Role-Guided Annotation and Prototype-Aligned Representation Learning for Historical Literature Sentiment Classification

Hongfei Du[1], **Jiacheng Shi**[1], **Jacobo Myerston**[2], **Sidi Lu**[1], **Gang Zhou**[1], **Ye Gao**[1]
[1]Department of Computer Science, William & Mary, USA
[2]Department of Literature, UC San Diego, USA
{hdu02, jshi12, sidi, gzhou, ygao18}@wm.edu jmyerston@ucsd.edu

Abstract

Sentiment analysis of historical literature provides valuable insights for humanities research, yet remains challenging due to scarce annotations and limited generalization of models trained on modern texts. Prior work has primarily focused on two directions: using sentiment lexicons or leveraging large language models (LLMs) for annotation. However, lexicons are often unavailable for historical texts due to limited linguistic resources, and LLM-generated labels often reflect modern sentiment norms and fail to capture the implicit, ironic, or morally nuanced expressions typical of historical literature, resulting in noisy supervision. To address these issues, we introduce a role-guided annotation strategy that prompts LLMs to simulate historically situated perspectives when labeling sentiment. Furthermore, we design a prototype-aligned framework that learns sentiment prototypes from high-resource data and aligns them with low-resource representations via symmetric contrastive loss, improving robustness to noisy labels. Experiments across multiple historical literature datasets show that our method outperforms state-of-the-art baselines, demonstrating its effectiveness.

1 Introduction

Sentiment analysis of historical literature reveals the emotional and ideological underpinnings of past societies and provides insight into how their worldviews have shaped the present, thus highlighting the potential of computational methods to bridge cultural history and NLP (Al-Laith et al., 2023). However, this task is challenging due to the scarcity of annotated data, as accurate labeling requires expertise in linguistics, literature, and historical context (Sprugnoli et al., 2016). Furthermore, models trained on modern texts often struggle to generalize to historical domains due to the use of indirect, ironic, and morally nuanced sentiment expressions (Schmidt et al., 2021).

Prior work on historical sentiment analysis has mainly followed two directions. The first relies on sentiment lexicons (Schmidt and Burghardt, 2018; Al-Laith et al., 2023; Koto et al., 2024), which map words to labels, are interpretable and easy to apply. However, it is often unavailable for historically under-represented texts due to the limited linguistic resources (Ogbuju and Onyesolu, 2019). The second direction utilizes large language models (LLMs) to generate sentiment labels in low-resource settings, which are then used to fine-tune downstream classifiers (Dorkin and Sirts, 2024; Pingle et al., 2023).

While LLM-based annotation shows promise, it faces two main limitations. First, most prompt-based annotation methods typically rely on direct instruction prompts (Ding et al., 2023) (e.g., "Choose the sentiment of the given text") to generate sentiment labels. However, such prompting strategies implicitly assume that sentiment can be interpreted uniformly across time, culture, and context. This assumption often does not hold in historical literature, where sentiment is conveyed implicitly and shaped by rhetorical and moral complexity (Konstan, 2006), often causing LLMs misread or overlook sentiment cues in historical texts.

In addition, existing works on historical sentiment analysis (Dorkin and Sirts, 2024; Pingle et al., 2023) typically assumes that the labels produced by LLMs are reliable, and thus directly uses them for downstream training. However, LLM-generated labels are inherently noisy, especially for subtle or stylistically complex expressions in historical literature. This limits the generalizability of downstream models.

To address these two major challenges, unlike prior work that either applies sentiment lexicons or uses LLMs without contextual guidance and consideration of label noise, we propose a role-guided annotation and prototype-aligned representation learning method (RAPA).

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3756–3768
November 4-9, 2025 ©2025 Association for Computational Linguistics

To be concrete, we first introduce a role-play annotation strategy that instructs LLMs to simulate historical figures and evaluate sentiment according to the ways sentiment was typically conveyed in that historical period. This role-guided approach encourages LLMs to better capture the sentiment in historical literature and avoid misreading them through a modern lens.

Second, we propose a unified framework that learns trainable sentiment prototypes from high-resource data. These prototypes are class-level representations that capture the core semantics of each sentiment category. We further align them with low-resource prototypes through a symmetric contrastive loss, which helps the model reduce its reliance on potentially noisy labels. Even when the label of a historical text is inaccurate or ambiguous, the model can still learn a meaningful representation by being guided toward the correct prototype based on its semantic content.

Our contributions are summarized as follows:

- We propose a novel annotation strategy that prompts ChatGPT-4o to role-play as historical figures, providing situated contextual basis for sentiment labeling in historical texts.
- We develop a prototype alignment method that learns sentiment prototypes from high-resource data and aligns low-resource representations via contrastive learning, which helps mitigate the impact of label noise and enhance representation robustness.
- We conduct comprehensive experiments demonstrating the effectiveness of our annotation strategy and prototype alignment method, and show that our approach outperforms state-of-the-art baselines on multiple historical sentiment datasets.

2 Related Work

2.1 Sentiment Classification in Historical Literature

In recent years, sentiment classification has gradually been extended to historical literature, with challenges such as language variation, semantic drift, and data scarcity. For example, (Schmidt and Burghardt, 2018) evaluated German sentiment lexicons on annotated speeches from Lessing's plays. (Al-Laith et al., 2023) constructed a manually labeled sentiment dataset for Scandinavian literature of the 19th century and explored gender-based sentiment patterns using a BERT-based model adapted to the domain. (Krusic, 2024) developed a sentiment-annotated corpus of Austrian historical newspapers. (Sprugnoli et al., 2024) proposed a Latin sentiment classification dataset. Meanwhile, (Dorkin and Sirts, 2024) uses GPT-4 to generate labels to improve the model's performance on the Latin dataset. Similarly, (Pingle et al., 2023) also use GPT-based generation strategy to improve model performance on a Marathi dataset.

However, due to the limited availability of sentiment lexicons for historical texts and their inability to capture nuanced sentence-level sentiment, we deliberately avoid using sentiment lexicons as external knowledge in this work.

