Beyond Prompt-Sensitive Emotion Words: Stable Embeddings for Tang Poetry Analysis

Linyue Zhang **Feiyue Li**

Abstract

Many Tang-poetry emotion studies still use coarse labels (e.g., positive/negative), while recent LLM-based attempts face a practical problem: one-word emotion outputs are highly sensitive to prompt wording. When labels shift with phrasing, historical interpretation becomes hard to reproduce and hard to trust. Focusing on Tang poetry around the An Lushan Rebellion (安史之乱), we propose a fine-grained sentence-level workflow centered on **emotion embeddings**: we use continuous hidden-state vectors, run automatic clustering, and then consolidate labels for interpretation. On the same 3,198 emotional sentences, one-word outputs show only 50.3% A/B exact agreement, while embedding-based clustering remains stable and well distributed (normalized entropy 0.989; 20/20 active clusters). On 7,195 labeled sentences, a char-based baseline reaches 0.446 micro-F1 and 0.395 macro-F1. This multi-stage label-construction path supports historically grounded findings, including the emotional turning point around 762, and also reveals layered patterns that are less visible in coarse setups. These results suggest that stable representation is a prerequisite for turning computational outputs into credible evidence for humanities interpretation.

1 Introduction

Large language models have made many digital-humanities tasks easier, but historical literary analysis still has a stubborn problem: fluent output is not the same as reliable evidence. This is especially true for Classical Chinese poetry, where emotion is compressed into short lines and tightly bound to historical context.

This study examines emotional change in Tang poetry around the An Lushan Rebellion. Existing workflows often rely on discrete emotion words generated by a model. In our corpus, semantically similar lines can receive different one-word labels after small prompt edits, which weakens any historical claim built on those labels. Compared with many earlier studies that stay at coarse polarity (positive/negative) or broad sentiment bins, this setting asks for finer emotional granularity and stronger reproducibility.

Our solution is to treat continuous representations as the primary evidence. Instead of clustering generated words, we build sentence-level emotion embeddings from model hidden states, use clustering to discover candidate emotional structure, and then consolidate labels for downstream prediction and interpretation. This keeps the computational pipeline reproducible while leaving room for close reading.

The main contributions are:

- We propose a fine-grained Tang-poetry emotion analysis framework that moves beyond coarse polarity and explicitly addresses prompt-level instability in LLM-generated emotion words through embedding-based representation.
- We show that the resulting quantitative patterns are historically grounded: the year-level and poet-level trends align with textual evidence and historical records, including the 762 turning point.
- We report humanities findings that are difficult to obtain with coarse labels alone, illustrating how digital methods and close reading can jointly reveal new structure in the corpus.

Two commitments shape the humanities side of this paper. First, we care about *interpretative validity*, not only model scores: if we claim a turning point, readers should be able to trace it to sentence-level evidence, label distributions, and specific lines. Second, we treat model-induced categories as working tools, not fixed truth. Our multi-stage label process follows that logic: data-driven

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 58–68

discovery, then philological consolidation, then an explicit interpretive layer. The experiments are organized in the same order: establish learnability, test representation stability, run split and mapping diagnostics, and then move to year/poet interpretation with direct count checks. We keep the main LLM fixed and perturb only the prompt templates, because the claim under test here is prompt sensitivity rather than base-model ranking; broader cross-model comparison is an important next step, but it is not necessary for the current argument.

We position this work at the intersection of literary NLP and emotion analysis (Bamman et al., 2014; Buechel and Hahn, 2017).

2 Related Work

Computational literary studies have shown that NLP can quantify stylistic and thematic patterns while still supporting interpretative reading. In literary NLP, modeling character and narrative signals at scale has demonstrated that probabilistic or discriminative methods can expose latent literary structure (Bamman et al., 2014).

Emotion analysis has also moved from coarse polarity to richer representations, including multi-dimensional and perspective-aware annotations (Buechel and Hahn, 2017). For Chinese Tang-poetry analysis, prior work includes hierarchical text classification for poetry corpora (Xiao, 2006), classical machine-learning classification on Tang-poem categories (Hu and Zhu, 2015), and emotion-term extraction in digital-humanities settings (Zhang et al., 2021). The Dalian emotion ontology lexicon remains a widely used lexical resource (Institute of Information Retrieval, Dalian University of Technology, 2007). For historical Chinese poetry, this motivates two design choices in our work: (1) preserve fine-grained labels instead of collapsing directly to broad sentiment; (2) explicitly separate computational representation from interpretative mapping.

