

HCASO: Hybrid Cross-Lingual Attention with Contrastive Alignment and Optimized Lexical Fusion for Bilingual E-Commerce Sentiment Analysis

Himaani Srivatsava and Gangadevi Ganesan*

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai 600089, India

hs0857@srmist.edu.in, gangadeg@srmist.edu.in

Abstract. Shared-encoder multilingual models that confound typologically different language representations scarcely support cross-lingual sentiment analysis. This article introduces HCASO (Hybrid Cross-Lingual Attention with Contrastive Alignment and Optimized Lexical Fusion for Sentiment Analysis) as a unified bilingual model, which consists of four elements: (i) language-specific English and Chinese ELECTRA encoders, (ii) cross-lingual attention that allows structured inter-language communication, (iii) a symmetric InfoNCE contrastive loss in a 256-dimensional space that allows the languages to achieve clearer alignment, and (iv) a two-stage gated lexical fusion module that adaptively combines TF-IDF and sentiment lexicon features with contextual encoder representations via a learned per-dimension mixing coefficient. The model is trained using cross-entropy, MSE intensity regression, and the contrastive objectives. Post-hoc temperature scaling was used to calibrate it. Compared to standard baselines, HCASO achieves the strongest overall balance of performance, with ROC-AUC of 0.9361, 0.8623, and 0.8474 on binary, ternary, and quinary MARC English-Chinese sentiment classification, respectively. HCASO performs best on ordinal consistency ($MSE = 0.0455$, $R^2 = 0.6356$, $Top-2 = 0.8211$) across settings, with XLM-RoBERTa-Base being slightly more accurate in terms of raw accuracy in quinary classification. Ablation experiments with McNemar tests confirmed that all the components - dual encoders ($p < 10^{-26}$ across all settings), lexical fusion ($p < 10^{-11}$ in the ternary setting), and contrastive alignment ($p = 0.028$ in the quinary setting) each contribute significantly. HCASO is a statistically validated alternative to the shared-encoder paradigm, which scales at 214.4M parameters and 10.89 GMACs.

Keywords: Cross-Lingual Sentiment Analysis, Chinese-English Bilingual NLP, Dual-Encoder Transformers, Contrastive Representation Alignment, Gated Lexical Fusion, E-Commerce Review Classification.

* Corresponding author

© 2026 Himaani Srivatsava and Gangadevi Ganesan. This is an open-access article licensed under the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>).

Reference: H. Srivatsava and G. Ganesan, “HCASO: Hybrid Cross-Lingual Attention with Contrastive Alignment and Optimized Lexical Fusion for Bilingual E-Commerce Sentiment Analysis,” *Complex Systems Informatics and Modeling Quarterly*, CSIMQ, no. 48, pp. 52–75, 2026. Available: <https://doi.org/10.7250/csimq.2026-48.03>

Additional information. Author ORCID iD: H. Srivatsava – <https://orcid.org/0009-0002-9431-9962>, G. Ganesan – <https://orcid.org/0000-0003-2339-3989>. PII S225599222600269X. Received: 16 April 2026. Accepted: 23 October 2026. Available online: 31 October 2026.

1 Introduction

Digital media and virtual communication within e-commerce environments have driven an exponential rise in user-generated content, most of all as online reviews or multimodal feedback [1], [2]. Media platforms and social media are now real-time digital feedback channels that represent public opinion. Consumer engagement in these channels has, in turn, become an index of customer satisfaction, brand reputation, and purchase intent, and increasingly influences purchase decisions in real time [3], [4], [5]. These opinion-driven signals directly affect data-driven companies' strategic decisions, product optimization, and competitive analytics [6], [7], [8].

Initially, lexicon-based approaches provided interpretability early on but lacked domain generalization and robustness [5], [7]. Traditional machine learning models such as SVM and Naive Bayes demanded massive feature engineering and had minimal cross-domain transferability [7], [9]. Switching to word representations through distributed means (Word2Vec, GloVe) facilitated semantic encoding [10], which was then succeeded by deep-learning models such as CNNs and Bi-LSTMs, utilizing contextual dependencies in both directions [7], [11]. BERT and RoBERTa models were also trained in many languages and showed significant gains over pre-transformer-based models on multilingual and bilingual e-commerce sentiment classifications [6], [7], [9].

Cross-lingual modeling ability was further improved by multilingual versions (e.g., mBERT, XLM-R) [1], [9], [12]. According to empirical evidence in various studies, transformer-based incremental models score 5 to 12 percent higher in F1-score relative to the traditional CNN/RNN baseline [5], [7], [10]. These models are, however, expensive in terms of computational overhead and complexity of memory, hindering real-time implementation [10], [13]. In addition, the efficiency and stability of fine-tuning large transformer models in low-resource settings remain inefficient and unstable [9].

E-commerce systems are becoming more and more multilingual and code-mixed, with linguistic diversity, morphology, and cultural idioms posing difficulties for sentiment identification [4], [6], [14], [15]. Semantic drift, lexical distribution shifts, and lack of labeled resources are common challenges with cross-lingual sentiment modeling [9], [16]. Translation-based techniques result in information loss, particularly in informal text and code switching [15], [16], [17].

Although a language-independent framework and a shared multilingual embedding space have been proposed in [9], the explicit cross-lingual semantic alignment mechanisms have not been studied [1], [12]. This disparity is more significant in under-resourced and Romanized script languages [14], [16]. Simultaneously, sentiment analysis has been developed to move beyond text-based processing (unimodal) to multimodal processing, with visual and acoustic cues being used [11], [18], [19]. The multimodal sentiment analysis (MSA) offers substantial benefits in terms of predictive performance using the complementary modality cues [18], [20].

Nevertheless, current fusion strategies suffer from several shortcomings, including intermodal imbalance, lack of synchronization, missing modalities, and high computational complexity [21]. Basic concatenation or shared projection models do not provide a highly effective model of modality complementarity on its own [20]. Such blind cross-modal fusion can even cause false correlations and poor robustness [22]. Hierarchical and validated semantic alignment across modalities has not been fully tackled, although cross-attention and cross-modal fusion mechanisms have been found superior in modeling interaction [18], [23].

In addition to document-level polarity, other finer-grained and aspect-based sentiment analysis (ABSA) has been identified as required to generate actionable e-commerce intelligence [24], [25]. The coarse-grained classification masks attribute-level dissatisfaction and domain-centered peculiarities [25], [26]. Aspect misalignment and context-distribution shift substantially reduce classification stability [27]. Although interactive attention models and hybrid transformer/recurrent models enhance contextual modeling [13], [28], there are no multilingual alignment, cross-domain adaptation, multimodal fusion, and computational efficiency solutions.

2 Literature Review

Transformer-based architecture has become the leading methodological paradigm of sentiment analysis, outperforming all traditional machine-learning and recurrent-neural baselines in a variety of domains and languages [10]. Initial comparative analyses have shown that contextual transformer encoders, like BERT, DistilBERT, and RoBERTa, outperform LSTM- and CNN-based baselines in binary e-commerce English language review classification achieving above 90% accuracy under carefully controlled and well-tuned preprocessing conditions [6], [10]. Hybrid models that learn on contextual embeddings in combination with either graph-based or sequential models also achieve significant improvements on aspect-level or document-level sentiment classification, e.g., combinations of BERT with BiGRU and Graph Convolutional Networks (GCN) achieve accuracies of 92.44–95.64% and an AUC of up to 82% on benchmark tasks such as SemEval-2014 and Amazon reviews [28]. These empirical findings have specified the representational strength of contextual encoders but have been exercised mainly in monolingual or in single-domain conditions, though the systematic study of bilingual interaction and hierarchical cross-lingual fusing of features has not yet been done in depth [1], [12].

Hierarchical transformers in the Chinese language have also been implemented in order to learn sentiment dynamics that are discourse-aware [3]. The HCIN model presented by Ding and colleagues, a dual-path RoBERTa-wwm-ext framework that was further enhanced with cross-attention, bilinear interaction, and gated fusion mechanisms, was specially developed to analyze the Chinese population’s opinion about delayed retirement policies [3].

On a curated Weibo dataset, consisting of 3,048 labeled samples and a larger dataset of 179,961 posts, HCIN was found to score extremely well on Macro-F1 under the controlled single-policy discourse conditions with a Fleiss Kappa of 0.73 [3]. Although HCIN already provides a positive example of how contextual interaction can be hierarchical, it is based only on the context of parent-child comments in a monolingual setting with no explicit investigation of the possibility of bilingual encoder segregation or alignment of cross-language representations [3]. An attention has also been given to bilingual e-commerce sentiment modeling. Shan et al. compared BERT and ELECTRA to BiLSTM and classical models on OnlineShopping10Cats, a dataset consisting of 62,000 Chinese product reviews translated into English [6]. According to the study, the ELECTRA achieved 98.09% accuracy and 97.11% F1 if the settings were clean-translated [6]. These figures are impressive, but they are dependent on a machine-translation pipeline, which implicitly presupposes the presence of semantic preservation in the languages [6]. It is in sharp contrast to the direct bilingual modeling strategies and also raises the question whether the linguistics-specific encoders, connected by the developed interaction means, are better equipped to preserve the contextual subtleties not reflected by the artifacts of translation.

These transformer variations with graph additions also provide architectural refinements to the monolingual sentiment analysis. As a demonstration of the effectiveness of structural encoding and Fourier graph operators, Zhang et al. have recommended the Fourier Graph Transformer Network (FGTNet) that achieves accuracies of up to 91.41% on Chinese review corpora [7]. However, these models focus on intra-linguistic structural associations and fail to evaluate bilingual fusion or interaction of cross-language features [7]. Similarly, multimodal transformer designs that multicast cross-attention have demonstrated extraordinary capabilities in task-to-task adjacent tasks. HSIF, a dual-stream transformer with bidirectional cross-attention in cryptocurrency trend forecasting, was introduced by Dashtaki et al. with a 97.48-accurate result in controlled conditions [23]. As much as HSIF highlights the effectiveness of cross-attention to the fusion of modalities, when applied to financial prediction as opposed to multilingual textual sentiment, the question of whether similar mechanisms of bidirectional attention can be adapted to multilingual textual encoding remains unaddressed [23].

Recent multilingual transformer studies have investigated both fine-tuning and prompt-based approaches. CMSA-mBERT, suggested by Nazir et al., offers Roman Urdu-English code-mixed sentiment classification with an accuracy of 79.94% and balanced precision-recall evaluation on

the CMDSA-24 dataset [1]. Ullah et al. proposed a prompt-based fine-tuning method [9] with XLM-RoBERTa to perform language-independent sentiment analysis across 8 languages and obtained an F1-score of 0.71 in a few-shot setting [9]. These studies point to the potential of multilingual pretraining, but they all use a single common encoder and do not explicitly estimate language-specific encoders interacting structurally with each other, and the space of representation is effectively shared and not explicitly connected by learned bidirectional attention [1], [9], [12].

Zhang et al. proposed a SKEP-based cross-domain Chinese aspect-based sentiment classification architecture, which is an adversarial GAN-based architecture with an 85.75% accuracy and small cross-domain improvements [25]. Despite the ability to change domains, as shown in this study, they focus intra-linguistically and fail to adopt bilingual encoder separation or cross-attention-based interaction as an important feature of the architecture [25].

Ensemble and hybrid methods have reached high benchmark levels of performance in controlled conditions within the greater context of research. The FastXCatStack by Hashmi and Yayilgan is built with FastText embeddings, XGBoost, and CatBoost, and it provides accuracies of up to 0.94 on Amazon Review datasets using various sentiment granularities [29]. Equally, BERF combines BERT embeddings and Random Forest probability scores on helpfulness, achieving an accuracy score and F1 of around 0.98 [8]. Although these ensemble systems provide strong predictive capability, they focus on classifier stacking over structural bilingual interaction or hierarchic feature fusion in transformer systems [8], [29].