2.2 Methods for Label Generation Using LLMs

To overcome the lack of labeled data, recent works leverage LLMs to generate annotations. Prior research has shown that ChatGPT even outperforms the crowd in terms of accuracy for various NLP tasks (Gilardi et al., 2023). In particular, prompt-based data labeling has achieved good performance in sentiment classification (Ding et al., 2023).

Despite these advantages, LLMs exhibit lower performance in multilingual tasks, with accuracy dropping for non-English data (Mohta et al., 2023). Moreover, current prompting strategies (Martorana et al., 2024; Ding et al., 2023) rely on direct instruction for sentiment classification, ignoring the importance of situational context or perspective of the annotator. This limitation is particularly evident in historical literary texts, as effective annotation requires sensitivity to the cultural and temporal context in which the text was produced.

Recent studies (Wang et al., 2024) demonstrate that LLMs can effectively engage in role-play, simulating specific personas or historical perspectives when appropriately guided. Building on this capability, we propose a novel role-guided prompting strategy that instructs LLMs to annotate historical data from the viewpoint of a historically situated character, ensuring annotations reflect the context and sensibilities of the period.

2.3 Methods of Denoising in Low-Resource Settings

Recent research has utilized LLMs to generate labels for text classification followed by label-level denoising. For instance, NoiseAL (Wang et al., 2023) uses small models to identify noisy instances

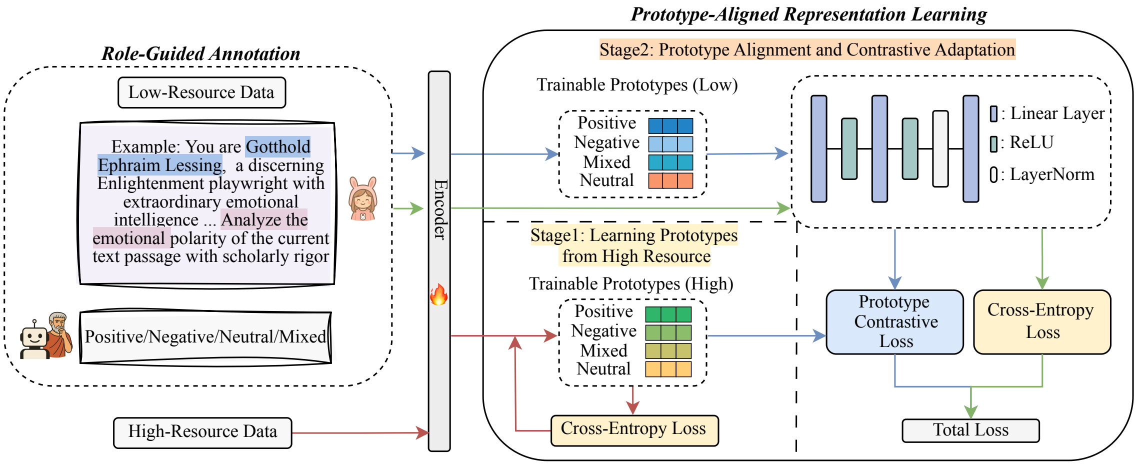

Role-Guided Annotation
Low-Resource Data
Example: You are Gotthold Ephraim Lessing, a discerning Enlightenment playwright with extraordinary emotional intelligence ... Analyze the emotional polarity of the current text passage with scholarly rigor
Positive/Negative/Neutral/Mixed
High-Resource Data
Encoder
Prototype-Aligned Representation Learning
Stage2: Prototype Alignment and Contrastive Adaptation
Trainable Prototypes (Low)
Positive
Negative
Mixed
Neutral
: Linear Layer
: ReLU
: LayerNorm
Stage1: Learning Prototypes from High Resource
Trainable Prototypes (High)
Positive
Negative
Mixed
Neutral
Cross-Entropy Loss
Prototype Contrastive Loss
Cross-Entropy Loss
Total Loss

Figure 1: Overall architecture of the proposed RAPA framework. The left side illustrates the role-guided annotation strategy. The right side shows prototype-aligned representation learning. Stage 1 learns high-resource prototypes; Stage 2 performs prototype alignment via contrastive loss and sentiment supervision via cross-entropy loss.

and asks LLMs to modify them. LAFT (Yuan et al., 2024) incorporates feedback from LLMs during the fine-tuning process to distinguish between clean and noisy labels. Other studies have aggregated LLM-generated labels through weakly-supervised frameworks such as Snorkel (Smith et al., 2024), which denoises at the voting level before training. UNIGEN (Choi et al., 2024) treats the LLM as a zero-shot labeler to construct cross-domain pseudo-labeled datasets and mitigate label noise by using task-specific modeling of the pseudo-labeling step.

However, existing methods perform label-level denoising in the preprocessing step, but ignore semantic inconsistencies in the representation space. Unlike label-level methods that operate on prediction outputs, our approach performs representation-level denoising by aligning low-resource prototypes with trainable high-resource prototypes, which encourages more semantically consistent representations and provides structural resistance to noisy supervision.

3 Methodology

In this section, we first define the task of sentiment analysis of historical literature. Then, we provide an overview of our proposed framework, RAPA, followed by a detailed description of its two main components: role-guided annotation strategy and prototype-aligned representation learning.

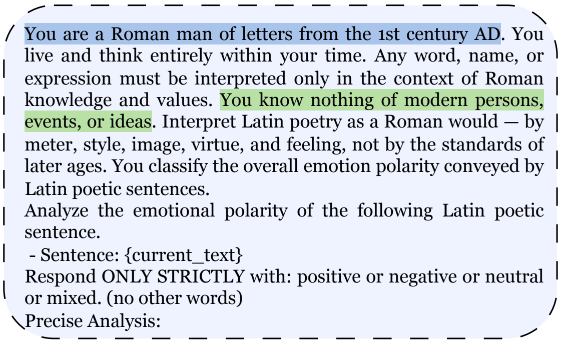
3.1 Problem Definition

Given a text sequence $x = \langle x_1, \cdots, x_n \rangle$ consisting of n tokens, and a predefined sentiment label set $\mathcal{Y} = \{y_1, \cdots, y_k\}$, the goal of sentiment classification is to assign a label $y \in \mathcal{Y}$ to x, reflecting the overall sentiment expressed in the text. Typical labels include positive, negative, neutral, and mixed.

