḥuma passages, we apply the following process:

1. Extract all passages from the given midrash collection, in our case *Yalkut Shimoni*.
2. Split long passages into segments of up to 50 words.
3. Run these segments through the style detection model.
4. Collect segments for which our model gives the highest score to the Tanḥuma class.
5. Run these segments through a text-reuse engine.
6. Keep only segments that do not have a well established source. (Our threshold was $\#n$-gram matches $\leq 0.2 \times \#n$-grams in query.[9])

IV RESULTS

As can be seen in Table 1, our baseline model achieves well over the random guess accuracy of 0.166 on the validation set, and achieves almost the same accuracy as the AlephBERT fine-

[6] This intertext engine was designed for this purpose by our partners at eLijah Lab, University of Haifa (https://elijahlab.haifa.ac.il/about-eng).

[7] https://lucene.apache.org/core.

[8] An 11th century dictionary of rabbinic literature by Nathan ben Jehiel of Rome.

[9] Subsequent experiments using different study cases demonstrated the need to adjust the threshold based on the specific task.

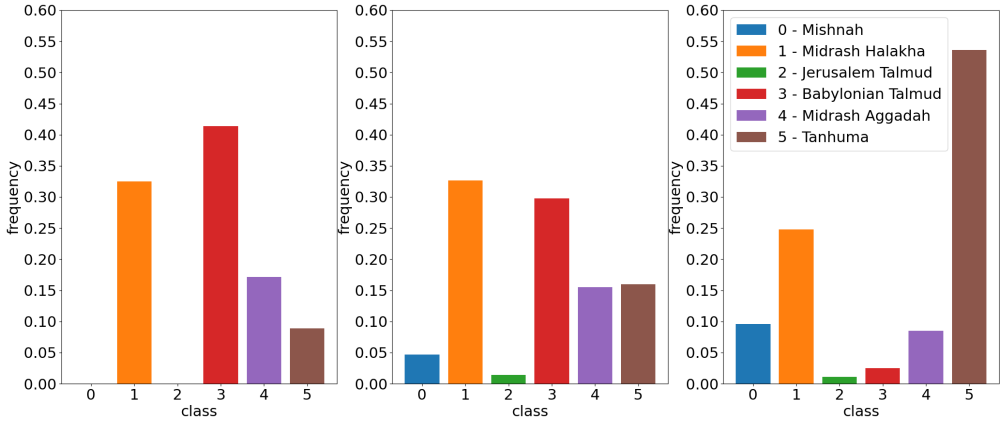

Figure 2: From left to right: (1) class frequencies for passages based on text reuse detection in *Yalkut Shimoni*; (2) predicted class frequencies for passages with high text reuse score; (3) predicted frequencies for passages with low reuse score.

tuned model. The BEREL-based model leads by a significant margin. However, we encountered multiple challenges when using this model for inference on paragraphs from *Yalkut Shimoni*:

1. The model's scores were not calibrated, most predictions were very close to 1.0 or 0.0, making it hard to experiment with different thresholds.
2. BEREL's accuracy on a small sample of paragraphs from *Yalkut Shimoni* was significantly lower than the corresponding validation accuracy. It seems that BEREL might have relied on some orthographic features that appeared in the training and validation sets but not in the new out-of-distribution text.
3. Transformer-based models are generally less interpretable, and have higher inference costs than classical ML models such as logistic regression.

For these reasons, we decided to use our baseline model for inference on *Yalkut Shimoni*.

In Figure 3, we can see that the the most common errors are mixing 'Tanḥuma' with 'Midrash Aggadah' which are indeed relatively similar genres. On the other hand, 'Babylonian Talmud' and 'Jerusalem Talmud' seem to be the most distinct classes, perhaps due to their extensive use of Aramaic in addition to Hebrew, each in its own unique dialect.

After taking the whole *Yalkut Shimoni* on the Pentateuch and following the process described in Section 3.4, we can analyze the prevalence of each class in the collection. As can be seen in Figure 2, the Babylonian Talmud is the most quoted class, while the Jerusalem Talmud is rarely, if ever, quoted. Our classifier gives a similar distribution to that of the text-reuse engine. However, when looking only at passages with low reuse score we see that the Babylonian Talmud rarely appears while 'Tanḥuma' becomes the most frequent predicted class by far, followed by 'Midrash Halakha.' This aligns with the fact that we know of lost works that belong to these categories, while the Babylonian Talmud was well preserved throughout the generations as the core text of the rabbinic tradition.