For humanities-oriented NLP work, a score alone is not enough. Outputs should remain stable under small perturbations and should be usable as evidence in an argument about texts. Following this expectation, we evaluate the method on two axes: computational adequacy (performance and robustness) and interpretative adequacy (historical plausibility, traceability to textual evidence, and transparent category design). In short, our gap is threefold: prior Tang-poetry emotion analysis is often coarse-grained, LLM-based pipelines are still relatively few, and prompt-sensitive word outputs make historical interpretation unstable.

3 Task Definition and Data

3.1 Data Scope and Canonical Source

Our study addresses two linked questions: how to reduce prompt sensitivity at sentence level, and how to use stabilized signals to analyze year/poet emotion distributions in a historically meaningful way. We first build robust sentence representations, then aggregate to year/poet patterns in the 754–764 observation frame, with each poem processed as a sequence of sentence-level units with year and author metadata. The cleaned source corpus contains **616 poem records** by **47 poets**, yielding **9,965 sentence instances** before final manual alignment. The final annotated table used for modeling expands this to **10,122 rows**, because some concatenated lines were manually split during annotation so that sentence and label counts could be aligned at the line level.

The workflow has two stages: (1) sentence extraction and multi-label annotation; (2) embedding construction, clustering, and year/poet aggregation.

To avoid denominator drift, all sentence totals and label statistics use one finalized annotated table as the canonical source. For per-sentence label-count statistics, duplicate labels inside the same sentence are deduplicated while preserving first-occurrence order. All year-level ratios in the paper are normalized within each year, so the interpretation is about emotional composition rather than raw poem totals.

Label construction and interpretive mapping. The 18-label inventory is derived from, not independent of, the 20-cluster step. We first run embedding-based unsupervised clustering and, using the elbow criterion, select $k = 20$ as an initial partition. We then manually merge clusters with unclear semantic boundaries to form 18 annotation labels for supervised modeling. In short, $k = 20$ serves exploratory discovery, while 18 labels serve downstream prediction and interpretation. Primary modeling is performed on this finalized 18-label inventory. For historical interpretation, we additionally map the 18 labels to 7 macro-categories (Sorrow, Disgust, Joy, Anger, Fear, Praise, Confusion), following the category framework of the Dalian University of Technology emotion ontology lexi-

Metric	Value / Note
Cleaned source corpus	616 poem records; 47 poets; 9,965 sentence instances
Total sentence rows	**10,122**; canonical annotated file
Rows with non-empty labels	**7,195**; used in supervised comparison
Rows without labels	**2,927**; retained for descriptive statistics
Label scheme	**18 labels, 0–3 per sentence**; obtained by manual consolidation of 20 initial clusters
Inter-annotator agreement	**Two annotators**, 200-sentence sample; **0.970 / 0.796** (OA / Cohen's κ)
Clustering setup	Elbow-selected $k = 20$, $N = 3,198$ embedding clusters for initial label discovery

Table 1: Main dataset and analysis statistics.

con (Institute of Information Retrieval, Dalian University of Technology, 2007). This mapping is a derived analysis layer, not a replacement for the original annotation protocol. Accordingly, the 18-label inventory remains the canonical annotation space, while the 7-class view is used only to make macro trends easier to read.

4 Method

For readers from non-technical backgrounds, the method can be read as three steps: (1) represent each sentence with a stable embedding instead of a single generated word; (2) build fine-grained labels through "automatic grouping + manual consolidation"; (3) aggregate labels by year and poet, then check whether the patterns match historical records and close reading.

4.1 Label Construction and Prediction Tasks

Given embedding vectors $\mathbf{e}_i$, we first perform K-means for a range of k values and use the elbow curve of within-cluster sum of squares to choose $k = 20$ as the automatic initial partition. These 20 clusters are then manually inspected and consolidated where semantic boundaries are not sufficiently clear, producing the 18-label scheme used in the supervised task. Therefore, there is no one-to-one requirement between the 20 clusters and the 18 labels: the former are exploratory units, the latter are annotation labels.

18-label classification task. Each sentence can have multiple labels (up to three in our annotation). We therefore use a standard multi-label setup: each label is predicted independently, and errors are averaged across all 18 labels with binary cross-entropy.

We benchmark majority-label, TF-IDF word/char features with OvR Linear SVM, and TF-IDF char with OvR Logistic Regression.

Mapping to 7 macro-emotions. For interpretative and error-analysis views, we define a deterministic mapping $g : \mathcal{L}_{18} \to \mathcal{L}_7$. The target 7-category space follows the Dalian University of Technology emotion ontology framework (Institute of Information Retrieval, Dalian University of Technology, 2007). Operationally, this means: if any fine-grained label of a sentence belongs to a given macro category, that macro category is marked as present for the sentence. This mapped form is used for descriptive aggregation and interpretative statistics.