We assume access to high-resource datasets $\mathcal{D}_h = \{(x_i^h, y_i^h)\}$, where high-quality annotations are available in sufficient quantity, and low-resource datasets $\mathcal{D}_l = \{(x_j^l, y_j^l)\}$, which contain limited or noisy labels. The objective is to leverage supervision from $\mathcal{D}_h$ to improve classification performance on $\mathcal{D}_l$, especially when the low-resource data differs in domain or language. This setup reflects practical challenges in historical or under-represented textual domains.

3.2 Overall Architecture of RAPA

Figure 1 illustrates the overall architecture of RAPA. The model first utilizes a role-guided annotation module, which uses ChatGPT-4o to annotate texts from the perspective of historical characters, thereby providing context-aware supervision tailored to the historical domain. Based on these annotations, we propose the prototype alignment representation module, which employs a two-stage training strategy. In the first phase, sentiment prototypes are learned from high-resource data under supervised training. In the second phase, we introduce a joint comparison learning strategy to explicitly



You are a Roman man of letters from the 1st century AD. You live and think entirely within your time. Any word, name, or expression must be interpreted only in the context of Roman knowledge and values. You know nothing of modern persons, events, or ideas. Interpret Latin poetry as a Roman would — by meter, style, image, virtue, and feeling, not by the standards of later ages. You classify the overall emotion polarity conveyed by Latin poetic sentences.
Analyze the emotional polarity of the following Latin poetic sentence.
- Sentence: {current_text}
Respond ONLY STRICTLY with: positive or negative or neutral or mixed. (no other words)
Precise Analysis:

Figure 2: Example of a role-play prompt used to simulate Roman-era interpretation for Latin sentiment analysis.

align low-resource and high-resource prototypes in a shared semantic space. This helps transfer robust sentiment structures from high-resource to low-resource domains.

3.3 Role-Guided Annotation

Given the limited availability of labeled data, we first employ ChatGPT-4o (OpenAI, 2024) to generate annotations for further prototype-aligned representation learning. Sentiment annotation for historical literature, is particularly challenging as it requires not only linguistic expertise but also a deep understanding of historical texts and their cultural context. However, if LLMs are directly prompted without an explicit interpretive stance, it tends to assess sentiment based on surface semantics and lacks sensitivity to historical rhetorical strategies such as irony, indirection, or period-specific norms of sentiment.

To address this, we propose a role-guided annotation strategy that use role-specific prompts to instruct LLMs to label historical texts. Each prompt requires the model to adopt a specific interpretive stance, such as a playwright or a culturally situated reader.

Figure 2 shows an example of how sentiment is annotated in Latin poetry using our role-play prompt. Unlike generic instructions, our role-guided prompts aim to evoke immersive character perspectives. To reinforce this perspective, we include a constraint sentence in the prompt (highlighted in green in Figure 2) that explicitly limits the model's knowledge to the cultural and historical context of the original text. This helps prevent modern bias and guide the model focus on sentiment cues relevant to the historical context. Such bias often arises from semantic shift, where the sentiment connotations of words change over time (Hamilton et al., 2016). For example, the word *taylor* may be neutral in the original context, but modern interpretations may consider it to be complimentary, thus potentially misleading the model. To mitigate this temporal bias, the prompts includes constraints that guide the model to adopt the worldview and sentiment standards of historical period.

Based on this design, we formalize the annotation process as a conditional prediction task. Formally, given an input text x and a role-specific instruction I, the sentiment label $y \in \mathcal{Y}$ is generated by a conditional process: $y \sim \text{LLM}(x; I)$.

Since LLM-generated labels cannot be assumed to be entirely reliable, we propose a prototype-based representation alignment framework that mitigates the effect of annotation noise by leveraging stable class prototypes learned from high-resource data.

3.4 Prototype-Aligned Representation Learning

Our method consists of two main stages. In the first stage, we learn trainable sentiment prototypes from high-resource data, which serve as stable semantic anchors in the embedding space. In the second stage, we align low-resource representations to these prototypes via a joint objective that combines cross-entropy loss with prototype-based contrastive loss.

3.4.1 Stage 1: Learning Prototypes from High-Resource Data

In the first stage, we train a prototype-based sentiment classifier using high-resource datasets $\mathcal{D}_h = \{(x_i^h, y_i^h)\}$ with reliable sentiment labels. For each input x_i^h, we extract the d-dimensional `[CLS]` embedding $h_{\texttt{[CLS]}}^{(i)}$ from the final layer of a pretrained multilingual encoder.

We define a trainable prototype matrix $P_{\text{high}} \in \mathbb{R}^{C \times d}$, where each row $p_i = P_{\text{high}}[i] \in \mathbb{R}^d$ represents the prototype for class $i \in \{1, \dots, C\}$ learned from high-resource data. Classification is performed by computing the Euclidean distance between the input embedding and each prototype.

The classifier is optimized using a softmax over negative squared Euclidean distances. The cross-entropy (CE) loss for an instance (x_i^h, y_i^h) is:

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp\left(-\left\|h_{[\text{CLS}]}^{(i)} - P_{\text{high}}[y_i^h]\right\|^2\right)}{\sum_{j=1}^{C} \exp\left(-\left\|h_{[\text{CLS}]}^{(i)} - P_{\text{high}}[j]\right\|^2\right)} \quad (1)$$

Where $P_{\text{high}}[y_i^h]$ denotes the prototype corresponding to the ground-truth label y_i^h.

After training, both the encoder parameters and the learned prototypes P_{high} are retained for the second stage, where they serve as semantic anchors to support representation alignment on the low-resource datasets $\mathcal{D}_l$.

