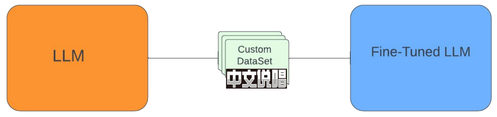
Fine-tuning LLMs for Chinese rap lyrics



Moshi Fu (Undergraduate studying Statistics, DS, & Mathematics)
Brendan C. Dowling (PhD Candidate in Chinese Language Sciences)
2025 Research Bazaar Poster
The University of Wisconsin–Madison

Abstract

We investigated embedding-based clustering of Chinese rap lyrics using LLM to support downstream tasks such as semantic analysis. Our pipeline involved using Qwen2 - 1.5B to convert lyrics to high dimensional embeddings. sim-cse based contrastive learning to let model learn our lyrics. Subsequent dimensionality reduction via UMAP and HDBSCAN clustering were performed. Our study wants to learn if small scale LLM can understand well for regionally distinctive language

Background

Objective: Unite our interdisciplinary experience to explore Chinese rap music topics based on textual features–as opposed to audio features.

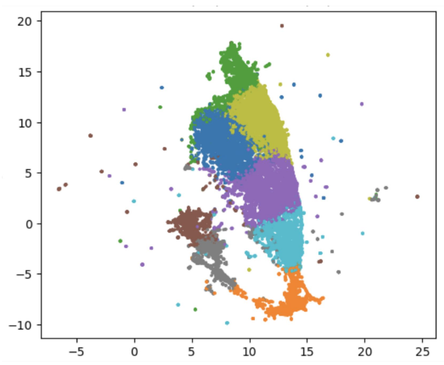
Dataset: Lyrics from 15000 rap songs and over 325 artists

Methodology

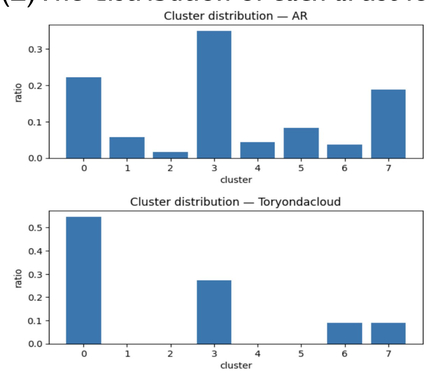
1. Data Preprocessing
   a. pandas and re
2. Transformer embedding conversion
   a. Embeddings by Qwen2 - 1.5B
3. Self supervised learning
   a. Contrastive learning using SimCse
4. Dimensionality Reduction
   a. UMAP
5. Clustering
   a. HDBSCAN
6. Evaluation
   a. Visualization
   b. silhouette score

Some Results of our downstream tasks

(1)Visulization of the cluster



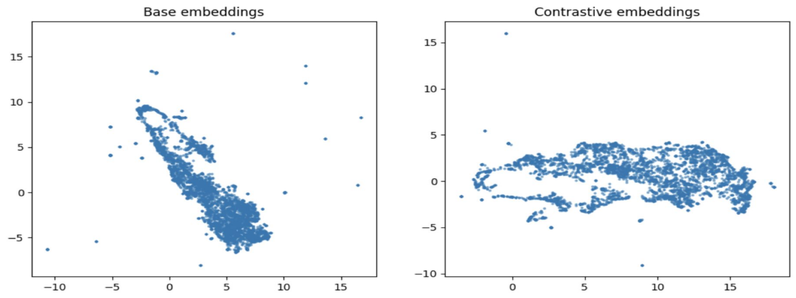
(2)The distribution of each artist for the 7 clusters



(3) Investigation of each cluster: what kind of emtion/ feeling does each cluster represent?

