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

A thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Master of Science

by

Shlomo Tannor

This research has been carried out under the supervision

of

Professor Nachum Dershowitz

Tel Aviv University

Blavatnik School of Computer Science

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 our method can be applied to uncover lost material from Midrash Tanḥuma.

Acknowledgments

I would like to express my deepest gratitude to my advisor Prof. Nachum Dershowitz who encouraged me to focus on interesting research problems, find a direction that would bring my different interests together and create something novel and useful while combining my love for computer science and Judaic studies. Thank you for pushing me when I needed it and for giving me the right amount of space and flexibility to follow my own path at the same time.

I would also like to thank Dr. Moshe Lavee from Haifa University who collaborated with us on this work and brought his expertise in midrashic literature. Our joint brainstorming was always a pleasure, and I always felt like I came out of our sessions with some new insights or understanding about Jewish studies, and the potential in leveraging computational tools to assist in this field. Even though we didn't always know where our project would end up, the joint ride and the product gave me great satisfaction.

Finally, I am grateful for my wife Dena who inspired and encouraged me to continue working with dedication even through major life events like the birth of our two children. Thanks for taking care of the zoners!

Table of Contents

List of Figures

List of Tables

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 me 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.

As the core part of this project I published a paper together with my advisor Professor Nachum Dershowitz and in collaboration with Dr. Moshe Lavee of Haifa University titled, "Style Classification of Rabbinic Literature for Detection of Lost Midrash Tanḥuma Material" [14]. An extended version of this paper is now under review for The Journal of Data Mining and Digital Humanities.

In this work, 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]

Since our initial publication, we have continued our joint work, delving into a more in-depth analysis of the tool's predictions on the Tanḥuma literature for the Book of Deuteronomy. We intend to publish our findings in a Judaic studies journal in the near future.

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

2

Related Work

In recent years, the application of natural language processing (NLP) and machine learning (ML) models to authorship attribution, plagiarism detection, and style classification has seen significant advancements, demonstrating their effectiveness in various tasks within stylometry and literary analysis. Dershowitz et al. [4] present an innovative method for automatic biblical source criticism, which examines preferences among synonyms and other stylistic attributes. This approach provides a foundation for using stylistic analysis in the context of religious texts.

Expanding on this, Akiva and Koppel [1] developed an unsupervised algorithm for decomposing multi-author documents. This work further supports the application of NLP and ML models in the field of authorship attribution and serves as a valuable reference for our study.

Additionally, Juola [7] provides an extensive review of authorship attribution techniques, discussing various methods and their applicability to different types of texts. This review offers a comprehensive understanding of the current state of the art in authorship attribution and style classification, which can inform the development of our methodology.

DeepMind, in collaboration with the University of Oxford, introduced Ithaca [2], a groundbreaking toolkit designed for the restoration and classification of ancient Greek epigraphs. This achievement highlights the immense potential of combining artificial intelligence with humanities research, inspiring our work on similar challenges within the field of Jewish studies.

Lastly, Siegal and Shmidman [13] applied computational tools to reconstruct Mekhilta Deuteronomy, a

lost halakhic midrash 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 work.

3

Methods and Results

3.1 Method

3.1.1 Dataset

Our training dataset was extracted from Sefaria's resources.[1] We use the raw text files and divide them into the following categories:

1. 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.

2. Midrash Halakhah – These midrash collections are dated to around the same time as the Mishnah but they are organized by the order of 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.

3. 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 century CE.

[1] `https://github.com/Sefaria/Sefaria-Export`

4. 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 century CE.

5. Midrash Aggadah – In this category we include early midrash works assumed to have been composed during the amoraic period (up to the 5th century CE) 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.

6. Midrash Tanḥuma – In this category we include later midrashic works assumed to have been composed in Palestine and which seem to reference what is known as Midrash Tanḥuma. The works included in training are: Midrash Tanḥuma, Midrash Tanḥuma Buber, and Deuteronomy Rabbah. Other works which we did not use during training but we partially associate with this category include Exodus Rabbah starting from Section 15[2] and Numbers Rabbah starting from Section 15.[3]

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 set.

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

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

3.1.2 Models

3.1.2.1 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 [9] 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.

3.1.2.2 AlephBERT

The next model we evaluate is AlephBERT [11] – 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 [15] for sequence classification, using the default parameters for three epochs.

