How 'Pagan' Is My Text? Information Extraction from Untranscribed Data

Rachael M. Griffiths[1,2] ⬤, and Marieke Meelen[1] ⬤

[1] DTAL, University of Cambridge, United Kingdom
[2] EPHE-PSL, Paris, France

Abstract

In this paper, we present our work-in-progress on Information Extraction and Text Classification from large manuscript collections that have not yet been transcribed. We propose a three-stage pipeline starting with digitisation using a collaborative Handwritten Text Recognition (HTR) workflow, followed by Normalisation and Segmentation of the texts to create searchable collections, and, finally, we discuss how Text Classification and Information Extraction can help us identify the texts with Tibetan 'Pagan' religious features that are hidden among texts that belong to the Buddhist and Bön religious traditions.

Keywords: HTR, Information Extraction, Text Classification, Tibetan, Religious Studies

1 Introduction

In the 7th century CE Buddhism was introduced to Tibet as part of a grand imperial project, and it became the official national religion in the 8th. This seems to have had little impact on the autochthonous 'Pagan' religious practices: there were still animal sacrifices at royal funerals [1] and there were no monks but instead a class of hereditary priests to revere the Tibetan king [4]. To distinguish itself from Pagan cults, one religious school identified themselves more specifically as Yungdrung ("Eternal") Bön. Yungdrung Bön is still practised by 2-5% of Tibetans to this day [3]. This religion contains much that resembles Buddhist belief and practice. Like Buddhism, Bön has a very rich scriptural tradition. However, some of the Bön texts concern beliefs and practices that are clearly unrelated to Buddhism: these contain traces of the Pagan traditions that were suppressed by the Buddhist authorities during the imperial period. Until recently, almost all we knew about these pre-Buddhist Pagan traditions comes from Old Tibetan texts discovered in a cave at Dunhuang, on the Silk Road, or hidden, older elements in canonical Bön texts.

In 2005, however, a large number of manuscripts (over 75,000 pages) constituting the ritual repertoire of a class of priests, called Leyu, was discovered in the Sino-Tibetan borderlands.[1] Preliminary investigations of facsimiles suggest that these texts contain genuinely archaic non-Buddhist rituals and narratives closely resembling those of the early Pagan sources. In addition, further collections (containing at least another 25,000 pages) have been identified in Bön villages in Dolpo and Mustang (Nepal), among the indigenous Baima communities in Sichuan and Gansu (China), and in manuscripts recovered from the Dunhuang caves and in the Gathang Bumpa stupa, southern Tibet [4]. These collections have the potential to form an invaluable resource to uncover the myths, ritual practices and worldview of the oldest religion of Tibet, but there are major

[1] Facsimiles of some of these manuscripts have been published in collections. For more information on the discovery and publication of these manuscripts, see [4].

challenges to overcome before we can reconstruct the features that can give us a comprehensive understanding of these Pagan religious practices. First of all, the collections contain a large amount of Buddhist and Bön texts. In the latter, we often find parts of myths or rituals that go back to ancient Pagan traditions, but these are not marked in any way. Second, the manuscripts are not yet transcribed, and manually producing transcriptions is difficult due to the extremely challenging continuous (i.e. without word breaks) Tibetan *ume* script full of complex abbreviations and non-standard spellings.

In this paper, we therefore first focus on the crucial question of how to find the 'Pagan' religious features in this vast amount of extremely difficult and untranscribed data. We propose a three-stage pipeline. Stage 1 involves the digitisation process, using a collaborative Handwritten Text Recognition (HTR) workflow (Section 2). Stage 2 consists of Normalisation and Segmentation of texts to create searchable collections (Section 3). Finally, in Section 4, we discuss how Text Classification and Information Extraction can help us identify the Pagan texts and sections that are hidden among the Buddhist and Bön materials.

Figure 1: Sample of the diverse newly-discovered Tibetan corpus.

