Perspectives – Interactive Document Clustering for Qualitative Data Analysis

Tim Fischer, Chris Biemann
Language Technology Group, Department of Informatics, University of Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Abstract

This paper introduces *Perspectives*, an interactive extension of a qualitative data analysis tool suite developed at our university, designed to empower Digital Humanities (DH) scholars to explore and organize large, unstructured document collections. *Perspectives* implements a flexible, aspect-focused document clustering pipeline with human-in-the-loop refinement capabilities. We showcase how this process can be initially steered by defining analytical lenses through document rewriting prompts and instruction-based embeddings, and further aligned with user intent through tools for refining clusters and mechanisms for fine-tuning the embedding model. The demonstration highlights a typical workflow, illustrating how DH researchers can leverage *Perspectives*'s interactive document map to uncover topics, sentiments, or other relevant categories, thereby gaining insights and preparing their data for subsequent in-depth analysis.

1 Introduction

The discourse analysis tool suite[1] (DATS), a platform developed at our university, supports Digital Humanities (DH) researchers in data analysis by providing access to computational methods for large-scale, multimodal data. Key functionalities include pre-processing, qualitative and quantitative analysis, and document retrieval (including keyword, metadata, and semantic search). However, DATS users often struggle to explore and organize extensive corpora. Keyword-based or semantic search functionalities are often insufficient and incompatible with a hermeneutic approach to conceptualization, and DATS lacks functionality for systematic structuring and a holistic overview. Organizing documents, however, is vital for qualitative analysis and advanced DATS features, such as timeline analysis.

Document clustering offers a promising solution, particularly for discovering themes within a corpus, a key interest of DH scholars. Modern clustering-based topic modeling (Grootendorst, 2022; Angelov and Inkpen, 2024; Reuter et al., 2024), which leverages pre-trained language model embeddings, has gained considerable traction not only because it offers advantages in simplicity, modularity, and performance, but also because it presents interesting opportunities for steerability and interactivity.

Unsupervised clustering methods can present a *take it or leave it* scenario for non-experts if the outputs do not align with their intent, rendering fully automatic methods ineffective. For DH scholars, who often approach corpora with specific research questions in mind, the ability to interactively manipulate clustering is of great interest. Such challenges are addressed by the Interactive Topic Modeling (ITM) community, which aims to achieve a controllable, collaborative human-machine process.

To address this gap, we introduce *Perspectives*, an interactive document clustering extension for DATS. Inspired by ITM and built on clustering-based neural topic modeling, we aim to make document clustering more accessible for DH scholars. The system is centered around an interactive document map to support exploratory workflows and implements a clustering pipeline that can be steered and refined in multiple ways. While offering visualization and analysis tools, its primary goal within DATS is to help users organize their collections into tagged datasets, ready for use with other analytical tools. The contributions of this paper are:

1. We introduce a flexible, aspect-focused clustering pipeline that combines (a) LLM-driven document rewriting to emphasize user-defined aspects, (b) instruction-steered embeddings to generate aspect-oriented document representations, and (c) few-shot fine-tuning of the embedding model to further align the representations with user intent.

[1] https://github.com/uhh-lt/dats

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 411–422
July 4, 2026

2. We evaluate the pipeline on multiple datasets, showing the effectiveness of aspect-oriented representations, document rewriting, and fine-tuning.

3. We present *Perspectives*, a human-in-the-loop document (HITL) clustering system that integrates the pipeline and provides an interactive 2D map for users to explore, validate, and directly manipulate clusters with established refinement operations.

2 Related Work

The task of identifying topics in texts has evolved from traditional probabilistic models, such as Latent Dirichlet Allocation (Blei et al., 2003). Recent approaches, termed clustering-based neural topic modeling or topic discovery (Wu et al., 2024), build on pre-trained language model embeddings. Top2Vec (Angelov, 2020) uses Doc2Vec (Le and Mikolov, 2014) embeddings, reduces their dimensionality with UMAP, clusters them using HDBSCAN, and identifies topics by finding words closest to cluster centroids. BERTopic (Grootendorst, 2022) follows a similar pipeline but employs a class-based TF-IDF mechanism to extract descriptive words for each identified cluster. Contextual Top2Vec (Angelov and Inkpen, 2024) refines this by representing documents with multiple segment vectors for fine-grained topic segmentation.

The topic modeling community has long recognized the need to make topic modeling more user-centric. Seed wordlists (Jagarlamudi et al., 2012), seed documents (Grootendorst, 2022), or seed sentences (Fang et al., 2023) can guide initial topic formation. Further, interactive topic modeling systems address alignment with user intent by providing operations such as merging, splitting, or deleting topics, as well as adding or removing documents/words from specific topics (Fang et al., 2023). Seelman et al. (2024) proposes to update topic models by embedding user-assigned word labels to topics. GPTopic (Reuter et al., 2024) leverages LLMs to build an accessible chat interface to explore and refine topics through natural language.