3.1.2.3 BEREL

The third model we evaluate is BEREL [12] – a Transformer model trained with a similar architecture to that of BERT-base [5] on rabbinic Hebrew texts.

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.

3.1.2.4 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 the 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.

3.1.3 Text Reuse Detection

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

For our text reuse detection, we use RWFS [10], a system designed for this goal by our partners at eLijah Lab, University of Haifa. RWFS uses fuzzy full-text search on windows of n-grams. The system is built on top of a Lucene index, 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. The match score for each retrieved document is given by the number of n-gram matches and the results are sorted accordingly 1.

3.1.4 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 [3, 8]. 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.[4] While the lost Tanḥuma material is explicitly cited in some works, it is often quoted without citation in various midrash anthologies.

To find candidates for "lost" Tanḥuma passages, we apply the following process 2:

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

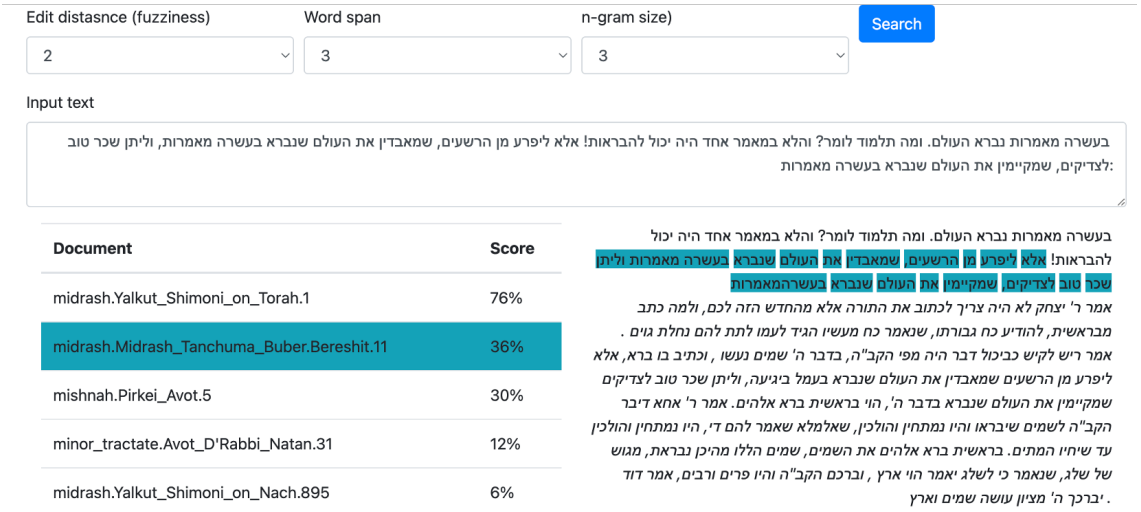

Figure 1: Our text reuse engine (RWFS) shows how a medieval midrash paragraph is reusing early material from various sources including Midrash Tanḥuma.

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 don't have a well established source. (Our threshold was #n-gram matches $\leq 0.2 \times$ #n-grams in query.)

3.2 Results

As can be seen in Table 3.1 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, however we came across multiple challenges when using this model for inference on paragraphs from Yalkut Shimoni:

Model	Validation Acc
Baseline	0.867
AlephBERT	0.879
BEREL	0.922
Morphological	0.560

Table 3.1: Model accuracy on validation set.

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've 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 4 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 extended use of Aramaic as opposed to Hebrew.

After taking the whole Yalkut Shimoni on the Pentateuch and following the process described in Section 3.1.4, we can analyze the prevalence of each class in the collection. As can be seen in Figure 3, 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 text-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 two 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 fifty items classified as Tanḥuma for manual labeling. For the labeling process, we had a midrash expert analyze these passages and look

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,[5] yielding an approximate precision of 0.6.

From Figure 5 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 0.8 by setting an appropriate decision threshold without a high cost to the recall.

3.2.1 Findings

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

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.[6]

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. [7]

Another notable finding is that some of the lost midrash collections known only from Ashkenaz (e.g. אבכיר, אספה, דברים זוטא) 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.

Finally, there were a number of paragraphs from ספרי דברים (Sifrei Devarim; a midrash halakha of the

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

[6] 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).