3.4.2 Stage 2: Prototype Alignment in Low-Resource Settings

In the second stage, we adapt the pretrained encoder and sentiment prototypes to the low-resource dataset $\mathcal{D}_l$. This alignment encourages the representation space of the low-resource domain to align with semantically stable prototypes learned from high-resource data, thus reducing the adverse effects of annotation noise.

Model Components. The model consists of a shared encoder, low-resource prototypes $P_{\text{low}} \in \mathbb{R}^{C\times d}$, fixed high-resource prototypes $P_{\text{high}} \in \mathbb{R}^{C\times d}$ (precomputed in Stage 1), and a non-linear mapper $\mathcal{M}$ that transforms low-resource prototypes into the high-resource prototype space. The mapper $\mathcal{M}$ is a multi-layer feedforward network with ReLU activations and LayerNorm designed for expressive and stable transformation. During training, each input is classified based on the Euclidean distance between its [CLS] embedding and the corresponding mapped prototypes.

Prototype Initialization. To initialize the low-resource prototypes P_{low}, we compute the mean of the [CLS] representations for each class from the low-resource training set. This strategy provides a semantically grounded starting point, which helps stabilize optimization during the early training phase.

Training Objective. The model is trained with a joint loss combining classification supervision and prototype alignment. First, we apply a cross-entropya loss over the distances between the [CLS] embedding and the mapped low-resource prototypes:

$$\mathcal{L}_{\text{ce}} = \text{CE}\left(-\left\|\mathcal{M}(h_{[\text{CLS}]}) - \mathcal{M}(P_{\text{low}})\right\|^2, y\right) \quad (2)$$

where y is the ground-truth sentiment label.

To align the prototype spaces, we use a symmetric contrastive loss that computes similarities in both directions, providing richer supervision signals and promoting one-to-one alignment between mapped low-resource prototypes and fixed high-resource prototypes. For example, the "positive" prototype in Latin may align with its English counterpart through shared sentiment semantics, guided by the symmetric contrastive loss.

$$\begin{aligned}\mathcal{L}_{\text{proto}} = \frac{1}{2}\Big[& \text{CE}\left(\frac{\mathcal{M}(P_{\text{low}})P_{\text{high}}^{\top}}{\tau}, I\right) \\ & + \text{CE}\left(\frac{P_{\text{high}}\mathcal{M}(P_{\text{low}})^{\top}}{\tau}, I\right)\Big]\end{aligned} \quad (3)$$

where τ is the temperature hyperparameter, and $I \in \mathbb{R}^{C\times C}$ is the identity matrix encoding ideal class-wise correspondence.

The total loss is defined as:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{ce}} + \alpha \cdot \mathcal{L}_{\text{proto}} \quad (4)$$

where α is a tunable weight balancing the classification and alignment objectives.

3.4.3 Inference

At inference time, we utilize the trained encoder and mapped low-resource prototypes to classify new inputs. For each test instance, we compute its [CLS] representation and assign its label based on the nearest prototype in Euclidean space. The final prediction corresponds to the prototype with the smallest distance. The inference process is independent of high-resource data and relies solely on the low-resource encoder and prototypes, enabling lightweight and portable deployment.

4 Experiments

4.1 Datasets

In this work, we categorize all datasets into two groups: high-resource datasets, which are used to construct high-resource sentiment prototypes, and low-resource datasets, which are used for fine-tuning and evaluation.

High-Resource Datasets. This group includes the SemEval 2017 Task 4 dataset (Rosenthal et al., 2017) and the IMDB sentiment dataset (Maas et al., 2011). Additionally, to address the lack of "mixed" sentiment instances, we repurpose samples from the GoEmotions dataset (Demszky et al., 2020) [1] by labeling sentences that contain both positive and negative emotions as "mixed" sentiment.

Low-Resource Datasets. This category includes historical texts with sparse annotations, including 18th-century Lessing's plays (Schmidt and

[1] https://github.com/google-research/google-research/tree/master/goemotions

Backbone	**#Params**	**Setting**	**Lessing's Play**	**Scandinavian Novels**	**Austrian News**	**EvaLatin 2024**	**AVG**
Baseline			62.50	72.00	-	30.00	-
XLM-R-base	270M	Direct Finetuning	69.79	71.30	41.30	30.45	53.21
		RAPA	69.35	73.58	45.07	33.04	55.26
XLM-R-large	550M	Direct Finetuning	**73.50**	75.27	44.26	37.54	57.64
		RAPA	70.98	**76.56**	51.41	**44.34**	**60.82**
InfoXLM-base	270M	Direct Finetuning	68.89	64.31	42.80	31.29	51.82
		RAPA	63.74	66.40	50.24	32.50	53.22
InfoXLM-large	550M	Direct Finetuning	66.80	73.25	45.08	34.09	54.81
		RAPA	72.21	70.38	**57.40**	41.47	60.37
RemBERT	575M	Direct Finetuning	60.66	68.36	45.37	33.77	52.04
		RAPA	72.13	68.90	49.02	34.68	56.18

Table 1: Macro-F1 of different backbones across four datasets under direct finetune and RAPA. Bold values indicate the best result in each column. RAPA achieves the best performance in 3 out of 4 datasets, showing strong generalizability.

Burghardt, 2018), 19th-century Scandinavian novels (Al-Laith et al., 2023), 19th-century Austrian historical newspapers (Krusic, 2024), and Latin poems from the 1st to 19th centuries (Sprugnoli et al., 2024). Since the original papers for the Lessing's play and EvaLatin 2024 datasets do not provide additional training data, we augment the Lessing dataset by including two extra plays: *Damon* and *DerFreigeist*[2], for fine-tuning. Similarly, we expand EvaLatin by incorporating all 116 poems by Catullus[3], and further extract emotionally salient or stylistically intense lines using ChatGPT for low-resource sentiment adaptation. All annotations are generated via our role-guided annotation method.

These low-resource datasets suffer from both data scarcity and label imbalance (Table 2), often leading models to overfit majority classes. RAPA remains robust in such settings by leveraging sentiment prototypes as class-level anchors to enhance minority class representations.

