Eng Int Syst (2026) 5: 593–606
© 2026 CRL Publishing Ltd

Engineering
Intelligent Systems

Application of Natural Language Processing Technology in Multilingual Sentiment Analysis and Sentiment Recognition

Yang Yuan*

School of Humanities and Communication, University of Sanya, Sanya 572022, Hainan, China

The recent increase in social media data in multiple languages has led to fragmented emotional utterances that contain dialects, slang, and emojis. Traditional NLP (Natural Language Processing) models are ineffective in detecting cross-lingual implicit emotions; thus, there is a need to develop a robust multimodal multilingual emotion recognition framework. This paper introduces a multilingual emotion understanding framework. First, the study utilizes XLM-R for cross-lingual semantic alignment and employs parallel corpus comparison learning to ensure the aligned encoding of emotional utterances in different languages within the same space. Second, a multimodal emotion symbol fusion mechanism is employed, utilizing CLIP and Emoji2Vec, to better represent the semantic meaning of emojis and images. Third, fastText and a dynamic dictionary mechanism are applied to adapt to dialects and slang, which facilitates the recognition of emotional utterances and non-standard texts. Fourth, adversarial training as part of transfer learning is used to increase cross-lingual robustness and emotion consistency. Lastly, BERT-CRF combined with an attention mechanism is utilized for joint recognition of polarities and fine-grained emotions, leading to better detection of sarcasm and implicit emotional understanding. Experiments show that the model achieves an accuracy of 89.2% in English emotion recognition across five languages, and the accuracy of dialect recognition in high-resource dialects is improved to 90.2% through a dynamic mechanism. With an overall accuracy of 82% and attention coverage of 69.3%, the framework's effectiveness in cross-language, multimodal, and non-standard texts was validated.

Keywords: Multilingual Emotion Recognition; Cross-lingual Semantic Alignment; Multimodal Emotion Fusion; Adversarial Transfer Training; Fine-grained Emotion Classification

1. INTRODUCTION

As online social spaces continue to expand globally, language interaction scenarios become increasingly complex, and text communication exhibits multi-layered differences, with language types, cultural backgrounds, and expression habits intertwined, resulting in a wide range of emotion-laden discourse [1–2]. Users across regions use dialects, internet slang, and pictorial symbols in their interactions, leading to more fragmented semantic transmission, deeper implicit tone, and semantic ambiguity in the emotion judgment process [3–4]. Emotion recognition research in multilingual environments is receiving significant attention due to the frequent shifts in public opinion dynamics in the digital world, the presence of behavioral tendencies, social attitudes, and group psychological states conveyed by emotional signals, and the research interest in emotion computing technology in public safety, business analysis, and psychological behavior observation scenarios [5–6]. The data expressing emotions contains repeatedly changing colors, and positive or negative stances are implied rather than expressed directly. The structures of irony and satire exhibit

*Corresponding AuthorEmail: yangyuan@sanyau.edu.cn

non-linear forms, making it difficult for traditional counting-based text methods to capture potential emotional tension. Moreover, cross-language differences lead to semantic mapping errors, highlighting research bottlenecks [7–8].

In the field of multilingual emotion recognition, natural language processing technology has gone through the bag-of-words model, word vectors, and pre-trained models. Researchers have designed emotion dictionary systems, integrated convolutional networks and recurrent structures to model multilingual texts, used corpus annotation and language rules to extract emotion features, expanded the feature space with multi-channel learning strategies, and tried to use contrastive modeling ideas to correct biases. In recent years, emotion analysis research has gradually shifted from traditional dictionary methods to deep learning models based on Transformer, which has significantly improved the accuracy and generalization ability of text emotion understanding [9]. On this basis, multimodal emotion analysis has become an important extension. For example, text, image and audio signals are integrated through context-sensitive multi-layer deep learning frameworks [10], and attention mechanisms are applied to capture cross-modal semantic inconsistencies to improve the detection of irony detection [11]. Furthermore, the Transformer-based fusion architecture has been widely used in multimodal emotion recognition, enhancing the robustness of the model by co-processing text, audio and visual cues [12], while bimodal systems focus on the integrated learning of speech and text embeddings to optimize recognition efficiency [13]. Related reviews systematically summarize the core progress and challenges in the field of sentiment analysis, emphasizing the importance of multimodal and cross-linguistic integration [14]. Meanwhile, lightweight cue distillation techniques in large language models have facilitated cross-linguistic instruction alignment [15], while the exploration of low-resource language processing applications has provided adaptive solutions for sentiment computing in multilingual environments [16]. Together, these works have driven the continuous evolution of sentiment analysis technology in multimodal, cross-linguistic, and low-resource scenarios. However, most methods are limited to a single-text modality, the semantics of emoticons and image descriptions are not well absorbed, the performance of low-resource language is poor, there is inadequate processing of complex emotional change structures, it is difficult to convey accurately any cross-cultural tone transitions, and structural stability still has potential for improvement.