The field of interactive (semi-supervised) clustering also acknowledges the value of integrating expert knowledge. While prior methods, such as hand-picked seed points (Basu et al., 2002) or pairwise constraints (Basu et al., 2002), offered control, they required extensive user feedback. Viswanathan et al. (2024) demonstrated that LLMs can guide semi-supervised clustering with minimal feedback by generating and encoding task-specific keyphrases for document representation.

Several systems facilitate the interactive exploration of document collections or topics. Nomic Atlas[2] is a notable commercial platform that offers high-performance 2D scatter-plot visualizations of document embeddings, enabling users to explore automatically extracted hierarchical topic clusters and interact via filtering and a chat assistant. Related academic systems innovate with HITL focus, such as direct integration into the labeling workflow (Seelman et al., 2024), tracking model changes (Fang et al., 2023), or RAG-based question-answering (Reuter et al., 2024).

Effective visualization helps explore and understand clusters. Standard techniques, such as top words, word clouds, document-cluster distributions, and 2D embedding projections (Wu et al., 2024), are all available in our extension.

While inspired by existing platforms, such as Nomic Atlas, *Perspectives* distinguishes itself by offering a user-centric HITL workflow tailored for DH: Users define the analytical lens before clustering and iteratively refine the outcomes afterward, with visualizations serving as a central component of this interactive sense-making process.

3 Envisioned Workflow

This section illustrates an idealized workflow using *Perspectives*, following Alice, a DH researcher investigating public discourse on climate change in Germany. Her goal is to understand major topics and the general sentiment towards climate change.

Alice uploads her corpus to DATS and opens the new *Perspectives* extension. To discover topics, she provides a *topic discovery* instruction to guide document embeddings (e.g., "Identify the topic"). The system generates an interactive 2D document map and a dashboard that provides a high-level summary, including clusters with LLM-generated names, keywords, and example documents, their relative frequencies, and a similarity matrix showing inter-cluster relationships. Alice navigates the map (see Figure 1), panning, zooming, and inspecting the visually distinct clusters. She identifies broad topics but notices that a theme like "Climate Protests" is missing.

Alice uses selection tools 1) to investigate documents of specific clusters. Linked statistic and information panels 2) provide aggregated details, such as frequent keywords and named entities. She

[2] atlas.nomic.ai

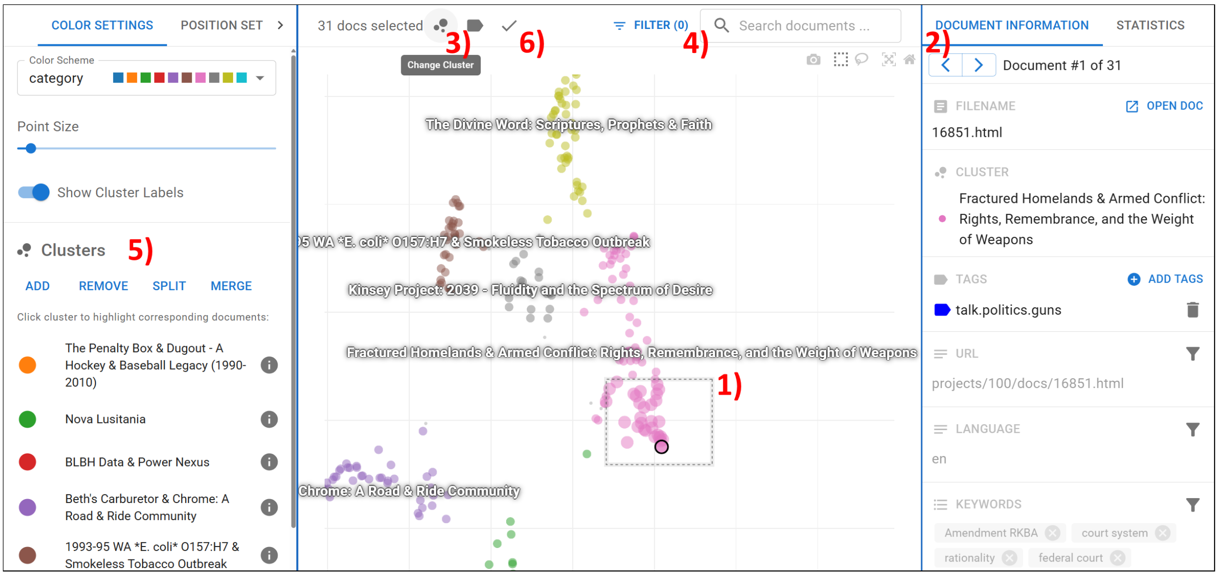

Figure 1: *Perspectives'* document map. Center: interactive scatter plot of documents colored by cluster. Hovering over a document previews its content. Left: settings & refinement operations, including the model fine-tuning operation. Right: statistics & information about selected documents. Top: toolbar with search & filtering. The UI is inspired by popular clustering interfaces such as Nomic Atlas, leveraging a familiar design to accelerate adoption.

spots misclassified articles, correcting them with the **Change Cluster** 3) function. To create her expected "Climate Protests" topic, she selects relevant documents (found via integrated search or filtering, 4) and uses the **Add cluster** function. If a topic is too broad, she uses the **Split Cluster** function to compute more granular sub-clusters. Conversely, she can **Merge Clusters** if several clusters represent a single cohesive theme 5).

