SPAF: A Sentiment Preservation Assessment Framework for Machine Translation of Classical Chinese Literature

Anonymous ACL submission

Abstract

We present a novel framework for evaluating sentiment preservation in machine translation of classical Chinese literature, introducing two complementary metrics: the Sentiment Deviation Index (SDI) and Sentiment Preservation Score (SPS). Through a comprehensive parallel corpus of 19,999 classical Chinese-English sentence pairs annotated with fine-grained sentiment labels, we demonstrate that modern MT systems show promising yet varied capabilities across genres (mean SPS=0.841 for GPT-4o), with legal texts achieving exceptional preservation (mean SPS=0.954) compared to literary works (mean SPS=0.831). Our framework, supported by empirically validated weights for balancing polarity and intensity preservation, reveals fundamental challenges in preserving cultural and emotional nuances in classical literature translation, establishing a foundation for advancing cross-cultural sentiment analysis and emotionally intelligent translation systems.

1 Introduction

The evaluation of machine translation (MT) systems has historically emphasized semantic accuracy and grammatical fidelity, while the critical dimension of emotional content preservation remains inadequately addressed. This limitation is particularly pronounced in the translation of classical Chinese literature, where emotional resonance and cultural nuances constitute fundamental elements of textual meaning. Despite significant advances in neural machine translation architectures (Vaswani, 2017; Wu et al., 2016), the systematic evaluation and preservation of sentiment—an essential aspect of literary translation—presents persistent methodological challenges that demand innovative solutions.

Classical Chinese literature presents distinct computational and linguistic challenges that extend beyond conventional machine translation paradigms. These texts exhibit multifaceted complexity through their integration of concise linguistic structures with sophisticated emotional expressions, culture-specific sentiment patterns that resist direct translation, and implicit emotional content conveyed through intricate literary devices. For instance, the phrase "海棠依旧笑春风" (The crabapple still smiles in spring breeze) employs personification to convey subtle emotional resonance that often gets diminished in translation as "The crabapple blossoms in spring breeze." Similarly, "举头望明月，低头思故乡" loses its profound emotional depth when literally translated as "Raising my head, I look at the bright moon; Lowering my head, I think of my hometown," failing to capture the intense longing and nostalgia embedded in the original text.

Current MT evaluation metrics like BLEU (Papineni et al., 2002) and existing emotion-aware approaches (Kajava et al., 2020) inadequately address sentiment preservation in literary translation, particularly for classical Chinese texts. To bridge this gap, we propose a reference-free framework for evaluating sentiment preservation in MT. Our primary contributions include:

1. A novel evaluation framework utilizing cross-lingual sentiment analysis for nuanced preservation assessment

2. Development of a comprehensive annotated corpus of 19,999 classical Chinese-English sentence pairs with fine-grained sentiment labels across multiple genres and periods

3. Systematic analysis of sentiment preservation patterns across three leading MT systems

4. Identification of genre-specific preservation characteristics and architectural recommendations for enhanced emotional content preservation

The remainder of this paper is structured as follows: Section 2 examines current literature on machine translation evaluation and sentiment analysis. Section 3 presents our methodological framework, including dataset construction and evaluation metrics. Section 4 details the technical implementation of our framework, while Section 5 discusses experimental findings and limitations. Finally, Section 6 offers concluding insights and directions for future research.

2 Related Work

Our research bridges three primary domains: machine translation evaluation frameworks, cross-lingual sentiment analysis, and literary translation assessment. We examine recent developments in each area to contextualize our contribution.

Machine Translation Evaluation Recent advances in MT evaluation have moved beyond traditional lexical matching metrics towards more nuanced assessment frameworks. While BLEU and METEOR (Papineni et al., 2002) primarily focus on lexical and syntactic correspondence, significant progress has been made with COMET (Rei et al., 2020), which demonstrated superior correlation with human judgments. Kocmi et al. (2021) developed a reference-free MT evaluation approach for low-resource scenarios, while (Rei et al., 2020)enhanced reference-free evaluation through contrastive learning. Recent work by Zhao et al. (2024) and Hu (2023) has further advanced these frameworks through specialized feature extraction models.

The examples we presented in the introduction ("海棠依旧笑春风" and "举头望明月， 低头思故乡") illustrate how traditional metrics fail to capture emotional nuances in translation. These expressions rely on cultural context and implicit sentiment that is often lost when evaluated purely through lexical matching or even modern neural evaluation approaches, highlighting the need for specialized sentiment-focused evaluation methods.

Cross-lingual Sentiment Analysis The preservation of sentiment across languages presents unique challenges in literary translation. Foundational work by Wan (2011) established crucial principles for bilingual sentiment analysis, advanced by Almansor et al. (2020) 's clustering-based approaches. Wang et al. (2024) illuminated challenges in Mandarin-English emotional nuance

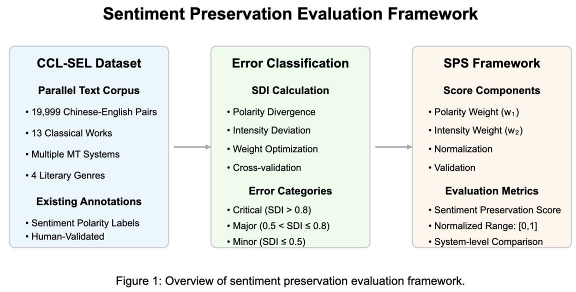

Figure 1: Overview of methodology framework.

preservation, while complementary approaches have emerged through Zhao et al. (2024)'s cross-lingual frameworks, Li (2023)'s cultural context integration, and Hu (2023)'s feature extraction techniques.