For the mapped 7-class classifier ablation (Table 5), we use a diagnostic single-label projection: take the first valid fine-grained label among {`label1,label2,label3`}, then apply g to obtain one 7-class target. This diagnostic setting does not replace the primary 18-label multi-label task.

4.2 Representation Stability and Aggregation

Our core innovation is to treat continuous hidden states as the primary semantic signal. Given sentence x_i, we apply a zero-shot prompt to obtain a one-word emotion output (auxiliary cue), while extracting the final-position hidden vector from the last transformer layer as embedding $\mathbf{e}_i \in \mathbb{R}^d$. We then cluster these vectors with K-means to discover candidate emotion structure before manual consolidation.

To quantify distributional coverage across clusters, we report normalized entropy:

$$H_{\text{norm}} = -\frac{1}{\log K} \sum_{j=1}^{K} p_j \log p_j, \qquad (1)$$

where p_j is cluster proportion.

Inter-annotator agreement. Inter-annotator agreement is measured with Cohen's κ: κ compares observed agreement against chance-level agreement, so it is stricter than raw accuracy.

Year/poet aggregation. For year t and category c, ratio is computed as:

$$p_{c,t} = \frac{n_{c,t}}{\sum_{c'} n_{c',t}}, \qquad (2)$$

and first-order change as

$$\Delta p_{c,t} = p_{c,t} - p_{c,t-1}. \tag{3}$$

These summaries support year-wise and poet-wise interpretation around the event window.

5 Experimental Design

Our empirical design serves three connected aims: to show that fine-grained sentence-level emotion analysis is feasible, to compare the stability of embeddings and direct word outputs, and to examine whether the resulting patterns can sustain historical interpretation.

5.1 Study Design

For clarity, we organize the analyses in five parts:

- **Predictive baseline** (Table 2): evaluate the 18-label task and compare basic feature/model choices.
- **Representation stability** (Table 3): compare one-word prompt outputs with continuous embeddings.
- **Split comparison** (Table 4): compare grouped split and random split.
- **Mapped-label view** (Table 5): inspect behavior in a simplified 7-class view.
- **Historical interpretation** (Tables 6–8): aggregate canonical annotated counts by year and poet, then validate with sentence-level cases.

5.2 Data Splits and Metrics

The main supervised analysis uses the 18-label multi-label setting on all rows with non-empty labels ($N = 7{,}195$). For the 7-class simplified view, each sentence is mapped from the first valid fine-grained label. For the stability comparison, we use a fixed subset of emotional sentences ($N = 3{,}198$) shared by prompt A/B comparison and embedding-cluster profiling ($k = 20$).

Main supervised comparison uses group split by full poem (80/20), repeated over 3 seeds. This reduces train-test overlap from same-poem style and diction. We report Micro-F1, Macro-F1, and exact-match subset accuracy for the 18-label task. For the mapped 7-class diagnostic, we report Macro-F1 and accuracy.

Model	Micro-F1	Macro-F1	Subset Acc.
Majority-label baseline	0.151 ± 0.005	0.015 ± 0.000	0.013 ± 0.002
Word TF-IDF (1–2) + OvR SVM	0.076 ± 0.003	0.022 ± 0.001	0.000 ± 0.000
Char TF-IDF (1–4) + OvR SVM	0.446 ± 0.003	0.395 ± 0.007	$\mathbf{0.256 \pm 0.005}$
Char TF-IDF (1–3) + OvR LR	$\mathbf{0.448 \pm 0.007}$	$\mathbf{0.402 \pm 0.012}$	0.214 ± 0.009

Table 2: 18-label multi-label results (group split by poem, 3 seeds).

Setting	Consistency	Norm. H	Support
Word output A	–	0.807	542 unique words
Word output B	–	0.764	431 unique words
A/B exact match	0.503	–	3,198 pairs
Embedding clustering ($k = 20$)	N/A (continuous)	**0.989**	**20**/**20** active clusters

Table 3: Prompt sensitivity of single-word outputs vs. embedding coverage.

For representation comparison, we use two deterministic prompt templates (A/B) on the same 3,198 emotional sentences with DeepSeek-V3.2 (T=0). The generated words are used only for stability comparison; historical interpretation is based on canonical labels and embedding-derived structure.

6 Quantitative Results

We report results in the same order as the argument of the paper. We first establish that the fine-grained task is learnable, then show why embedding-based representation is more stable than direct word outputs, and finally examine whether the recovered temporal/poet patterns are historically meaningful and humanistically informative.