With all topics identified, Alice improves their definitions by reviewing representative documents and marking them with the **Accept Cluster Assignment** function 6). Having provided minimal feedback, she triggers the **Refine Model** process, which fine-tunes the embedding model. The map reloads, showing more distinct and coherent clusters. Satisfied with her thematic map, Alice explores sentiment. She creates a new perspective, this time providing a *sentiment analysis* instruction and an optional document rewriting prompt (e.g., "Summarize sentiment towards climate change"). This generates a new map visualizing sentiment-based clusters, which she can explore and refine using the same interactive tools.

Finally, Alice exports her refined topics and sentiments from *Perspectives* as tags back into her main project, enabling further analysis tools within the DATS ecosystem.

4 Interactive Clustering

As illustrated by the workflow, *Perspectives* facilitates interactive document clustering, empowering users to guide analysis towards specific aspects and iteratively refine categorizations. This process begins with an aspect-focused clustering pipeline and continues through HITL refinement operations. The pipeline is illustrated in Figure 2.

We propose two strategies to define the analytical lens and initially steer the clustering process (green). First, we employ alternative document representations, in which an LLM rewrites documents based on a prompt, thereby modifying the content for clustering. This is a one-time process per perspective. Second, we use instruction-finetuned embedding models (Su et al., 2023) that generate aspect-oriented embeddings from an instruction, capturing specific semantic facets.

We continue with established steps (orange): reduce dimensionality with UMAP for improved clustering, then use HDBSCAN to group documents. Cluster representations include (1) keywords from c-TF-IDF, (2) LLM-generated names and descriptions seeded by those keywords, (3) the cluster centroid (mean of document embeddings), and (4) representative documents identified via cosine similarity to the centroid.

Next, document embeddings (reduced to 2D using UMAP) are presented as colored dots on a map. Here, users can iteratively post-process the clus-

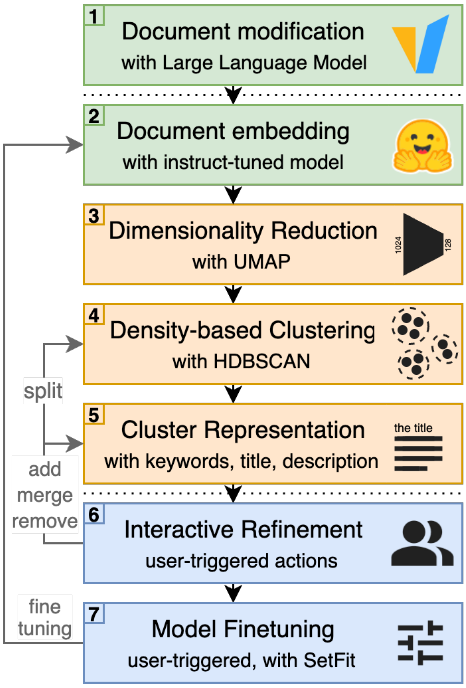

Figure 2: The proposed interactive clustering pipeline. The initial clustering process is guided by providing rewriting and embedding instructions (green) to focus the document representations on user-defined aspects. The established core pipeline (orange) identifies clusters and builds various textual representations. Users can post-process the clustering (blue) through refinement operations (grey), triggering some steps of the pipeline.

tering with several HITL refinement capabilities (blue; grey arrows). Whenever clusters are manipulated (i.e., documents are added or removed), their representations (keywords, name, description, centroid, and representative documents) are recomputed. The refinement operations are:

Change Cluster Documents are reassigned to an existing cluster or marked as an outlier.

Add Cluster (Documents) Selected documents form a new cluster.

Add Cluster (Text) Provided name and description are embedded to define the cluster centroid. Documents are automatically reassigned if their cosine similarity to the new centroid exceeds their similarity to their current cluster's centroid.

Merge Clusters Two clusters are combined into a single new cluster.

Remove Cluster The cluster is deleted. Its documents' embeddings are compared to the centroids of all remaining clusters using cosine similarity and assigned to the most similar one.

Split Cluster HDBSCAN is applied only to the documents within the cluster, yielding a set of more granular sub-clusters, and the original cluster is removed.

Additionally, we propose fine-tuning the model to further align the document representations with user intent. It is inspired by the findings of Thielmann et al. (2024). They demonstrated that with few (1-5) labeled documents per class, significant improvements in topic coherence can be achieved. Perspectives operationalizes this by allowing users to *accept cluster assignments*. These labeled examples are used to create a contrastive dataset to fine-tune the embedding model with the SetFit[3] library. This adjusts the embedding space, pulling similar documents closer and pushing dissimilar ones apart (details are explained in Section 5.4 and Appendix Section A.2). After fine-tuning, the entire dataset is re-embedded and re-clustered, resulting in an updated map that reflects the improved cluster structure.

HDBSCAN inherently identifies outlier documents, visualized as distinctively colored (small, grey) points on the map. They are considered during refinement operations. For instance, when a new cluster is added, outliers can be assigned to it if their similarity is sufficient.