Recent studies have attempted to use large-scale pre-trained models to construct semantic layers at a deep level, build cross-linguistic semantic alignment mechanisms, and add adversarial strategies to improve the robustness of language transfer. robustness. Due to previous research efforts, attention-enhanced classification frameworks and multi-label recognition structures have emerged, as have multimodal semantic fusion modes. In terms of cross-linguistic semantic understanding, sentence transformers have been applied to English-Urdu lexical similarity calculation, providing an important foundation for semantic alignment in multilingual sentiment analysis [17]. At the same time, the comprehensive development of image-to-text conversion technology has provided technical support for visual-language interaction in multimodal sentiment analysis, especially in the generation of multilingual descriptions, where significant progress has been made [18]. To address the challenges of low-resource language processing, researchers have developed deep learning architectures for specific languages such as Uzbek [19] and achieved interpretable cross-linguistic sentiment analysis by means of deep context representation and attention mechanisms [20]. In low-resource scenarios, transfer learning methods based on multilingual BERT (Bidirectional Encoder Representations from Transformers) have been successfully applied to sentiment analysis tasks in Austronesian languages [21]. Label-efficient sentiment analysis technology has gradually attracted attention as a way to further improve the efficiency of the model [22], while cross-language sentence pair interaction feature capture model based on pseudo-corpus and multilingual embedding provides a new solution for sentiment computing of low-resource languages [23]. Despite the rich research progress, the stability of fine-grained sentiment classification in weak language texts is insufficient, online language is rapidly evolving, and there is an abstraction bias in the identification of irony, metaphor and double tone. Furthermore, there is a lack of adaptive enhancement strategies in cross-cultural sentiment interpretation, and the limitations of previous research approaches are clearly evident in complex contexts [24–25].

This study constructs an adaptive multimodal multilingual emotion learning framework for multilingual emotion-recognition tasks. It designs a cross-lingual semantic alignment structure, sets up a pipeline for fusion of emoji and image semantics, applies dialect and slang extension modules, employs adversarial transfer thinking to consolidate cross-lingual generalization performance, and constructs a joint structure of polar emotions and fine-grained labels, focusing on the parsing process of irony and implicit semantics. The research process involves corpus alignment preprocessing, embedding structure setting, construction of an emotion label system, contrastive constraint training, and hierarchical emotion analysis. The aim is to enhance the accuracy of emotion signal expression parsing in complex contexts, explore the evolutionary characteristics of emotion meaning sharing structures in cross-cultural contexts, and provide theoretical contributions and practical reference value for research on multilingual social interaction semantics.

2. DESIGN AND IMPLEMENTATION OF MULTILINGUAL EMOTION UNDERSTANDING MECHANISM

2.1 Semantic Alignment Encoding Based on Cross-Language Large Model

To solve the problems of uneven and incomplete distributions of the semantic space and cross-linguistic semantic mapping bias in multilingual emotion expression, this study creates a common semantic space using a cross-linguistic pre-trained model, and applies a layer-wise alignment strategy under parallel corpus supervision to guarantee the even distribution

of emotion expression across languages. To eliminate feature biases stemming from the differences among languages, a cross-linguistic-coded model XLM-R (XLM-RoBERTa) was used for the backbone structure, into which parallel sentence pairs were fed. The semantic distance differences were combined together through a contrastive training mechanism. The emotional semantic expression is constrained at the same time at the sentence level, word vector layer, and attention layer to create both the implicit emotion symbols and explicit emotion vocabulary together. The model is constructed by aligning Transformer parameters and a common sub-word embedding structure to project heterogeneous texts (e.g., Russian, Arabic, Chinese, and English) into a common vector representation matrix for density space. The convergence direction is stable across distributed language gradient signals, and cross-linguistic transfer effects can be strengthened through an end-to-end approach [26–27].