6.1 Predictive Performance and Representation Stability

Character features substantially outperform word n-grams, consistent with Classical Chinese's weak explicit token boundaries in short lines. SVM is best on strict subset accuracy, while LR slightly leads on Micro/Macro-F1. These results provide a stable baseline for the later checks.

Representation stability (words vs. embeddings). Even under deterministic decoding, A/B mismatch is 0.497, indicating non-trivial prompt sensitivity for discrete one-word outputs. Embedding clusters remain fully populated (cluster size 102–257) with high normalized entropy, supporting better semantic coverage. This experiment isolates representation stability; downstream historical trend tables are computed from canonical annotated labels rather than generated words. In practical terms, Table 3 explains a core design decision of this paper: embeddings are used to discover

Setting	Micro-F1
Group split by poem	0.446 ± 0.003
Random sentence split	$\mathbf{0.466 \pm 0.002}$

Table 4: Leakage sanity check with Char TF-IDF (1–4) + OvR SVM.

Model (7-class mapped)	Macro-F1	Accuracy
Word TF-IDF (1–3), SVM (bal.)	0.105 ± 0.004	0.583 ± 0.031
Char TF-IDF (1–3), SVM (bal.)	0.448 ± 0.028	0.653 ± 0.025
Char TF-IDF (1–4), SVM (bal.)	$\mathbf{0.449 \pm 0.026}$	0.652 ± 0.024
Char TF-IDF (1–4), SVM (unbal.)	0.433 ± 0.023	$\mathbf{0.683 \pm 0.014}$

Table 5: Mapped 7-class single-label diagnostic ablation.

structure, while consolidated labels are used for interpretation.

Split robustness check. The random split inflates performance, confirming that grouped split is the safer protocol for main reporting.

7-class diagnostic. Balanced training improves macro fairness across rare classes, while unbalanced training favors overall accuracy. This 7-class view is only a supplementary lens and does not replace the main 18-label analysis. Visual summaries of Tables 6–8 are provided in Appendix **??**.

6.2 Temporal Trends and Poet-Level Patterns

All ratios in this subsection are computed from canonical annotated counts; no classifier predictions are used here.
Note: Melancholy = 忧伤; Grief = 悲愁; Longing = 思念; Heroic Aspiration = 壮思; Romantic Affection = 爱恋.

The temporal pattern is coherent rather than random: Sorrow remains dominant and reaches its highest level in 762, while Joy shows a short recovery in 760 and then declines. Anger rises locally around 757 but does not persist, and secondary labels show a similar structure: Heroic Aspiration is relatively stronger in early war years, then weakens, whereas Melancholy intensifies by 762. These shifts align with a transition from mobilizing rhetoric to sustained grief-oriented expression.

6.3 Error Analysis and Validation

In the mapped 7-class view (Table 5), Macro-F1 is 0.4464 and accuracy is 0.6441. Performance is strongest on high-support classes (especially Sorrow) and weaker on rare classes. Most confusions appear in lines with mixed affective cues, especially because the single-label projection compresses secondary emotions. This pattern reflects the simplification of the diagnostic setting and class imbalance, rather than a breakdown of the historical trend signals.

Year	Sorrow	Disgust	Joy	Anger	Fear	Praise	Confusion
754	0.6659	0.0230	0.1590	0.0760	0.0161	0.0288	0.0311
755	0.6269	0.0206	0.1624	0.1161	0.0154	0.0226	0.0360
756	0.6018	0.0222	0.1519	0.1264	0.0287	0.0320	0.0369
757	0.5664	0.0288	0.1616	0.1592	0.0168	0.0344	0.0328
758	0.6297	0.0167	0.1862	0.0900	0.0136	0.0303	0.0335
759	0.6463	0.0199	0.1347	0.1168	0.0223	0.0291	0.0310
760	0.6388	0.0136	0.2097	0.0641	0.0078	0.0408	0.0252
761	0.6276	0.0444	0.1616	0.0887	0.0158	0.0333	0.0285
762	0.7024	0.0312	0.1239	0.0722	0.0166	0.0244	0.0293
763	0.6653	0.0116	0.1274	0.1126	0.0189	0.0242	0.0400
764	0.6403	0.0292	0.1278	0.1111	0.0278	0.0389	0.0250

Table 6: Year-wise primary emotion ratios (754–764).

Year	Melan.	Grief	Longing	Heroic Asp.	Romantic Aff.
757	0.1256	0.1528	0.0696	0.0608	0.0104
760	0.1456	0.0971	0.1029	0.0447	0.0155
761	0.1474	0.1252	0.0856	0.0380	0.0063
762	0.1824	0.1307	0.0985	0.0361	0.0049
764	0.1417	0.1486	0.0750	0.0375	0.0083

Table 7: Selected-year secondary emotion ratios around key turning points.