Systematic reviews also demonstrate that although research in multilingual sentiment analysis continues to grow, most studies are based on translation pipelines or multilingual transformers using only a single encoder that lacks the structure of cross-language interaction [16]. Mabokela et al. mention that about 35% of many multilingual models rely on machine translation, but 61–62% have deep-learning structures that do not explicitly represent cross-lingual alignment other than through shared embeddings [16]. This fact indicates that there is a paradigm difference between pretraining in multilingual and explicit bilingual interaction mechanisms within transformer models [16].

Altogether, the body of previous research shows that there have been the following five key research gaps:

1. Lack of explicit cross-linguistic semantic correspondence in multilingual transformer-based e-commerce systems [1], [9], [12], [16].
2. Not strong in low-resource and code-mixed environments, especially when lexical and cultural variation is present [14], [15].
3. Excessive use of transformer architectures, which are computationally intensive and therefore do not scale well to deployment [10], [13].
4. Poor hierarchical multimodal fusion processes, which lead to imbalances between modalities and spurious correlations [20], [21], [22].
5. Lack of fine-grained cross-domain and context-sensitive sentiment modeling, including in interpretability and transferability [25], [27], [29].

Even though transformer-based models achieve much better contextual modeling [6], [7] in multilingual and cross-domain e-commerce settings, they have limited performance because they do not have explicit cross-lingual alignment mechanisms [1], [12]. Moreover, the current multimodal fusion techniques have disadvantages such as intermodal disproportion and computational overhead [21], [23].

In order to address these limitations, the Hybrid Cross-Lingual Attention with Contrastive Representation Alignment and Lexical Fusion (HCASO) framework is presented in this article; it is a unified architecture that incorporates the language-specific transformer encoders, the interaction of cross-lingual attention, the alignment of representations through contrastive representation, and the gated infusion of lexical features to predict customer sentiment.

The framework utilizes the following components:

1. A dual-encoder system with language-specific ELECTRA transformers for English and Chinese that allows language-specific contextual representation learning without the constraints of shared tokenization.
2. A cross-lingual attention process, where each of the languages encodes the CLS tokens in a way that their explicit encoding into each language considers the other language context and creates a structured semantic interaction between the two encoder streams.
3. A contrastive alignment task that draws English and Chinese representations of similar sentiment polarity nearer in a shared projection space without the need to have parallel corpora.
4. An adaptive gated lexical fusion module to combine projected TF-IDF and sentiment lexica of the gated contextual representation with a learned per-fusion dimensional mixing coefficient, without destroying discriminative surface-level cues as well as deep contextual representations.

While HCASO does not propose a new architectural primitive, it is important to note that the essential element is the principled integration of these four components, and the statistical isolation of the effects of each component through ablation (Section 5.4).

3 Methodology

From the perspective of representation learning, cross-lingual sentiment classification involves the simultaneous resolution of two inconsistent goals: those of aligning together semantically equivalent sentiment expressions throughout typologically various languages, while retaining the contextual nuances, which the shared encoder smooths over. This natural trade-off is particularly apparent in the case of English and Chinese, which have fundamental differences in the morphological composition, tokenization granularity, and definition of word boundaries and structures in syntactic organization. The proposed HCASO framework solves this challenge by a well-designed integration of language-specific contextual encoding, two-way cross-lingual attention interaction, contrastive representation alignment, and adaptive lexical feature fusion.

Formally, suppose $D = \{(x_i, y_i)\}_{i=1}^N$ is a bilingual corpus of customer reviews, where $y_i \in \{0, 1, \dots, C-1\}$ denotes a discrete sentiment label. Three classification granularities are considered: Binary ($C = 2$), Ternary ($C = 3$), and Quinary ($C=5$). Each instance carries the continuous sentiment intensity label target $s_i = y_i/(C-1) \in [0,1]$, utilized as supervision for an auxiliary regression objective that guarantees ordinal consistency within the learned representation space. An associated language identifier $\ell_i \in \{0,1\}$ directs the instance to the corresponding language-specific encoder.

3.1 HCASO Architecture Overview

HCASO architecture has five sequential functional components. Initially, language-specific contextual embeddings are given by a dual encoder representation module based on independently parameterized transformers. Subsequently, both encoder outputs are brought into the same semantic subspace using a shared projection head to allow for cross-linguistic contrastive alignment. This is followed by a gating mechanism based on a sigmoid that is used to selectively modulate the contextual signal before fusion. Next, there is an adaptive lexical fusion gate, which combines projected TF-IDF and lexicon-derived features with the gated contextual representation with a learned per-dimension mixing coefficient. Finally, categorical sentiment probabilities as well as a continuous estimate of the sentiment intensity are generated from the fused representation using dual prediction heads. The end-to-end pipeline is shown in Figure 1.

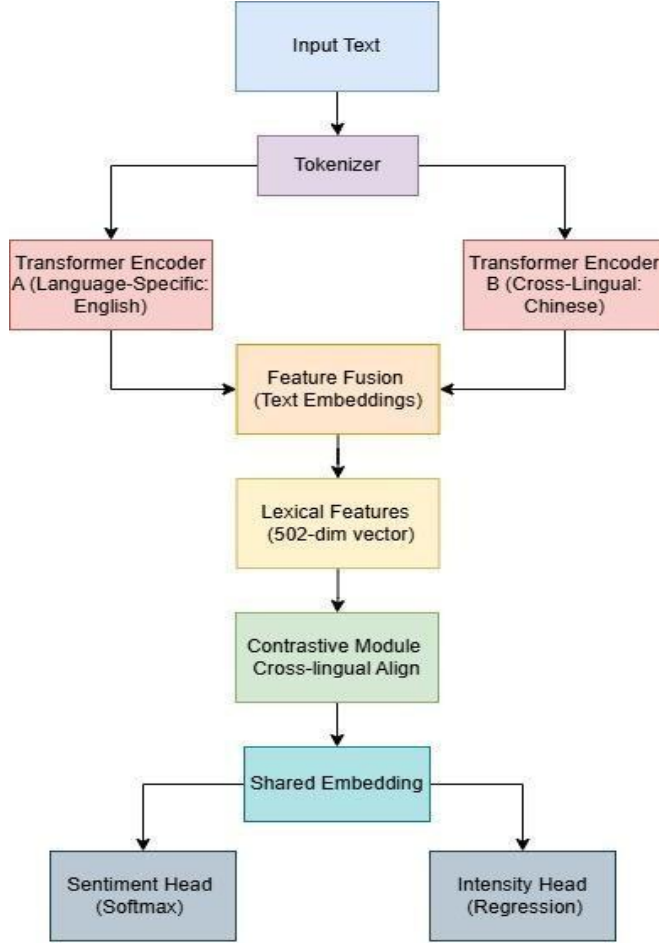


Figure 1. Overview of HCASO architecture

3.2 Dual Encoder Representation

Language-specific contextual embeddings are extracted with two ELECTRA-based discriminator encoders with different parameters. The CLS token representations of an English input x_{en} and a Chinese input x_{zh} are computed as:

$$h_{en} = \text{Encoder}_{en}(x_{en}) \quad (1)$$

$$h_{zh} = \text{Encoder}_{zh}(x_{zh}) \quad (2)$$

where $h_{en}, h_{zh} \in \mathbb{R}^{768}$. The English encoder is adapted from google/electra-base-discriminator, while the Chinese encoder is based on hfl/chinese-electra-base-discriminator. Both encoders are end-to-end fine-tuned using a small learning rate of 2×10^{-5} with respect to the task-specific layers, as in standard discriminative fine-tuning for transformers.

The empirical and structural rationale for using separate encoders follows from typological differences between the two languages. English uses whitespace-delimited tokenization and an analytic subject-verb-object grammatical structure, while Chinese requires character-level segmentation (which we implement here using the Jieba word segmentation tool) and a topic-prominent grammar and little morphological variation.

A single encoder trained on both languages must be able to deal with incompatible tokenization schemes in a common embedding space, inevitably compressing the language-specific distributional characteristics. By feeding a corresponding encoder with each input, HCASO retains the full representational capacity of each pretrained contextual model before two representations interact cross-lingually. At runtime, each input in a batch is routed to its corresponding language encoder $\ell_i \in \{en, zh\}$. A mixed-language batch is processed by running a forward pass through

each encoder separately on the samples belonging to its language. The resulting outputs are scatter-masked, and then assigned back into their original batch positions to form a single universal batch tensor $\mathbf{h} \in \mathbb{R}^{B \times 768}$.

3.3 Shared Projection Space and Cross-Lingual Alignment

To support the meaningful comparison and contrasting learning between English and Chinese representations, two encoder outputs are projected into a shared 256-dimensional subspace and normalized to unit norm by a 2-layer projection head with ReLU nonlinearity and L2 norm, as follows:

$$z = \frac{w_2 \sigma(w_1 h)}{\|w_2 \sigma(w_1 h)\|} \quad (3)$$

where $W_1 \in \mathbb{R}^{768 \times 768}$ and $W_2 \in \mathbb{R}^{768 \times 256}$. The unit norm constraint confines all embeddings to the sphere on a 256-dimensional hypersphere, which guarantees that similarity is only a function of angular relations, and which is also invariant to scale: a prerequisite for the symmetric InfoNCE objective described in Section 3.7.

Bidirectional cross-lingual attention is added to enable the representation of each language to use context from the other:

$$z_{\text{en}} = \text{Attention}(h_{\text{en}}, h_{\text{zh}}) \quad (4)$$

$$z_{\text{zh}} = \text{Attention}(h_{\text{zh}}, h_{\text{en}}) \quad (5)$$

This bidirectional formulation ensures that both English and Chinese embeddings are enriched with the other language’s contextual signals and vice versa to provide a structured interaction channel between the two language streams. Notably, this interaction is done at the CLS token level rather than across full sequences and averts the quadratic computational complexity of full cross-attention when cross-attending over 128-token full sequences.

3.4 Lexical Feature Representation

In addition to contextual embeddings, a 502-dimensional lexical feature vector is built to capture the distributional patterns at the surface level as well as sentiment-bearing lexical cues which may not be well represented by these contextual encoders, especially in the case of short or noisy review texts. Term frequency-inverse document frequency weighting is computed over a vocabulary of 500 terms, fitted only on the balanced training subset to avoid information leakage.

The resulting 500 dimension TF-IDF vector is enriched with 2 scalar sentiment polarity features, an English surface level sentiment score (the difference of count of positive and negative keyword occurrence), and a Chinese character level sentiment score (based on a curated lexicon of 10 positive {好, 很好, 喜欢, 满意, 棒, 优秀, 开心, 推荐, 值得, 方便} and 10 negative {差, 不好, 失望, 糟糕, 讨厌, 垃圾, 慢, 问题, 坏, 不满意} sentiment terms). The space dimensions of the 502 features are therefore projected into the encoder’s hidden space by a learned linear transformation:

$$f_{\text{lex}} = W_{\text{lx}} \tilde{\mathbf{x}}_{\text{tfidf}} \quad (6)$$

where $W_{\text{lx}} \in \mathbb{R}^{502 \times 768}$. This hybrid lexical representation helps to be sensitive to the frequency patterns of domain-specific words and, moreover, to the sentiment vocabulary in language that varies in the English and Chinese inputs.

3.5 Gated Fusion Mechanism

Direct fusion through simple addition or concatenation of contextual and lexical features could lead to domination by the high-dimensional transformer features, reducing the impact of the lexical features. To address this, HCASO uses a two-stage adaptive gating mechanism. The contextual representation is first elementwise modulated by sigmoid gate g to downsample dimensions with little importance to the context of the situation:

$$g = \sigma(W_g h) \quad (7)$$

$$z_{\text{gate}} = g \odot h \quad (8)$$

where $W_g \in \mathbb{R}^{768 \times 768}$ and $\odot$ denotes element-wise (Hadamard) multiplication. This gating operation is used as a learned filter before the fusion operation in order to selectively preserve informative features from the context before the fusion operation.