2.1.1 Cross-Language Semantic Encoding and Parameter Sharing Mechanism

In the encoding stage, the texts of each language are uniformly processed into sub-word level sequences and input into the XLM-R model for Transformer encoding. The multi-head self-attention structure captures the internal dependencies across contexts. The model parameters adopt a cross-language sharing strategy. Instead of setting independent language encoders, the model achieves semantic abstraction consistency by sharing weights, thereby avoiding feature fragmentation caused by excessive segmentation of language granularity. The corpus input consists of parallel sentence pairs, which are encoded separately to obtain hidden layer vectors. Then, the global sentiment vector representation matrix is generated by the sentence vector mean pooling and CLS label extraction strategy. The training process sets semantic distance constraints and uses cosine similarity to construct loss. The goal is to improve the similarity of the vectors of synonymous sentiment expressions and expand the difference between the vectors of different sentiments [28]. This is done with equation 1:

$$L_{align} = 1 - cos\left(h_{src}, h_{tgt}\right) \tag{1}$$

where, h_{src} and h_{tgt}represent the source language and target language embeddings, respectively. The optimization employs AdamW and a segmented learning rate decay strategy, dynamically adjusting the gradient signal based on batches of corpus data to converge the model towards a continuous distribution of cross-linguistic representations. After multiple rounds of training, a cross-lingual emotion encoder is obtained, whose output can converge the same emotional semantics to adjacent positions in texts of different languages, improving the alignment granularity of implicit emotion expressions. To control model bias, a data shuffling mechanism is constructed to shuffle language labels, and a cross-lingual dropout strategy is added to increase the proportion of random language masking, enabling the model to maintain stable semantic discrimination accuracy even in the absence of explicit language markers.

2.1.2 Contrastive Learning-Driven Reinforcement of Emotional Semantic Consistency

A contrastive learning module is applied based on the encoding structure. This module takes parallel corpus sentiment labels and unlabeled cross-sentence combinations as input to form sentiment positive and negative sample pairs, and uses a dual-tower structure to extract alignment vectors. Positive samples come from semantically equivalent and sentimentally consistent texts, while negative samples come from texts with opposite sentiment directions or semantically fragmented texts, thereby constructing a similarity-reduction and difference-expansion mechanism. During training, a temperature parameter is established to control the gradient response amplitude, making the fine-grained sentiment feature clustering process more stable and preventing irrelevant words and phrases from misleading the model. The corpus sampling strategy applies random mixing of cross-cultural scenarios and weighted extraction of weak-resource language samples, enabling the encoder to maintain a stable convergence trend outside of high-frequency language-dominated environments, forming a robust cross-linguistic sentiment representation.

The embedded vectors output in this stage form a word-sentence, two-layer emotion structure in a unified space. The low-dimensional semantic matrix captures the distribution trends of emotion words and metaphorical words, while the high-dimensional sequence vectors focus on short-term emotion fluctuations. Contrastive training combined with a parameter freezing mechanism partially preserves deep parameters within the weight update interval, adjusting emotion projection only in the shallow and mid-level layers. This prevents drastic oscillations in the semantic space, resulting in stable cross-linguistic emotion alignment capabilities. The final model-generated encoded representation integrates multilingual emotion features and maintains consistency in the shared space.

Figure 1(a) shows the distribution of several languages—English, Chinese, Spanish, Arabic, and Russian—, in the emotion semantic space. The horizontal and vertical axes represent semantic dimension 1 and semantic dimension 2, respectively. The clustering trajectories of different languages in the space indicate that the initial offset between languages and the trend of the emotional semantics tend to be concentrated after alignment. Each emotion category point (joy, anger, sadness, fear, surprise) gradually shows convergence and clustering in several language regions, indicating that the cross-language semantic alignment model effectively reduces the emotional expression bias caused by language differences. In Figure 1(b), the horizontal axis represents the number of training rounds, and the vertical axis represents the changes in the similarity of positive sample emotion pairs and the similarity of negative sample pairs during the training process. The similarity of positive samples gradually increases, while the similarity of negative samples decreases, proving that the contrastive learning mechanism achieves a stable improvement in discrimination ability in multilingual emotion alignment, assisting the model to obtain more robust performance in terms of recognizing cross-language emotion expression.