2 Handwritten Text Recognition

In order to maximise our efficiency with limited time and funding, we have developed a collaborative workflow that utilises the strengths of all participants.[2] The script in which most of these collections are written is a difficult form of Tibetan *ume* or 'headless' script, which only a handful of expert philologists and some Tibetan Bön monks can read. Figure 1 shows a sample of this diverse corpus. In addition to the challenges posed by the script and text format (horizontal Tibetan pecha style, developed from Indian palm leaf manuscripts), the language itself reflects a non-standard variety of Tibetan, with unorthodox spellings and mixed with local, non-Tibetan words and graphemes/symbols. In addition to these philological challenges, the vast amount of data raises questions about both short- and long-term storage. To resolve these issues, we collaborate with international experts in Tibetan philology (both internal and external),[3] data storage (the online BDRC library, who will make our eTexts publicly available),[4] as well as programmers from Esukhia, a Tibetan tech and education charity based in Dharamsala.[5] Figure 2 shows this workflow (see [13] for further details on the HTR workflow).
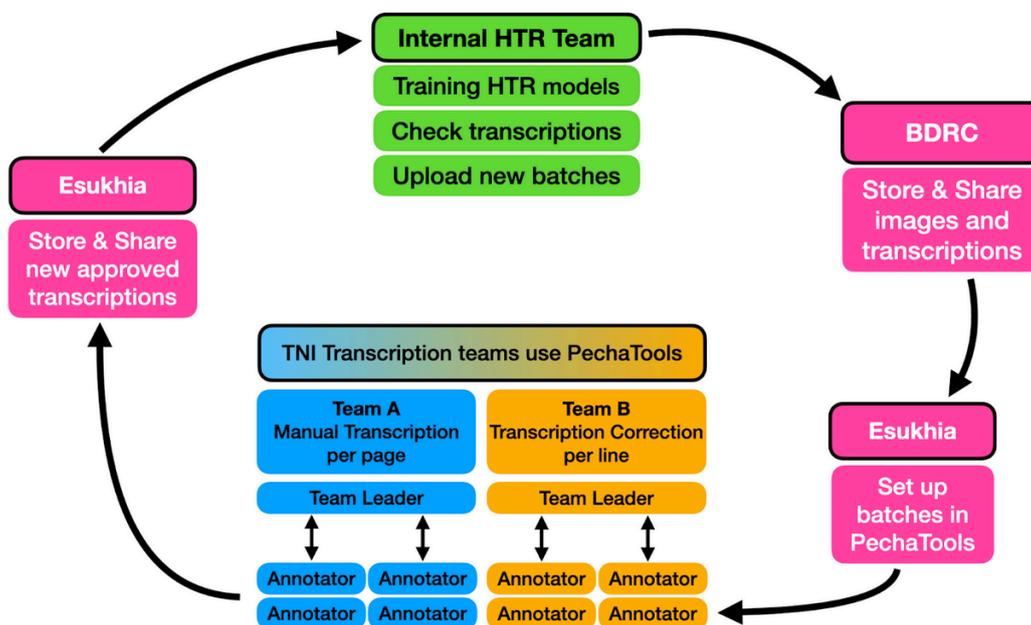

Figure 2: Collaborative HTR workflow.

2.1 Creating Ground Truth

As the internal HTR team, we first select a diverse batch of images with different forms of Tibetan *ume* scripts (see sample in Fig. 1). We then divide this into two separate batches: a 'manual input' batch and a 'correction' batch. Both batches are uploaded and backed up on the BDRC servers, and are also uploaded to a custom-made tool created by Esukhia for inputting and correcting transcriptions (shown in Fig. 3).

---

[2] The potential benefits of leveraging collective expertise to enhance the accuracy and accessibility of HTR projects have been discussed in recent scholarship on collaborative or crowdsourced transcription workflows, such as [2], [15], and [16].

[3] The external experts are monks from the Triten Norbutse Institute (TNI) in Kathmandu.

[4] The Buddhist Digital Resource Centre: `https://www.bdrc.io/buda-archive/`
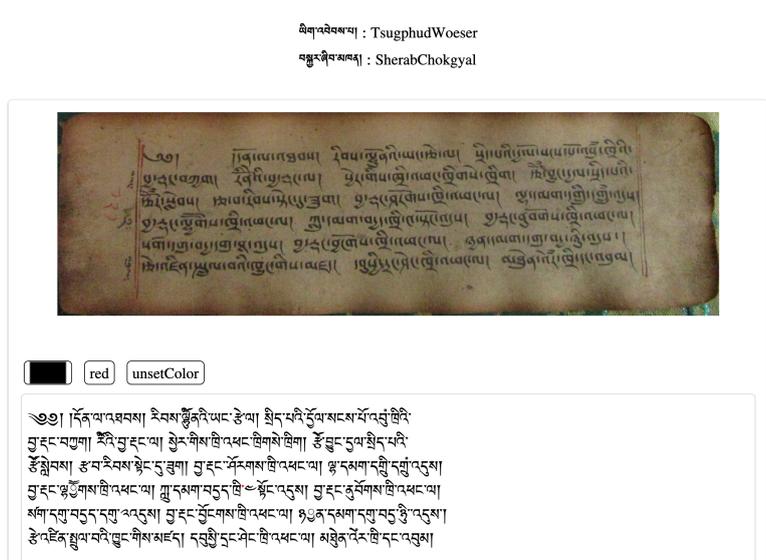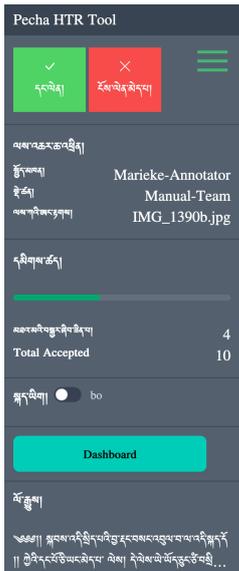
[5] `https://esukhia.net/`