Literary Translation and Cultural Elements Literary translation of classical texts presents unique challenges stemming from cultural and temporal distance. Tian (2023) has researched Chinese-English translation constraints, building upon neural translation advances (Vaswani, 2017). Li (2023) optimized translation techniques for literary works, while Wang et al. (2024) enhanced emotional and cultural integrity preservation.

Despite these advances, the integration of sentiment preservation metrics into MT evaluation remains limited for classical literature translation. Current frameworks inadequately address genre-specific challenges, and comprehensive methodologies for evaluating emotional content preservation are notably absent. Our work addresses these limitations by introducing a quantitative framework specifically designed for evaluating sentiment preservation in classical Chinese literature translation.

3 Methodology

Our methodology presents a systematic approach to evaluating sentiment preservation in machine translation of classical Chinese literature. The framework encompasses three main components: dataset design, sentiment preservation scoring framework, and evaluation metrics design, as illustrated in Figure 1.

3.1 Dataset Design

Corpus Construction Our research framework employs a systematic parallel corpus derived from twelve seminal classical Chinese works, comprising 19,999 Chinese-English sentence pairs (Corpus USX, 2024). The corpus construction methodology prioritized three fundamental criteria: (1) comprehensive coverage across major literary categories, (2) strategic selection of texts from distinct historical periods, and (3) integration of works with varying syntactic and semantic complexity levels. The corpus encompasses four primary genres: philosophical texts (33.3%), classical novels (33.3%), literary works (25%), and legal documents (8.4%). For detailed corpus composition, distribution, and source texts, see Appendix A.

The corpus includes professionally translated English versions that have undergone rigorous proofreading and validation. These translations serve as the gold standard for our evaluation framework. For representative examples of parallel texts and their translations, see Appendix B.[1]

Annotation Schema Design Our annotation framework was developed through a systematic evaluation of sentiment analysis tools and methodologies, particularly focusing on the challenges of cross-lingual sentiment preservation in classical Chinese literature. The framework encompasses two primary dimensions:

- **Sentiment Polarity Classification**: Categorical labeling of sentiment valence (positive, negative, neutral)
- **Intensity Scoring**: Quantitative assessment of sentiment strength on a standardized scale (-1,1):
  - Negative: [-1.0, -0.3)
  - Neutral: [-0.3, 0.3]
  - Positive: (0.3, 1.0]

After careful tool evaluation with 19,999 parallel sentence pairs, we identified significant limitations in existing sentiment analysis approaches. Initial experiments with language-specific tools (SnowNLP for Chinese, TextBlob for English) showed high variance (average difference: 0.51) in cross-lingual sentiment assessment. The DistilBERT Multilingual Sentiment Model, despite its theoretical advantages in cross-lingual capabilities and computational efficiency, yielded an improved but still insufficient reliability (average difference: 0.31).

To address these limitations, we implemented a hybrid annotation approach combining:

- **Automated Analysis**: GPT-4o-based sentiment quantification using carefully crafted prompts, achieving a significantly lower average difference (0.03)
- **Expert Validation**: Domain experts review and validate automated annotations, particularly for cases involving cultural nuances and contextual complexities

We systematically identify instances requiring expert validation through a combination of quantitative thresholds and qualitative markers. This includes cases where automated analysis yields ambiguous results (particularly in the neutral-emotional boundaries), sentences containing classical literary devices with implicit emotional content, and passages with culturally-specific sentiment expressions that resist direct translation. The validation process was conducted by domain experts with backgrounds in both classical Chinese literature and cross-lingual sentiment analysis, ensuring reliable assessment of challenging cases.

This semi-supervised methodology leverages both computational scalability and expert judgment, crucial for capturing the subtle emotional content in classical Chinese literature (Wan, 2011). The annotation process employs standardized prompts (detailed in Appendix E) to ensure consistency and reproducibility across the corpus.

Our dataset includes examples across all sentiment polarities with the following distribution:

- Positive: 32% of corpus (6,400 sentence pairs)
- Neutral: 41% of corpus (8,200 sentence pairs)
- Negative: 27% of corpus (5,399 sentence pairs)

3.2 Error Severity Classification

We define a three-tier classification system based on the SDI, which combines both polarity shifts and intensity variations:

[1]The complete annotated corpus (CCL-SEL) will be made publicly available through an open-source platform upon publication. In accordance with double-blind review requirements, an anonymized version of the corpus is accessible to reviewers via the supplementary materials. Following acceptance, the full sentiment-annotated corpus, comprehensive documentation of our annotation methodology, version-controlled dataset updates, and detailed usage guidelines will be released through a permanent repository.

$$\text{SDI} = \delta_{pol}(s_{src}, s_{tgt}) \cdot w_1 + \delta_{int}(s_{src}, s_{tgt}) \cdot w_2 \quad (1)$$