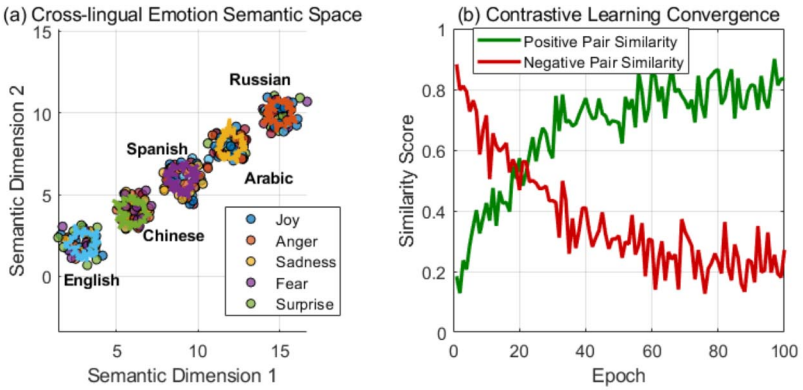


Figure 1 Schematic diagram of multilingual semantic alignment and model convergence. Figure 1(a) Visualization of cross-lingual emotion semantic space distribution. Figure 1(b) Contrastive learning similarity

2.2 Multimodal Emotion Symbol Fusion Mechanism

This section discusses the semantic disconnection and ambiguity propagation issues of textual emotion expression, emoji expressions, and visual emoji expressions resulting from the implementation of a fusion structure. In particular, a cross-modal pre-trained model and a specialized emotion symbol vector model are used to embed visual information and emoji sequences into the text emotion inference pathway, such that text emotion vectors and symbol vectors are converged into the same latent space while using an emotion bias compensation signal. Overall, the process accepts raw text, emoji sequences, and image inputs, arranging the input sequence information structure through a timestamp and dialogue order alignment strategy. An emotion symbol attention gating structure is injected into the hidden layer of Transformer to mitigate the influence of pure text weights. During training, a semantic synchronous sampling mechanism is constructed from inputs containing visual expressions and emojis. Emojis convey various encoded meanings in real-time situations, so their expressions can be mapped by weighted connections to fixed-dimensional tensor spaces to allow various input signal modalities to enter consistent encoding channels through batch normalization preprocessing. Furthermore, the training objective is to progressively minimize the visual-text emotion difference distance and inhibit misjudgment-producing tendencies resulting from the absence of semantics in emojis.

2.2.1 Emoji and Text Shared Vector Injection Strategy

The emoji representation uses the Emoji2Vec vector extraction structure, which maps every emoji symbol to a fixed-dimensional emotion vector. The symbol vector is written into a predetermined slot into the text encoding sequence through the sub-word masking insertion mechanism and emoji weights are pruned by means of gated projection. This structure does not have a separate channel; instead, it loads symbol biases into the word vector sequence so that implicit markers such as irony, disgust, and pleasure which emojis convey are positioned alongside the text. During the forward pass, an independent attention distribution matrix is created for each emoji. Sparsity constraints are established by the probability of co-occurrence of the symbol in the emotion corpus which causes symbolic expressions to yield differential aggregation states in ironic, sarcastic and exaggerated expressions. The training incorporates a symbol perturbation strategy, which randomly removes some emojis and inserts pseudo-symbol vectors in order to encourage some stability in inference bias in the absence of some symbols [29–30]. Along with this, a symbol consistency term is added in the loss function so that the text emotion output aligns with the trend of emoji weight. This equation is:

$$L_{emo} = \| f_{text}(x) - f_{emoji}(e) \|_2 \tag{2}$$

where, $f_{text}(x)$ represents the text sentiment vector, and $f_{emoji}(e)$ represents the emoji sentiment vector. A segmented freezing strategy is used for parameter training, limiting the gradient backwards magnitude in the first layer of word embedding and encoding. This prevents the symbol injection of bad sections from generating a linear steady-state response rather than drastic perturbations, and uses a periodic weight drift correction mechanism to reduce shifts in sentiment direction.