5 Evaluation

We evaluate the effectiveness of our additions to the established clustering pipeline, which aims to be steerable towards certain aspects. We analyze initial guidance through (1) document rewrites and (2) instruction-tuned embeddings, as well as our proposed refinement operation of (3) model fine-tuning with few labeled data.

5.1 KNN Accuracy Metric

The goal of *Perspectives* is to assist users in obtaining a solid document categorization using the 2D document map. While topic modeling inspired our work, we refer to the extensive evaluation of Angelov and Inkpen (2024), which demonstrated the superiority of clustering-based topic modeling approaches over traditional and neural methods. Instead, we evaluate KNN accuracy, a standard metric for clustering interfaces. A good visualization should group documents with the same label and yield high classification accuracy within the visualization space (Pham and Le, 2021). A KNN classifier is trained to classify texts based on their 2D coordinates, and accuracy is reported. It mim-

[3] huggingface.co/docs/setfit

ics user behavior by deriving classes from dense clusters and can be seen as a proxy for end-to-end evaluation. Note that this metric operates directly on the two-dimensional embeddings and does not account for the HDBSCAN clustering.

5.2 Datasets

To evaluate the effectiveness of steering and guidance, we selected various datasets that cover text materials (news articles, books, social media, and songs) and tasks (topic discovery, stance detection, frame analysis, and bias detection) similar to real-world research projects conducted by our project partners who actively use DATS.

Amazon Product Reviews[4] consists of 2,500 sampled reviews with 5 *star* ratings and 15 *product* categories. Spotify Songtexts[5] comprises 5,000 sampled song lyrics with 10 *genres* and 5 *emotions*. 20 Newsgroups includes 3,600 sampled posts on 20 *topics*. GVFC (Liu et al., 2019) contains 1300 news article summaries with 9 *frames*. Blurbs (Remus et al., 2019) consists of 1,200 sampled German book *blurbs* with 8 main genres. Israel-Palestine (Ali et al., 2025) comprises 10,000 Reddit comments with 3 *stances*. News Bias[6] contains 8,500 sampled news articles with 3 political *biases*. For detailed descriptions, see Appendix Section A.1.

5.3 Models

For evaluation, we exclusively use open-licensed models that run locally. This approach ensures data privacy for our primary users (academic researchers) handling sensitive information. It also ensures the system is deployable in resource-constrained environments, accommodating users with limited computing resources. We use Gemma 3 (27B-instruct) (GemmaTeam, 2025) for LLM-driven document rewriting and multilingual-e5-large-instruct (500M) (Wang et al., 2024) for the instruction-tuned embedding model, selected due to its strong performance on the MTEB benchmark (Muennighoff et al., 2023). Further model details are provided in Appendix A.3.

5.4 Experiment Setup

First, we evaluate the impact of document rewriting and task-specific embeddings. We use straightforward prompts to generate summaries and keyphrases of the original documents (see Appendix Table 2). Then, we use short instructions to compute task-specific document embeddings (see Appendix Table 3). This results in 6 scenarios: original text (*text*), summary (*summ*), keyphrases (*keyp*), embedded with (+*inst*), and without instructions. Second, we evaluate the impact of fine-tuning the embedding model with few labeled examples (*2, 4, 8, 16-shot*), aiming to align the embedding space better with validated clusterings. We use SetFit to tune the embedding model for one epoch on increasing amounts of examples sampled randomly from the training data. The few labeled training examples are used to create a contrastive dataset with an equal number of positive and negative pairs. Instead of updating all model parameters, we train LORA adapters (Hu et al., 2022) with a cosine similarity loss to speed up training and reduce hardware requirements. The KNN accuracy metric is computed on document embeddings reduced to two dimensions with UMAP. Few-shot experiments are repeated 10 times with different training sets. The results are averaged across runs to mitigate fluctuations. All experiments were conducted on a single A100 GPU (80GB). All parameters, prompts, and training details are explained in Appendix A.2. The code to reproduce the results is available[7].

5.5 Results

Table 1 presents our evaluation results (see Appendix Figure 3 for a detailed breakdown. Consistent with previous work, we observe that embeddings generated with instructions (+*inst*) outperform those without in 8 out of 9 datasets we evaluated. More interestingly, we observe that document rewrites (keyphrases or summaries) prove beneficial in 6 out of 9 datasets: For example, *summary+inst* improves product categorization accuracy by 24.46 points, stance detection by 17.55 points, and bias detection by 5.33 points. For all other aspects, *text+inst* yields the best performance. Table 1 also highlights our best-performing few-shot results. These scores were achieved using the best-performing *zero-shot* configurations, for example *summary+inst* for product, frames, stance, and bias, *keyphrases+inst* for genre, and *keyphrases* for blurbs. While fine-tuning the embedding model with a few labeled examples consistently improved KNN accuracy, the gains were generally minor,