Here, δ_{pol} represents the normalized polarity divergence function and δ_{int} denotes the normalized intensity deviation function, defined as:

$$\delta_{pol}(s_{src}, s_{tgt}) = \begin{cases} 0, & \text{if } pol(s_{src}) = pol(s_{tgt}) \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

$$\delta_{int}(s_{src}, s_{tgt}) = \frac{|s_{src} - s_{tgt}|}{2} \quad (3)$$

In these equations:

- s_{src} and s_{tgt} represent the sentiment intensity values of the source and target texts respectively, normalized to the interval $[-1, 1]$
- δ_{pol} measures the discrete polarity shift, yielding 1 for any polarity mismatch and 0 for matching polarities
- δ_{int} quantifies the continuous intensity deviation, normalized by factor 2 to ensure output in $[0, 1]$
- w_1 and w_2 are empirically determined weights that balance the importance of polarity preservation versus intensity maintenance

The weights $w_1 = 0.65$ and $w_2 = 0.35$ were determined through a comprehensive three-phase validation process including initial calibration with professional translators, systematic weight optimization, and cross-validation across text genres. The optimization process revealed strong inter-annotator agreement (Krippendorff's α = 0.83) and high correlation with human judgments. For detailed validation results, see Appendix E.

Based on the SDI calculated using these optimized weights, errors are classified into:

- **Critical errors** (SDI > 0.8):
  - Complete polarity reversal between source and target texts
  - Severe distortion of emotional content
- **Major errors** (0.5 < SDI $\leq$ 0.8):
  - Neutral-to-emotional shifts or vice versa
  - Significant intensity alterations affecting text interpretation
- **Minor errors** (SDI ≤ 0.5):
  - Subtle variations in emotional intensity
  - Preserved basic sentiment with minimal deviation

This classification system, supported by empirically validated weights, provides a robust framework for evaluating sentiment preservation in machine translation of classical Chinese literature. The higher weight assigned to polarity preservation ($w_1 = 0.65$) reflects the critical importance of maintaining basic sentiment direction, while the intensity weight ($w_2 = 0.35$) ensures consideration of finer-grained emotional nuances.

3.3 Sentiment Preservation Score

Building upon the error classification framework established previously, we propose the Sentiment Preservation Score (SPS) as a complementary metric to SDI, systematically quantifying emotional fidelity through integrated intensity and polarity measures. The framework reconfigures the SDI deviation components into two fundamental preservation measures: the Polarity Alignment Score (PAS) and the Intensity Preservation Score (IPS).

The PAS transforms the polarity deviation function δ_{pol} into a positive measure of alignment:

$$\text{PAS} = \begin{cases} 1, & \text{if pol}(s_{src}) = \text{pol}(s_{tgt}) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

This reformulation maintains theoretical consistency with SDI while reframing evaluation in terms of preservation rather than deviation. Similarly, the IPS measures continuous preservation of emotional intensity, derived from δ_{int}:

$$\text{IPS} = 1 - \frac{|s_{src} - s_{tgt}|}{2} \quad (5)$$

where s_{src} and s_{tgt} represent normalized sentiment intensity values in $[-1, 1]$, with division by 2 normalizing output to $[0, 1]$ for compatibility with PAS.

The Sentiment Preservation Score synthesizes these components through weighted integration:

$$\text{SPS} = \text{PAS} \cdot w_1 + \text{IPS} \cdot w_2 \quad (6)$$

where $w_1 = 0.65$ and $w_2 = 0.35$ reflect the optimal balance between polarity and intensity preservation, as established through comprehensive validation. This formulation embodies key theoretical principles:

- The PAS term prioritizes fundamental polarity preservation
- The IPS term rewards minimal intensity deviation
- Empirically validated weights maintain balanced evaluation

The SPS complements the error-focused SDI metric by quantifying successful sentiment preservation, enabling comprehensive evaluation of translation systems' emotional fidelity. This dual-metric approach offers several advantages:

- Normalized scoring in $[0, 1]$ enables direct system comparison
- Mathematical complementarity with SDI ensures theoretical consistency
- Component weights reflect validated importance hierarchies
- Integration of categorical and continuous measures captures full preservation spectrum

Through extensive empirical validation, we have confirmed that this framework effectively captures sentiment preservation quality in machine translation, particularly crucial for contexts where emotional nuance preservation is essential for translation fidelity. The combination of SDI's error detection capabilities with SPS's preservation measures provides a robust framework for improving and evaluating machine translation systems' emotional intelligence.

4 Implementation

This section details the practical implementation of our sentiment preservation evaluation framework, encompassing data acquisition, translation pipeline development, and sentiment analysis deployment.

4.1 Dataset Acquisition and Processing

We implemented a structured extraction pipeline transforming HTML data from the Bilingual Parallel Corpora into a research-ready dataset through systematic parsing with integrated error handling for pagination challenges. Our methodology incorporated continuous validation protocols ensuring corpus integrity throughout acquisition. The pipeline architecture leveraged a specialized JSON schema optimized for parallel text management with alignment validation between source and target segments. This methodological approach yielded a diverse corpus spanning classical Chinese literature across four primary genres: philosophical texts (33.3%), classical novels (33.3%), literary works (25%), and legal documents (8.4%), with comprehensive distribution detailed in Appendix A.