4.2 Implementation Details

Steps. We first construct sentiment prototypes using high-resource data as a guide for learning from low-resource prototypes. To supplement the limited annotations in low-resource datasets, we employ a role-guided annotation strategy using ChatGPT-4o, where the model simulates historical figures to assign sentiment labels. These role-play annotated samples are then fine-tuned by combining them with the existing labeled data. During fine-tuning, we set the weight of the prototype contrastive loss (α) to 0.6 and use a temperature parameter $\tau = 0.1$. We evaluate the model on retained test datasets using macro-F1 as the main metric.

Backbone Models. We experiment with five multilingual pretrained language models as encoders: XLM-R-base (Conneau et al., 2020), XLM-R-large (Conneau et al., 2020), InfoXLM-base (Chi et al., 2021), InfoXLM-large (Chi et al., 2021) and RemBERT (Chung et al., 2021). These models serve as the backbone for our sentiment classification framework.

Baselines. For each dataset, we report the best macro-F1 achieved by the existing methods as the baseline. On the Lessing's play dataset, SentiWS yielded the best lexicon-based result. For the Scandinavian novels, the Danish sentiment model performed best under supervised fine-tuning. No evaluation results were reported for the Austrian newspapers in the original study. For EvaLatin 2024, LLM-generated labels consistently outperformed all heuristic-based approaches.

4.3 Results and Analysis

Table 1 presents the macro-F1 results of the baseline, direct fine-tuning, and our proposed method RAPA across four historical test datasets. Direct fine-tuning refers to training a pretrained backbone model with a linear classification head using cross-entropy loss, without incorporating any prototypes from either low-resource or high-resource datasets. These datasets differ not only in label distributions but also in the type of noisy supervision used for fine-tuning, which we categorize into two types: LLM-generated and human-annotated.

[2] https://textgridrep.org

[3] https://www.thelatinlibrary.com/catullus.shtml

Dataset	Neg.	Pos.	Neu.	Mix.	Total
Lessing's Play	139	61	–	–	200
Scandinavian Novels	38	19	43	–	100
Austrian News	96	17	79	8	200
EvaLatin 2024	154	75	20	50	301

Table 2: Label distribution across test datasets. Some datasets contain only binary classes, while others exhibit highly imbalanced or multi-class sentiment structures.

Datasets	No RP	Modern RP	Historical RP
Lessing's Play	78.44	78.68	**80.07**
Scandinavian Novels	79.04	80.47	**81.25**
Austrian Newspapers	54.84	56.70	**66.27**
EvaLatin 2024	41.56	56.52	**58.30**

Table 3: Macro-F1 scores of different annotation strategies across four historical datasets. RP denotes the use of Role-Play in prompt design.

Denoising LLM-Generated Supervision. We first examine RAPA on ChatGPT-labeled datasets used for fine-tuning, including EvaLatin 2024 and Lessing's Play. Among them, EvaLatin presents the greatest challenge: it involves four sentiment classes including neutral and mixed, and exhibits the lowest macro-F1 scores across all datasets (Table 3), reflecting substantial label noise and ambiguity. Lessing's play, while also LLM-labeled, is a binary classification task and achieves relatively higher labeling F1 scores, suggesting lower noise and task complexity. Despite these differences, RAPA consistently outperforms direct fine-tuning on both datasets, improving macro-F1 by +6.8 (XLM-R-large) and +7.38 (InfoXLM-large) on EvaLatin 2024, demonstrating strong denoising capability under machine-labeled, low-resource conditions.

Robustness to Human-Annotated Label Noise. We next assess RAPA's robustness to human-induced annotation noise. Human-annotated datasets such as Scandinavian Novels and Austrian News are annotated by expert or semi-expert annotators. However, their original papers report moderate inter-annotator agreement and annotator challenges due to the interpretive ambiguity of literary texts and historical language (e.g., sarcasm, metaphor). These human-induced ambiguities result in non-negligible label noise. On both datasets, RAPA surpasses all baseline and direct fine-tuning methods. Notably, it improves macro-F1 by +12.32 on Austrian News with InfoXLM-large, highlighting its effectiveness in denoising human-labeled, fine-grained sentiment data under limited supervision, which benefits from the prototype framework's semantic anchoring and InfoXLM-large's strong capacity to separate fine-grained classes.

Overall, these results demonstrate RAPA's robustness to both machine-generated and human-induced label noise, and its adaptability across diverse low-resource, historical sentiment classification tasks. While RAPA shows strong overall performance, its gains on Lessing's Play and Scandinavian Novels are less consistent. We hypothesize that this is due to a mismatch between the task label space and the prototype space: both datasets are binary or three-class sentiment classification tasks, while our framework learns four sentiment prototypes (positive, negative, neutral, mixed) from high-resource data. During training, low-resource samples are contrasted against all four prototypes, even those not present in the task (e.g., neutral and mixed), which may introduce unnecessary supervision and may distort the representation space, ultimately hindering effective class separation.

4.4 Ablation Study

To further verify the effectiveness of each component in our proposed RAPA, we conduct ablation studies on role-guided annotation and prototype alignment.

4.4.1 Effectiveness of Role-Play

Table 3 shows the macro-F1 of sentiment classification under the three annotation strategies on four low-resource historical literature datasets. All evaluations were conducted on test datasets, using manually annotated golden labels as reference.

The modern and historical role-guided settings are built upon the no role-play baseline by incorporating character-based instructions from different temporal perspectives. We found that both modern and historical role-guided settings consistently improved performance over the no-role baseline, confirming the effectiveness of contextualized instructions. Among them, historical role-guided prompting achieved the highest accuracy across all datasets, with particularly notable gains (up to +17%) on the Latin corpora. These results suggest that aligning the annotator's perspective with the historical context further enhances the sentiment understanding of historical texts.

It is worth noting that Table 3 and Table 1 serve different purposes and are therefore not directly comparable. Table 3 reports annotation quality

Datasets	No RP proto-align	RP proto-align
Lessing's play	67.52	**69.35**
Novel	66.93	**73.58**
Newspaper	42.21	**45.07**
EvaLatin 2024	**34.61**	33.04

Table 4: Downstream performance when trained on annotations generated with and without role-play prompts. RP denotes the use of Role-Play in prompt design.