[7] It seems for example that Yalkut Shimoni on the books of the prophets and Midrash Hagadol tend to rework early material more extensively than Yalkut Shimoni on the Torah.

tannaitic period) that were detected by our classifier as midrash Tanḥuma. One such paragraph (Sifrei Devarim 26) contained some notable phrases associate with Tanḥuma and other later midrashic works including זהו שׁאמר הכתוב ("As it is said in scripture") and הקדוֹשׁ ברוך הוא ("The holy one, blessed be He").[8] 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.

3.3 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 will display the scores for each of the classes along with features' (unigrams, bigrams and trigrams) corresponding contributions (Figure 6).

Additionally, as can be seen in Figure 7, we display the contribution of the various parts of the text to the prediction in a more convenient way, by highlighting the important features in the text.

[8] As opposed to the prevalent use of המקום (lit. "The Place") in the tannaitic period for example.

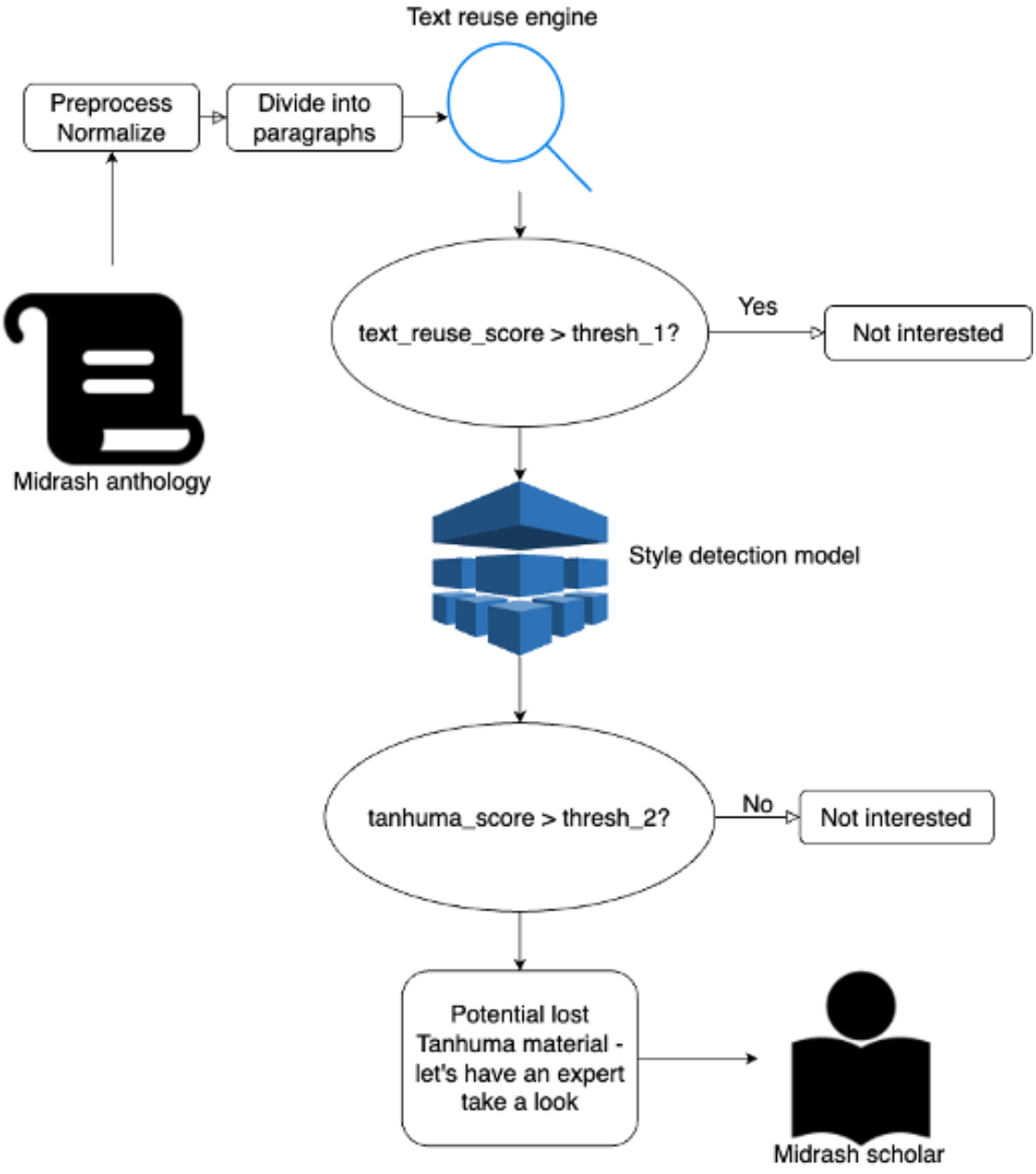

Figure 2: Full flow of our algorithm for detecting lost Tanḥuma material.