A second gating mechanism then computes a gating adaptive mixing coefficient $\alpha \in (0,1)^{768}$, conditioned together with the gated contextual and lexical representations:

$$\alpha = \sigma(W_f[z_{\text{gate}}; f_{\text{lex}}]) \quad (9)$$

$$z_{\text{fused}} = \alpha \odot z_{\text{gate}} + (1 - \alpha) \odot f_{\text{lex}} \quad (10)$$

where $W_f \in \mathbb{R}^{1536 \times 768}$. The per-dimension formulation of α makes it possible to control fine-grained feature integration as different dimensions possess different levels of lexical and contextual sources' informativeness. Though the fused representation vector $z_{\text{fused}} \in \mathbb{R}^{768}$ is used for the unified input at the following prediction modules.

3.6 Sentiment Prediction Heads

Two prediction modules are performed on z_{fused} . The first head of classification is comprised of a two-layer MLP ReLU activation/dropout regularization ($p = 0.3$) and is expressed as a categorical distribution of the sentiment over C classes:

$$\hat{y} = \text{Softmax}(W_c z_{\text{fused}} + b_c) \quad (11)$$

$$\hat{s} = W_r z_{\text{fused}} + b_r \quad (12)$$

$$p = \text{Softmax}(z) \quad (13)$$

$$\hat{y} = \text{argmax}_c p_c \quad (14)$$

where $W_c \in \mathbb{R}^{C \times 384}$ and $W_r \in \mathbb{R}^{1 \times 768}$. The regression head opens an auxiliary supervision signal to enforce the ordinal relation in the embedding space, which causes the representations in adjacent sentiment categories to be close to each other.

3.7 Contrastive Representation Learning

To encourage language-invariant representations of sentences within the shared projection space, a symmetric within-batch contrastive objective is used between English and Chinese samples. Let L2-normalized representations be z_i and z_j , the cosine similarity is given at $\tau = 0.07$ temperature as shown in Equation 15.

$$\text{sim}(z_i, z_j) = \frac{z_i z_j^T}{\tau} \quad (15)$$

The corresponding symmetric InfoNCE loss for some mini-batch is given by:

$$L_{contrastive} = -\log \frac{\exp(\text{sim}(z_i, z_j))}{\sum_k \exp(\text{sim}(z_i, z_k))} \quad (16)$$

Within each mini-batch, projected English representations z_{en} and Chinese representations (z_{zh}) are matched based upon their index up to the count (both ways), and the symmetrical loss is calculated in both directions. For the cases when batches are monolingual because of sampling, the term of contrast is removed without affecting the gradient updates and therefore retains the training stability for imbalanced language distribution.

3.8 Training Objective

The overall objective for training involves using a weighted linear combination of classification, regression and contrastive loss:

$$L_{cls} = -\sum_{c=1}^C y_c \log y^c \quad (17)$$

$$L_{reg} = \frac{1}{N} \sum_{i=1}^N (s^i - s_i)^2 \quad (18)$$

$$L_{total} = L_{cls} + 0.4L_{reg} + \lambda L_{contrastive} \quad (19)$$

The regression weight of 0.4 compensates for the difference in scale between cross-entropy loss and mean squared error. The contrastive weighting factor λ is a tuneable hyperparameter which is investigated in the range of $\lambda \in 0.05, 0.1, 0.2, 0.4$. Class imbalance is handled by a frequency inverse weight in the form of a cross-entropy loss, and the class weights are normalized so that the sum equals 1. Additionally, a `WeightedRandomSampler` is used at the data loading stage to ensure that there are balanced class distributions in mini-batches.

4 Experimental Setup

All the experiments were carried out on the Multilingual Amazon Reviews Corpus (MARC) [30], a large-scale benchmark consisting of product reviews of six languages and five star-rating categories. In this work, only the English and Chinese subsets are chosen for evaluation, inspired by their large typological and structural divergence in lexical and grammatical terms such as morphological complexity, word token granularity, and grammatical organization.

Such an isolated bilingual setup facilitates easier observation of the effects of cross-lingual interaction before expanding the system to allow scale to more general multilingual scenarios. Accordingly, all experiments are done within the context of a controlled bilingual evaluation framework formulated with samples of English and Chinese filtered from the MARC corpus.

To balance the classes and have stable training under GPU memory limitations, a stratified subset with up to 5,000 samples per sentiment class per language is taken from the MARC training split. The former one-to-five star ratings are encoded into class membership as follows: In the case of binary classification, ratings ≤ 2 are assigned to class 0 (negative), and ratings ≥ 4 are assigned to class 1 (positive), where reviews with three stars are omitted. For ternary classification, ratings ≤ 2 , $=3$, and ≥ 4 are considered as classes 0, 1, and 2, respectively. For quinary classification, labels are directly mapped, where the ratings are assigned 0–4.

4.1 Data Preprocessing

All the textual inputs are first lowercased and then cleaned using regular expression filtering to remove URLs and user mentions. The non-alphanumeric characters outside the CJK Unicode block (`\u4e00 - \u9fff`) are removed, thus preserving the Chinese characters but removing the extra punctuation.

For English texts, stopwords removal is done using NLTK English stopwords list. Chinese texts are divided into words using the Jieba word segmentation method, which also adds explicit whitespace delimiters between the tokens before TF-IDF vectorization. This approach is standard processing in Chinese natural language processing.

Each input sample is created by combining the titles of the reviews and body with a whitespace token to form a single text field. A TF-IDF vectorizer with 500 maximum vocabulary terms is trained only on the balanced training set of samples to prevent data leakage. The resulting sparse representation of the features is then combined with 2 scalar features representing sentiment polarity with a final lexical feature vector of 502 dimensions. The vectorizer vocabulary, English keywords list, and the Chinese sentiment lexicon are fixed during the inference. No test set adaptation is performed.

4.2 Baseline Models

Four categories of baseline models are examined for all classification settings to give performance references between different levels of representational capacity.

Traditional machine learning baselines are trained on the 500-dimensional TF-IDF features using Multinomial Naive Bayes and Logistic Regression with L2 regularization and 1000 iterations, and a Linear Support Vector Machine (SVM). The SVM model is then further wrapped with a three-fold CalibratedClassifierCV to obtain calibrated probability estimates for the ROC-AUC and ECE evaluation.

Lightweight neural baselines such as a one-dimensional CNN and GRU-based RNN are included. The CNN used has 1 convolutional layer (kernel size of 3, 100 filters)/adaptive max pooling, and the GRU model has limited hidden states (128). Both models share similar 128-size embedding layers and are trained for 3 epochs with the Adam optimizer and learning rate of 10^{-3} and evaluated using the complete settings of the metrics suite.

The **CNN baseline** follows Equation 20 and Equation 21, and the **RNN baseline** follows Equation 22 and Equation 23.

$$h = \text{ReLU}(W_{\text{conv}} * x + b) \quad (20)$$

$$h_{\text{pool}} = \max(h) \quad (21)$$

$$h_t = \text{GRU}(x_t, h_{t-1}) \quad (22)$$

$$h_{\text{final}} = h_T \quad (23)$$

Compact transformer baselines include DistilBERT-multilingual (distilbert-base-multilingual-cased, 135.3M parameters), BERT-Tiny (<https://huggingface.co/prajjwal1/bert-tiny>, 4.4M parameters), and BERT-Mini (<https://huggingface.co/prajjwal1/bert-mini>, 11.2M parameters) which are finetuned with sequence classification head using AdamW, 3×10^{-5} , and a batch size of 30, for 2 epochs [31], [32].

Cross-lingual transformer baselines include mBERT and XLM-RoBERTa-Base that are trained (3 epochs) with AdamW and a learning rate of 2×10^{-5} , cosine annealing, and 10% linear warmup. These models are pivotal comparators in the sense that they are a representation of the dominant paradigm of multilingualism in the form of a single encoder, against which the proposed dual-encoder HCASO architecture is compared.

4.3 Implementation Details

All experiments are implemented in PyTorch and run on CUDA-enabled GPU with mixed-precision training using torch.cuda.amp.GradScaler. The maximum sequence length is set to 128 tokens. Training is done with batch size 16, with WeightedRandomSampler applied at the data

loader level. Encoder parameters are learned at 2×10^{-5} with weight decay 0.01, and non-encoder parameters are optimized at 5×10^{-5} . Parameter groups are divided into decay and non-decay groups (bias and LayerNorm weights). All experiments are executed with a fixed random seed 42 with deterministic CUDA operations for reproducibility.

The model is trained with a maximum of 10 epochs with AdamW and a differential learning rate algorithm. Cosine annealing schedule with 10% linear warmup is used during the entire training. Gradient norms are bounded to a maximum of 1.0. Early stopping with patience of 3 is used based on validation macro F1, and the best-performing checkpoint is saved.

4.4 Post-Hoc Calibration

Probability calibration is performed through temperature scaling [33]. A scalar temperature parameter T is optimized on the validation set by minimizing cross-entropy loss of scaled logits using the L-BFGS optimizer for 50 iterations.

$$p = \text{Softmax}\left(\frac{z}{T}\right) \quad (24)$$

All probability-based metrics such as ECE and ROC-AUC are calculated using temperature-scaled outputs. The Expected Calibration Error is computed using $M = 15$ equal-width bins.

$$ECE = \frac{\sum_{m=1}^M |B_m|}{n} |acc(B_m) - conf(B_m)| \quad (25)$$

4.5 Ablation Study Design

To systematically quantify the contribution of each architecture component, the following four ablation configurations are defined:

1. The full HCASO model represents the complete proposed architecture.
2. The No Lexical Fusion variant replaces the lexical feature input with a zero vector, eliminating all TF-IDF and lexicon signals.
3. The No Contrastive Alignment configuration disables the contrastive learning objective by forcing the loss term to return None, equivalent to setting $\lambda = 0$.
4. The Single Encoder variant replaces the Chinese encoder with one shared with the English encoder, resulting in no language-specific encoding experience.

All four configurations are trained in the same experimental settings for 2 epochs using the AdamW optimizer at 2×10^{-5} . McNemar’s test was used to assess statistical significance between pairwise comparisons between the full model and each ablation variant. Given paired prediction vectors, a 2×2 contingency table is constructed to classify concordant and discordant prediction outcomes. The formula for the continuity-corrected test statistic is:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \quad (26)$$

McNemar’s test was used to determine statistical significance for the four ablation configurations; all results are presented in Section 5.4, along with Table 4.

5 Results and Discussion

The confusion matrices (Figure 2) show the strong diagonal dominance across all classification granularities. In the binary case, off-diagonal misclassification is minimal and is symmetric for the balanced sampling strategy. In the ternary configuration, the most prominent confusion can be attributed to the neutral class being misclassified as either “positive” or “negative”, signifying the inherent ambiguity of the mid-point sentiment category. In the quinary setting, the misclassifications

are clustered between adjacent ordinal classes, especially between two-star and three-star, and between four-star and five-star ratings.