[4] kaggle.com/datasets/mexwell/amazon-reviews-multi
[5] kaggle.com/datasets/devdope/900k-spotify
[6] kaggle.com/datasets/articoder/news-bias-dataset
[7] https://github.com/uhh-lt/perspectives

	emotion	genre	product	stars	topic	frames	blurbs	stance	bias
text	45.10	27.20	36.72	47.54	71.17	59.78	73.84	51.19	47.20
+inst	**50.60**	27.50	55.76	**58.86**	**71.24**	65.00	73.25	48.21	48.51
keyphrases	49.15	34.85	54.64	46.34	65.70	60.34	**76.73**	61.63	50.76
+inst	49.51	**34.90**	60.64	49.68	65.40	58.31	76.66	63.14	50.69
summary	47.69	32.59	48.16	52.14	60.62	64.34	76.46	62.47	48.41
+inst	48.04	32.64	**61.18**	54.02	61.91	**65.91**	74.80	**68.74**	**52.53**
2-shot	50.70	36.63	62.98	59.71	71.36	65.10	76.79	70.17	52.72
4-shot	50.92	36.60	63.54	60.24	71.49	65.59	76.74	70.56	53.03
8-shot	51.64	36.66	63.62	60.09	71.58	66.92	76.92	71.06	53.09
16-shot	52.07	37.02	64.04	61.27	72.15	67.27	77.85	71.51	54.09

Table 1: Evaluation of document rewriting, instruction-based embeddings, and few-shot fine-tuning. Top: unsupervised (*zero-shot*) results. Bottom: few-shot results using the best configuration (e.g, *text+inst* for emotion & topic).

in the range of 2-3 points, when compared to the *zero-shot* setting. Thielmann et al. (2024), using 1-5 labeled examples per class, observed more significant improvements in topic coherence. Still, our results suggest that incorporating more labeled examples could further enhance performance.

Our findings guided the implementation of *Perspectives*: Task-specific instructions are mandatory for creating a perspective, as they consistently yield superior performance. The impact of document rewrites proved to be dataset-dependent. Therefore, providing a document-rewriting prompt is optional but recommended. To streamline the user experience, we integrated templates for common tasks, such as emotion, sentiment, and topic clustering.

6 Tool Suite Integration

The *Perspectives* frontend is built with React, uses Plotly.js[8] for efficient, interactive scatter plots, and MUI[9] for all other material design components. A FastAPI[10] (Python) backend provides the functionalities through the asynchronous task manager RQ[11]. We implement GPU-powered RQ workers, which ensure UI responsiveness during intensive computations, such as embedding, clustering, and model fine-tuning. We integrate our proposed clustering pipeline into DATS using the same parameters as in our experiments.

Fast response times are critical for interactive systems. We achieve high performance for our LLM-driven functionalities (like name generation and document rewriting) by combining vLLM and LiteLLM[12]. This setup leverages batch processing, streaming, and prefix caching for rapid inference. While the initial, optional document rewriting process may incur some latency depending on corpus size, it is an acceptable, one-time operation. Subsequent refinement operations are significantly faster, as their processing time scales with the number of clusters. Furthermore, the embedding model fine-tuning is also quick, requiring only one training epoch on a small set of labeled examples.

Using modern libraries such as Sentence Transformers and vLLM, switching models is straightforward. This ensures that DATS stays up to date with recent advances in NLP and provides users with state-of-the-art models.

7 Conclusion

This paper introduced *Perspectives*[13], an interactive extension to the open-source DATS platform designed to help DH scholars explore and organize large document collections. We demonstrated an aspect-focused clustering pipeline that leverages LLM-based document rewriting and instruction-steered embeddings. Central to *Perspectives* is HITL refinement, providing a suite of operations to manipulate clusters and fine-tune the embedding model with minimal input. This iterative approach, visualized on an interactive 2D map and supported by a dashboard, empowers users to create meaningful categorizations. By integrating these capabilities, *Perspectives* provides a powerful yet accessible tool that enables domain ex-

[8]https://plotly.com/
[9]https://mui.com/
[10]https://fastapi.tiangolo.com/
[11]https://python-rq.org/
[12]www.litellm.ai, docs.vllm.ai
[13]Demo: https://dats.ltdemos.informatik.uni-hamburg.de/

perts to leverage their knowledge effectively. It transforms document clustering into a collaborative sense-making activity, prepares data for further analysis within DATS, and provides a valuable tool for user-centered DH research. Future work will expand the UI's functionality, extend the scope of the extension to include multimodal data (images), and incorporate more labeled examples for few-shot scenarios.

Limitations

During initial testing of *Perspectives* with our project partners, a challenge in the model fine-tuning mechanism was identified: The **Refine Model** operation requires re-embedding and re-clustering of the entire dataset, resulting in the loss of spatial understanding and disruption of the user's mental map as the 2D projection recalculates. To mitigate this, we are currently developing a refinement history component. This feature, presented as an interactive timeline at the bottom of the map, stores and allows navigation between all previous map states (e.g., pre- and post-refinement views). We animate the transition between any two states, which helps users reconstruct their spatial understanding of the data. This feature also addresses a request from project partners, allowing them to revert undesirable refinement operations.