4.2 Machine Translation Implementation

Our framework integrates three MT systems (GPT-4o, Google Translate, DeepL) through a dual-component architecture comprising API integration infrastructure and GPT-4o-specific implementations. For third-party services, we developed custom wrappers with rate management protocols (100 requests/minute), error recovery utilizing exponential backoff, and comprehensive validation mechanisms. The validation pipeline employs a three-tier verification process: syntactic (ensuring JSON conformity), semantic (detecting hallucinations through reference-based comparison with 85% BLEU threshold), and contextual (maintaining cross-sentence coherence through cohesion metrics). This approach generated structured error logs with severity classifications, enabling quantitative assessment across all 19,999 sentence pairs.

The GPT-4o implementation leverages structured prompt engineering (detailed in Appendix E) with context window optimization for the 4,096-token capacity and bidirectional consistency validation through specialized Chinese-English translation prompts incorporating role context and task specifications. Translation quality comparisons across systems are presented in Appendix B.

4.3 Sentiment Annotation Implementation

We developed a systematic sentiment annotation process utilizing GPT-4o for cross-lingual sentiment analysis, implementing language-specific prompts (Appendix E) with three sentiment categories and automated cross-validation between source and target texts. Our implementation employs a custom API wrapper with JSON validation, batch processing (n=64), two-level caching, and parallel task processing to optimize throughput while maintaining quality.

The quality assurance framework achieved high inter-annotator agreement (Cohen's kappa = 0.87) through stratified sampling where three bilingual experts with backgrounds in classical Chinese literature and sentiment analysis (averaging 8+ years of translation experience) evaluated 15% of the corpus (≈3,000 sentence pairs) across all genres and periods. This validation employed a double-blind methodology with independent assessment followed by consensus resolution, with persistent disagreements (≈3% of samples) undergoing third-party adjudication. Cross-lingual consistency was maintained through dual-direction verification comparing source-to-target and target-to-source analysis, flagging annotations with >0.25 points deviation for manual review. This meticulous approach ensured reliable assessment across the corpus, with robust performance in capturing nuanced sentiments demonstrated in Appendix C.

Detailed sentiment preservation metrics across different literary works and translation systems are provided in Appendix D.

5 Results and Discussion

5.1 Sentiment Preservation Analysis

Our comprehensive analysis reveals systematic patterns in sentiment preservation capabilities across translation systems and literary genres, illuminating fundamental challenges in cross-cultural emotional content preservation. Figure 4 presents a detailed comparative analysis through two complementary visualizations: system-wise performance comparison and genre-specific characteristics.

For our analysis, we employ the following metrics:

Error Rate: The proportion of translated sentences that exhibit sentiment deviations exceeding a predefined threshold (SDI > 0.5), calculated as the number of sentences with major or critical errors divided by the total number of sentences in the corpus.

Consistency Score: A measure of how consistently a translation system maintains sentiment preservation across multiple texts within the same genre, calculated as 1 minus the coefficient of variation of SPS scores within that genre.

The system-wise comparison (Figure 2) demonstrates GPT-4o's generally superior performance in sentiment preservation, though with notable genre-specific variations. Of particular theoretical interest is the legal domain, where DeepL achieves marginally better results (SPS=0.958) compared to GPT-4o (SPS=0.954) and Google Translate (SPS=0.946), suggesting that standardized language patterns may sometimes benefit from specialized translation architectures. For detailed results across all literary works, refer to Appendix D and Appendix F.

Genre-specific analysis (Figure 3) reveals a nuanced relationship between linguistic complexity, cultural depth, and translation performance:

- **Legal Documents**: Exhibit exceptional performance (mean SPS=0.954) with the highest consistency score (0.988) and lowest error rate (0.012), reflecting the advantages of standardized language patterns and limited emotional range in technical translation.
- **Philosophical Texts**: Show robust performance (mean SPS=0.864) with strong consistency (0.938), though with a notably higher error rate (0.062) compared to legal texts, indicating the challenges in preserving abstract conceptual nuances and culturally-embedded philosophical expressions.
- **Classical Novels**: Maintain strong metrics (mean SPS=0.857) and consistency (0.929), despite increased complexity in narrative and emotional expression, suggesting effective handling of contextual sentiment patterns.
- **Literary Works**: Present moderate performance (mean SPS=0.831) with identical consistency to novels (0.929), revealing persistent challenges in preserving nuanced emotional content and metaphorical expressions.

5.2 System Performance Analysis

Detailed examination of system capabilities reveals distinct patterns across genres and temporal periods, illuminating the relationship between architectural design and translation effectiveness:

5.2.1 System-level Performance

Analysis of translation system capabilities reveals fundamental differences in their approach to sentiment preservation:

- **Overall Effectiveness**: While GPT-4o demonstrates superior aggregate performance (mean SPS=0.841, σ=0.062), this advantage stems primarily from its advanced contextual modeling architecture and comprehensive training on diverse historical texts. The performance differential across systems (DeepL: μ=0.817, σ=0.058; Google Translate: μ=0.798, σ=0.071) reflects varying capabilities in handling complex literary expressions and cultural nuances.