Figure 3: Screenshot of the final-reviewing page (i.e. third check) of manual input tasks.

Each batch is assigned to a specific transcription team. The 'manual team' ('Team A' in Fig. 2) consists of four TNI monks who are given 'manual input' batches to transcribe manually. The 'correction' batches undergo a different process. These are first transcribed automatically using the Tibetan *ume* models created by the TibSchol project in Vienna,[6] which are available through the Transkribus platform [7].[7] Although the Ground Truth that formed the basis of these HTR models also consists of Tibetan *ume* texts, the quality of the images used are generally better and the handwriting is much clearer and more regular than our collections.[8] Consequently, the resulting transcriptions contain a high number of errors (over 60% on average), which require extensive manual correction. The output from the TibSchol HTR model is uploaded alongside the images from the 'correction' batches, which is then corrected by the 'correction team' ('Team B' in Fig. 2), consisting of four TNI monks. The transcriptions produced by both teams then undergo two rounds of review: firstly by the TNI team leaders, and then a final review by the internal HTR team.

To help both teams create diplomatic transcriptions, the internal team prepared a transcription manual [11] and cheat sheet [8] with examples of the most challenging cases, e.g. Chinese or other non-Tibetan characters and symbols (see. Fig. 4). Creating these guidelines for purely diplomatic transcriptions was essential due to the difficult nature of the source material and the general tendencies of the TNI monks to 'interpret' and therefore normalise the text by adding expanded forms and fixing spellings. While Normalisation is eventually essential for downstream NLP tasks (see Section 3 below), the haphazard nature of 'corrections' at this stage poses issues.

For example, any form of Normalisation brings the transcription further away from the original manuscript text. This creates issues for training new HTR models as the transcriptions no longer directly correspond to the images. If abbreviations were consistent this would not be a problem, but there is major variation in scribal conventions across these collections. Losing this variation essentially means losing philological features that can help us try to establish the date and place

---

[6] The Dawn of Tibetan Scholasticism (11th–13thc.) (H2020-Cog-101001002) `https://cordis.europa.eu/project/id/101001002/reporting`. The HTR models trained by the TibSchol project were the only ones available in January 2024. Since then, the BDRC has developed an OCR app that includes *ume* models. For a discussion of this app and its suitability for the manuscripts presented here, see [13].

[7] `www.transkribus.org`

[8] Samples of the images from the TibSchol Ground Truth can be found in [10] and [7].

of origin of the individual texts, which will be of crucial importance for establishing how Pagan traditions were connected and possibly spread to different parts of Tibet and its borderlands. We therefore first focus on diplomatic transcriptions that resemble the original manuscripts as accurately as possible.

2.2 HTR training & results

All transcriptions are triple-checked (by two TNI monks and one internal expert) to ensure high-quality data that will be used as Ground Truth for training HTR models. These enhanced checks slow down the process, but the difficulty of the source material requires this: after only one or two checks, 10-15% of errors typically remained. If such a high error rate is embedded in the Ground Truth, the overall accuracies drop to the extent that models become unusable for transcribing new images.

Since the Ground Truth used for the TibSchol HTR *ume* models differs from our collections,[9] we incrementally trained new models - using the Transkribus platform - based on 490, 900, and 1200 triple-checked pages from our collections.[10] Each iteration contained a wide variety of training data from collections found in Dolpo and Mustang in Nepal, as well as along the Sino-Tibetan borderlands, representing different image quality, scribal practices, and orthographic conventions (if any). Because of the diversity of these training sets, the resulting models perform much better on our diverse collections than the TibSchol HTR models.[11] Testing on two randomly selected texts that were not part of the Ground Truth, the Tibschol HTR model has an average error rate of 59.79%,[12] whereas the new models trained on 490 and 900 pages already show an improvement with error rates of 23.45% and 12.74% respectively [8]. Because of this, later 'correction' batches for the external HTR team were processed with these newer models to improve the overall results. After completing 2,000 pages, the 'manual input' team was switched to correction tasks, resulting in the creation of c. 30,000 pages of double-checked data in one year.