To evaluate our classifier on the target task, we sampled a random set of 50 items classified as Tanḥuma for manual labeling. A midrash expert analyzed these passages and looked them up in the early print edition of *Yalkut Shimoni*, which tends to include citations in the margins. Sections that were ascribed to Yelammedenu and sections that were recognized as being typical Tanḥuma material were labeled as "positive," while all other passages were labeled "negative." Out of these items, 22 were cited as Yelammedenu, while an additional 8 were recognized as

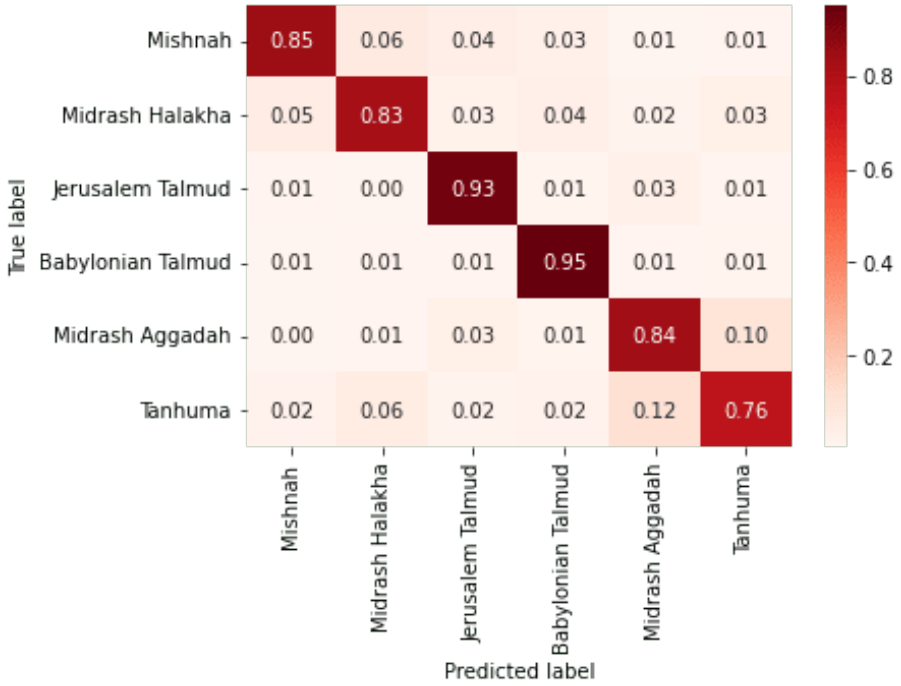

Figure 3: Confusion matrix for baseline model, normalized by row.

typical Tanḥuma material from lost sources,[10] yielding an approximate precision of 60%.

From Figure 4, we see that the precision grows monotonically with the decision threshold, indicating that the model is useful in recovering lost Tanḥuma material. Furthermore, we see that we can achieve a precision of approximately 80% by setting an appropriate decision threshold without a high cost to recall.

4.1 Findings

Using the methodology we described to investigate thoroughly the makeup of *Yalkut Shimoni* on Deuteronomy, there were some interesting findings that arose, as well as some questions.

A systematic expert examination of all results for the first half of Deuteronomy (approximately 10% of the Pentateuch) revealed that all known citations of Yelammedenu, ranging from 100 to 600 words, had at least one sub-paragraph of 50 words recognized as part of the genre. In most cases, the majority of the citation was also identified as part of the genre. In practical terms, this means that every block of text was correctly identified, resulting in a 100% recall rate.

Interestingly, one paragraph that was detected as "lost Tanḥuma" material was actually cited as *Deuteronomy Rabbah* in the early print version of *Yalkut Shimoni*. However, our version of *Deuteronomy Rabbah* had a very low text reuse match for this paragraph. This result raises the question of whether the author of *Yalkut Shimoni* had a different version of the text from what we have.[11]

Another question that rises from this phenomenon is the extent to which the midrash collators rephrase and reorganize the early material they work with as opposed to copying full sections.[12]

[10] These latter items are perhaps the more exciting find as they have previously been unidentified.

[11] We do know of one alternative version to the text that was prevalent in Spain in the 13th c. This is known as "Deuteronomy Rabbah [Lieberman]."

[12] It seems, for example, that *Yalkut Shimoni* on the Prophets and the Yemenite anthology, *Midrash Hagadol*,