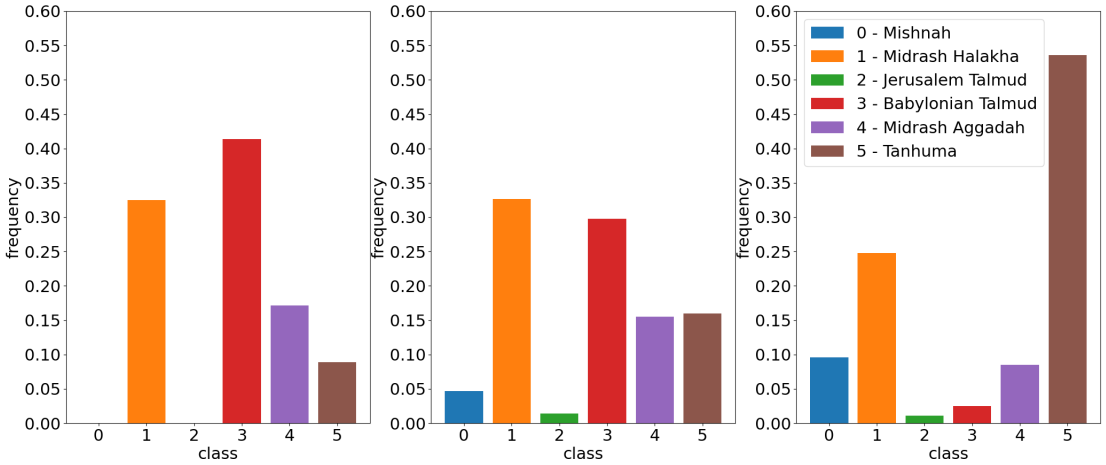

Figure 3: From left to right: (1) class frequencies for passages based on text reuse detection in Yalkut Shimoni; (2) predicted class frequencies for passages with high text reuse score; (3) predicted class frequencies for passages with low text reuse score.

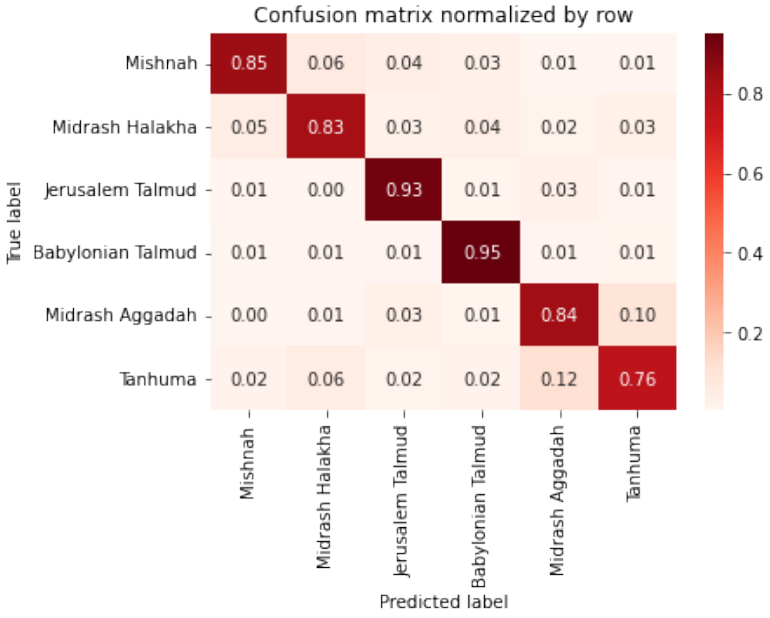

Figure 4: Confusion matrix for baseline model.