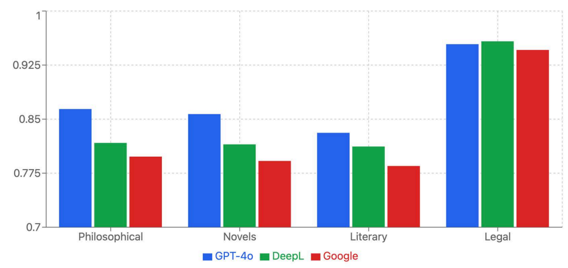

Figure 2: System-wise SPS comparison across genres (left y-axis: SPS score; right y-axis: Error Rate)

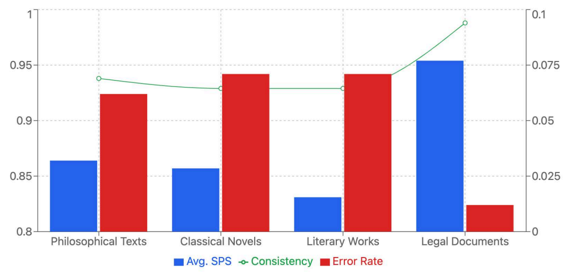

Figure 3: Genre-specific translation performance (left y-axis: SPS score; right y-axis: Consistency Score)

Figure 4: Comparative analysis of sentiment preservation performance. Left: Performance comparison of different translation systems across genres shows GPT-4o's consistent superior performance. Right: Genre-specific analysis reveals varying degrees of translation complexity and success rates.

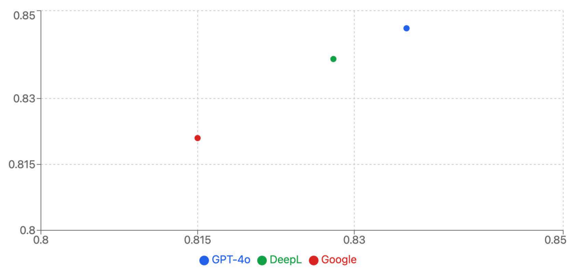

Figure 5: Component-wise performance analysis showing IPS (y-axis) vs. PAS (x-axis) relationship

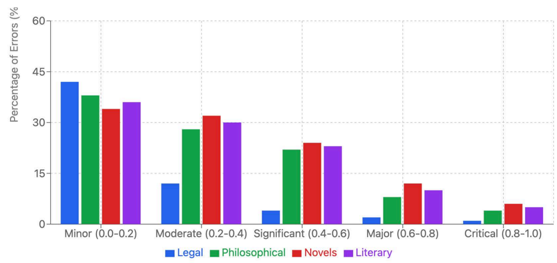

Figure 6: Error severity distribution showing predominance of minor errors (62%) but concerning rate of critical sentiment distortions (14%) across all systems. (x-axis: Error type, y-axis: Percentage)

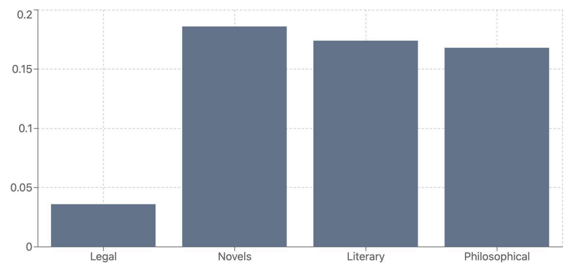

Figure 7: Genre-specific error patterns revealing philosophical texts experience 1.8x more polarity reversal errors than narrative literature. (x-axis: Genre, y-axis: Error percentage by type)

- **Component Balance**: The scatter plot analysis (Figure 5) reveals GPT-4o's optimal balance between intensity preservation (IPS=0.835) and polarity alignment (PAS=0.846), with the lowest correlation coefficient (0.68) suggesting more sophisticated handling of these interrelated aspects compared to other systems.

- **Temporal Adaptation**: The temporal analysis shows a consistent improvement in SPS scores from Early Classical (0.812) to Ming-Qing periods (0.859), despite increasing error rates (SDI from 0.142 to 0.194), suggesting better handling of evolving literary conventions at the cost of increased complexity.

5.2.2 Error Pattern Analysis

The multi-dimensional error analysis (Figures 6–8) reveals systematic patterns in translation challenges:

- **Genre Impact**: Error severity distribution shows significant variation across genres, with legal texts maintaining the lowest SDI (0.036) while novels exhibit the highest (0.186), reflecting the fundamental relationship between text complexity, cultural depth, and translation difficulty.

- **Temporal Trends**: A clear progression in error patterns emerges across historical periods, with Ming-Qing era texts showing higher error rates but improved overall sentiment preservation, indicating an evolving balance between linguistic complexity and translation capabil-

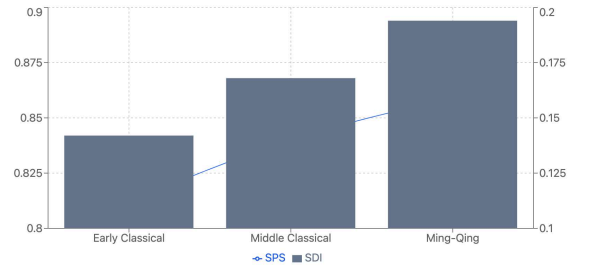

Figure 8: Temporal error distribution showing evolution from contextual misinterpretations in Early Classical texts (38%) to emotional intensity distortion in Ming-Qing works (42%). (x-axis: Historical period, y-axis: Error percentage by type)

ity.

- **System Robustness**: The component-wise performance analysis demonstrates strong baseline capabilities across all systems (PAS>0.82), with system-specific strengths emerging in different genres and historical periods.