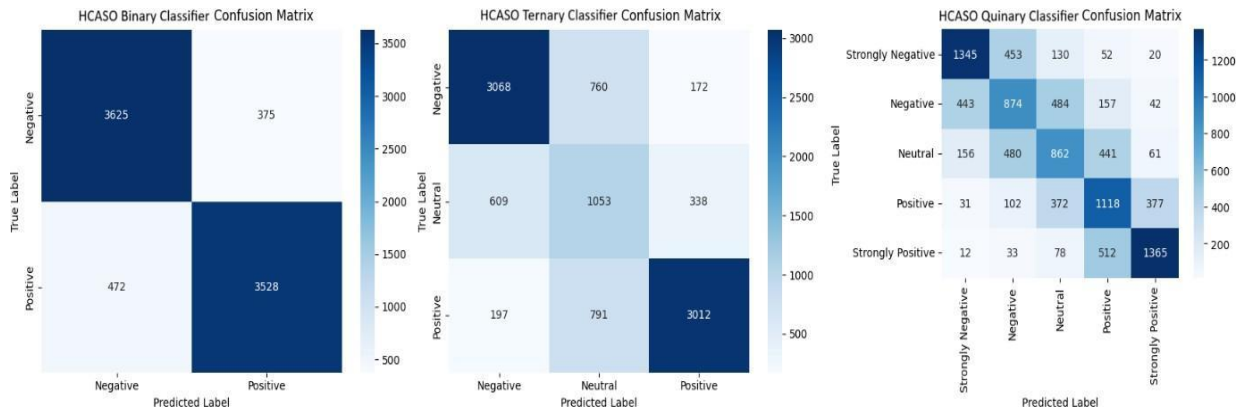


Figure 2. HCASO confusion matrices for binary, ternary, and quinary classification

Table 1 shows an overall evaluation of HCASO for performing binary, ternary, and quinary sentiment classification on the MARC English-Chinese test set. In binary and ternary configurations, HCASO outperforms all the models evaluated. Notably, the ternary Top-2 accuracy of 0.8908 means that even when the Top-1 prediction is incorrect, the true label appears in the top-2-ranked predictions in roughly 89% of cases.

Table 1. HCASO test set performance for binary, ternary, and quinary sentiment classification. ECE↓ and MSE↓ are lower-is-better. Top-2 not applicable (-) for binary classification.

Metric	Binary	Ternary	Quinary
Accuracy	0.8941	0.7133	0.5564
Macro F1	0.8941	0.6792	0.5572
Weighted F1	0.8941	0.7235	0.5572
Micro F1	0.8941	0.7133	0.5564
Precision	0.8944	0.6838	0.5594
Recall	0.8941	0.6822	0.5564
MCC	0.7885	0.5630	0.4458
ROC-AUC	0.9361	0.8623	0.8474
Cohen κ	0.7883	0.5603	0.4455
Top-2 Accuracy	-	0.8908	0.8211
ECE↓	0.0784	0.0698	0.1104
MSE↓	0.1049	0.0882	0.0455
R ²	0.5803	0.5592	0.6356

In the quinary setting, XLM-RoBERTa-Base performs slightly better on the classification accuracy (0.5605 vs 0.5564) while HCASO is better on the Macro F1 (0.5572 vs 0.5559), as well as on the ordinal consistency metrics: MSE (0.0455 vs 0.0558), R² (0.6356 vs 0.5538) and Top-2 accuracy (0.8211 vs 0.8206). The ROC-AUC metric shows a relatively robust measure across all three configurations (0.9361, 0.8623, and 0.8474, respectively), suggesting that HCASO retains strong probabilistic-ranking ability regardless of increasing task complexity.

5.1 Cross-Language Performance Analysis

Accuracy and macro F1 scores are, respectively, given in Table 2 for English and Chinese subsets. In the binary setting, the difference in accuracy between the two language subsets is 1.33 percentage points, and in the quinary setting the difference in accuracy is 2.72 percentage points (macro-F1 gap: 3.51 points), a relatively small difference despite the greater task difficulty.

Table 2. Language-Specific Evaluation of HCASO (English vs. Chinese)

Classifier	English Accuracy	Chinese Accuracy	English F1	Chinese F1	Gap
Binary	0.9008	0.8875	0.9007	0.8875	0.0133
Ternary	0.7158	0.7108	0.6912	0.6649	0.0050
Quinary	0.5700	0.5428	0.5740	0.5389	0.0272

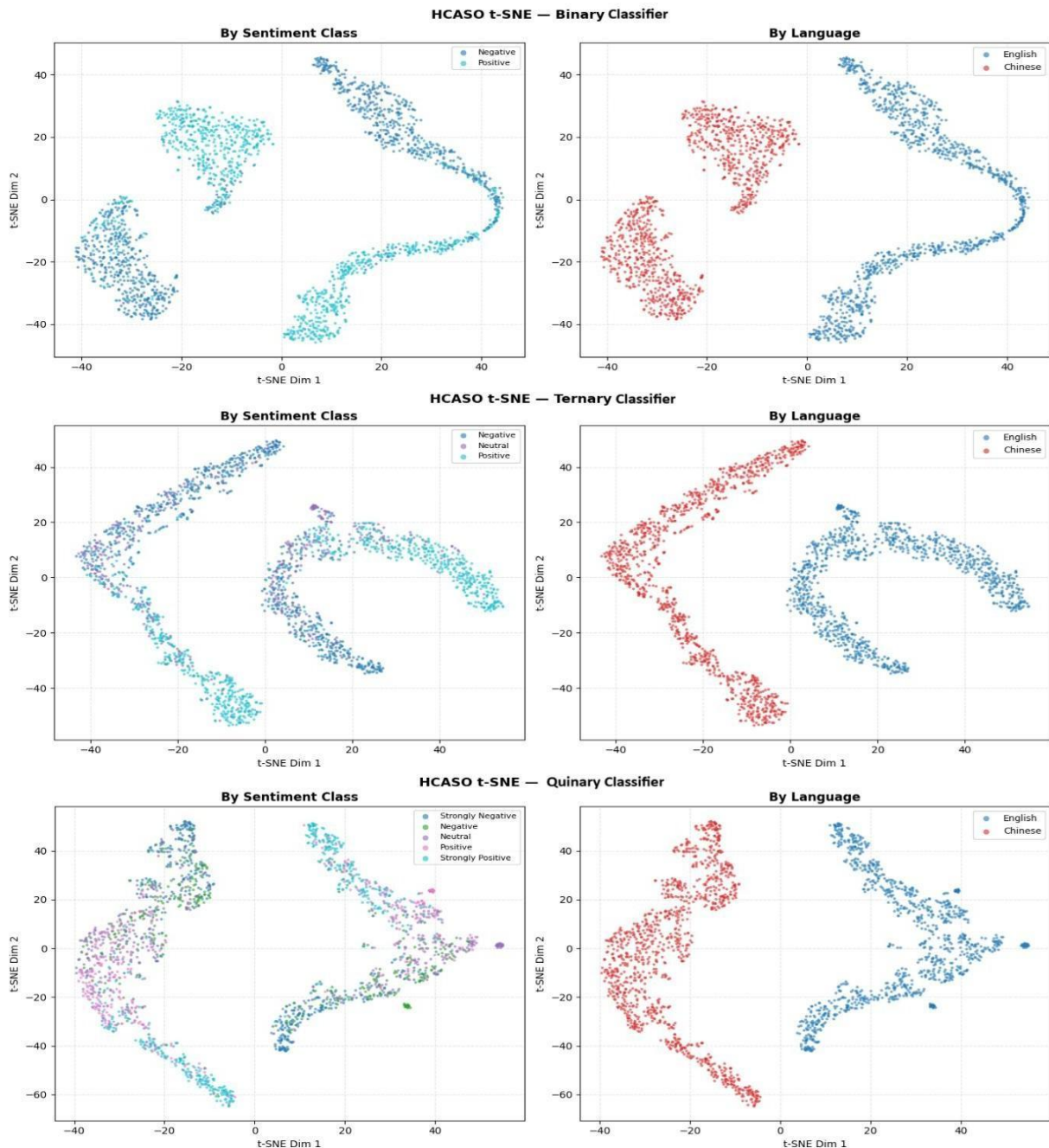


Figure 3. t-SNE projection of HCASO shared representation space

The t-SNE visualization of the shared 256-dimensional projection space (Figure 3) gives qualitative support for these results. The representations across the languages cluster according to the sentiment polarity and not on the linguistic origin. English and Chinese samples associated with the same class of sentiment are mixed within shared clusters, as opposed to forming language-specific groupings. This finding indicates that the contrastive alignment objective can be applied in addressing cross-lingual distributional shifts. A partial overlap on intermediate classes of sentiment is observed in the ternary and quinary cases, particularly on the neutral or the boundary classes, as expected from the ambiguity observed in confusion matrices.

5.2 Comparison with Baseline Models

The **traditional ML baselines** (Naive Bayes, Logistic Regression, and Linear SVM) achieve reasonable binary accuracy (see Table 3), but significantly drop off in the quinary setting, which reinforces the idea that bag-of-words representations cannot capture fine-grained distinctions of ordinal sentiment. **Neural baselines (CNN, RNN)** achieve some improvement over traditional ML; however, they cannot overcome the representational gap without using pre-trained contextual embeddings for bilingual multi-class classification. With a single-encoder design, DistilBERT is the best among **compact transformers** (Table 3) but is outperformed by HCASO by 2.66 and 1.50 percentage points on binary and ternary accuracy, respectively.

Table 3. Comparative evaluation of all baseline models and HCASO across binary, ternary, and quinary sentiment classification on the MARC English–Chinese test set.

Binary classifiers													
Baseline	Accu racy	Macro F1	Weight ed F1	Mirco F1	Precisi on	Recall	MCC	ROC- AUC	κ	ECE↓	MSE↓	R ²	
Naive Bayes	0.8061	0.8061	0.8061	0.8061	0.8062	0.8061	0.6123	0.8909	0.6122	0.0815	0.1939	0.2245	
Logistic Regression	0.8111	0.8111	0.8111	0.8111	0.8113	0.8111	0.6224	0.8960	0.6222	0.0357	0.1889	0.2445	
Linear SVM	0.8086	0.8086	0.8086	0.8086	0.8086	0.8086	0.6173	0.8930	0.6173	0.0184	0.1914	0.2345	
CNN	0.8287	0.8278	0.8278	0.8287	0.8359	0.8287	0.6646	0.9127	0.6575	0.0590	0.1713	0.3150	
RNN	0.8309	0.8309	0.8309	0.8309	0.8309	0.8309	0.6618	0.9099	0.6618	0.0422	0.1691	0.3235	
BERT-Tiny	0.6897	0.6895	0.6895	0.6897	0.6904	0.6898	0.3801	0.7756	0.3795	0.0897	0.3103	-0.2410	
BERT-Mini	0.7461	0.7457	0.7457	0.7461	0.7478	0.7461	0.4939	0.8354	0.4922	0.0224	0.2539	-0.0155	
DistilBERT	0.8651	0.8651	0.8651	0.8651	0.8652	0.8651	0.7303	0.9349	0.7303	0.0488	0.1349	0.4605	
mBERT	0.8742	0.8742	0.8742	0.8742	0.8744	0.8742	0.7486	0.9397	0.7485	0.0444	0.1258	0.4970	
XLM-R	0.8849	0.8849	0.8849	0.8849	0.8849	0.8849	0.7698	0.9484	0.7697	0.0481	0.1151	0.5395	
HCASO	0.8941	0.8941	0.8941	0.8941	0.8944	0.8941	0.7885	0.9361	0.7883	0.0784	0.1049	0.5803	
Ternary classifiers													
Baseline	Accu racy	Macro F1	Weigh ted F1	Mirco F1	Preci sion	Recall	MCC	ROC- AUC	κ	Top-2	ECE↓	MSE↓	R ²
Naive Bayes	0.6287	0.6031	0.6396	0.6287	0.6087	0.6079	0.4371	0.8023	0.4339	0.8561	0.1080	0.1631	0.1845
Logistic Regression	0.6344	0.6048	0.6436	0.6344	0.6085	0.6079	0.4419	0.8051	0.4398	0.8643	0.0239	0.1609	0.1954
Linear SVM	0.6357	0.6029	0.6428	0.6357	0.6048	0.6049	0.4402	0.8021	0.4391	0.8675	0.0518	0.1641	0.1794
CNN	0.5781	0.5743	0.6041	0.5781	0.6385	0.5976	0.4196	0.8097	0.3892	0.8639	0.1406	0.1401	0.2994
RNN	0.6052	0.5976	0.6345	0.6052	0.6464	0.6102	0.4445	0.8179	0.4219	0.8768	0.1069	0.1244	0.3782
BERT-Tiny	0.5382	0.4777	0.5268	0.5382	0.4836	0.4841	0.2675	0.6862	0.2647	0.7591	0.1214	0.2658	-0.3291
BERT-Mini	0.5838	0.5379	0.5833	0.5838	0.5380	0.5379	0.3490	0.7486	0.3490	0.8024	0.0456	0.2112	-0.0561
DistilBERT	0.6853	0.6532	0.6963	0.6853	0.6587	0.6526	0.5213	0.8481	0.5183	0.8984	0.0562	0.1179	0.4105
mBERT	0.6887	0.6583	0.7013	0.6887	0.6652	0.6623	0.5289	0.8510	0.5252	0.9024	0.0529	0.1119	0.4406
XLM-R	0.7108	0.6770	0.7206	0.7108	0.6806	0.6801	0.5587	0.8619	0.5563	0.9115	0.0495	0.1028	0.4862
HCASO	0.7133	0.6792	0.7235	0.7133	0.6838	0.6822	0.5630	0.8623	0.5603	0.8908	0.0698	0.0882	0.5592

