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

Shlomo Tannor
School of Computer Science
Tel Aviv University
shlomotannor@mail.tau.ac.il

Nachum Dershowitz
School of Computer Science
Tel Aviv University
nachum@tau.ac.il

Moshe Lavee
Department of Jewish History
Haifa University
mlavee@univ.haifa.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 recently released pretrained Transformer models for Hebrew. Additionally, we demonstrate how our method can be applied to uncover lost material from Midrash Tanḥuma.

1 Introduction

Midrash anthologies are multi-layered works that consist of text in multiple languages, composed by different authors spanning different generations and locations. The midrash collator often merges and quotes various earlier sources, sometimes paraphrasing previous material. These complex processes can make it hard for scholars to clearly separate and detect the different sources which the collection is composed of. Identifying sections which originate in one source or another can shed light on many scholarly debates and help researchers gain a better understanding of the historical development of the rabbinic corpus.

The ability to analyze and classify rabbinic texts in an automated way has tremendous potential. Placing old manuscripts, uncovering lost material that is quoted in later works (e.g. parts of Midrash Tanḥuma, *Mekhilta Deuteronomy*), and determining authorship or dating of a text are examples for such uses. This great potential motivated us to turn to current state-of-the-art natural language processing (NLP) methods to determine whether we can currently solve any such high-impact problem.

We propose a system for classification of rabbinic literature by detecting unique stylistic patterns in the language of the text. Additionally, we demonstrate how our classifier can be used to uncover lost midrashic material that is quoted in later works. As a test case, we apply our method to detect lost sections of the Midrash Tanḥuma that are quoted in the *Yalkut Shimoni*.[1]

2 Related Work

Work from recent years on authorship attribution and plagiarism detection has demonstrated the effectiveness of stylometry and literary style classification in general.

Dershowitz et al. (2015) perform automatic biblical source criticism by looking at preferences among synonyms and other stylistic attributes. Siegal and Shmidman (2018) used computational tools to help reconstruct the lost *Mekhilta Deuteronomy*. They start off with a list of candidate texts, and the main problem they focus on is removing quotes or near-quotes of existing material from other sources. Ithaca (Assael et al., 2022) is an impressive toolkit for restoration and classification of ancient Greek epigraphs.

[1] A medieval midrash anthology from the 13th century CE.

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 42–46
November 20, 2022.

3 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:

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 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 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 which 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 which make up what is referred to as Tanḥuma-Yelammedenu Literature. The works included in training are: *Midrash Tanḥuma*, *Midrash 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"` in order 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.

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

[3] See "Exodus Rabbah," *Encyclopaedia Judaica*, for the rationale behind this division.

[4] See "Numbers Rabbah," *Encyclopaedia Judaica*, for the rationale behind this division.

In addition to the potential benefit of using a model that was pretrained on similar text to that of the target domain, BEREL also uses a modified tokenizer that doesn't split up acronyms which 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 (https://morph-analysis.dicta.org.il/). We then train a logistic regression model over an aggregation of all morphological features that appear in a given paragraph.

3.3 Text Reuse Detection

To achieve our end goal of detecting lost midrashic material, we combine our style classification model with a filtering algorithm based on text-reuse detection. For reuse detection, we use RWFS (Schor et al., 2021), a system designed for this goal using fuzzy full-text search on windows of n-grams. For our corpus of texts 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. The match score for each retrieved document is given by the number of n-gram matches and the results are sorted accordingly.

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*.[5] While lost Tanḥuma material is explicitly cited in some works, it is often quoted without citation in various anthologies.

[5]An early dictionary for rabbinic literature from the 11th century CE.

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

Table 1: Model accuracy on validation set.

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

1. Extract all passages from the given midrash collection, in our case we used *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 $\leq 0.2 \cdot \#n$-grams in query.)

4 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-tuned model. The BEREL-based model leads by a significant margin, nevertheless, we choose to use our baseline model for inference on *Yalkut Shimoni* due to its more calibrated scores, and its higher explainability.[6]

In Figure 2, we can see that the the most common errors are mixing 'Tanḥuma' with 'Midrash Aggadah.' 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.

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 1, the Babylonian Talmud is the most quoted class, while the Jerusalem Talmud is rarely, if ever, quoted. Our classifier gives a similar distribution to

[6]For logistic regression, the model weights correspond directly to an n-gram's contribution to the score given to a specific class.