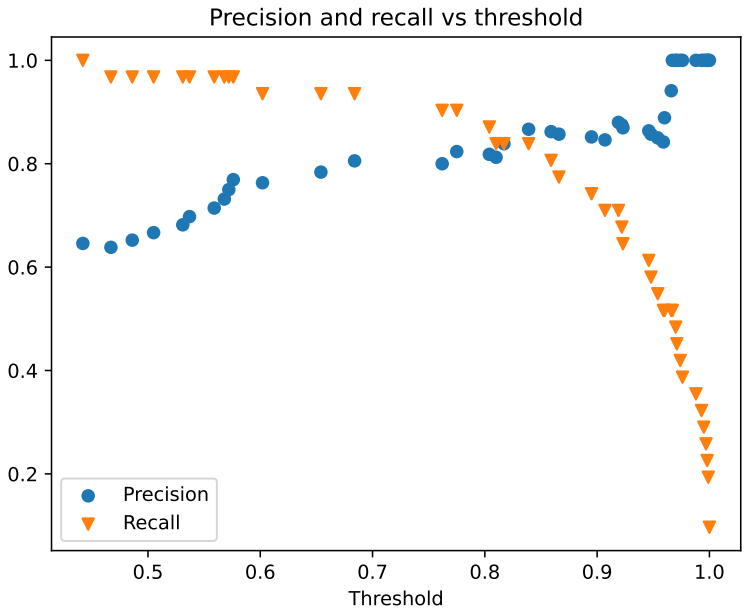

Figure 5: Precision and recall for lost Tanḥuma detection.

Enter a paragraph here

ילמדנו רבינו מהו להציל תיק הספר עם הספר מפני הדליקה בשבת

Explained as: linear model

y=Mishnah (probability 0.139, score 0.032) top features		y=Halacha (probability 0.084, score -0.473) top features		y=Yerushalmi (probability 0.180, score 0.290) top features		y=Bavli (probability 0.097, score -0.323) top features		y=Aggadah (probability 0.067, score -0.694) top features		y=Tanchuma (probability 0.433, score 1.169) top features	
Contribution?	Feature	Contribution?	Feature	Contribution?	Feature	Contribution?	Feature	Contribution?	Feature	Contribution?	Feature
+0.323	מפני	+0.053	רבינו	+0.356	מהו	+0.109	רבינו	+0.131	הספר	+0.429	מהו
+0.100	עם	+0.032	עם	+0.064	מפני הדליקה	+0.012	להציל	+0.056	עם	+0.355	ילמדנו
+0.089	בשבת	+0.019	מפני			+0.003	הדליקה	-0.008	מהו להציל	+0.178	ילמדנו רבינו
+0.042	תיק	-0.008	מהו להציל	+0.064	בשבת	-0.009	מפני הדליקה				
+0.036	להציל			+0.053	מפני				ילמדנו	+0.112	רבינו

Figure 6: An example of our application's output on a typical Midrash Tanḥuma text.

y=Tanchuma (probability 0.433, score 1.169) top features

Contribution?	Feature
+1.169	Highlighted in text (sum)

ילמדנו רבינו מהו להציל תיק הספר עם הספר מפני הדליקה בשבת

Figure 7: Significant features are highlighted in the text to provide an explanation that's easier to process.

4

Discussion

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.

There is currently an initiative for developing a digital library of Tanḥuma-Yelammedenu Literature, and we believe our work will be of high value to midrash researchers working on collecting various Tanḥuma sources along with related material and potential lost material belonging to this genre. The tools and classifiers we created in this research will be released and available to those working on such projects.

Additionally, our method can be expanded and applied to many more open questions in Jewish studies. One future direction is exploring the Baraitot[1] that appear in the Babylonian Talmud and the Jerusalem Talmud and their relationship to each other and to other tanaitic sources.

Finally, another promising direction would be to apply our method to the many unorganized manuscripts that have been found in collections like the Cairo Geniza and classify them automatically. This effort would require dealing carefully with noisy text with errors originating in handwritten text recognition.

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

Bibliography

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

[2] 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, Mar 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04448-z. URL https://doi.org/10.1038/s41586-022-04448-z.

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

[4] 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.

[5] 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.

[6] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794, 2022.

[7] Patrick Juola. Authorship attribution. Foundations and Trends® in Information Retrieval, 1:233–334, 03 2008. doi: 10.1561/1500000005.