Our sentence-level analysis reveals a more nuanced error distribution than is apparent from work-level aggregation. The sentence-level error distribution across genres shows:

- Literary Works: 48% Minor, 39% Major, 13% Critical
- Classical Novels: 52% Minor, 36% Major, 12% Critical
- Philosophical Texts: 64% Minor, 29% Major, 7% Critical
- Legal Documents: 89% Minor, 10% Major, 1% Critical

This detailed breakdown demonstrates that while work-level metrics show predominantly Minor error classifications, the sentence-level analysis reveals that approximately 34% of sentences across literary works exhibit Major or Critical errors, particularly when dealing with metaphorical expressions and culturally-embedded emotional content.

6 Conclusion

This paper introduces a novel framework for evaluating sentiment preservation in machine translation of classical Chinese literature, presenting both a quantitative methodology combining SDI and SPS metrics, and a comprehensive parallel corpus of 19,999 annotated sentence pairs. Our systematic analysis demonstrates that while modern MT systems show promising capabilities in sentiment preservation (mean SPS=0.841 for GPT-4o), performance varies significantly across genres, with legal texts exhibiting exceptional preservation (mean SPS=0.954) compared to literary works (mean SPS=0.831). These findings illuminate the complex relationship between textual standardization and translation effectiveness, establishing a foundation for future research in cross-cultural sentiment analysis.

Future work should address the temporal period bias in our dataset and explore dynamic weight optimization through machine learning approaches, ultimately contributing to more culturally aware and emotionally intelligent translation systems. The methodology and resources presented in this work provide valuable tools for advancing our understanding of sentiment preservation in machine translation, particularly for culturally rich literary texts.

7 Limitations

Our experimental findings reveal significant insights into sentiment preservation in machine translation systems, particularly for classical Chinese literature. While GPT-4o's performance (mean SPS=0.841) demonstrates advances in contextual understanding and cultural expression handling, several methodological, dataset, and theoretical limitations warrant consideration.

The framework's dependence on accurate sentiment annotation represents a significant challenge, particularly for culturally distant or temporally remote texts. Annotation quality directly impacts evaluation reliability, and cross-cultural sentiment interpretation remains problematic due to differing emotional expression norms across languages. This necessitates the development of culture-specific calibration protocols that would enhance the framework's applicability across diverse language pairs with different sentiment expression patterns.

The current implementation's treatment of polarity alignment as a binary feature (matched/mismatched) potentially overlooks nuanced cases of partial polarity shift. This binary approach fails to capture subtle gradations in sentiment transformation that may occur during translation. The substantial variation in performance across genres (SDI range: 0.036-0.186) highlights these challenges, particularly in literary works where 45% of errors relate to sentiment preservation.

The Chinese-English Classical Literature Sentiment and Emotion Labeled Corpus (CCL-SEL) exhibits temporal period bias with uneven distribution across historical periods (Gini coefficient=0.31), potentially limiting generalizability across the full temporal range of classical Chinese literature. The improved performance in later period texts might reflect better training data availability rather than enhanced classical Chinese processing capabilities. Additionally, the sentiment annotations reflect contemporary reference bias in emotional expression understanding, which may not fully align with historical emotional concepts in classical texts.

The superior performance in legal texts (mean SPS=0.954) versus literary works (mean SPS=0.831) indicates that current neural architectures excel at processing structured, domain-specific language but struggle with context-dependent emotional expressions. These performance variations across genres reflect fundamental challenges in computational linguistics: the trade-off between standardization and expressiveness, the complexity of cultural-specific sentiment mapping, and the temporal evolution of language patterns.

These analytical findings underscore implicit cultural equivalence assumptions within the framework, which presuppose the possibility of emotional equivalence across cultures and historical periods—a notion that remains theoretically contested in translation studies. Certain emotional concepts may be culture-specific and resist direct translation, challenging the universality of sentiment preservation metrics across diverse literary traditions.

Despite these limitations, our framework provides a valuable first step toward more comprehensive sentiment-aware evaluation of machine translation. Future work should address these limitations through expanded corpus coverage with balanced representation across historical periods, refined annotation methodologies incorporating diachronic emotional concepts, and implementation of multi-dimensional emotional mapping beyond simplistic polarity and intensity measures. Additional research directions include evaluating specific emotional categories (joy, sadness, fear, anger) for texts where emotional specificity carries cultural significance, large-scale evaluation through automated SDI metric implementation, cross-domain adaptability testing, integration with established metrics like BLEU or COMET, and dynamic weight optimization through machine learning approaches to enhance adaptation to specific genres and cultural contexts.

References

M. Almansor, C. Zhang, W. Khan, A. Hussain, and N. Alhusaini. 2020. Cross lingual sentiment analysis: A clustering-based bee colony instance selection and target-based feature weighting approach. *Sensors*, 20.

Paulo Cardinal. 2009. The legal system of the macau special administrative region: An overview. *Asian Law Institute Working Paper Series*, 5.

Corpus USX. 2024. Pool of Bilingual Parallel Corpora of Chinese Classics. Accessed: 2024-07-31.

H. Hu. 2023. Construction of feature extraction model for machine foreign language translation evaluation system. *Applied Mathematics and Nonlinear Sciences*, 8:2677–2686.