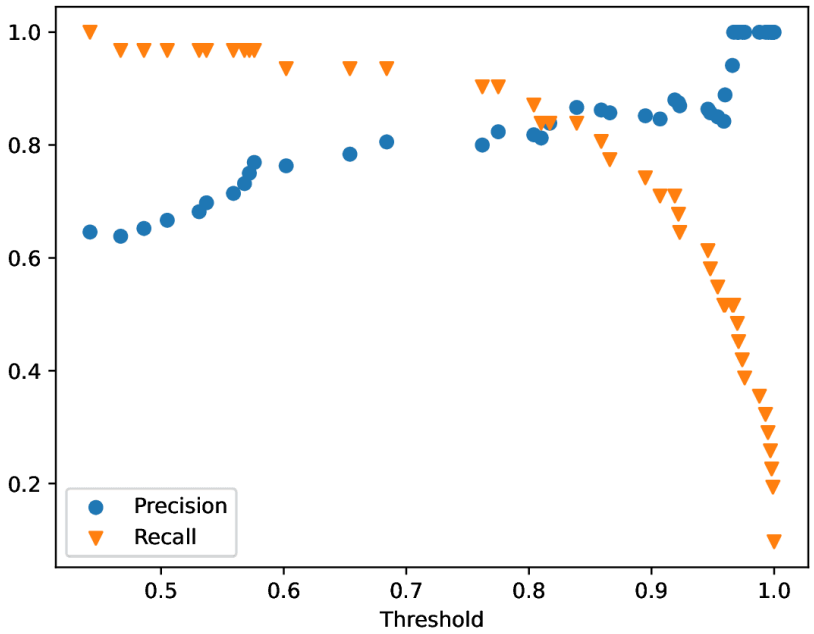

Figure 4: Precision and recall as function of the decision threshold for lost Tanḥuma material.

Another notable finding is that some of the lost midrash collections known only from Ashkenaz (e.g. דברים זוטא, אספה, אבכיר) got a very high score for Tanḥuma style. This might hint that there is a stronger connection between these works and the Tanḥuma literature than previously thought, and perhaps they should be considered as part of the same genre as Tanḥuma in some contexts. [13]

Finally, there were a number of paragraphs from *Sifre Deuteronomy*, a midrash halakha of the tannaitic period, that were detected by our classifier as Tanḥuma. One such paragraph (*Sifre Deuteronomy* 26) contained some notable phrases associate with Tanḥuma and other later midrashic works including זהו שאמר הכתוב ("As it is said in Scripture") and הקדוש ברוך הוא ("The holy one, blessed be He").[14] As it turns out, in one of the manuscripts (Vatican manuscript 32) some of these terms do not appear. This phenomenon might suggest that over the course of time some terms from later periods such as the Tanḥuma literature might have made their way into our current versions of earlier texts.

V USER TOOLS

In order to provide access to our model's predictions and corresponding explanations, and turn our research into a tool that can assist midrash scholars, we built an interactive application based on the open-source Streamlit platform to wrap our model's inference process. Given an input paragraph, the app displays the scores for each of the classes along with features' (unigrams, bigrams and trigrams) corresponding contributions (Figure 5).

Additionally, as can be seen in Figure 6, we display the contribution of the various parts of the

tend to rework early material more extensively than does *Yalkut Shimoni* on the Pentateuch.

[13] The strong correlation of these texts with the Tanḥuma genre has been validated by the only comprehensive study of these texts, as documented by Geula [2006]. The fact that our method has highlighted a largely unrecognized phenomenon within the Humanities field underscores its significant practical value for scholars.

[14] As opposed to the prevalent use of המקום (lit. "The Place") in the tannaitic period, for example, as a metonym for God.

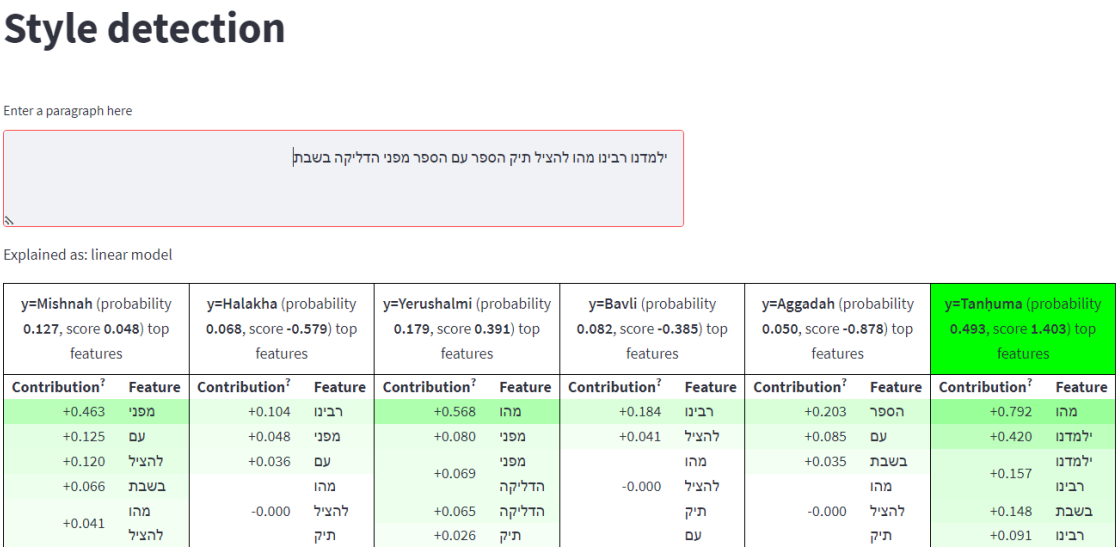

Figure 5: An example of our application's output on a typical *Midrash Tanḥuma* text.