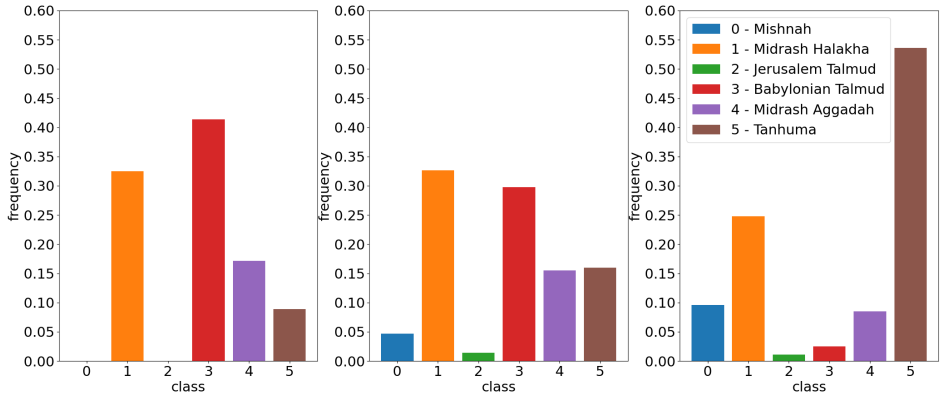

Figure 1: From left to right: (1) class frequencies for passages based on text reuse detection; (2) predicted class frequencies for passages with high text reuse score; (3) predicted frequencies for passages with low reuse score.

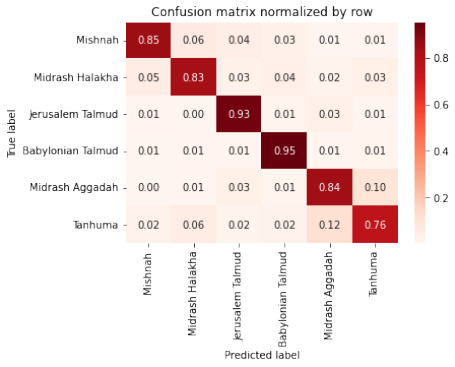

Figure 2: Confusion matrix for baseline model.

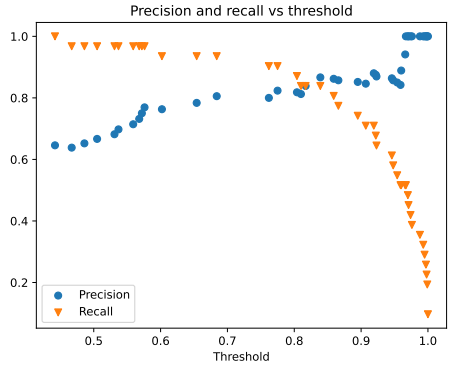

Figure 3: Precision and recall for lost Tanḥuma.

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 for manual labeling a random set of 50 items classified as Tanḥuma. 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 (ילמדנו) 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 typical Tanḥuma material from lost sources,[7] yielding an approximate precision of 60%.

From Figure 3, 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.

5 Discussion and Future Work

Our results for detecting Tanḥuma sections in *Yalkut Shimoni* demonstrate that our method can be a useful tool for researchers working on recovering lost rabbinic material.

We are planning a digital library of Tanḥuma-Yelammedenu literature and believe our work will be of high value to researchers working on detecting lost material of this genre. We intend to

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

run our classifier on additional collections such as *Midrash HaGadol* for which we don't currently have ground truth labels to help uncover additional lost Tanḥuma passages.

Our method can be expanded and applied to many more open questions in Jewish studies. An obvious direction involves applying it to other lost midrashic material. Another is exploring the Baraitot[8] that appear in the Babylonian Talmud and the Jerusalem Talmud and their relationship to each other. Also promising would be to apply it to the many fragmentary manuscripts that have been found in collections like the Cairo Geniza. This would require dealing carefully with noisy text with errors originating in handwritten text recognition.

Acknowledgments

This research was supported in part by grant no. 1188 from the Israeli Ministry of Science and Technology.

References

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

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

Idan Dershowitz, Navot Akiva, Moshe Koppel, and Nachum Dershowitz. 2015. Computerized source criticism of biblical texts. *Journal of Biblical Literature*, 134(2):253–271.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Ronit Nikolsky and Arnon Atzmon. 2021. *Studies in the Tanhuma-Yelammedenu Literature*. Brill.

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. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

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

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. *Computing Research Repository*, arXiv 2208.01875.

Michal Bar-Asher Siegal and Avi Shmidman. 2018. Reconstruction of the Mekhilta Deuteronomy using philological and computational tools. *Journal of Ancient Judaism*, 9(1):2–25.

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. 2020. 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. Association for Computational Linguistics.

[8] A tannaitic tradition not incorporated in the Mishnah, see: "Baraita," *The Jewish Encyclopedia*.