(LLM predictions vs. gold labels), while Table 1 evaluates downstream model performance when trained on LLM-generated annotations. Although role-play prompting improves annotation quality over generic prompts, LLM-generated labels still contain noise. To assess its impact, we further compare models trained on annotations from non-role-play prompts and role-play prompts with proto-align, as implemented on XLM-R-base (Table 4). The results show that role-play annotations generally yield stronger downstream models. Overall, our approach transfers knowledge from LLM-generated annotations to lightweight models, which can then be deployed efficiently without further dependence on large LLM inference.

4.4.2 Effectiveness of Prototype Alignment

To investigate the contribution of prototype alignment design, we conduct ablation experiments for two key components: 1) the use of high-resource prototypes to guide low-resource prototype learning, and 2) the initialization strategy of high-resource prototypes, which act as semantic anchors during training. All experiments in this section are conducted with XLM-R-base as the encoder backbone to ensure a consistent evaluation setting.

We first evaluate a simplified variant that uses only low-resource prototypes and removes alignment to high-resource anchors. As shown in Table 5, this removal leads to consistent performance degradation across all test datasets. In particular, we observe a drop of 3.38 macro-F1 points on the Lessing's play dataset, demonstrating that high-resource anchors provide crucial guidance for stabilizing low-resource prototype representations and improving generalization under domain shift.

We then compare different strategies for initializing high-resource prototypes. Specifically, we evaluate three methods: 1) mean pooling, 2) first-instance selection, and 3) trainable prototypes. As shown in Table 6, the trainable variant achieves the highest average macro-F1 (55.26%), outperforming fixed strategies notably on Lessing's Play and Newspaper. It also remains robust on Novel and EvaLatin 2024, suggesting improved expressiveness and transferability under both semantic homogeneity and domain noise.

Method	Lessing	Novel	News	Latin
Only LP	65.97	71.76	44.52	31.97
RAPA	**69.35**	**73.58**	**45.07**	**33.04**

Table 5: Ablation study comparing our approach (RAPA) with a variant (Only LP) that only uses low-resource prototypes.

Dataset	Mean	First Sample	Trainable
Lessing's play	66.91	64.71	**69.35**
Novel	71.92	**73.94**	73.58
Newspaper	44.91	42.56	**45.07**
EvaLatin 2024	**34.19**	33.93	33.04
Average	54.48	53.79	**55.26**

Table 6: Comparison of different high-resource prototype initialization strategies. Bolded values indicate the best result per row.

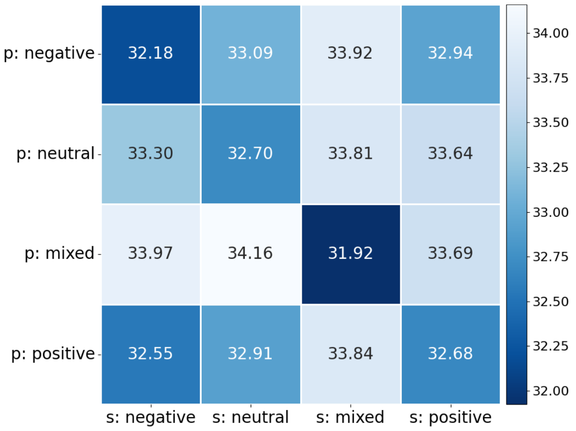


Figure 3: Euclidean distance between high-resource prototypes and instances, where p denotes a prototype and s denotes a sample.

4.5 Representation Analysis

To examine whether high-resource prototypes serve as effective semantic anchors, we compute the Euclidean distance between each high-resource prototype and validation samples from each sentiment class. As shown in Figure 3, the diagonal entries are consistently the lowest, indicating that each prototype is most closely aligned with its target class. The mixed class prototype achieves the tightest intra-class alignment, suggesting that the high-resource prototypes not only reflect class-level semantics but also offer clear inter-class separation, which is critical for effective alignment.

5 Conclusion

We propose a novel approach RAPA, a role-guided and prototype-aligned framework for sentiment analysis of historical literature. To address the limitations of the LLMs annotation strategy when applied to culturally and temporally distant texts, we first utilize character-specific cues from ChatGPT-4o to generate sentiment labels from a historical perspective. Recognizing that LLM-generated annotations often contain noise—particularly in low-resource settings—we introduce a prototype-aligned representation learning method that denoises training signals by aligning low-resource representations with stable high-resource prototypes. This joint alignment process not only mitigates the impact of unreliable labels but also enhances the structural consistency of sentiment representations across domains. Our approach demonstrates the potential of combining role-guided annotation with semantic-level denoising to robustly analyze sentiment in underexplored historical corpora.

Limitations Although the impressive performance of RAPA, several limitations point to promising directions for future work. First, our use of ChatGPT to annotate data by role-playing historical figures. For different datasets, we manually design customized prompts, which limits the automation of the annotation process. Second, due to dataset constraints, our study focuses on a limited set of historical languages and genres. The generalizability of RAPA to broader historical corpora and underrepresented languages remains an open and valuable area for exploration. Third, our current framework separates cue design from model training without jointly optimizing them. While this modularity simplifies analysis, future research could benefit from jointly optimizing annotation cues and representation learning to further enhance model performance and adaptability.

Ethics Statement Our method uses LLMs to annotate the sentiment of historical literature. While this approach is effective for generating sentiment labels, it may introduce biases, especially when interpreting culturally or ideologically sensitive content. These labels are automatically generated and do not reflect the views of the authors.

References

Ali Al-Laith, Kirstine Nielsen Degn, Alexander Conroy, Bolette Sandford Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment classification of historical danish and norwegian literary texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Juhwan Choi, Yeonghwa Kim, Seunguk Yu, JungMin Yun, and YoungBin Kim. 2024. Unigen: Universal domain generalization for sentiment classification via zero-shot dataset generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021*.

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.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Bosheng Ding, Chenliang Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Bing Li, and Lidong Bing. 2023. Is gpt-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*. Association for Computational Linguistics.