[8] Ronit Nikolsky and Arnon Atzmon. Studies in the Tanhuma-Yelammedenu literature. Brill, 2021.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Solomon Tannor, Nachum Dershowitz, and Moshe Lavee. Style classification of rabbinic literature for detection of lost midrash tanhuma material. In Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, pages 42–46, Taipei, Taiwan, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.nlp4dh-1.6.

[15] 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.

Appendix A

Unsuccessful Research Approaches and Challenges

The Talmud in tractate Berakhot 8b states:

לוחות ושברי לוחות מונחות בארון

The tablets (of the Covenant) and the broken tablets are placed in the Ark of the Covenant in the Temple. The broken tablets are not cast away simply because they are broken, but they are preserved in the holiest of places to remind us that our failed attempts and past experiences shape us and teach us much of what we know. In light of this teaching we will share some research approaches that have been tried that did not bear fruit yet.

Early on in our journey we started out with an approach to automatically analyse the structure of the core rabbinic text – the Mishnah.

The Mishnah text is generally organized by topic and divided into 6 sedarim (orders) which in turn contain 63 tractates. However, in the Mishnah there are also many local groups of adjacent passages that are grouped together because of some unifying theme or a shared context. Identifying such groups of passages could help researchers better understand the process of the editing of the Mishnah back in the 3rd century CE.

The objectives and questions we stated back then for this research direction were as follows:

1. Creating "objective" means for analyzing the organization of Tannaitic literature:

   (a) Are the different orders and tractates well-defined entities? Do they each have unique characteristics?

   (b) Can we detect literary units that do not follow the division into sedarim and tractates?

   (c) Analyzing patterns of thematic instability: When and why does it happen? Which topics are related to each other? Etc.

2. Leveraging automation to reveal additional insights into the structure of the Mishnah that have not yet been discussed widely.

In order to assist in this problem we trained models to detect the seder and the tractate of a given passage. We were able to reach 75% validation accuracy on the task of detecting the "order" of the passage and 60% validation accuracy on the task of detecting the tractate when using random 90-10 splits. The accuracy on the tractate classification problem is surprisingly high considering there are 63 different tractates in the Mishnah. This result might hint that the division into tractates is a more significant division than the higher level "order" division.

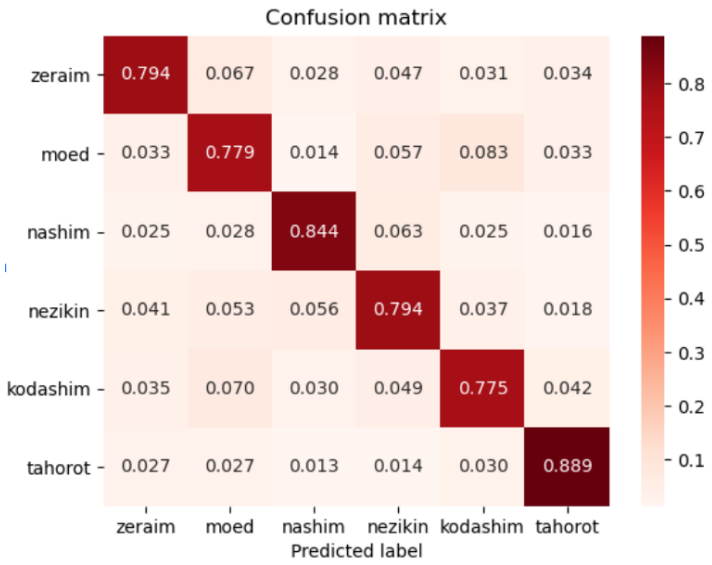

Figure 6: Confusion matrix for seder classification using an English translation to the Mishnah.

Looking at the confusion matrix in Figure 6, we saw that both in the original Hebrew and using an English translation of the text our model performed better on some orders than on others. We can conclude for example that the seder of "Taharot", which deals with ritual impurities and the order of "Zeraim" which deals with agriculture tend to be more self-contained than the the order of "Moed" which discusses the various holidays and calendar year.

Furthermore, using an English translation of the Mishnah for this task instead of the original Hebrew boosted the accuracy significantly (using the same model architecture). However, we suspect this is due to the verbosity of the translation that tends to add contextual information rather than translate the Hebrew word for word.