Our experimental evaluation focused on a single LLM for document rewriting and a single embedding model. Different choices of LLMs or embedding models could yield varying performance outcomes. Furthermore, the few-shot training approach is sensitive to the selection of training samples. To mitigate this high dependency, we conducted 10 runs with different sampled training sets. However, an optimal combination of training examples may not have been discovered. Our experiments also did not involve extensive prompt engineering for document rewriting, as we limited the prompts to summarization and keyphrase generation. Exploring other prompts could potentially lead to further improvements.

References

Hasin Jawad Ali, Ajwad Abrar, S. M. Hozaifa Hossain, and M. Firoz Mridha. 2025. Social media polarization during conflict: Insights from an ideological stance dataset on Israel-Palestine Reddit comments.

Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics.

Dimo Angelov and Diana Inkpen. 2024. Topic Modeling: Contextual Token Embeddings Are All You Need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, Miami, FL, USA. ACL.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2002. Semi-supervised Clustering by Seeding. In *Proceedings of the Nineteenth International Conference on Machine Learning*, page 27–34, San Francisco, CA, USA. Morgan Kaufmann Publishers.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. ACL.

Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and Rob Procter. 2023. A User-Centered, Interactive, Human-in-the-Loop Topic Modelling System. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 505–522, Dubrovnik, Croatia. ACL.

GemmaTeam. 2025. Gemma 3 Technical Report.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with c-TF-IDF.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, Online.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, Avignon, France. ACL.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, page 1188–1196, Beijing, China. JMLR.org.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. ACL.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. ACL.

Dang Pham and Tuan M. V. Le. 2021. Neural Topic Models for Hierarchical Topic Detection and Visualization. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 35–51, Cham. Springer International Publishing.

Steffen Remus, Rami Aly, and Chris Biemann. 2019. GermEval 2019 Task 1: Hierarchical Classification of Blurbs. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Arik Reuter, Anton Thielmann, Christoph Weisser, Sebastian Fischer, and Benjamin Säfken. 2024. GP-Topic: Dynamic and Interactive Topic Representations.

Kyle Seelman, Mozhi Zhang, and Jordan Boyd-Graber. 2024. Labeled Interactive Topic Models.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. In *Findings of the Association for Computational Linguistics*, pages 1102–1121, Toronto, Canada. ACL.

Anton F. Thielmann, Christoph Weisser, and Benjamin Säfken. 2024. Human in the Loop: How to Effectively Create Coherent Topics by Manually Labeling Only a Few Documents per Class. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 8395–8405, Torino, Italia. ELRA and ICCL.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large Language Models Enable Few-Shot Clustering. *Transactions of the Association for Computational Linguistics*, 12:321–333. Place: Cambridge, MA Publisher: MIT Press.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

A Appendix

A.1 Dataset Details

Amazon Product Reviews Customer reviews organized by product category and star rating (proxy for sentiment). We use English split and consider reviews of the 15 most frequent categories. We sample 500 reviews per star for testing.

Spotify Songtexts Song lyrics categorized by *emotion* and *genre*. We consider the top 10 genres and top 5 emotions for English song texts. We sample 500 song texts per genre for testing.

20 Newsgroups A topic modeling dataset comprising 18000 newsgroups posts on 20 topics. We use 20% of the articles for testing.

GVFC The Gun Violence Frame Corpus comprises news article headlines, annotated into 9 different frames. We use an extension, which includes three-sentence summaries of 1300 articles.

German Blurbs This dataset consists of 20k blurbs categorized into 8 main genres, each with its own fine-grained sub-genres. We sample 1200 blurbs for testing. We consider the main genre.

Israel-Palestine A stance detection dataset of 10k Reddit comments related to the Israel-Palestine conflict, categorized as Israel, Palestine, Neutral.

News Bias A dataset 8.5k news events retrieved from AllSides. It provides three views (left, center, right) for each news event. We sample 10k documents from the top 8 topics for testing.

A.2 Experiment Parameters & Libraries

Document Modification We use the instruction-tuned Gemma 3 (27B) model. The vLLM and LiteLLM libraries handle this process for batch inference, with the *max tokens to generate* parameter set to 2048; other sampling parameters, such as *temperature* and *top p*, remain at their default values. Prompts used for document modification are listed in Table 3.

Document Embedding We use the Sentence Transformers library with *batch size* set to 32 and *normalize embeddings* set to true to generate aspect-oriented embeddings with the instruction-tuned *multilingual-e5-large-instruct* (500M) embedding model. Instructions used for document embedding are listed in Table 2

Dimensionality Reduction A GPU-accelerated variant of UMAP from the cuML library reduces the dimensionality of resulting embeddings. The *distance metric* is cosine similarity, *N neighbors* is 15, and *min distance* is 0.0. For evaluation, *N components* is 2. For production, it is 128.

KNN Accuracy Metric The k-nearest neighbors accuracy metric is computed using the sklearn library via 5-fold cross-validation, with the reported accuracies averaged across folds.

Density-based Clustering We use HDBSCAN for density-based clustering. The *min samples* is set to 40, *distance metric* is set to euclidean.

Cluster Representation c-TF-IDF extracts 50 keywords, Gemma 3 generates title & description using constrained generation to ease parsing.