Table 3. Continued

Quinary Classifiers													
Baselines	Accur acy	Macro F1	Weigh ted F1	Mirco F1	Precisi on	Recall	MCC	ROC- AUC	κ	Top-2	ECE↓	MSE↓	R ²
Naive Bayes	0.4753	0.4678	0.4678	0.4753	0.4655	0.4753	0.3451	0.7852	0.3441	0.7247	0.0967	0.1014	0.1891
Logistic Regression	0.4681	0.4616	0.4616	0.4681	0.4587	0.4681	0.3358	0.7891	0.3351	0.7231	0.0253	0.1014	0.1892
Linear SVM	0.4699	0.4602	0.4602	0.4699	0.4578	0.4699	0.3387	0.7828	0.3374	0.7121	0.0348	0.1040	0.1678
CNN	0.4887	0.4816	0.4816	0.4887	0.4827	0.4887	0.3624	0.8030	0.3609	0.7435	0.1039	0.0856	0.3151
RNN	0.4956	0.4849	0.4849	0.4956	0.4905	0.4956	0.3725	0.8062	0.3695	0.7564	0.0476	0.0738	0.4094
BERT-Tiny	0.3448	0.2891	0.2891	0.3448	0.3280	0.3448	0.1980	0.6940	0.1810	0.5989	0.0434	0.1915	-0.5317
BERT-Mini	0.4223	0.4140	0.4140	0.4223	0.4130	0.4223	0.2792	0.7465	0.2779	0.6709	0.0389	0.1255	-0.0041
DistilBERT	0.5414	0.5383	0.5383	0.5414	0.5371	0.5414	0.4271	0.8384	0.4268	0.7978	0.0553	0.0638	0.4892
mBERT	0.5510	0.5457	0.5457	0.5510	0.5442	0.5510	0.4395	0.8421	0.4388	0.8015	0.0472	0.0616	0.5069
XLM-R	0.5605	0.5559	0.5559	0.5605	0.5541	0.5605	0.4512	0.8486	0.4506	0.8206	0.0601	0.0558	0.5538
HCASO	0.5564	0.5572	0.5572	0.5564	0.5594	0.5564	0.4458	0.8474	0.4455	0.8211	0.1104	0.0455	0.6356

Bold = best per metric. ECE↓ and MSE↓ are lower-is-better. In the quinary setting, XLM-RoBERTa-Base leads on hard classification metrics while HCASO leads on ordinal consistency (MSE↓, R², Top-2) and calibration (ECE↓). Top-2 is not applicable for binary classification. κ = cohen kappa, XLM-R = XLM-RoBERTa-Base.

Against the **strongest baselines (mBERT and XLM-RoBERTa-Base)** HCASO leads in binary and ternary accuracy and macro F1 (Table 3). In the quinary setting, XLM-RoBERTa-Base has a slight edge over raw accuracy; however, HCASO achieves better on ordinal consistency (higher R², lower MSE), indicating that its performance advantage becomes more pronounced as tasks become more challenging.

5.3 Calibration Analysis

ROC curves in all three settings indicate good discriminative power. The curves of the two classes are found to be very close to the upper left corner in the binary setting, with a micro-AUC of 0.9361. The neutral class has the lowest AUC under the ternary setting because this class is semantically ambiguous, but the overall macro-AUC (0.8623) is still far from the random baseline. In the quinary setting, the per-class AUC shows a clear ordinal pattern in that extreme classes (1-star, 5-star) outperform middle classes, with an overall macro-AUC of 0.8474, demonstrating the stability of the probabilistic ranking of the classes, even under fine-grained classification.

A pattern closely related, but distinct, can be seen with precision–recall (PR) analysis. In the binary case, the average precision (AP) is high and stable for both classes (AP = 0.933). The neutral class has a low AP of 0.484, while the negative and positive classes are still high (0.847 and 0.868, respectively), in the ternary setting. In the quinary setting, the two extreme classes are comparatively strong (AP = 0.705 and 0.701 for strongly negative and strongly positive, respectively), while the intermediate classes are much weaker (AP = 0.479, 0.451 and 0.530 for negative, neutral and positive, respectively), thus confirming the weakness of HCASO in terms of precision-recall trade-off for semantically overlapping intermediate sentiment classes. The full ROC and PR curves for all three settings are given in the Appendix (Figures A1–A2).

The reliability diagrams (Figure 4) illustrate model confidence in relation to the actual accuracy at different settings. ECE values are 0.0784 (binary setting), 0.0698 (ternary setting), and 0.1104 (quinary setting). The ternary classifier is the most calibrated and closely follows the ideal diagonal. The binary classifier is overconfident in a mid-confidence band, with accuracy dropping to 0.39 around confidence 0.65, before it gets back on track, while the quinary classifier is underconfident at low to moderate confidence levels and accurate at high confidence levels – a

characteristic that matches the difficulty of assigning distinct probability mass across five ordinal classes.

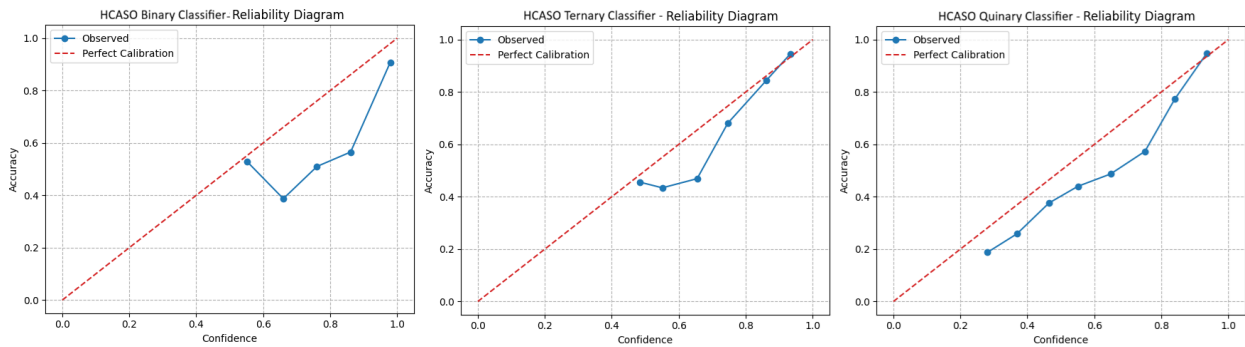


Figure 4. Reliability diagrams for HCASO (binary, ternary, and quinary classifiers)

5.4 Ablation Study Results

The ablation results reflected in Table 4 indicate that the Single Encoder variant leads to the largest performance drop for all settings (binary accuracy drops by -3.95 percentage points, ternary by -5.34% , and quinary by -5.78%), as evidenced by the McNemar test results $p < 10^{-26}$ for all settings, confirming language-specific encoding as the most significant architectural factor. No Lexical Fusion variant makes the second largest drop, particularly in the ternary classification, where the accuracy drops from 0.7042 to 0.6752 ($p < 10^{-11}$). The No Contrastive Alignment variant results in a smaller but still significant drop in the quinary setting ($p = 0.028$), suggesting that the benefit of contrastive alignment increases as granularity of the task increases. Full McNemar contingency heat maps are included in the Appendix (Figure A3).

Table 4. Ablation study comparison of Full HCASO with three ablated variants

Ablated variant	Accuracy	Macro F1	Weighted F1	Micro F1	Precision	Recall	MCC	ROC-AUC	κ	ECE↓	MSE↓	R ²
Binary Classifier												
HCASO	0.8857	0.8856	0.8856	0.8857	0.8877	0.8857	0.7735	0.9474	0.7715	0.0496	0.0935	0.6259
No L.F	0.8857	0.8857	0.8857	0.8857	0.8860	0.8857	0.7718	0.9366	0.7715	0.0702	0.1053	0.5787
No C.A	0.8881	0.8881	0.8881	0.8881	0.8883	0.8881	0.7765	0.9397	0.7762	0.0645	0.0999	0.6000
S.E	0.8462	0.8459	0.8460	0.8462	0.8487	0.8462	0.6949	0.9198	0.6925	0.0639	0.1174	0.5305
Ternary Classifier												
HCASO	0.7042	0.6745	0.7176	0.7042	0.6837	0.6789	0.5542	0.8555	0.5493	0.1382	0.0882	0.5592
No L.F	0.6752	0.6569	0.6954	0.6752	0.6816	0.6686	0.5285	0.8536	0.5149	0.1238	0.0889	0.5554
No C.A	0.6960	0.6663	0.7068	0.6960	0.6787	0.6739	0.5446	0.8548	0.5374	0.1454	0.1014	0.4929
S.E	0.6508	0.6237	0.6650	0.6508	0.6462	0.6294	0.4804	0.8271	0.4710	0.0942	0.1034	0.4830
Quinary Classifier												
HCASO	0.5491	0.5501	0.5502	0.5491	0.5532	0.5491	0.4368	0.8443	0.4364	0.1652	0.0478	0.6176
No L.F	0.5482	0.5401	0.5401	0.5482	0.5409	0.5482	0.4371	0.8377	0.4352	0.1842	0.0535	0.5723
No C.A	0.5388	0.5417	0.5417	0.5388	0.5494	0.5388	0.4246	0.8401	0.4235	0.1507	0.0488	0.6097
S.E	0.4913	0.4811	0.4811	0.4913	0.4810	0.4913	0.3659	0.8073	0.3641	0.0715	0.0579	0.5369
<i>Bold = best per metric. ECE↓ and MSE↓ are lower-is-better. L.F = Lexical Fusion, C.A = Contrastive Alignment, S.E = Single Encoder. Acc = accuracy, Wt.F1 = weighted F1, κ = cohen Kappa.</i>												

The pattern of tradeoffs varies for each classification granularity. In the binary setting, the No Contrastive Alignment variant slightly outperforms Full HCASO on all raw classification metrics

(Accuracy, Macro F1, MCC, Precision, Recall), whereas Full HCASO still performs best on the ROC-AUC and on all of the calibration/ordinal metrics (ECE, MSE, R^2).