2.2.2 Image Semantic Coupling and Attention Competition Mechanism

The CLIP (Contrastive Language-Image Pre-training) model is used to encode image representations by extracting vectors of visual semantics and inserting tags of visual emotion fragments in dialogue order. After visual features enter the Transformer, a cross-modal attention competition module is activated to separate the visual and linguistic weights within the attention matrix. Also, an upper limit is placed on visual weights based on the distribution of emotion tags to ensure that visual signals play a supporting, not dominant, role. The model incorporates an emotion resonance gate in the hidden unit layer, filtering visual sub-features aligned with the text direction and suppressing noisy image content, thus protecting the emotion carrier from background visual interference. For the training process, image occlusion and blurring perturbation strategies are applied, and negative sample matching is established to ensure that the compressed

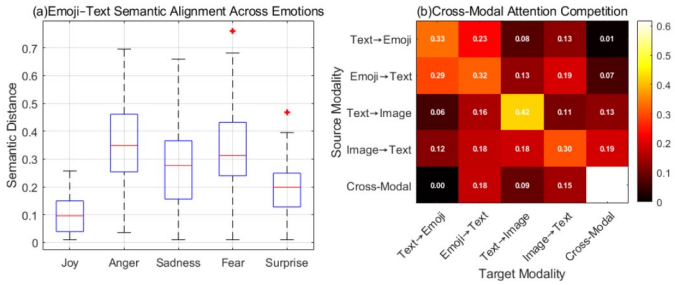


Figure 2 Visualization of key mechanisms in multimodal emotion fusion. Figure 2(a). Distribution of Emoji-text semantic distance under different emotion categories. Figure 2(b). Heatmap of cross-modal attention competition.

vectors output from image sequences without emotional symbolism are stably concentrated in a neutral distribution region. The multimodal loss structure uses a weight adaptive strategy, reducing the proportion of visual loss in the early stages of text-based training to allow the encoder to establish a stable text emotion center of gravity, and then gradually increasing the proportion of visual channel weights to allow the model to learn cross-modal emotion association patterns. In the output stage, text, emoji and visual vectors are fused into a vector stack through a linear layer. Residual connections are used to suppress gradient oscillations and maintain the continuous convergence trend of the emotion direction [31–32].

Figure 2 illustrates the performance of the multimodal emotion recognition model in terms of emotion expression consistency and cross-modal information competition mechanism. In Figure 2(a), the horizontal axis represents the five basic emotions (Joy, Anger, Sadness, Fear, and Surprise), and the vertical axis represents the semantic distance distribution between Emojis and text, indicating the degree of fit between text content and accompanying emojis. The results show that the Joy category has the smallest semantic distance, indicating higher consistency between Emojis and text when users express positive emotions. The Anger and Fear categories have a larger semantic distance, indicating stronger implicit emotions or semantic shifts when expressing negative emotions, posing a greater challenge to the multimodal model. Figure 2(b) presents the cross-modal attention weight competition relationship, with the horizontal axis representing the target modality and the vertical axis representing the source modality. A higher proportion of attention is observed in the Cross-Modal region, indicating that the model generally tends to dynamically seek information fusion across different modalities. Simultaneously, the Text? image weight is high, indicating that the interaction between text and image emotional cues is a key channel.

2.3 Dialect and Slang Adaptive Sentiment Extension Module

This section constructs adaptive expansion paths for dialects and slang in non-standard language expression scenarios. The model input contains a large number of regional terms, online variant words, and expanded spoken text. Directly mapping these to a general vocabulary would result in sentiment drift and polarity-splitting in the encoding results. Therefore, a dynamic vocabulary supplementation strategy and a sub-word-level modeling process are applied to ensure stable vector representations of new and variant words. After word segmentation and sub-word slicing of the original text, high-frequency out-of-vocabulary (OV) words and variant fragments are collected and clustered according to the posting user, interacting object, and regional tags to construct a dialect phrase candidate set and a slang expansion pool. During processing, a continuously updated context window is established for each candidate word, extracting surrounding sentiment words and syntactic features to generate the initial semantic position of OV words. Time-slice rolling statistics are used to eliminate short-term traffic noise, maintaining a steady-state supplementation rhythm during the vocabulary expansion process.