Interactive Refinement The split operation uses the same HDBSCAN parameters, other operations have no parameters.

Model Fine-tuning We use the SetFit library to fine-tune the embedding model in two stages.

For the first stage (one epoch), ten distinct few-shot training sets are created per dataset and shot count (2, 4, 8, 16) by randomly sampling *num shots* examples per class. They are constructed as contrastive datasets of document pairs, where a pair is deemed positive if both belong to the same class and negative otherwise. Positive pairs are oversampled to balance the dataset, which enables substantial training even with limited labeled samples. We use rank-stabilized LoRA adapters with a *rank* of 24, an *alpha* of 8, and a *dropout* of 0.1, drastically reducing the number of trainable parameters. The training objective is to minimize cosine similarity loss.

The second stage (16 epochs) involves training a differentiable classification head on the updated embedding model. We set the *end-to-end training* parameter to true to further train the embedding model. Since our focus is solely on the document embeddings, the classification head is discarded after training.

A.3 Model Details

We utilized two primary models for the evaluation:

Gemma 3 (27B-instruct) for Document Rewriting This is a *lightweight*, instruction-tuned LLM from Google. It was trained on 14T tokens, features a large 128k context window, and supports over 140 languages.

multilingual-e5-large-instruct (500M) for Document Embedding This instruction-tuned embedding model was selected as one of the best-performing models on the clustering subtask of the MTEB benchmark (Muennighoff et al., 2023). It uses mean pooling as its pooling strategy. It is initialized from XLM Roberta (Conneau et al., 2020), continually trained on a mixture of multilingual datasets, and supports 100 languages.

<table>
<tr><th>Dataset</th><th>Instruction</th></tr>
<tr><td>emotion</td><td>Identify the main emotion expressed in the given (summary of a | keyphrases of a) song text</td></tr>
<tr><td>genre</td><td>Identify the main genre of the given (summary of a | keyphrases of a) song text</td></tr>
<tr><td>stars</td><td>Identify the sentiment of the given (summary of an | keyphrases of an) Amazon review</td></tr>
<tr><td>product</td><td>Identify the category of an Amazon product based on the review (summary | keyphrases)</td></tr>
<tr><td>topic</td><td>Identify the topic or theme of the given (summary of a | keyphrases of a) news article</td></tr>
<tr><td>frame</td><td>Identify the framing of the given (summary of a | keyphrases of a) news article</td></tr>
<tr><td>blurb</td><td>Identifiziere das Genre, dass durch die (Zusammenfassung | Schlüsselbegriffe) des Klappentextes beschrieben wird.</td></tr>
<tr><td>stance</td><td>Identify the stance towards the Israel-Palestine conflict of the given (summary | keyphrases) of a news article</td></tr>
<tr><td>bias</td><td>Identify the political leaning of the given (summary of a | keyphrases of a) news article</td></tr>
</table>

Table 2: Instructions provided to the embedding model. We adopted them to resemble those of the original authors.