In the ternary setting, Full HCASO leads on every classification metric as well as on MSE and R^2 ; its only shortfall is calibration, where Single Encoder achieved the lowest (better) ECE. In the quinary setting, Full HCASO is better on almost every measure with only two marginal discrepancies: No Lexical Fusion variant achieves a slightly higher MCC (0.4371 vs. 0.4368) and Single Encoder variant attained a much lower ECE (0.0715 vs. 0.1652), although it is the weakest variant regarding all other aspects.

We interpret this last result with caution: The apparent calibration advantage of Single Encoder in the quinary setting is probably due to systematically lower confidence scores in all predictions, and not due to better calibrated uncertainty estimates, because the discriminative performance of Single Encoder on the same task is comparatively weak (accuracy 0.4913, R^2 0.5369). The advantage of Full HCASO is more pronounced in ternary and quinary settings, and in the ordinal/ranking metrics in general, than in all settings and metrics.

Per-class F1 analysis (Figure 5) shows that extreme sentiment classes 0 ($F1 = 0.647$) and 4 ($F1 = 0.710$) consistently have better performance than intermediate classes ($F1$ range 0.466–0.499), in line with the already known results in ordinal sentiment classification.

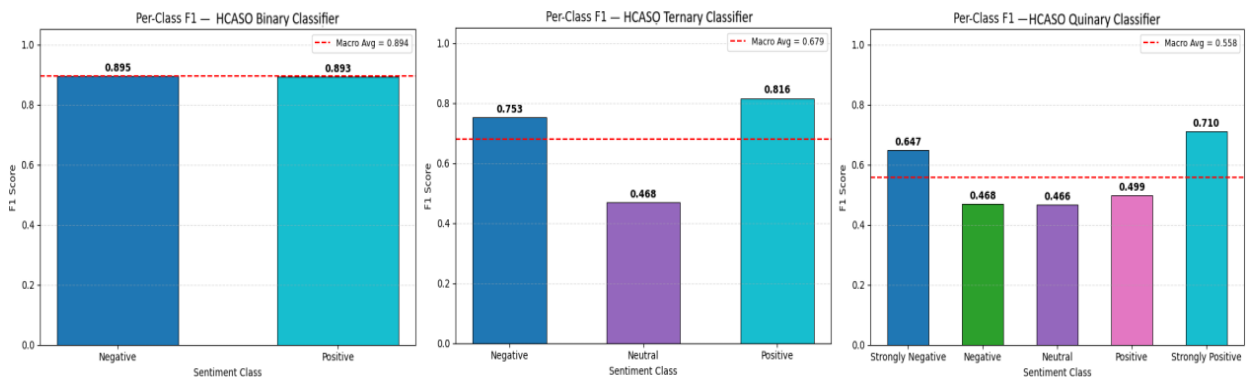


Figure 5. Per-Class F1 scores for three different granularity classifications

5.5 Model Complexity and Efficiency

HCASO runs at 10.89 GMACs per forward pass, and the task-specific components are less than 2 million parameters. This means that the performance gains are not by scaling the parameters but rather by designing the architecture. Inference time analysis (Figure 6) shows that for the HCASO time complexity is about 8.42 seconds per 1,000 samples with 2.7x overhead for XLM-RoBERTa-Base. This trade-off is seen for the steady performance improvements.

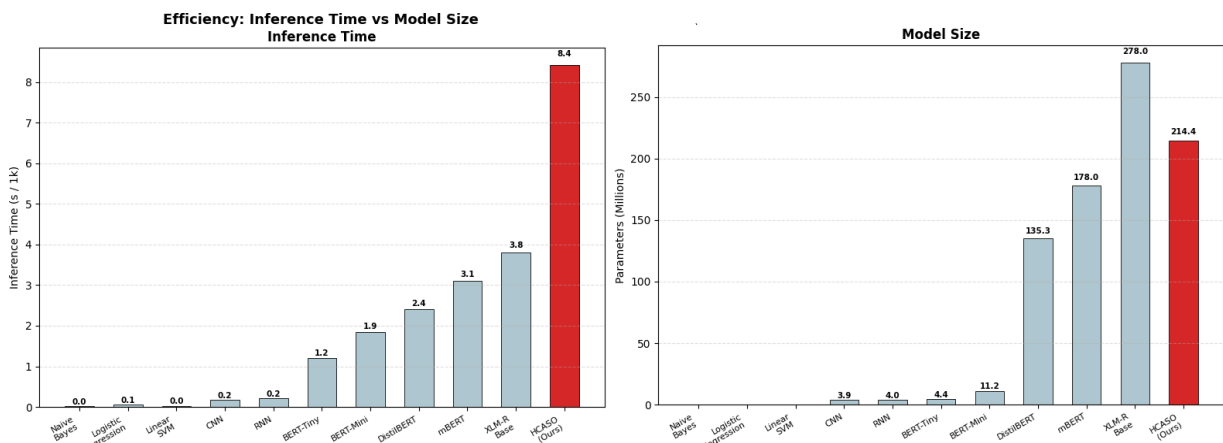


Figure 6. Inference time/1000 samples (left) and number of model parameters (right)

5.6 Robustness and Domain Generalization

On this subset for robustness evaluation, HCASO achieves clean-input accuracies of 0.8941, 0.7105, and 0.5551 (binary, ternary, and quinary, respectively), the latter two of which fall slightly short of the numbers reported in Table 1 as they are measured on a separate robustness-test run, rather than the main test set. These drop to 0.8678, 0.6663, and 0.5166 on the 20% word dropout (2.94%, 6.22%, and 6.94%, respectively), as shown in Figure 7. The cross-domain analysis of the long-review subsample (>200 characters) shows a mixed effect, with an increase in binary accuracy from 0.8941 to 0.9041 ($n = 886$), and a modest decrease in ternary and quinary accuracy from 0.7105 to 0.6950 ($n = 1,128$) and 0.5551 to 0.5452 ($n = 1,128$), respectively. This pattern is broadly supportive of the perspective that longer reviews tend to have more complex argumentative organization and hedged sentiment, making fine-grained (ternary/quinary) sentiment classification slightly more difficult, but coarse-grained binary polarity classification relatively unaffected.

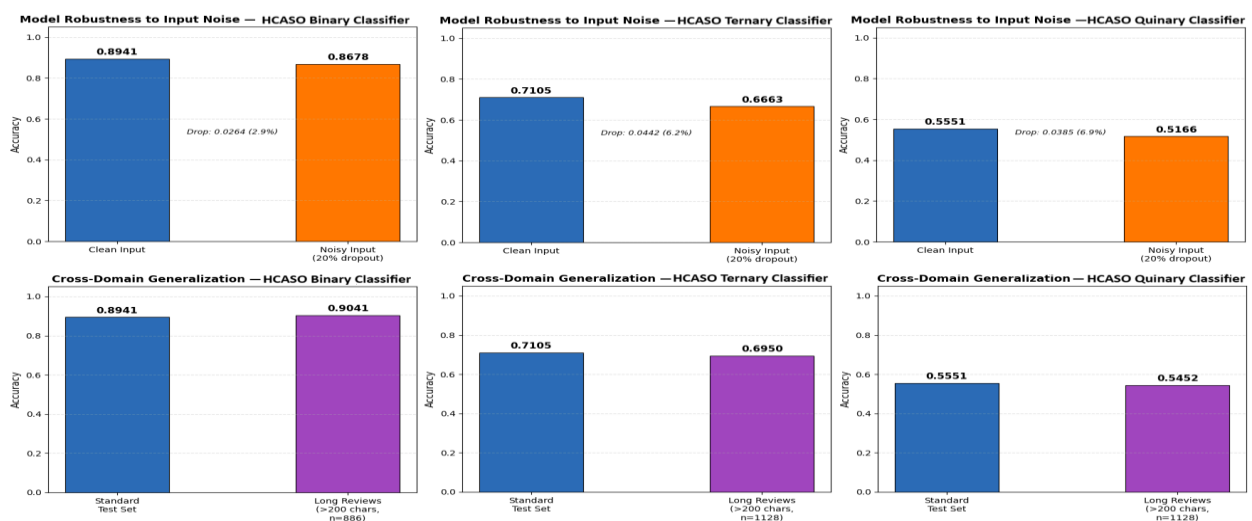


Figure 7. Robustness and cross-domain generalization charts for binary, ternary, and quinary HCASO

5.7 Error Analysis

Two systematic patterns of errors were revealed by the high-confidence error analysis (Figure 8). First, there is an overall mix-up between corresponding ordinal classes: two-star reviews are anticipated as one-star, four-star as five-star – the natural semantic indistinctness of each ordinal polarity delimiting. Second, it has a number of high-confidence cross-polarity errors, which are associated with both short and vocabulary-strained texts where the lexical branch and encoder are both offering uninformative representations.

5.8 Hyperparameter Sensitivity

Validation weighted F1 peaks consistently at $\lambda = 0.20$ for all three modes of classification at 0.8895, 0.7281, and 0.5539 for binary, ternary, and quinary configurations, respectively, (Figure 9) – before declining at $\lambda = 0.40$. The variation of the performance, using different binary/ternary/quinary granularities, is 1.20, 2.82, and 2.33 percentage points, respectively confirming that the value of $\lambda = 0.20$ is not an arbitrary value.

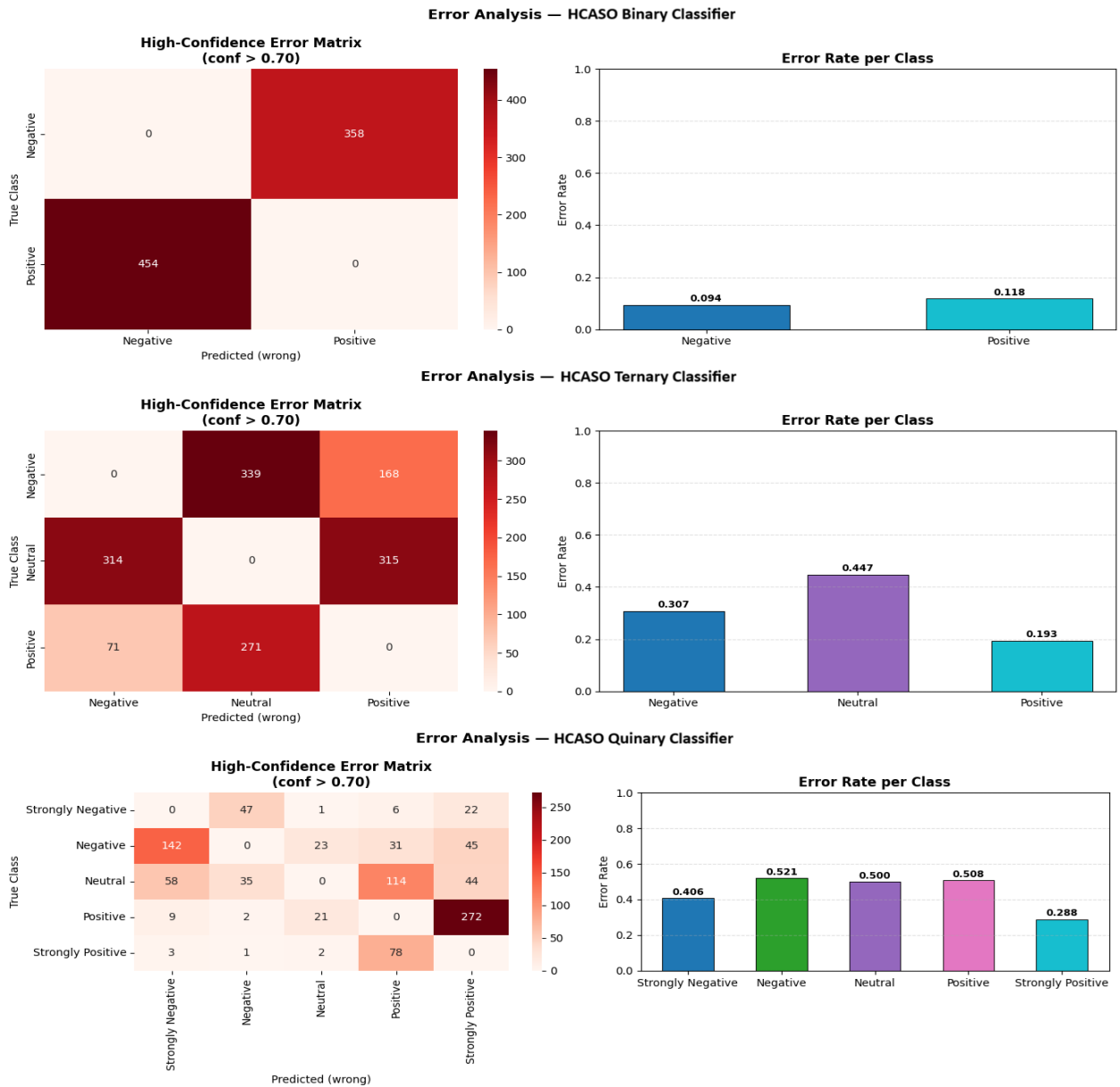


Figure 8. High confidence error analysis. The confusion matrix with predictions greater than 0.70 confidence and erroneous labels is depicted in a heatmap (left side). Error rate per class is shown on the right side.