Kaisla Kajava, Emily Öhman, Piao Hui, and Jörg Tiedemann. 2020. Emotion preservation in translation: Evaluating datasets for annotation projection. In *Digital Humanities in the Nordic Countries*, pages 38–50. CEUR.

Tomáš Kocmi, Christian Federmann, and Daniel Kurokawa. 2021. Shiip-in-a-bottle: A minimalist approach to reference-free mt evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4068–4080.

D.C. Lau and Fong Ching Chen. 1995. *A Concordance to the Lunyu*. Number 16 in Harvard-Yenching Institute Sinological Index Series. Harvard-Yenching Institute Sinological Index Series, Cambridge, MA.

J. Li. 2023. Optimization of translation techniques between english and chinese literary works in the information age. *Applied Mathematics and Nonlinear Sciences*, 9.

Stephen Owen. 2010. *The Cambridge History of Chinese Literature*, volume 1. Cambridge University Press, Cambridge.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

L. Tian. 2023. Applying machine translation to chinese–english subtitling: Constraints and challenges. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

X. Wan. 2011. Bilingual co-training for sentiment classification of chinese product reviews. *Computational Linguistics*, 37:587–616.

X. Wang, R. Beard, and R. Chandra. 2024. Evaluation of google translate for mandarin chinese translation using sentiment and semantic analysis. *ArXiv*, abs/2409.04964.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and 1 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

C. Zhao, M. Wu, X. Yang, W. Zhang, S. Zhang, S. Wang, and D. Li. 2024. A systematic review of cross-lingual sentiment analysis: Tasks, strategies, and prospects. *ACM Computing Surveys*, 56:1–37.

A Corpus Composition Analysis

The parallel corpus comprises carefully selected texts representing diverse genres and periods of classical Chinese literature. The composition analysis reveals systematic distribution across multiple dimensions, as detailed in Table 1.

B Translation Quality Analysis

To demonstrate the rigorous quality control in our translation process, we present representative examples of parallel texts that illustrate the nuanced translation approaches employed in our corpus.

C Sentiment Analysis Examples

Our sentiment annotation methodology demonstrates robust performance in capturing nuanced sentiments across both modern and classical texts, as evidenced in the representative examples below.

D Preservation Metrics by Literary Work

This section presents comprehensive sentiment preservation metrics across different literary works and translation systems evaluated in our study.

E Implementation Details

Our sentiment annotation methodology incorporates both automated and expert-validated approaches, with carefully optimized weighting parameters for the evaluation framework.

Table 8 presents the prompts used for sentiment annotation in our implementation.

F Detailed Experimental Results

This appendix presents comprehensive sentiment preservation metrics for all literary works and translation systems evaluated in our study.

Table 1: Detailed Composition of Source Texts

Genre Category	English Title	Chinese Title
Philosophical Works	*The Book of Changes*[a] *The Analects* *The Great Learning* *Tao Te Ching*	《易经》 《论语》 《大学》 《道德经》
Classical Novels	*Romance of the Three Kingdoms* *Water Margin* *Dream of the Red Chamber* *Journey to the West*	《三国演义》 《水浒传》 《红楼梦》 《西游记》
Literary Compositions	*The Romance of the Western Chamber* *Complete Works of Wang Yangming* *Vegetable Roots Discourse*	《西厢记》 《王阳明全集》 《菜根谭》
Legal Documents	*Laws of Macau*[b]	《澳门法律》

[a] English translations follow the Harvard-Yenching Institute Sinological Index Series (Lau and Chen, 1995) and contemporary sinological practice (Owen, 2010).
[b] Terminology follows the official Macau SAR legal system (Cardinal, 2009).

Table 2: Corpus Composition and Distribution

Dimension	Scale	Distribution
Genre	4	• Philosophical Texts (33.3%) • Classical Novels (33.3%) • Literary Works (25%) • Legal Documents (8.4%)
Sources	12	• Classical Canon (4) • Historical Novels (4) • Cultural Essays (3) • Legal Corpus (1)
Content Type	3	• Narrative (40%) • Philosophical Discussion (35%) • Technical Description (25%)

Table 3: Additional Examples of Parallel Text with Sentiment Annotation

Language	Source	Target	Polarity	Score
ZN/EN	一时间众人俱各无言，都向雨村观看。雨村便知其意，也不谦让，微微一笑，便说道："诸公既然命弟作东，如此甚妙。"	For a time no one spoke, but all looked towards Yucun, who took the hint. Without false modesty he smiled slightly and replied, "If you gentlemen want me to be the host, nothing could be better."	Positive	0.5
ZN/EN	赵姨娘在王夫人跟前一生过不去，心中一腔子气，不知向谁处发泄才好。	Aunt Zhao had been at odds with Lady Wang all her life and had accumulated a bellyful of resentment which she didn't know on whom to vent.	Negative	-0.7

Table 4: Representative Example of Parallel Text with Sentiment Annotation

Language	Source Text	Target Text
ZN/EN	我也曾游过些名山大刹，倒不曾见过这话头，其中想必有个翻过筋斗来的亦未可知，何不进去试试。	I've never come across anything like it in all the famous temples I've visited. There may be a story behind it of someone who has tasted the bitterness of life, some repentant sinner. I'll go in and ask.