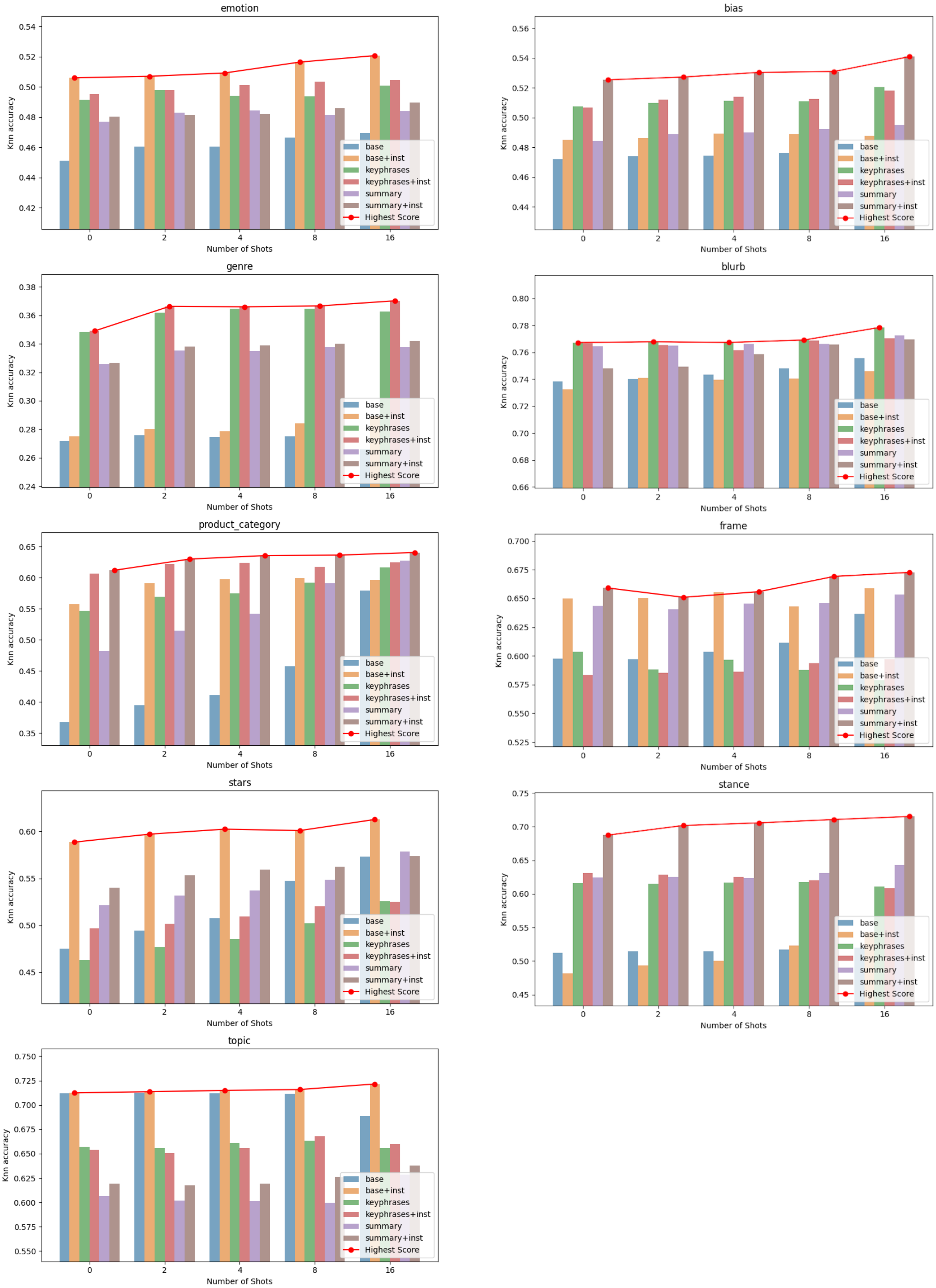

Figure 3: Detailed evaluation with increasing number of labeled examples per class.

Dataset	Prompt
emotion-summary	Write a concise summary (maximum 5 sentences) that focuses on the emotional tone of the following song lyrics. Analyze the lyrics to determine the main emotion being conveyed and describe how it is expressed. Conclude with an emotion categorization:
emotion-keyphrases	Generate keyphrases (max 5 phrases) that describe the emotional tone of the following song lyrics. Focus on phrases that reflect the emotional tone, mood, or feelings expressed in the lyrics.
genre-summary	Write a concise summary (maximum 5 sentences) that focuses on the genre of the following song lyrics. Analyze the lyrics to determine the musical genre, referencing stylistic elements, themes, or influences. Conclude with a genre categorization:
genre-keyphrases	Generate keyphrases (max 5 phrases) that describe the musical genre of the following song lyrics. Focus on phrases that reflect the genre's characteristics, style, or influences.
stars-summary	Write a concise summary (maximum 5 sentences) that focuses on the sentiment of the following Amazon review. Analyze the review to determine the main sentiment being conveyed and describe how it is expressed. Conclude with a sentiment categorization:
stars-keyphrases	Generate keyphrases (max 5 phrases) that describe the sentiment of the following Amazon review. Focus on phrases that reflect the sentiment, mood, or feelings expressed in the review.
product-summary	Write a concise summary (maximum 5 sentences) that focuses on the product categorization of the following Amazon review. Analyze the review to determine the discussed product's categorization, referencing its features, type, or purpose. Conclude with a categorization:
product-keyphrases	Generate keyphrases (max 5 phrases) that describe the product of the following Amazon review. Focus on phrases that reflect the product's type or category.
topic-summary	Write a concise summary (maximum 5 sentences) that focuses on the topic or theme of the following news article. Analyze the article to determine the main topic or theme being discussed. Conclude with a topic categorization:
topic-keyphrases	Generate keyphrases (max 5 phrases) that describe the topic of the following news article. Focus on phrases that reflect the main topic being discussed.
blurbs-summary	Schreibe eine prägnante Zusammenfassung (maximal 5 Sätze), die sich auf das Genre des folgenden Klappentextes konzentriert. Analysiere den Text, um das Genre zu bestimmen. Schließe mit einer allgemeinen Genre-Kategorisierung ab.
blurbs-keyphrases	Generiere Schlüsselbegriffe (max 5 Begriffe), die das Genre des folgenden Klappentextes beschreiben. Fokusiere dich auf Begriffe, die das allgemeine Genre widerspiegeln.
stance-summary	Write a concise summary (maximum 5 sentences) that focuses on the stance of the following Reddit post. Analyze the post to determine whether it is Pro-Israel, Pro-Palestine, or Neutral. Conclude with a stance categorization:
stance-keyphrases	Generate keyphrases (max 5 phrases) that describe the stance of the following Reddit post. Focus on phrases that reflect the stance being discussed.
bias-summary	Write a concise summary (maximum 5 sentences) that focuses on the political leaning of the following news article. Analyze the article to determine the main political framing. Conclude with a political categorization:
bias-keyphrases	Generate keyphrases (max 5 phrases) that describe the political leaning of the following news article. Focus on phrases that reflect the political framing.

Table 3: Prompts used to rewrite the original documents. We use the same prompts for topics and frames.