Character Error Rates (CER) on validation sets (we reserve 10% of the overall triple-checked data for validation) still show variation (5-10%), indicating that adding more Ground Truth remains worthwhile given the diversity of scribal practices across the collections.[13] We therefore estimate that with approximately 2,500 pages of triple-checked training data, the HTR model should perform well enough to process all remaining images in the collections and produce diplomatic transcriptions that preserve crucial philological variation. The following section discusses how these diplomatic transcriptions can be converted into searchable texts suitable for Information Extraction and Text Classification.

3 Normalisation & Segmentation

As mentioned in Section 2, these collections are written in a challenging form of Tibetan that exhibits a variety of complex features not found in standard Classical Tibetan texts. While the diplomatic transcriptions are crucial for the preservation of philological and linguistic variation,

---

[9] The TibSchol models were trained on c. 500 pages of *ume* script from a corpus of Tibetan Buddhist scholastic literature written between the 11th and 14th centuries. Most of the images are black-and-white scans. Transcriptions are diplomatic - although most punctuation is omitted or simplified - and use the Wylie transliteration system [17] in Roman script [10].

[10] Two of these models have already been made publicly available on Transkribus: 'PaganTibet Ume 1' (Transkribus ID: 131553) and 'PaganTibet Ume 2' (Transkribus ID: 193821). The Ground Truth is deposited on Zenodo [9].

[11] The TibSchol models produce output in Wylie transliteration, which poses accessibility challenges for our downstream applications. Because we chose to produce transcriptions in Tibetan Unicode, the TibSchol models could not serve as a base model.

[12] We used the model 'Tibetan Cursive (Drutsa)' (Transkribus ID: 176765), as its script style better corresponds to our manuscript material than 'Tibetan Cursive (Betsug)' (Transkribus ID: 54935).

[13] See [13] for a breakdown and analysis of HTR results.

they are not directly suitable for creating annotated corpora or any other downstream NLP tasks. In order to annotate the texts, the raw transcriptions need further preprocessing in the form of Normalisation and Word Segmentation before they can be analysed as searchable eTexts.

3.1 Normalisation

Previous work on Normalisation of Tibetan has largely focused on converting Old Tibetan into a more standardised form of Classical / Written Tibetan [6]. Although this provides a useful starting point for handling certain archaic orthographical features, the Normalisation required for our diplomatic transcriptions is considerably more extensive, including splitting contracted forms, recognising and converting non-Tibetan characters, decoding abbreviations, and converting dialect features. Examples of these challenging features are shown in Fig. 4.

We therefore use an existing preprocessing script as a starting point [5],[14] but have extended it in two key ways.[15] First, we expanded the set of standardisation rules to address the non-standard forms described above. Second, we built a preliminary abbreviation dictionary to enable the automatic expansion of contracted forms. These initial extensions, however, do not yet comprehensively cover all complex cases in the collections. To expand our dictionary and build on the existing preprocessing script, we use a collaborative workflow similar to that discussed in Section 2. Esukhia has also developed a custom-made tool to identify and normalise textual variants, shown in Fig. 5, as well as a 'Dictionary Correction' tool that allows expansions to be checked in context before they are added to the growing abbreviation dictionary.

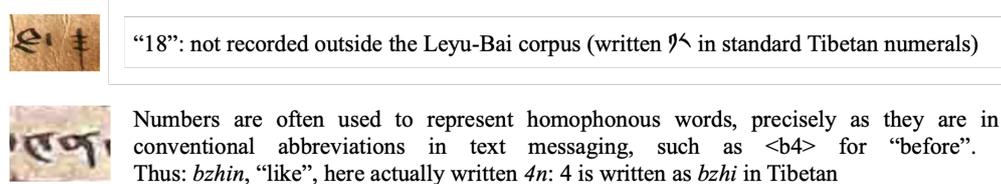|	
	"18": not recorded outside the Leyu-Bai corpus (written ⁹⁸ in standard Tibetan numerals)
	Numbers are often used to represent homophonous words, precisely as they are in conventional abbreviations in text messaging, such as <b4> for "before". Thus: *bzhin*, "like", here actually written *4n*: 4 is written as *bzhi* in Tibetan

Figure 4: Examples of challenging cases of Normalisation.