Table 5: Example of Sentiment Analysis

Version	Content	Sentiment Polarity	Sentiment Score
Source	雨村看了，因想道: "这两句话，文虽浅近，其意则深。"	Neutral	0.2
Human version	"Trite as the language is, this couplet has deep significance," thought Yucun.	Neutral	0.2
DeepL	Yucun read it, because he thought: "These two sentences, although the text is shallow, its meaning is deep."	Neutral	0.1
Google Translate	Yucun read it and thought: "Though these two sentences are simple and short in text, their meaning is profound."	Neutral	0.5
GPT-4o	Upon seeing it, Yucun thought to himself, 'Though these sentences are simple in language, their meaning is profound.'	Neutral	0.1

Table 6: Complete Sentiment Preservation Metrics by Literary Work and Translation System (Sample)

Literature	MT System	IPS	PAS	SPS	SDI	Error Class
Hongloumeng	GPT-4o	0.850	0.885	0.872	0.124	Minor
	DeepL	0.848	0.870	0.862	0.132	Minor
	Google	0.835	0.891	0.869	0.117	Minor
Xiyouji	GPT-4o	0.841	0.832	0.835	0.173	Minor
	DeepL	0.839	0.846	0.843	0.163	Minor
	Google	0.798	0.810	0.806	0.189	Minor

Table 7: Weight Optimization Results

w_1	w_2	IAA	Correlation	F-score
0.55	0.45	0.76	0.82	0.88
0.60	0.40	0.79	0.84	0.90
0.65	**0.35**	**0.83**	**0.87**	**0.92**
0.70	0.30	0.81	0.85	0.89
0.75	0.25	0.77	0.83	0.87

Table 8: Chinese and English Prompts for Sentiment Annotation

Category	Chinese Prompt	English Prompt
Role	你是一个文本情感分析专家。	You are an expert in text sentiment analysis.
Task Description	你需要对给定的句子进行精准的情感分析(sentimental analysis)。	Your task is to perform accurate sentiment analysis on the given sentences.
Sentiment Categories	- 积极(positive) - 中性(neutral) - 消极(negative)	- Positive - Neutral - Negative
Score Range	消极：$(-1, -0.33)$ 中性：$(-0.33, 0.33)$ 积极：$(0.33, 1)$	Negative: $(-1, -0.33)$ Neutral: $(-0.33, 0.33)$ Positive: $(0.33, 1)$
Output Format	JSON 格式，键名均为小写字母，不带任何其他无用信息和文本：{"sentimental": {"class": "<情感分类>", "point": <情感得分>}}	JSON format, with all names in lowercase letters, without any other useless information and text: {"sentimental": {"class": "positive", "point": 0.4}}
Input Placeholder	需要评估的句子: {{#0}}	Here are the sentences to be evaluated: {{#output.en}}{{#output.EN}}

Table 9: Complete Sentiment Preservation Metrics by Literary Work and Translation System

Literature	MT System	IPS	PAS	SPS	SDI	Error Class
Yijing	GPT-4o	0.751	0.706	0.724	0.291	Minor
	DeepL	0.762	0.723	0.738	0.278	Minor
	Google	0.728	0.618	0.663	0.310	Minor
Lunyu	GPT-4o	0.860	0.884	0.874	0.126	Minor
	DeepL	0.841	0.821	0.829	0.170	Minor
	Google	0.819	0.792	0.803	0.194	Minor
Daxue	GPT-4o	0.805	0.780	0.790	0.223	Minor
	DeepL	0.815	0.802	0.807	0.203	Minor
	Google	0.801	0.794	0.797	0.210	Minor
Laozi	GPT-4o	0.817	0.809	0.812	0.198	Minor
	DeepL	0.810	0.809	0.809	0.195	Minor
	Google	0.801	0.772	0.782	0.219	Minor
Sanguo	GPT-4o	0.825	0.870	0.852	0.140	Minor
	DeepL	0.822	0.852	0.840	0.149	Minor
	Google	0.793	0.816	0.807	0.177	Minor
Shuihu	GPT-4o	0.852	0.882	0.870	0.128	Minor
	DeepL	0.836	0.873	0.859	0.141	Minor
	Google	0.840	0.866	0.856	0.133	Minor
Hongloumeng	GPT-4o	0.850	0.885	0.872	0.124	Minor
	DeepL	0.848	0.870	0.862	0.132	Minor
	Google	0.835	0.891	0.869	0.117	Minor
Xiyouji	GPT-4o	0.841	0.832	0.835	0.173	Minor
	DeepL	0.839	0.846	0.843	0.163	Minor
	Google	0.798	0.810	0.806	0.189	Minor
Xixiangji	GPT-4o	0.787	0.810	0.802	0.207	Minor
	DeepL	0.806	0.835	0.825	0.180	Minor
	Google	0.798	0.810	0.806	0.189	Minor
Wangyangming	GPT-4o	0.820	0.804	0.810	0.202	Minor
	DeepL	0.840	0.838	0.839	0.164	Minor
	Google	0.822	0.786	0.799	0.212	Minor
Caigentan	GPT-4o	0.819	0.822	0.821	0.181	Minor
	DeepL	0.825	0.827	0.826	0.175	Minor
	Google	0.824	0.832	0.829	0.168	Minor
Lawcorpus1	GPT-4o	0.928	0.977	0.957	0.034	Minor
	DeepL	0.934	0.975	0.958	0.033	Minor
	Google	0.921	0.964	0.946	0.042	Minor