We also experimented with various encodings of the text that preserve only certain aspects of the text in order to identify the main criteria used for the organization of the Mishanh. Keeping only the 100 most common tokens reduced the validation accuracy from 0.74 to 0.39, preserving only rabbis names reduced the accuracy further to 0.2, indicating that most of the rabbis are featured in discussions across a multitude of topics across all tractates.

We were able to achieve a slightly increase to 0.76 validation accuracy using a finetuned version of AlephBERT that we trained using the masked language modeling task on a large selection of early rabbinic texts before training the model on the downstream classification task.

One notable experiment contained splitting by tractate rather than randomly selecting a 90-10 split of passages. In this setting, for each tractate we trained a designated model that saw all other tractates during training and finally we used the unseen tractate as the validation set. These models' accuracy fell to $\approx$40% from 75% on the validation set which supports our conjecture that most of the tractates are relatively self-contained

Using a self-supervised topic modeling algorithm (on the English translation of the Mishnah), based on BERTopic [6] (a combination of BERT embeddings, UMAP dimension reduction and DBSCAN clustering), we tried to perform the same tasks using only the cluster attributed to each passage. The algorithm generated 86 distinct topics which were enough to obtain 0.61 accuracy on the seder classification problem and 0.38 accuracy on the tractate classification problem. These relatively high accuracy results demonstrate the extent to which tractates and orders can indeed be distinguished by topic.

Finally, after we have seen that the Mishnah generally is divided into self-contained units based on topic,

	Topic	Count	Name
0	-1	1445	-1_said_him_must_have
1	0	160	0_altar_chamber_gate_cubit
2	1	136	1_shabbat_bone_band_ba
3	2	107	2_town_courtyard_courtyards_wall
4	3	104	3_judges_court_ruling_executed
...	...	...	...
81	80	11	80_festival_curry_mourners_comb
82	81	11	81_drained_blood_hatat_sake
83	82	10	82_tithes_exempt_liable_exempted
84	83	10	83_sela_copper_silver_coins
85	84	10	84_grain_consuming_produce_flour

Figure 7: Unsupervised generated topics using BERTopic.

how can we detect those regions in the Mishnah where a division into clear-cut topics was replaced by an editing based on context, theme or desire to not cut and paste an existing oral tradition but preserve the original organic structure?

By applying our models for seder classification and tractate classification on the full text of the mishnah divided into passages we can detect anomalies in the prediction patterns. By examining such anomalies and "mistakes" made by the model we were able to automatically detect many instances of "thematic instability", in which the redactor of the Mishnah preferred another consideration over the clean topic division.

The results of this research direction were hard to quantify and formulate as a significant discovery at the stage we were at, and so we decided to take the toolkit we developed, refine it further and apply it to related problems including the detection of lost Midrash Tanḥuma literature.

תקציר

אוספי מדרשים הינם יצירות רבניות מורכבות הכוללות טקסט במספר שפות, אשר עברו תהליכים ארוכים ולא יציבים של העברה בעל־פה ובכתב. זיהוי המקור של פסקה נתונה באוסף כזה היא משימה לא קלה הנתונה לויכוח בין חוקרים, אך משימה זו חשובה ביותר להבנת הפסקה ויחסה לטקסטים אחרים בקורפוס הרבני.

על מנת לסייע בפתרון בעיה זאת, אנו מציעים מערכת סיווג של הספרות הרבנית על־סמך סגנון הטקסט, תוך שימוש בהתפתחויות האחרונות בתחום של עיבוד שפה טבעית עבור טקסטים בעברית. בנוסף, אנו מדגימים איך המתודה שלנו יכולה לסייע בזיהוי קטעים אבודים השייכים לספרות התנחומא־ילמדנו.

הפקולטה למדעים
מדויקים ע"ש ריימונד
ובברלי סאקלר
אוניברסיטת תל אביב

סיווג סגנונם של טקסטיים רבניים לשם זיהוי קטעים אבודים מספרות מדרש תנחומא

חיבור זה הוגש כחלק מהדרישות לקבלת התואר
"מוסמך אוניברסיטה" – M.Sc.

על ידי

שלמה טנור

העבודה נכתבה בהדרכתו של

פרופ' נחום דרשוביץ

בית הספר למדעי המחשב ע"ש בלווטניק
אוניברסיטת תל אביב