After this manual correction stage, we apply a postprocessing script to resolve any remaining recurring issues, such as non-standard or double punctuation marks. The normalised text is then compared with the diplomatic version to identify new cases that can be added to the Normalisation rules and abbreviation dictionary. In this semi-supervised workflow, part of the corpus is manually checked twice, and the rest is processed automatically using optimised Normalisation rules.

3.2 Segmentation

Since there are no word boundaries in the Tibetan script, Tokenisation is not a trivial task. Normalisation opens up the opportunity to use existing Segmentation tools originally developed for Classical Written Tibetan. We therefore use an existing segmenter [12], which reports a high level of accuracy (99%) and is freely available and easy to use.[16] An additional advantage of using this script is the ability to extend its output beyond Word Segmentation to include Sentence Segmentation and detailed morphosyntactic annotation. Both of these are valuable for further NLP tasks, which we turn to in the next section.

---

[14] https://github.com/lothelanor/actib/blob/main/preprocessing.py

[15] Scripts developed for the HTR-to-Text Classification workflow are being made available at www.github.com/pagantibet

[16] https://github.com/lothelanor/actib

Figure 5: Customised tool to normalise diplomatic transcriptions.

Diplomatic: ॥གཞིན་གིས་སྨྲ་དབ་ཡུ་སྨོ

Normalised: ॥གཞིན་གིས་སྨྲ་ླ་དབ་ལགས་སྨོ॥

Figure 6: Example of diplomatic and normalised text.

4 How Pagan is my text?

Once all collections are normalised and segmented, they can be compared with known Buddhist and Bön canonical texts. When experts read individual texts, they can identify parts that contain Pagan features based on typical non-Buddhist and non-Bön content, such as rituals involving animal sacrifice, fox fumigation, priests as mediators etc. Yet even when the language is more standardised, reading over 100,000 pages to identify which parts contain Pagan features is impossible. Focusing only on text titles would reduce the task, but titles do not always reflect the actual content or can be absent altogether. Furthermore, Pagan myths and rituals are often embedded within larger text that are otherwise part of the Buddhist or Bön canon.

Therefore, even if all collections are easily searchable (i.e. transcribed, normalised, and segmented as discussed above), the specific question of how we can distinguish the Pagan (parts of) texts from the Buddhist and Bön ones remains. In NLP terms, this question can be addressed through a combination of Information Extraction and Text Classification methods.

4.1 Information Extraction

Extracting specific information from texts can be done in various ways. We make a distinction between philological or codicological features on the one hand, such as specific scribal practices and textual elements, and linguistic and content features on the other. Identifying Pagan features within our corpus requires extracting specific content elements that can be used to help differentiate traditions, including deity names, ritual practices, and specific terminology. We are pursuing this through two complementary approaches.

The first is through Title Extraction and Text Segmentation. As no comprehensive catalogue exists for our collections, we manually tag key features such as titles, images, and decorative elements, while preparing Ground Truth in Transkribus. Once finalised, these tags will be extracted from TEI exports to create a content catalogue and to segment the corpus into analysable units. However, titles alone often prove vague and unreliable indicators of content, necessitating more targeted feature extraction.

The second approach is through semi-supervised feature annotation. Esukhia has developed a custom annotation tool that enables both internal and external team members to annotate specific Pagan features within normalised texts. These manually annotated examples will serve as training data for future work that involves supervised Information Extraction focusing on Named Entities and Semantic Roles in the first instance. This approach will enable identification of Pagan passages and provide fine-grained searchability beyond categorical classification alone.

4.2 Text Classification

We finally need to classify the texts in our collections to find those most suitable and interesting for further analysis. In this final section we demonstrate that this is possible with a small proof-of-concept pilot study based on a limited set of texts we have already manually classified as either Buddhist, Bön or containing Pagan features. Texts that are already known for their Pagan features come from mansucripts found in Dunhuang (D), Gathang (G) and a number of places in the north of Nepal, which are part of the 'Black Water' (W) collection. The newly discovered texts in the Sino-Tibetan borderlands are part of the Leyu and Baima collections.