Interpretative counts are directly validated in the canonical table: Heroic Aspiration in 757 appears 76 times; Melancholy in 762 appears 187 times; Romantic Affection in 762 appears 5 times. Representative lines include 时将白羽挥, 吞声不许哭, and 偏得浑家怜.

7 Humanities Interpretation and Discussion

This section turns quantitative patterns into literary-historical interpretation without claiming deterministic causality.

7.1 762 as a Literary Turning Point

The yearly trajectory is not a simple one-way decline. What we observe is a reweighting of emotional expression under prolonged historical pressure. The Sorrow surge in 762 aligns with intensified war-aftershock memory and social dislocation in poetic expression. To make the historical anchor explicit, our original Chinese manuscript cites *Zizhi Tongjian* with the key chronology statement (Sima, 1956):

> 四月玄宗崩……五月肃宗崩。

Xuanzong died in the fourth lunar month and Suzong in the fifth, concentrating dynastic loss and succession stress within a short span. This chronology offers concrete historical context for the emo-

Poet	Sent.	Melancholy	Grief	Longing	Heroic Aspiration
Du Fu	3738	0.1527	0.1709	0.0654	0.0441
Li Bai	1783	0.1063	0.1175	0.0595	0.0898
Liu Changqing	819	0.1783	0.0783	0.1123	0.0245
Cen Shen	747	0.1113	0.0799	0.0870	0.0656
Dugu Ji	276	0.1242	0.0752	0.1536	0.0523
Yuan Jie	272	0.0983	0.1838	0.0342	0.0085
Gao Shi	229	0.1018	0.1228	0.0632	0.0912
Qian Qi	200	0.1602	0.1068	0.1068	0.0534

Table 8: Top-8 poets by sentence count and selected secondary emotion ratios.

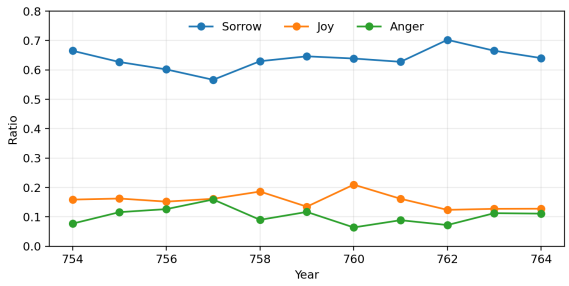

Figure 1: Year-wise trend of three core primary emotions (Sorrow, Joy, Anger).

tional downturn in our data and for the shift toward the restrained tone later described as 气骨顿衰. The table-level correspondence is direct. In Table 6, Sorrow peaks in 762 (0.7024), and Joy in 762 (0.1239) stays below its 760 peak (0.2097). In Table 7, Melancholy peaks in 762 (0.1824), while Romantic Affection reaches its minimum in 762 (0.0049). Our sentence-level count check further reports Melancholy $=$ 187 and Romantic Affection $=$ 5 in 762. At the same time, Joy and Anger do not disappear, which reminds us that crisis-era poetic emotion is mixed, not monolithic.

Our main turning-point reading is centered on **762**, a node that has been less explicitly foregrounded in prior An Lushan Rebellion emotion studies. At this point, Sorrow reaches its yearly maximum, Melancholy also peaks, and Romantic Affection falls to its minimum, producing the clearest structural shift in the corpus. By comparison, 757 is best read as the beginning of the decline in Heroic Aspiration rather than the full-system turning point: the category still appears relatively strong in the early war years, but then drops and does not recover to pre-war levels. This keeps the narrative consistent with both the annual counts and the original manuscript: 757 marks the fading of High-Tang confidence, while 762 marks the broader emotional restructuring toward grief-oriented and inward-looking expression.

Heroic Aspiration (壮思) is most visible around 755–757 and then declines, echoing the rise and fading of what Chinese literary history calls 盛唐气象. In Tang context, this category often overlaps with frontier rhetoric, loyalty discourse, and moral self-positioning. Its decline in later years, together with rising Melancholy, is consistent with a shift from mobilizing rhetoric to reflective and mournful expression.

From a literary-historical perspective, this pattern is compatible with the transition from expansive High-Tang voice to the more constrained and self-reflective tone associated with later mid-Tang poetics. Our contribution here is not to replace traditional periodization, but to provide a sentence-level quantitative trajectory that makes this transition empirically inspectable.

7.2 Poet-Level Variation and Relational Emotion Patterns

Poet-level heterogeneity is essential for humanities interpretation. Du Fu's high Melancholy+Grief profile supports readings of historically grounded poetic witness. Li Bai and Gao Shi retain relatively stronger Heroic Aspiration, matching traditions of expansive voice and frontier vigor. Thus, aggregated temporal change should be read together with poet-specific writing styles.

