Style Classification of Rabbinic Literature for Detection of Lost Midrash Tanhuma 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

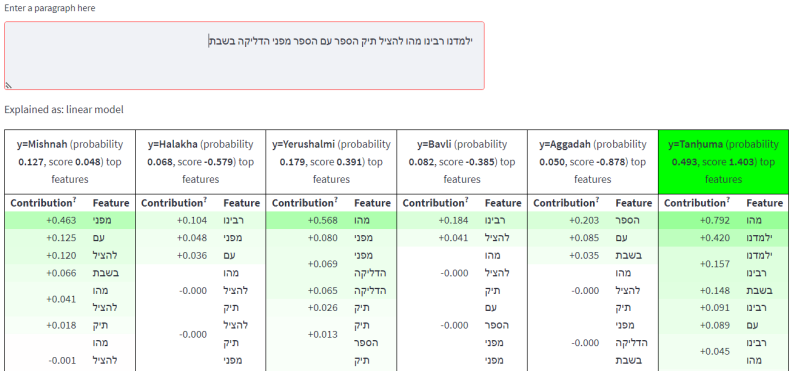
Abstract

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 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, Tanhuma-Yelammedenu, that has been preserved in later anthologies. We suggest a novel approach to solve this task that combines a text-reuse engine with a style classification model.

For our style classification model we created a dataset of rabbinic texts divided into six broad categories: Mishnah, Midrash Halakha, Babylonian Talmud, Jerusalem Talmud, Classic Midrash Aggadah, and Tanhuma. We then trained and evaluated multiple models on their ability to detect the class of an unseen text. The models we tested include logistic regression over n-grams (our baseline), Hebrew BERT based models and a model that examines only morphological data. When applied to unknown paragraphs from Yalkut Shimoni we were able to achieve over 60% precision for detecting lost Tanhuma material based on expert evaluation.

The method we propose has proved effective in detecting and analyzing stylistic elements in midrashic texts, uncovering patterns and features that are pivotal in hypothesizing the texts' origins. Additionally, this research underscores the potential of computational text analysis in advancing the study of ancient texts, offering a novel perspective in the exploration of midrashic literature's stylistic richness and diversity.



Enter a paragraph here

ילמדנו רבינו מהו להציל תיק הספר עם הספר מפני הדליקה בשבת

Explained as: linear model

y=Mishnah (probability 0.127, score 0.048) top features		y=Halakha (probability 0.068, score -0.579) top features		y=Yerushalmi (probability 0.179, score 0.391) top features		y=Bavli (probability 0.082, score -0.385) top features		y=Aggadah (probability 0.050, score -0.878) top features		y=Tanhuma (probability 0.493, score 1.403) top features	
Contribution?	Feature	Contribution?	Feature	Contribution?	Feature	Contribution?	Feature	Contribution?	Feature	Contribution?	Feature
+0.463	מפני	+0.104	רבינו	+0.568	מהו	+0.184	רבינו	+0.203	הספר	+0.792	מהו
+0.125	עם	+0.048	מפני	+0.080	מפני	+0.041	להציל	+0.085	עם	+0.420	ילמדנו
+0.120	להציל	+0.036	עם	+0.069	מפני הדליקה		מהו	+0.035	בשבת	+0.157	ילמדנו רבינו
+0.066	בשבת		מהו			-0.000	להציל		מהו		
+0.041	מהו להציל	-0.000	להציל	+0.065	הדליקה		תיק	-0.000	להציל	+0.148	בשבת
			תיק	+0.026	תיק		עם		תיק	+0.091	רבינו
+0.018	תיק מהו	-0.000	להציל תיק	+0.013	תיק הספר	-0.000	הספר		מפני	+0.089	עם
-0.001	להציל		מפני		תיק		מפני	-0.000	הדליקה	+0.045	רבינו מהו
							מפני		בשבת		

Example of our style detection algorithm on a typical Tanhuma passage