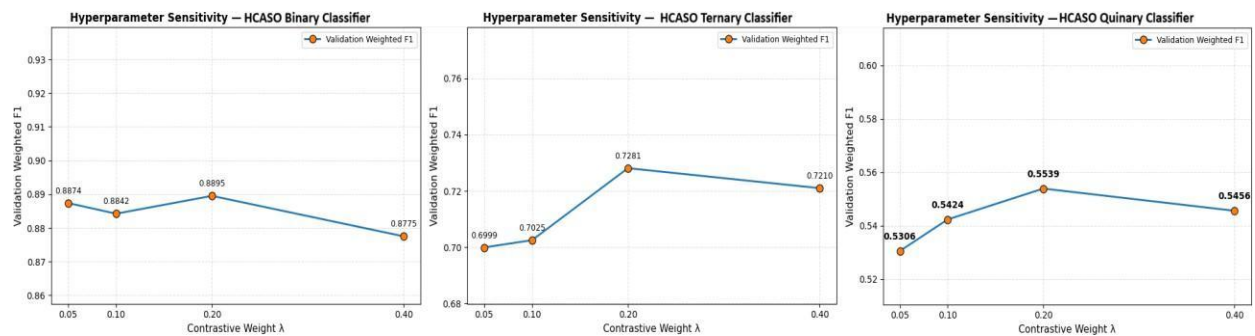


Figure 9. Sensitivity analysis of hyperparameter contrastive weight λ in binary, ternary, and quinary configurations

6 Conclusion

This article introduced a unified framework, named HCASO (Hybrid Cross-Lingual Attention with Contrastive Alignment and Optimized Lexical Fusion), for bilingual sentiment analysis to overcome the representational weakness of shared-encoder multilingual methods.

6.1 Strengths

1. HCASO’s contribution is the principled integration of four mechanisms: language-specific encoding, cross-lingual attention, contrastive alignment, and gated lexical fusion, and statistically isolating the effects of each one of them separately with ablation.
2. This central hypothesis, that dedicated per-language encoding with structured cross-lingual interaction is better than implicit language mixing in shared multilingual encoders, was confirmed in the binary, ternary, and quinary classification of the MARC English-Chinese corpus with statistical validation at each step of the ablation study.
3. In the binary and ternary settings, HCASO is the best performer on every metric.
4. It fares better on every metric than the others in the quinary setting, with the exception of raw accuracy (0.5605 compared to 0.5564 for XLM-RoBERTa-Base).
5. HCASO is consistently best on other properties that are important for downstream tasks that depend on confidence estimates or ranked outputs rather than just the top-1 label: Ordinal consistency (MSE, R^2) and Calibration (ECE).

6.2 Limitations

1. The advantage of HCASO is not as broad as it might appear on the surface in all metrics and regimes but rather is strongest in the higher difficulty (quinary) and calibration-sensitive regimes.
2. The ablation results follow a similar pattern, with the No Contrastive Alignment variant slightly outperforming Full HCASO on a few raw classification measures in binary and ternary settings, and Full HCASO consistently outperforming on the measures of calibration and ordinal consistency across the three granularities.
3. Evaluation is restricted to a stratified subset of up to 5,000 samples per class per language, due to GPU memory constraints, rather than the full MARC corpus, and to a single, typologically distant language pair (English-Chinese).
4. These results cannot be generalized to other language pairs, lower-resource languages, unbalanced real-world data, or truly multilingual (more than two-language) settings.
5. The evaluation does not account for code-switched and intra-sentence mixed-language text that is often found in current e-commerce reviews: the dual-encoder routing mechanism requires that all inputs belong to one of the supported languages, and its behavior for code-mixed input has not been tested.
6. HCASO has ~214.4M parameters and uses 10.89 GMACs per forward pass. This allows for batch sentiment analysis of offline data, and may result in latency limitations in real-time, high-volume deployments without compression or hardware acceleration.

6.3 Future Work

Multilingual and Cross-domain scaling: The training of the entire MARC corpus and the testing of the rest of the four MARC languages (German, French, Japanese, Spanish) would allow for a systematic investigation of the scaling of dual-encoder benefits according to typological distance. Related directions comprise zero-shot transfer to new text domains (e.g., social media, forums, or customer-service text) and extension to less-resourced language pairs, with limited labeled data.

Noise-robust training and code-switching: To extend HCASO to intra-sentence code-mixed English–Chinese text, we make the following three concrete changes: (1) a mixed-language tokenization strategy to segment and tag spans by language before sending them to a router; (2) sub-sentence encoder routing, instead of relying on a single dominant-language assumption per input; and (3) joint contrastive batches constructed from code-mixed samples, such that the alignment objective learns context-sensitive partial correspondence to align spans instead of whole-sentence correspondence. This, together with the noise-injected training (typos, dropped tokens, translation artifacts), would address a practical deployment gap that the current bilingual-separation design does not cover.

The **code-mixed English–Chinese** intra-sentence input was **informally tested** using the deployed demonstration system, and the results indicate that the dual-encoder routing does not fail completely, possibly because each token is still sent to a reasonable encoder depending on the script. The following observation is anecdotal, namely, from ad hoc interactions with the demo interface rather than a controlled evaluation, and it should not be taken as a measure of code-mixed robustness; nonetheless, it does inspire the systematic code-switching evaluation outlined above.

Efficiency: parameter-efficient fine-tuning such as LoRA, adapters, and prefix tuning can help decrease both the training and inference costs and retain the benefits of language-specific encoding.

Data availability

The source code implemented in Google Colab notebooks (<https://github.com/himaanisrivatsava/HCASO>) and the dataset (<https://www.kaggle.com/datasets/mexwell/amazon-reviews-multi>) for all the reported experiments are provided for reproducibility and further extension of the work.

References

- [1] M. K. Nazir, C. N. Faisal, M. A. Habib, and H. Ahmad, “Leveraging Multilingual Transformer for Multiclass Sentiment Analysis in Code-Mixed Data of Low-Resource Languages,” *IEEE Access*, vol. 13, pp. 7538–7554, 2025. Available: <https://doi.org/10.1109/ACCESS.2025.3527710>
- [2] H. Huang, A. A. Zavareh, and M. B. Mustafa, “Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions,” *IEEE Access*, vol. 11, pp. 90367–90382, 2023. Available: <https://doi.org/10.1109/ACCESS.2023.3307308>
- [3] W. Ding, H. Bhuiyan, and M. K. Hossain, “SCOPE: A Transformer-based framework for analyzing sentiment of Chinese public toward delayed retirement,” *IEEE Access*, vol. 14, pp. 25146–25166, 2026. Available: <https://doi.org/10.1109/ACCESS.2026.3664185>
- [4] E. Kahya Özyirmidokuz, B. Molu Elmas, and E. A. Stoica, “AI-Based Sentiment Analysis of E-Commerce Customer Feedback: A Bilingual Parallel Study on the Fast Food Industry in Turkish and English,” *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 4, pp. 294–294, 2025. Available: <https://doi.org/10.3390/jtaer20040294>
- [5] M. F. Almufareh, N. Jhanjhi, N. A. Khan, S. N. Almuayqil, M. Humayun, and D. Javed, “BertSent: Transformer-Based Model for Sentiment Analysis of Penta-Class Tweet Classification,” *IEEE Access*, vol. 12, pp. 196803–196817, 2024. Available: <https://doi.org/10.1109/ACCESS.2024.3515836>
- [6] S. Shan, J. Sun, and R. M. C. Macawile, “Examining Customer Satisfaction through Transformer-Based Sentiment Analysis for Improving Bilingual E-Commerce Experiences,” *IEEE Access*, vol. 13, pp. 51027–51041, 2025. Available: <https://doi.org/10.1109/ACCESS.2025.3551666>
- [7] Y. Zhang, Q. Ma, Y. Zhang, and Y. Liu, “Sentiment Analysis for E-Commerce Product Reviews Using a Fourier Graph Transformer Network,” *IEEE Access*, vol. 13, pp. 194127–194138, 2025. Available: <https://doi.org/10.1109/ACCESS.2025.3632518>
- [8] A. Ishtiaq, K. Munir, A. Raza, N. A. Samee, M. M. Jamjoom, and Z. Ullah, “Product Helpfulness Detection with Novel Transformer Based BERT Embedding and Class Probability Features,” *IEEE Access*, vol. 12, pp. 55905–55917, 2024. Available: <https://doi.org/10.1109/ACCESS.2024.3390605>