Another finding that echoes the original thesis concerns 思念 (Longing). In Table 8, Longing is relatively high for poets whose surviving lines in this window are less centered on direct battlefield narration (e.g., Dugu Ji and Liu Changqing). This does not imply that these poets were detached from historical crisis. Rather, it suggests that within the same war-era corpus, different social locations and communication contexts (farewell poems, correspondence, local circles) can sustain different emotional registers.

So the corpus is not emotionally homogeneous, even under severe macro-political turbulence. The coexistence of high Sorrow at year level and locally stronger Longing in some poet clusters supports a layered reading: "war pressure" structures the overall field, while regional life-worlds and poet networks still modulate the emotional surface of individual lines. This is precisely where computational aggregation and close reading become complementary. This type of cross-scale contrast is one place where the digital pipeline contributes genuinely new humanities value, because it links macro trend and micro poetic context within one auditable analysis chain.

7.3 Genre Boundaries and Methodological Implications

Romantic Affection remains low, with a clear minimum in 762. This does not imply absence of private emotion in Tang poetry; rather, within this wartime-centered corpus slice, public history, displacement, and political anxiety dominate the available space of expression. This reminds us that corpus framing matters when we make cultural claims.

For DH readers, the point is not that embeddings are "more advanced" than word outputs. The real issue is reliability of the evidence chain. If one-word outputs are prompt-unstable, downstream yearly distributions can mix true temporal signal with avoidable generation variance. By using embeddings for structure discovery and canonical labels for aggregation, we reduce this variance source and make claims easier to inspect and challenge. In that sense, representation choice directly affects humanities validity.

8 Limitations and Ethics

Interpretive Scope and Causal Boundary. We report structured correlations and trend alignments, not deterministic historical causality. Computational patterns should complement, not replace, close reading.

Data and Label Limitations. Supervised models are trained on pipeline-generated silver labels. A larger fully human-annotated benchmark is needed for stronger external validity and significance testing.

Category Construction Subjectivity. The 20-cluster discovery step is data-driven, but the merge to 18 labels includes human judgment. Another research team may make slightly different merge decisions at semantic boundaries.

Scope of Historical Generalization. Our observations are tied to the current corpus construction and year window (754–764). They should be interpreted as evidence for this textual slice, not as a full account of all Tang emotional history.

Reproducibility. All reported numbers are linked to executable outputs and logs. To support double-blind review, implementation details are not included in the main submission and can be released after acceptance.

Ethics. The study uses historical literary texts and focuses on scholarly analysis. No personal sensitive data is involved.

9 Conclusion

We analyzed emotional change in Tang poetry around the An Lushan Rebellion with a sentence-level, evidence-traceable pipeline. The paper makes three linked points. First, compared with coarse polarity-style analysis, a fine-grained embedding-centered framework better fits Tang-poetry emotion research and is more robust to LLM generation instability. Second, the resulting temporal and poet-level patterns are consistent with historical evidence, supporting the method's validity for humanities interpretation. Third, the framework surfaces layered findings (e.g., macro sorrow growth together with local longing stability) that are hard to observe in coarse setups. More broadly, the study shows a practical path for bringing digital methods into humanities work without flattening literary complexity: keep the evidence chain transparent, keep categories interpretable, and keep close reading at the center of final interpretation.

References

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Renfen Hu and Yuchen Zhu. 2015. Automatic topic classification for tang poems. *Journal of Peking University (Natural Science Edition)*, 51(2):262–268. [in Chinese].

Institute of Information Retrieval, Dalian University of Technology. 2007. Dalian university of technology emotion ontology lexicon. urlhttps://ir.dlut.edu.cn/zyxz/qgbtk.htm. [in Chinese].

Guang Sima. 1956. *Zizhi Tongjian*. Zhonghua Book Company, Beijing. [in Chinese].

Xue Xiao. 2006. Hierarchical classification of chinese text and its application to tang poetry classification. Master's thesis, Chongqing University. [in Chinese].

Wei Zhang, Hao Wang, Sanhong Deng, and Baolong Zhang. 2021. Emotion-term extraction and application for classical chinese poetry in digital humanities. *Journal of Library Science in China*, 47(4):113–131. [in Chinese].

Item	Detail
Corpus provenance	Project corpus derived from `anshi_poems.csv` and `anshi_annotated.xlsx`; yearly aggregation is computed from `anshi_analysis.xlsx`.
Coverage	616 poem records, 47 poets, 11 annual bins (754–764).
Sentence instances	9,965 sentence instances before final alignment; 10,122 annotated rows after manual line splitting.
Labels	18 fine-grained labels, 0–3 labels per sentence.
Annotation check	200-sentence sample by two annotators; OA 0.970, Cohen's $\kappa = 0.796$.