	cluster	chunks	hits	hit_积极	ratio_积极	hit_消极	ratio_消极	hit_愤怒	ratio_愤怒	hit_思念	ratio_思念	hit_励志	ratio_励志	hit_派对	ratio_派对	hit_城市/意象	ratio_城市/意象	hit_爱情	ratio_爱情	suggested_label
0	0	4753	6943	1301	0.187383	315	0.045369	478	0.068846	369	0.053147	766	0.110327	1239	0.178453	2301	0.331413	174	0.025061	城市/意象
1	1	1839	257	28	0.108949	13	0.050584	2	0.007782	6	0.023346	11	0.042802	162	0.630350	30	0.116732	5	0.019455	派对
2	2	2706	4308	824	0.191272	319	0.074048	155	0.035980	272	0.063138	431	0.100046	486	0.112813	1668	0.387187	153	0.035515	城市/意象
3	3	5426	13870	2976	0.214564	1027	0.074045	634	0.045710	957	0.068998	1761	0.126965	2262	0.163086	3705	0.267123	548	0.039510	城市/意象
4	4	1570	553	80	0.144665	8	0.014467	7	0.012658	10	0.018083	35	0.063291	98	0.177215	307	0.555154	8	0.014467	城市/意象
5	5	2185	384	75	0.195312	20	0.052083	15	0.039062	17	0.044271	38	0.098958	85	0.221354	125	0.325521	9	0.023438	城市/意象
6	6	5410	15554	4020	0.258454	1947	0.125177	654	0.042047	1455	0.093545	2168	0.139385	1138	0.073164	3144	0.202134	1028	0.066092	积极
7	7	2944	4284	810	0.189076	349	0.081466	165	0.038515	324	0.075630	564	0.131653	863	0.201447	1009	0.235528	200	0.046685	城市/意象

(4) comparison between and after the contrastive learning(only sample 5000 embeddings)



(5)Some examples of top words/ top artists for each cluster

	cluster	rank	term	score
0	0	1	斗地主	0.000958
1	0	2	哈哈哈	0.000743
2	0	3	斗地主 斗地主	0.000678
3	0	4	扒皮	0.000590
4	0	5	哈哈哈 哈哈哈	0.000500
5	0	6	芳草地	0.000479
6	0	7	帅到 朋友	0.000456
7	0	8	扒皮 扒皮	0.000444
8	0	9	帅哥	0.000429
9	0	10	疆要	0.000421
10	0	11	帅到	0.000398
11	0	12	拐子	0.000363
12	0	13	淅沥沥	0.000362
13	0	14	超哥	0.000348
14	0	15	钱生	0.000345

cluster_top_artists_v2

cluster	artist	percentage
0	壳子	3.32
0	YOUNG	2.36
0	谢帝	2.1
0	Bridge	1.73
0	李尔新	1.66
0	Mengzi	1.58
0	宝石老舅	1.56
0	大傻 (Damnshine)	1.37
0	王齐铭	1.37
0	AR	1.37
1	王嘉尔	6.25
1	Higher Brothers	4.08
1	天府事变CDREV	3.37
1	艾福杰尼	2.94
1	MULA SAKEE	2.56
1	班比 Bambii	2.45
1	马思唯	2.18
1	乃万	2.12
1	Turbo飙亮	2.12
1	新街口组合	2.01
2	GAI	3.66
2	宝石老舅	3.4
2	胡彦斌 TIGER HU	3.33

Discussion

- Even though we enhanced 3 incoherent clusters to 7 semantically meaningful clusters, the Silhouette score decreases as expected. Any way do avoid that?
- The top words for each cluster is still some very common words like: me, you, we etc. Even though I cleaned them up before performing downstream tasks, is it necessary to clean them up before converting them into embeddings?
- What other data augmentation methods for lyrics are recommended under this context?

Conclusion

- **Conclusion**: The Qwen2-1.5B model can only cluster our lyrics to 3 clusters which is somehow meaning less. By using contrastive learning, we get 7 more meaningful clusters comparing to the baseline. And by manually checking the performance of the downstream tasks, the clusters indeed show the meaningful result to some extent.
- **Future Work**
  - Find more ways to do data augmentation.
  - Consider labeling data manually
  - Consider larger input data (we use 60,000 lines, but we have 600,000+ in total!)
- **Final goal**: Fine-Tuning LLMs for Regional Styles and Custom Flows: a Chinese Rap Lyrics Generator

References & Acknowledgements

1. Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., & Tian, H. (2021). Qwen2-1.5B Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation.
2. Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates.
3. SimCSE Framework for Contrastive Learning. Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. Published in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Acknowledgements: Chris Endemenn, Elaine Wu, Finn Kuusisto, Lao Kai, Ryan Bemowski, GPUs, OpenAI GPT-4o