- [9] F. Ullah et al., “Prompt-based fine-tuning with multilingual transformers for language-independent sentiment analysis,” *Scientific Reports*, vol. 15, article 20834, 2025. Available: <https://doi.org/10.1038/s41598-025-03559-7>
- [10] I. Duru and A. S. Sunar, “Transformer and Pre-Transformer Model-Based Sentiment Prediction with Various Embeddings: A Case Study on Amazon Reviews,” *Entropy*, vol. 27, no. 12, pp. 1202–1202, 2025. Available: <https://doi.org/10.3390/e27121202>
- [11] V. Naveen and S. Sountharajan, “A Lightweight Multimodal Framework with Dilated CNN Audio Modelling and SE-Enhanced MobileNetV2 for Real-Time Sentiment Analysis of Customer Product Reviews,” *IEEE Access*, vol. 13, pp. 217825–217845, 2025. Available: <https://doi.org/10.1109/ACCESS.2025.3648722>
- [12] N. Raychawdhary, S. Bhattacharya, C. Seals, and G. Dozier, “Empowering Sentiment Analysis in African Low-Resource Languages Through Transformer Models and Strategic Language Selection,” *IEEE Access*, vol. 13, pp. 147859–147873, 2025. Available: <https://doi.org/10.1109/ACCESS.2025.3599480>
- [13] N. Kai Ning Loh, C. Poo Lee, T. Song Ong and K. Ming Lim, “MPNet-GRUs: Sentiment Analysis With Masked and Permuted Pre-Training for Language Understanding and Gated Recurrent Units,” *IEEE Access*, vol. 12, pp. 74069–74080, 2024. Available: <https://doi.org/10.1109/ACCESS.2024.3394930>
- [14] M. Akash Rahman, M. Begum, T. Mahmud, M. S. Hossain, and K. Andersson, “Analyzing Sentiments in eLearning: A Comparative Study of Bangla and Romanized Bangla Text Using Transformers,” *IEEE Access*, vol. 12, pp. 89144–89162, 2024. Available: <https://doi.org/10.1109/ACCESS.2024.3419024>
- [15] E. Hashmi, S. Y. Yayilgan, and S. Shaikh, “Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers,” *Social Network Analysis and Mining*, vol. 14, article 86, 2024. Available: <https://doi.org/10.1007/s13278-024-01245-6>
- [16] K. R. Mabokela, T. Celik, and M. Raborife, “Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape,” *IEEE Access*, vol. 11, pp. 15996–16020, 2023. Available: <https://doi.org/10.1109/ACCESS.2022.3224136>
- [17] M.S.U. Miah, M. M. Kabir, T. B. Sarwar, M. Safran, S. Alfarhood, and M. F. Mridha, “A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM,” *Scientific Reports*, vol. 14, article 9603, 2024. Available: <https://doi.org/10.1038/s41598-024-60210-7>
- [18] G. Mu, C. Chen, X. Li, J. Li, X. Ju, and J. Dai, “Multimodal Sentiment Analysis of Government Information Comments Based on Contrastive Learning and Cross-Attention Fusion Networks,” *IEEE Access*, vol. 12, pp. 165525–165538, 2024. Available: <https://doi.org/10.1109/ACCESS.2024.3493933>
- [19] S. A. M. Zaidi, S. Latif, and J. Qadir, “Enhancing Cross-Language Multimodal Emotion Recognition with Dual Attention Transformers,” *IEEE Open Journal of the Computer Society*, vol. 5, pp. 684–693, 2024. Available: <https://doi.org/10.1109/OJCS.2024.3486904>
- [20] C. Peng, C. Zhang, X. Xue, J. Gao, H. Liang, and Z. Niu, “Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification,” *Tsinghua Science and Technology*, vol. 27, no. 4, pp. 664–679, 2022. Available: <https://doi.org/10.26599/TST.2021.9010055>
- [21] K. Zhao, M. Zheng, Q. Li, and J. Liu, “Multimodal Sentiment Analysis – A Comprehensive Survey from a Fusion Methods Perspective,” *IEEE Access*, vol. 13, pp. 64556–64583, 2025. Available: <https://doi.org/10.1109/ACCESS.2025.3554665>
- [22] K. R. Jerripothula, A. Rai, K. Garg, and Y. S. Rautela, “Feature-Level Rating System Using Customer Reviews and Review Votes,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 5, pp. 1210–1219, 2020. Available: <https://doi.org/10.1109/TCSS.2020.3010807>
- [23] S. M. Dashtaki, M. H. Chagahi, A. Bahadori, B. Moshiri, Md. J. Piran, and A. Montazeri, “HSIF: A Transformer-Based Cross-Attention Framework for Cryptocurrency Trend Forecasting via Multimodal Sentiment–Market Fusion,” *IEEE Access*, vol. 13, pp. 156600–156612, 2025. Available: <https://doi.org/10.1109/ACCESS.2025.3605522>
- [24] H. Fan and J. Chen, “Position Perceptive Multi-Hop Fusion Network for Multimodal Aspect-Based Sentiment Analysis,” *IEEE Access*, vol. 12, pp. 90586–90595, 2024. Available: <https://doi.org/10.1109/ACCESS.2024.3404261>
- [25] Y. Zhang, C. Zhu, and Y. Xie, “Fine-Grained Sentiment Analysis of Cross-Domain Chinese E-Commerce Texts Based on SKEP_Gram-CDNN,” *IEEE Access*, vol. 11, pp. 74058–74070, 2023. Available: <https://doi.org/10.1109/ACCESS.2023.3296447>
- [26] C. Guo and X. Zhang, “Intelligent Prediction of Cross-Border E-Commerce Customer Satisfaction Using Deep Learning Embeddings,” *IEEE Access*, vol. 12, pp. 173268–173278, 2024. Available: <https://doi.org/10.1109/ACCESS.2024.3494776>

- [27] S. Nayab, M. K. Hanif, R. Talib, and M. U. Sarwar, "Aspect-Context Level Information Extraction via Transformer Based Interactive Attention Mechanism for Sentiment Classification," *IEEE Access*, vol. 11, pp. 57683–57692, 2023. Available: <https://doi.org/10.1109/ACCESS.2023.3279396>
- [28] M. R. R. Rana, A. Nawaz, S. U. Rehman, M. A. Abid, M. Garayevi, and J. Kajanová, "BERT-BiGRU-Senti-GCN: An Advanced NLP Framework for Analyzing Customer Sentiments in E-Commerce," *International Journal of Computational Intelligence Systems*, vol. 18, article 21, 2025. Available: <https://doi.org/10.1007/s44196-025-00747-1>
- [29] E. Hashmi and S. Y. Yayilgan, "A robust hybrid approach with product context-aware learning and explainable AI for sentiment analysis in Amazon user reviews," *Electronic Commerce Research*, vol. 25, pp. 5139–5171, 2025. Available: <https://doi.org/10.1007/s10660-024-09896-5>
- [30] P. Keung, Y. Lu, G. Szarvas, and N. A. Smith, "The Multilingual Amazon Reviews Corpus," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4563–4568. Available: <https://doi.org/10.18653/v1/2020.emnlp-main.369>
- [31] P. Bhargava, A. Drozd, and A. Rogers, "Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics," in *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, 2021, pp. 125–135. Available: <https://doi.org/10.18653/v1/2021.insights-1.18>
- [32] M. Ni, Z. Sun, and W. Liu, "Reversible jump attack to textual classifiers with modification reduction," *Machine Learning*, vol. 113, pp. 5907–5937, 2024. Available: <https://doi.org/10.1007/s10994-024-06539-6>
- [33] C. Guo, G. Pleiss, Y. Sun, and K. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning, PMLR*, 2017, pp. 1321–1330.

Appendix

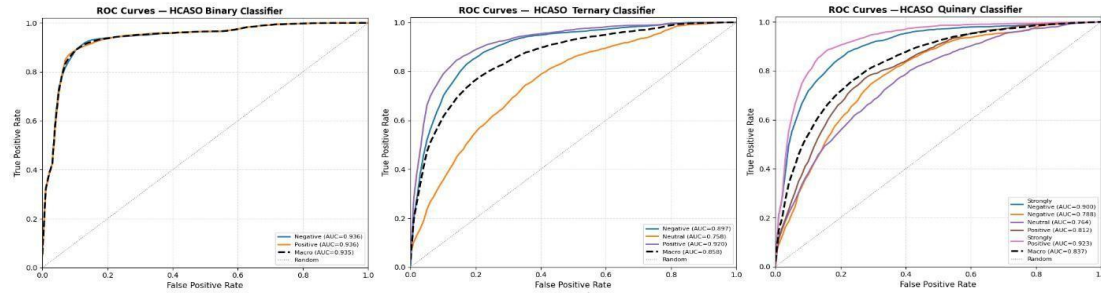


Figure A1. ROC curves for binary, ternary, and quinary versions of HCASO

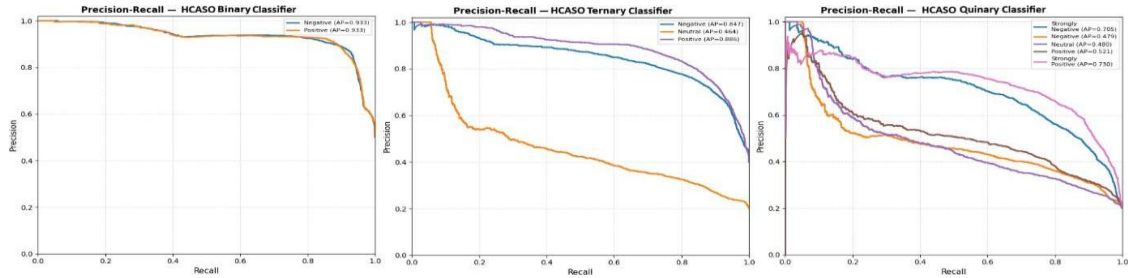


Figure A2. Precision-recall curves for binary, ternary, and quinary versions of HCASO

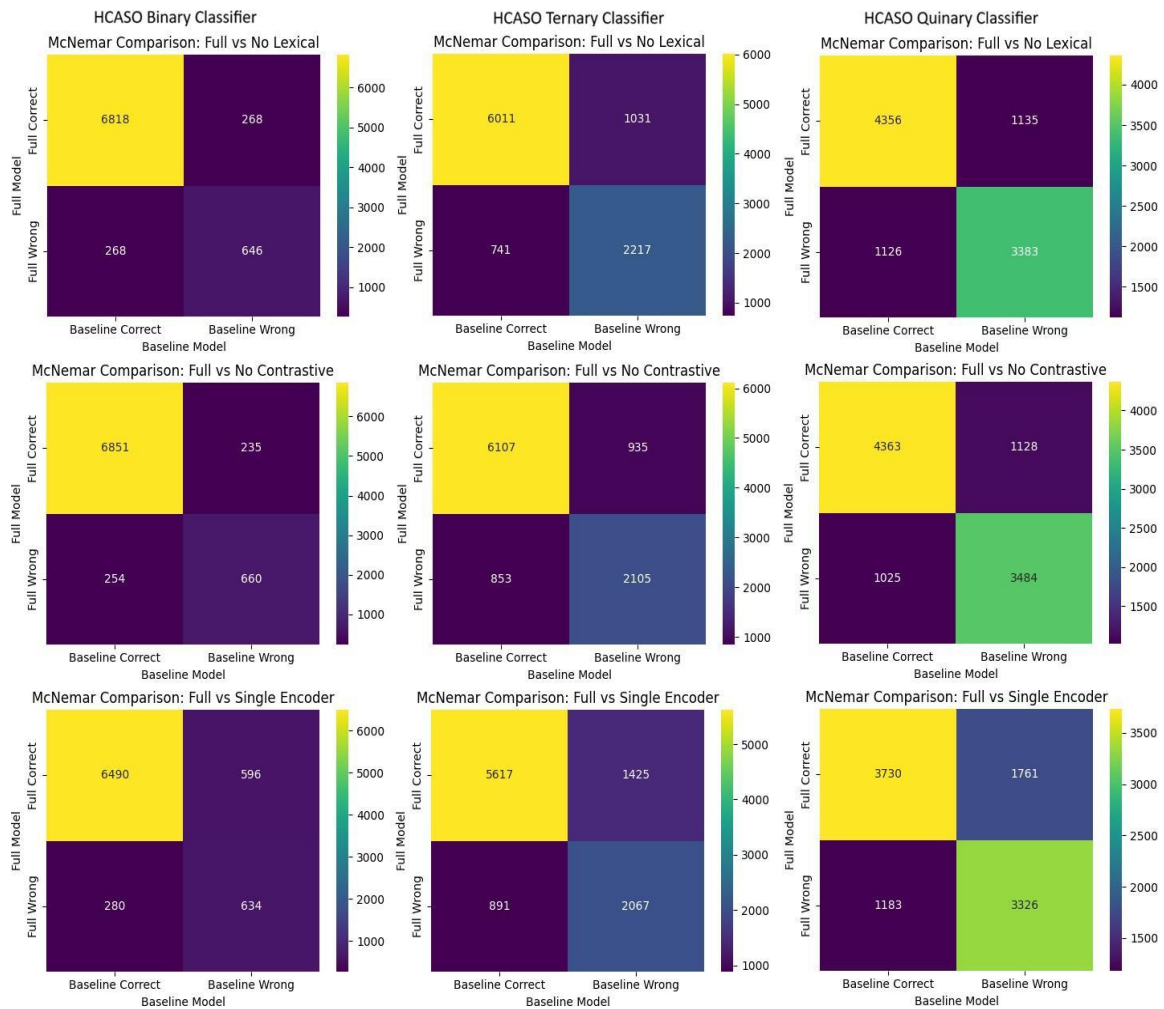


Figure A3. McNemar test contingency heatmaps for pairwise significance testing