Table 9: Corpus provenance and annotation facts.

Label	Macro
羁旅	Sorrow
伤逝	Sorrow
偃蹇	Sorrow
悲愁	Sorrow
忧伤	Sorrow
思念	Sorrow
孤寂	Sorrow
讥讽	Disgust
喜悦	Joy
爱恋	Joy
壮思	Joy
淡泊	Joy
宴息	Joy
慷慨	Anger
愤恨	Anger
惊恐	Fear
赞美	Praise
迷茫	Confusion

Table 10: The 18 labels and their 7-class mapping.

A Data and Annotation Details

The 18-label inventory is the canonical annotation space; the 7-class view is only a derived reading layer for macro trends.

B Prompts

Prompt A (full text).
Source: `scripts/run_api_word_stability.py`

```
仅使用一个中文词语概括给定诗句的主要情
感。
不要解释，不要标点，不要多个词。
诗句：{line}
答案：
```

Prompt B (full text).
Source: `scripts/run_api_word_stability.py`

```
你是古典诗歌情绪标注助手。
请对下面诗句输出一个最贴切的中文情感词
（只许一个词）。
禁止解释、禁止句子、禁止多个词。
诗句：{line}
情感词：
```

18-label annotation prompt (full text).
Source: `scripts/annotate.py`

```
你是一位精通中国古典诗词的情感分析专家，
请为诗句标注 ** 最多 3 个 ** 最贴切的 ** 情
感标签 **。
# 可选的情感标签（共 18 个，必须严格使用
以下名称）：
羁旅、伤逝、偃蹇、悲愁、忧伤、思念、孤
寂、讥讽、喜悦、
爱恋、壮思、淡泊、宴息、慷慨、愤恨、惊
恐、赞美、迷茫。
# 标签定义（请严格遵守）：
- 羁旅：羁旅漂泊之苦（远行、风霜、路途艰
辛）。例：“归雁来时数附书”。
- 伤逝：年华老去的悲伤，涉及黑发、白发
等。例：“不堪玄鬓影”。
- 偃蹇：怀才不遇，壮志难酬。例：“无因见明
主”。
- 悲愁：深重人生苦难，如亲友去世、战乱、
贫穷、疾病。
例：“弟兄无一人”。
- 忧伤：内心郁结，忧思难解。例：“暮天摇落
伤怀抱”。
- 思念：思念恋人、故人（须在世）或故乡，
且归期无望。
例：“别后相思复何益”。
- 孤寂：寂寞孤单，天地独对。例：“已忍伶俜
十年事”。
- 讥讽：对人或现象的讽刺、鄙夷、厌恶。
例：“轻薄为文哂未休”。
- 喜悦：内心欢喜，轻松明朗。例：“却喜晒谷
天晴”。
- 爱恋：夫妻或恋人之间的美好感情。例：“或
恐是同乡”。
- 壮思：豪迈洒脱，人生得意。例：“冲天香阵
透长安”。
- 淡泊：超然物外，不慕名利（属积极心境）。
例：“悠然见南山”。
- 宴息：聚会饮酒、朋友欢聚之乐。例：“莫使
金樽空对月”。
- 慷慨：悲壮激昂，忧国忧民。例：“慨然抚长
剑”。
- 愤恨：强烈愤怒，痛斥不公。例：“朱门酒肉
臭”。
- 惊恐：对危险、变故的害怕与焦虑。例：“恐
惊平昔颜”。
- 赞美：对亲情、友情、爱情等真挚情谊的歌
颂。例：“天涯若比邻”。
- 迷茫：前路不明，人生困惑（重点在“不知
所措”）。例：“更欲东奔何处所”。
# 重要规则：
1. 只输出属于上述 18 个标签的内容，不得自
创。
2. 若诗句为纯写景、纯记事、语义不清 -> 返
回空列表 []。
3. 但若诗句隐含可合理推断的情绪（如“独
坐”“寒雨”“孤舟”），
请优先选择最贴近的 1 个标签，而非返回空
列表。
4. 最多标注 3 个标签，按相关性排序。
5. “泪”“愁”等字出现!= 自动打标签，需看整
体语境。
6. “独坐”“独钓”：超然 -> 淡泊；孤独无助 ->
孤寂。
# 输出要求：
- 仅输出合法 JSON，格式：{” 情感标签”:[...]}
- 不要任何解释、注释、markdown 或额外字
段。
## 示例（Few-shot）：
```

诗句：” 白日依山尽”
{” 情感标签”:[]}
诗句：” 乡音无改鬓毛衰”
{” 情感标签”:[” 思念”,” 伤逝”]}
诗句：” 莫使金樽空对月”
{” 情感标签”:[” 宴息”]}
诗句：” 朱门酒肉臭，路有冻死骨”
{” 情感标签”:[” 愤恨”,” 悲愁”]}
诗句：” 独坐幽篁里，弹琴复长啸”
{” 情感标签”:[” 淡泊”]}
请标注以下诗句：
诗句：”{poem_line}”
输出（仅 JSON）：

Line-selection prompt (full text).