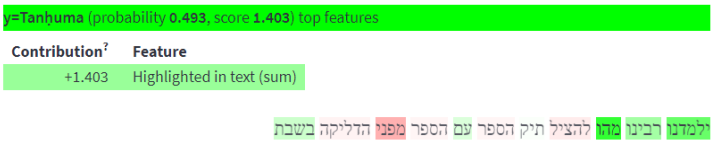

Figure 6: Significant features are highlighted in the text to provide an explanation that is easier to process.

text to the prediction in a more convenient way by highlighting the important features in the text.

Conclusion

In conclusion, our method for detecting Tanḥuma sections in *Yalkut Shimoni* showcases its potential as a valuable tool for scholars involved in the recovery of lost rabbinic material. This is particularly relevant in the ongoing initiative to develop a digital library of Tanḥuma-Yelammedenu Literature, where our work has significant implications. The tools and classifiers we have developed in this study will prove useful for midrash researchers engaged in compiling various Tanḥuma sources and discovering related and potentially lost material of this genre. These resources are publicly accessible,[15] fostering their widespread future use in related projects.

Beyond its immediate application, our method offers potential for wider use in Jewish studies. One exciting future direction is to examine the baraitot[16] that appear in the Babylonian Talmud and in the Jerusalem Talmud. Investigating their interrelationships and connections with other tannaitic sources could provide fresh insights into these traditions.

Additionally, the suggested method could be applied to the many unorganized and unstudied manuscripts discovered in collections such as the Cairo Geniza,[17] paving the way for their automatic classification. Despite the challenges posed by the noisy text created by current

[15] Online at https://github.com/shlomota/Midrash-Style-Classification
[16] A tannaitic tradition not incorporated in the Mishnah, see: "Baraita," *The Jewish Encyclopedia*.
[17] Online at https://fgp.genizah.org.

engines for handwritten text recognition, the potential benefits to the academic community in terms of improved access and understanding of these vital documents are considerable.

This research represents a significant advancement in the use of computational tools for analyzing Jewish literature and traditions. As we continue to refine and expand our methodologies, we anticipate further contributions to the discovery and development of innovative tools, which will undoubtedly enhance our understanding of Jewish textual traditions.

References

Navot Akiva and Moshe Koppel. A generic unsupervised method for decomposing multi-author documents. *J. Assoc. Inf. Sci. Technol.*, 64:2256–2264, 2013.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283, March 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04448-z. URL https://doi.org/10.1038/s41586-022-04448-z.

Marc Bregman. *The Tanhuma-Yelammedenu Literature: Studies in the Evolution of the Versions*. Gorgias Press, 2003.

Idan Dershowitz, Navot Akiva, Moshe Koppel, and Nachum Dershowitz. Computerized source criticism of biblical texts. *Journal of Biblical Literature*, 134(2):253–271, 2015. ISSN 00219231. URL http://www.jstor.org/stable/10.15699/jbl.1342.2015.2754.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Christopher Forstall and Walter Scheirer. *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*. Springer, Cham, Switzerland, January 2019. ISBN 978-3-030-23413-3. doi: 10.1007/978-3-030-23415-7.

Amos Geula. *Lost Aggadic Works Known only from Ashkenaz: Midrash Abkir, Midrash Esfa and Devarim Zuta*. Phd thesis, Hebrew University of Jerusalem, 2006.

Patrick Juola. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1:233–334, March 2008. doi: 10.1561/1500000005.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1356–1364, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1136.

Ronit Nikolsky and Arnon Atzmon. *Studies in the Tanhuma-Yelammedenu Literature*. Brill, 2021. URL https://classics-at.chs.harvard.edu/classics18-schor-raziel-kretzmer-lavee-kuflik/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Uri Schor, Vered Raziel-Kretzmer, Moshe Lavee, and Tsvi Kuflik. Digital research library for multi-hierarchical interrelated texts: from 'Tikkoun Sofrim' text production to text modeling. *Classics@18*, 2021.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 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, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.4. URL https://aclanthology.org/2022.acl-long.4.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. Introducing BEREL: BERT embeddings for rabbinic-encoded language. *Computing Research Repository*, arXiv 2208.01875, 2022. doi: 10.48550/ARXIV.2208.01875. URL https://arxiv.org/abs/2208.01875.

Michal Bar-Asher Siegal and Avi Shmidman. Reconstruction of the Mekhilta Deuteronomy using philological and computational tools. *Journal of Ancient Judaism*, 9(1):2–25, 2018. doi: https://doi.org/10.30965/21967954-00901002. URL https://brill.com/view/journals/jaj/9/1/article-p2_2.xml.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.