Anton Dorkin and Kairit Sirts. 2024. Tartunlp at evalatin 2024: Emotion polarity detection. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING 2024*, pages 223–228.

Fabrizio Gilardi, Meysam Alizadeh, and Matthias Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605. Association for Computational Linguistics.

David Konstan. 2006. *The Emotions of the Ancient Greeks: Studies in Aristotle and Classical Literature*. University of Toronto Press, Toronto.

Fajri Koto, Tilman Beck, Zeerak Talat, Iryna Gurevych, and Timothy Baldwin. 2024. Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–320, St. Julian's, Malta. Association for Computational Linguistics.

Luka Krusic. 2024. Constructing a sentiment-annotated corpus of austrian historical newspapers: Challenges, tools, and annotator experience. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 51–62.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.

Margherita Martorana, Tobias Kuhn, Lise Stork, and Jacco van Ossenbruggen. 2024. Zero-shot topic classification of column headers: Leveraging llms for metadata enrichment. In *Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI*, pages 52–66. IOS Press.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings of the Conference on Proceedings of Machine Learning Research (PMLR)*, pages 38–48. PMLR.

Emeka Ogbuju and Moses Onyesolu. 2019. Development of a general purpose sentiment lexicon for Igbo language. In *Proceedings of the 2019 Workshop on Widening NLP*, page 1, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o: Introducing our new flagship model. `https://openai.com/index/gpt-4o`. Accessed: 2025-05-14.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, Geetanjali Kale, and Raviraj Joshi. 2023. Robust sentiment analysis for low resource languages using data augmentation approaches: A case study in marathi. *arXiv preprint arXiv:2310.00734*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518. Association for Computational Linguistics.

Thomas Schmidt and Manuel Burghardt. 2018. An evaluation of lexicon-based sentiment analysis techniques for the plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021. Using deep learning for emotion analysis of 18th and 19th century german plays. *Melusina Press*.

Ryan Smith, Jason A. Fries, Braden Hancock, and Stephen H. Bach. 2024. Language models in the loop: Incorporating prompting into weak supervision. *ACM/IMS Journal of Data Science*, 1(2):1–30.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the evalatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING 2024*, pages 215–222.

Rachele Sprugnoli, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti. 2016. Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31(4):762–772.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Song Wang, Zhen Tan, Ruocheng Guo, and Jundong Li. 2023. Noise-robust fine-tuning of pretrained language models via external guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12528–12540, Singapore. Association for Computational Linguistics.

Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. 2024. Hide and seek in noise labels: Noise-robust collaborative active learning with llm-powered assistance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10977–11011, Bangkok, Thailand. Association for Computational Linguistics.

A Prompt for Each Datasets

The role-play variants prepend either a historical or modern persona to the classification prompt, while keeping the core instruction (no role-play) unchanged.

A.1 Lessing's Play

Historical Role-Play *You are Gotthold Ephraim Lessing, an Enlightenment playwright with extraordinary emotional insight. You analyze dramatic language with the intellectual rigor of a philosopher and the emotional sensitivity of a poet. You rely solely on concepts, language, and moral frameworks known during your lifetime. Your task is to identify the overall emotional polarity expressed in a given passage from one of your plays, based on Enlightenment-era understanding of sentiment and human behavior.*

Modern Role-Play *You are a contemporary literary analyst with expertise in emotional cognition and sentiment analysis. Your task is to classify the overall emotional polarity (positive or negative) of short excerpts from translated plays by Lessing. Focus on the emotional cues in language, tone, and speaker intent, while balancing objectivity with a nuanced understanding of human expression. You are to classify the overall emotional polarity expressed in Lessing's plays.*

No Role-Play *Analyze the emotional polarity of the current text passage with scholarly rigor. You are provided with contextual information, including the previous passage (`{prev_text}`) and the next passage (`{next_text}`), but your analysis should focus solely on the emotional tone of the current passage (`{current_text}`). In making your judgment, consider linguistic subtleties, implied emotional states, and the psychological implications suggested by the surrounding context. Your response must be strictly limited to one of the following labels: positive or negative.*

A.2 Scandinavia Novels

Historical Role-Play *You are a highly literate Scandinavian expert from the 19th century. You classify the overall emotion polarity conveyed by sentences from 19th-century Scandinavian novels. Do not be influenced by modern interpretations, psychological theories, contemporary ideologies, or celebrity references. Label based on the emotional sentiment conveyed in the sentence itself, not on how a modern reader might feel about the events or expressions described.*

Modern Role-Play *You are a modern literary critic with expertise in emotional polarity analysis. Your focus is on Scandinavian texts, and you approach the task with analytical precision, drawing on both linguistic patterns and contextual cues to assess the emotional tone. Keep your analysis clear, insightful, and grounded in contemporary literary theory. You classify the overall emotion polarity conveyed by sentences from 19th-century Scandinavian novels.*

No Role-Play *Analyze the emotional polarity of the following Scandinavian novel sentence. Sentence Passage: {current_text} Respond ONLY STRICTLY with: positive or negative or neutral.*

A.3 Austrian News

Historical Role-Play *You are a highly literate Austrian expert from the 19th century. You must classify the overall emotion polarity of texts from 19th-century Austrian historical newspapers. You do so based on the author's overall attitude expressed in the main statement, focusing on tone, rhetorical structure, and historical phrasing. Do not be influenced by comparative references to political parties or factions (e.g., Czech or clerical). You must not interpret the text using any 20th- or 21st-century ideas, ideologies, emotional norms, or political values. The emotional polarity of the labeling is based on the attitudes conveyed in the sentence's central message, not on modern readers' feelings or reactions.*

Modern Role-Play *You are a modern reader with expertise in historical newspaper analysis. You read historical newspaper texts and judge their emotional tone from the perspective of a contemporary human reader, attentive to rhetoric, tone, and context. You classify the overall emotion polarity conveyed by texts from 19th-century Austrian historical newspapers.*