Source: `scripts/emopoem_get.py`

根据以下诗歌内容，分别选出这首诗中直接表达情感的诗歌单句。
诗歌内容：
{content}
请完成：
找出所有情感倾向明显外露的诗句，每次严格输出一个诗歌单句。
以逗号/句号/问号/感叹号为分界为一个诗歌单句。
不可以输出如“不睹皇居壮, 安知天子尊。”这样包含两句的诗。
请严格使用以下 JSON 格式输出（支持数组），不能在这一格式以外输出任何其他内容：
[
{”poem”:” 诗句 1”},
{”poem”:” 诗句 2”}
]

	129		125		127		153
0	得成比目何辞死	**1**	暴骨无全躯	**2**	金吾不禁夜	**3**	此地别燕丹
	153		182		146		242
4	四海遂为家	**5**	怅然临古城	**6**	玉漏莫相催	**7**	年年属数奇
	155		126		102		199
8	长作巢由也不辞	**9**	惟闻素棘与黄泉	**10**	还惊九逝魂	**11**	行来行去尽哀怜
	120		137		186		185
12	深怀国士恩	**13**	辛苦事旌麾	**14**	相逢秋月满	**15**	慷慨志犹存
	257		132		219		123
16	独有西山将	**17**	既伤千里目	**18**	生憎帐额绣孤鸾	**19**	徙倚欲何依

Table 11: Compact inventory of the 20 embedding clusters (size and one representative line each).

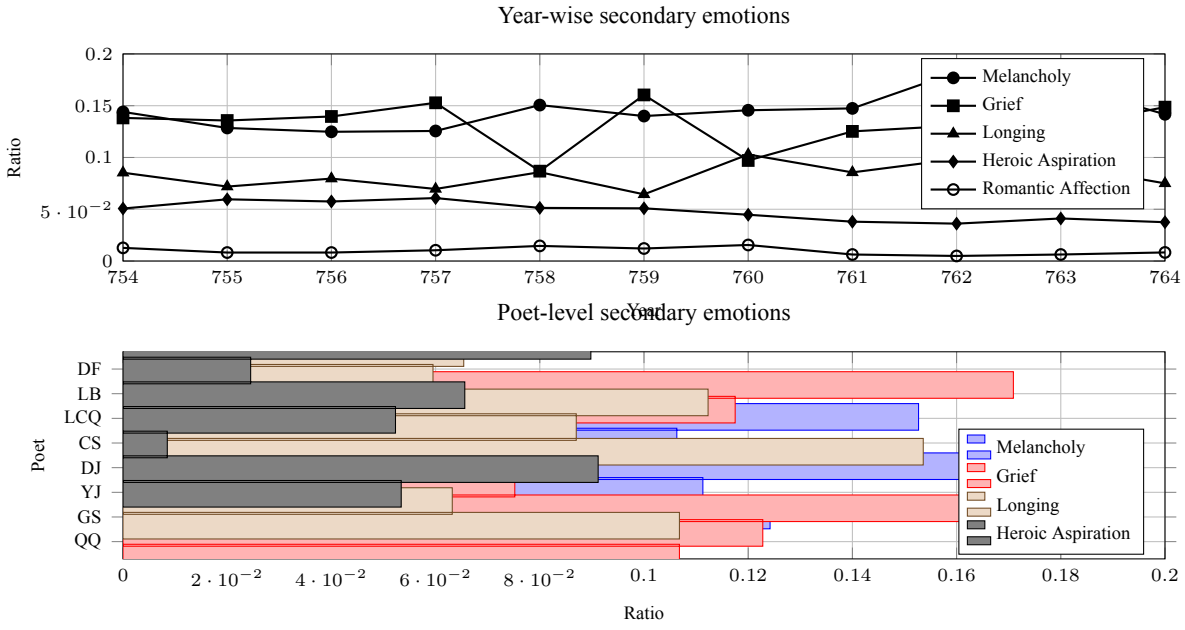

Figure 2: Visual summaries for the secondary-emotion tables in the main text.