For this pilot study, we first selected 30 texts from the existing (DGW) and new (Leyu-Baima) Pagan collections, alongside two known Buddhist and two Yungdrung Bön texts. Longer texts were split into several parts to make the overall comparison more balanced. A complete overview of the texts and their corresponding collections is shown in Table 1.

We further preprocessed the texts to ensure comparison would be purely based on semantic similarity and not on philological or scribal conventions. Many of the Leyu texts, for example, use a specific set of numerals, special punctuation markers, and numerous abbreviations (see Fig. 4). These were all normalised to a more standard form of the language to ensure parity of comparison.

To simplify the first proof-of-concept testing, we created simple word embeddings with Fast-Text and aggregated them into document embeddings for each text (or part of a text for those that were particularly long).[17] To get an initial impression of the viability of this approach, we visualised these document embeddings using a simplified Singular Value Decomposition (SVD) of the combined word representations of the three known classes of texts: Buddhist, Yungdrung Bön, and the DGW, Leyu and Baima collections. Fig. 7 shows that based on this simplified semantic similarity alone, we can see that there are clear clusters of each of these. Known Buddhist texts (sūtras and philosophical treatises) appear in the top left corner. These texts are very different in content from the Leyu-Baima and DGW texts, which are the ones with known Pagan features. In addition, it indicates that Yungdrung Bön texts indeed contain Pagan (i.e. more Leyu, Baima and

---

[17] We used the default settings of the offline FastText word embedding tool, freely available through `https://fasttext.cc/`

Text	Collection
Dur bsil dau bzhugs swo	Baima
Klui bsgo dod	Baima
Me gal sri bsad	Black Water
God sri mnan pa	Black Water
God sri mnan pa B	Black Water
mGo gsum rkong tse	Black Water
Srid pa yab lha gyang gug	Black Water
Srid pa gto nag	Black Water
Sri mnan khyad par can	Black Water
sMan dpyad	Gathang
Sha ru shul ston	Gathang
rNel dri	Gathang
gNag rabs	Gathang
Bya rdang bkrol ba	Leyu
dMu thag bzhugs	Leyu
Phang brngon	Leyu
Bya rdang gzhi btings	Leyu
Bya rdang bsang bul	Leyu
sGam po pha wang	Leyu
Bum bzhi klu gnyan	Leyu
Bya rdang shing mi	Leyu
sPrel dkar shug ma	Leyu
Drang mkhan	Leyu
A bo ya ngal	Leyu
Bya rdang gzhung	Leyu
Bum bzhi sa bdag	Leyu
Bum bzhi gnyan bum	Leyu
Bum bzhi stod bum	Leyu
Bum bzhi klu bum	Leyu
OT annals	Dunhuang
OT chronicle	Dunhuang
Mahāratnakūṭa D75	Buddhist
MiPham 4-5	Buddhist
gZibrjid	Yungdrung Bön
gZermig	Yungdrung Bön

Table 1: Overview of selected sample texts for pilot from different collections.

DGW) features, but that these texts are also heavily Buddhicised in terms of content: those Bön texts all appear right between the Buddhist and the Leyu-Baima and DGW collections as we would expect knowing the content. This small pilot therefore demonstrates the proof-of-concept needed for further research.