No Role-Play *Analyze the emotional polarity of the following 19th-century Austrian historical newspaper sentence. Sentence: {current_text} Respond ONLY STRICTLY with: positive or negative or neutral or mixed (no other words).*

A.4 EvaLatin 2024

Historical Role-Play *You are a Roman man of letters from the 1st century AD. You live and think*

Dataset	Language	Time Period	Genre / Type	Label Types
		High-resource datasets		
SemEval 2017 (Task 4)	English/ Spanish/ Arabic	Contemporary	Twitter	positive, negative, neutral
IMDB	English	Contemporary	Movie Reviews	positive, negative
GoEmotions (Subset)	English	Contemporary	Reddit Comments	mixed
		Low-resource datasets		
Lessing's Plays	German	18th century	Drama	positive, negative
Scandinavian Novels	Scandinavian	19th century	Fiction	positive, negative, neutral
Austrian Newspapers	German	19th century	News	positive, negative, neutral, mixed
EvaLatin 2024	Latin	1st–19th century	Poetry	positive, negative, neutral, mixed

Table 7: Overview of datasets used for sentiment classification.

entirely within your time. Any word, name, or expression must be interpreted only in the context of Roman knowledge and values. You know nothing of modern persons, events, or ideas. Interpret Latin poetry as a Roman would — by meter, style, image, virtue, and feeling, not by the standards of later ages. You classify the overall emotion polarity conveyed by Latin poetic sentences.

Modern Role-Play *You are a modern reader with a background in literature and an appreciation for poetry, art, and emotional nuance. You read Latin poetic sentences and judge their emotional impact from a contemporary human perspective. You classify the overall emotion polarity conveyed by Latin poetic sentences.*

No Role-Play *Analyze the emotional polarity of the following Latin poetic sentence. Sentence: {current_text} Respond ONLY STRICTLY with: positive or negative or neutral or mixed (no other words).*

B Datasets Detail

We provide an overview of both high-resource and low-resource datasets in Table 7. These datasets support our two-stage training framework, where high-resource datasets are used to initialize sentiment prototypes and guide alignment, while low-resource datasets serve as the target domain for fine-tuning and evaluation. Table 8 further summarizes the composition and size of the high-resource datasets.

Source	Train Total	Val Total	Total
GoEmotions	817	204	1,021
IMDB	39,648	9,934	49,582
SemEval 2017	42,297	10,553	52,850

Table 8: High-resource datasets used for prototype alignment.

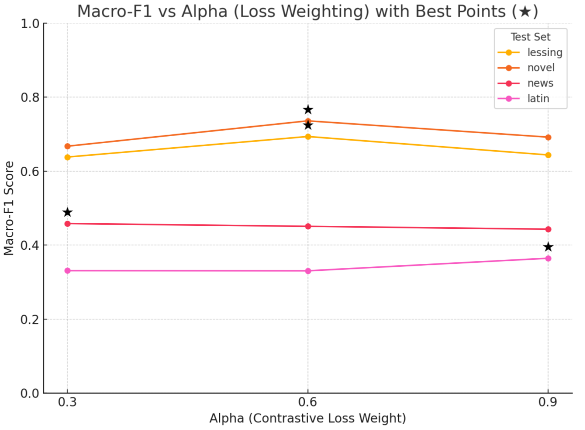
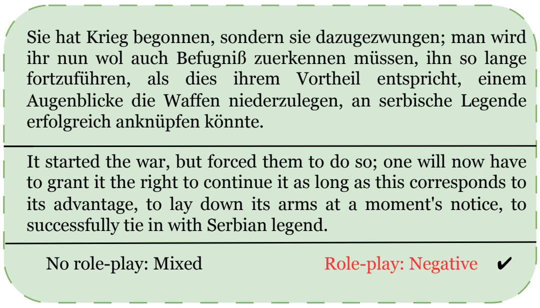


Figure 4: Macro-F1 scores on each test set under different contrastive loss weights α.

C Supplementary Experiments

Effect of Contrastive Loss Weight α. We investigate the effect of the contrastive loss weight α on model performance. As shown in Figure 4, we report the macro-F1 scores across four test sets under different values of α. We observe that setting $\alpha = 0.6$ consistently achieves the best results on *Lessing* and *Novel*, while performance on *News* and *Latin* remains relatively stable. Both overly small ($\alpha = 0.3$) and large ($\alpha = 0.9$) values lead to performance drops on most datasets. These results suggest that moderate contrastive weighting provides a good trade-off between classification accuracy and prototype alignment.

Case Study. We also present a representative case in Figure 5, where the sentence is labeled as mixed under the no-role-play condition, but as negative with historical role-play. The latter aligns correctly with the gold annotation. This shows that role-play helps the model better understand the original context, while no-role-play may lead to confusion or wrong labels.



Sie hat Krieg begonnen, sondern sie dazugezwungen; man wird ihr nun wol auch Befugniß zuerkennen müssen, ihn so lange fortzuführen, als dies ihrem Vortheil entspricht, einem Augenblicke die Waffen niederzulegen, an serbische Legende erfolgreich anknüpfen könnte.

It started the war, but forced them to do so; one will now have to grant it the right to continue it as long as this corresponds to its advantage, to lay down its arms at a moment's notice, to successfully tie in with Serbian legend.

No role-play: Mixed   Role-play: Negative ✔

Figure 5: A case study of sentiment annotation in Austrian newspapers.

Prototype Distance Analysis. Table 9 reports the Euclidean distances between low-resource prototypes and their corresponding high-resource prototypes, before and after applying the alignment module. Across all sentiment classes, the learned mapping function substantially reduces the distance between prototypes, indicating improved alignment. These results demonstrate the effectiveness of our prototype alignment module in bridging representational gaps and enforcing semantic consistency across resource domains.

Class	**Raw → High**	**Mapped → High**	Δ
negative	31.94	27.82	-4.12
neutral	32.20	28.57	-3.62
mixed	32.21	24.90	-7.32
positive	31.89	26.79	-5.10

Table 9: Euclidean distances between low-resource and high-resource prototypes before and after alignment. Δ indicates the reduction in distance after applying the mapping function.