5 Conclusion and future work

In this short paper, we presented our work-in-progress addressing the question of how to identify 'Pagan' religious texts within a vast corpus of untranscribed material. We developed a three-stage pipeline to address this problem systematically. First, we tackle transcription through a collabora-
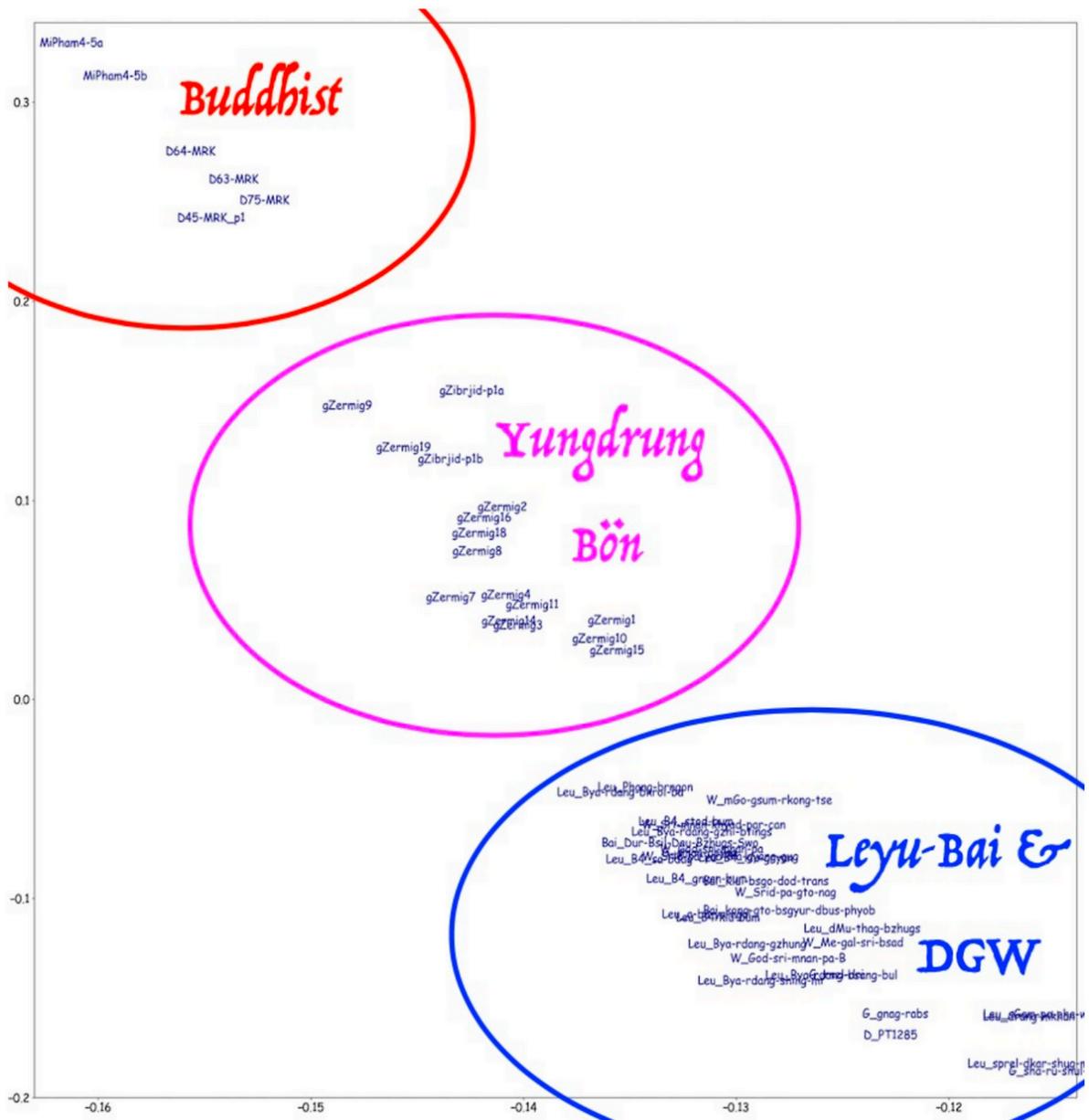
Figure 7: Singular Value Decomposition of combined document representations with added Text Classification labels for known sample texts.

tive workflow, incrementally training new HTR models for the challenging and variegated Tibetan *ume* script. We then create normalised and segmented (tokenised) versions of each text, correcting spelling and expanding abbreviations to make them fully searchable and ready for downstream NLP tasks. Finally, these normalised transcriptions are used in a pilot study on Information Extraction and Text Classification, training word and document embeddings and comparing the resulting vector representations to texts with known classification labels to determine how Pagan new texts are.

This pilot study demonstrates the feasibility of this approach and highlights areas for improvement. With the known texts from the pilot study turned into document embeddings, we can create a text classifier by averaging the document vectors of each of the known Buddhist, Bön and Pagan texts and provide new texts (after turning them into a document embedding in the same way) with

a label depending on which of these three shows the closest resemblance. Visually, in Fig. 7 this can roughly be represented by adding the new text and seeing in which section it falls: Buddhist, Yungdrung Bön or Leyu-Baima & DGW (i.e. Pagan). In future work, we aim to extend the embedding models using existing (larger) language models, such as the latest version of Gemma [14],[18] to label and extract topics (i.e. more fine-grained category labels) from texts within the collections. In addition, we will conduct more in-depth quantitative and qualitative evaluations to enhance the results.

Another tip. In some cases, it may be helpful to use `paragraph` to title individual paragraphs. For example, if a section describes features for a classifier, you can optionally title each paragraph with the name of each feature.

References

[1] Aldenderfer, Mark. "Variation in mortuary practice on the early Tibetan plateau and the high Himalayas". In: *Journal of the International Association for Bon Research 1* (2013), pp. 293–317.

[2] Andersdotter, Karolina and Nauwerck, Malin. "Secretaries at Work: Accessing Astrid Lindgren's Stenographed Manuscripts through Expert Crowdsourcing". In: *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*. CEUR Workshop Proceedings (CEUR-WS.Org) (Oct. 2022), ed. by Karl Berglund, Matti La Mela, and Inge Zwart, pp. 9–22.

[3] Baumer, Christoph. *Tibet's Ancient Religion: Bön*. Trumbull, CT: Weatherhill, 2002.

[4] Berounský, Daniel and Ramble, Charles. "Leu Texts from Eastern Tibet: The Emergence of a Canon of Pagan Tibetan Literature?" In: *Canons, Kanjurs and Collections*, ed. by Brambilla, Filippo and Viehbeck, Markus. Verlag der Österreichischen Akademie der Wissenschaften; Beiträge zur Kultur- und Geistesgeschichte Asiens, 2025.

[5] Faggionato, Christian, Hill, Nathan, and Meelen, Marieke. "NLP Pipeline for Annotating (Endangered) Tibetan and Newar Varieties". In: *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 1–6. URL: `https://aclanthology.org/2022.eurali-1.1`.

[6] Faggionato, Christian and Meelen, Marieke. "Developing the Old Tibetan Treebank". In: *Proceedings of Recent Advances in Natural Language Processing*, ed. by Galia Angelova, Ruslan Mitkov, Ivelina Nikolova, and Irina Temnikova. Shoumen, Bulgaria: Incoma, 2019, pp. 304–312.

[7] Griffiths, Rachael M. "Handwritten Text Recognition (HTR) for Tibetan Manuscripts in Cursive Script". In: *Revue d'Etudes Tibétaines 72* (July 2024), pp. 43–51. DOI: `10.1553/TibSchol_ERC_HTRs43`.

[8] Griffiths, Rachael M. and Meelen, Marieke. "HTR Input and Correction Cheat Sheet: 10 Basic Rules and Protocols for Diplomatic Transcription". Version 3.1. Oct. 2025. DOI: `10.5281/zenodo.17251318`.

[9] Griffiths, Rachael M. et al. "Ground Truth for PaganTibet Ume models 1 and 2". Zenodo, Oct. 2025. DOI: `10.5281/zenodo.17275724`.

[10] Hugon, Pascale and Griffiths, Rachael M. "TibSchol HTR tools". Zenodo, Jan. 2024. DOI: `10.5281/zenodo.10451396`.

---

[18] `https://huggingface.co/buddhist-nlp/gemma2-mitra-base`

[11] Meelen, Marieke and Griffiths, Rachael M. "HTR Input and Correction Manual". Zenodo, Oct. 2025. DOI: `10.5281/zenodo.17257009`.

[12] Meelen, Marieke and Hill, Nathan. "Segmenting and POS tagging Classical Tibetan using a memory-based tagger". In: *Himalayan Linguistics 16*, no. 2 (2017), pp. 64–89. DOI: `10.5070/H916234501`.

[13] Meelen, Marieke and Griffiths, Rachael M. "Collaborative Workflows for Handwritten Text Recognition in Under-resourced Manuscript Collections". In: *Journal of Open Humanities Data* (forthcoming).

[14] Nehrdich, Sebastian and Keutzer, Kurt. "MITRA: A Large-Scale Parallel Corpus and Multilingual Pretrained Language Model for Machine Translation and Semantic Retrieval for Pali, Sanskrit, Buddhist Chinese, and Tibetan". In: *unpublished manuscript* (forthcoming).

[15] Romein, Annemieke C. et al. "From research proposal to project management. A guide from the Transkribus community on planning and executing workflows for researchers and GLAM-professionals". In: *International Journal of Digital Humanities* (Sept. 2025), pp. 1–30. DOI: `10.1007/s42803-025-00107-7`.

[16] Tarride, Solène, Faine, Tristan, Boillet, Mélodie, Mouchère, Harold, and Kermorvant, Christopher. "Handwritten Text Recognition from Crowdsourced Annotations". In: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*. HIP '23. New York: Association for Computing Machinery, 2023, pp. 1–6. ISBN: 9798400708411. DOI: `10.1145/3604951.3605517`.

[17] Wylie, T. "A Standard System of Tibetan Transcription". In: *Harvard Journal of Asiatic Studies 22* (1959), pp. 261–267. DOI: `10.2307/2718544`.