2.3.1 Sub-word Modeling and Fine-grained Word Vector Injection Process

The system adopts the FastText sub-word modeling structure, which segments out-of-vocabulary words into continuous character segments, encodes them separately, and aggregates them in the embedding layer, so that the model can produce a continuous transition expression for morphological variants. In the prediction stage, a vector is generated for each sub-word segment and written into the main encoding channel in a weighted aggregation manner. The weights are determined based on dialect frequency, contextual consistency and network domain confidence factor [33]. In the training stage, the text is truncated and reconstructed, and some sub-word segments are randomly masked and replaced with fake noise segments, so that the model can generate an emotional direction compensation response to the segment error. This mechanism prevents the model from having an emotional void due to the lack of new words, and allows variant words to gradually form a stable polarity direction. The vector training target constructs the sub-word aggregation loss term as shown in equation (3):

$$L_{sub} = \sum_{i} \| E(w_i) - \frac{1}{k} \sum_{j=1}^{k} E(s_{ij}) \|_2 \tag{3}$$

where, $E(w_i)$ is the target word vector, and $E(s_{ij})$ is the corresponding sub-word vector. The above constraints exert

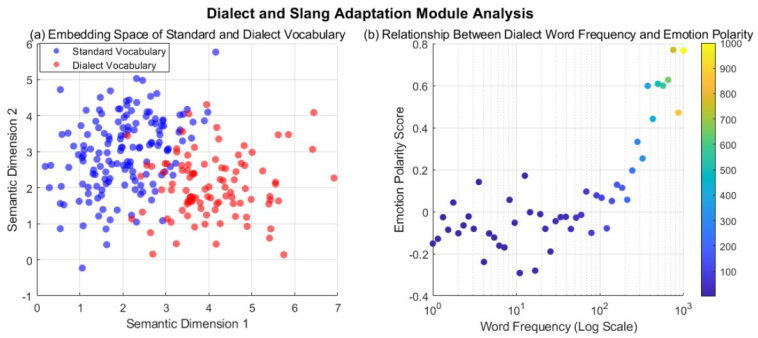


Figure 3 Visualization of the adaptive expansion mechanism of dialects and slang. Figure 3(a). Semantic vector space distribution of standard words and dialect words. Figure 3(b). Relationship between dialect word frequency and emotional polarity.

a stable pull on the short segments from long words, avoiding polarity drift and maintaining the continuous trajectory of word vectors in the emotion space.

2.3.2 Dynamic Dictionary Construction and Emotion Bias Regulation Mechanism

The dynamic dictionary maintains out-of-vocabulary words and slang terms in an incremental manner, marks the initial polarity of newly emerging emotion-dominant words and sets a signal attenuation coefficient to prevent short-term burst words from causing large-scale bias. The dictionary update cycle is triggered by a window counter, and the survival threshold of the term is judged by the accumulation of semantic similarity, and the term with deteriorated confidence is eliminated. In the model fine-tuning stage, a specific attention channel is established for the dictionary, and a dedicated weight matrix is set so that dialect and slang word vectors obtain independent attention convergence paths. In backpropagation, the gradient amplitude of the extended words is limited to avoid disturbing the general vocabulary. In the output coupling stage, the extended word vectors are injected into the main emotion vector in a residual manner, so that the number of non-standard expressions in the final prediction is controlled and does not damage the overall encoding direction [34]. This structure ensures that the emotion features of dialect expressions and network language are corrected step by step with the corpus update cycle, forming a stable growth mechanism, so that the model can maintain the ability to absorb emerging expressions and extend the emotion polarity in long-term operation.

Figure 3 illustrates two key features of the dialect and slang adaptive sentiment extension module. In Figure 3(a), a comparison of the scatter distribution of standard and dialect words in the two-dimensional semantic embedding space shows that the dialect word clusters have a significant shift in space relative to the standard word clusters, indicating that dialect sentiment expression has linguistic uniqueness and requires targeted modeling for alignment and enhancement. In Figure 3(b), the horizontal axis represents word frequency (logarithmic scale), and the vertical axis represents sentiment polarity score. The scatter color changes with word frequency, and the results show that high-frequency dialect words gradually tend towards a more defined polarity range, while low-frequency dialect words have greater polarity fluctuations, demonstrating the important role of dynamic dictionary and sub-word modeling in capturing the features of sparse dialect sentiment.

2.4 Cross-linguistic Emotion Transfer and Adversarial Enhancement Mechanisms

This stage of the study is designed to counter domain bias and language noise interference in the transfer of emotion expression across languages. It develops a multilingual transfer pathway that can align the emotion expression from a high-resource language to a low-resource language. During data pre-processing, the data sources are defined by language labels, and the primary and target languages are defined in the order they will be input. Then, the primary and target language sources input the data into a shared encoder that ensures both the source language text and target language text are upgraded in the same parameter space by the gradient. At run-time for the model, frozen shallow semantic extractor units maintain a relatively stable low-level embedding structure and bias the transfer of errors within these encoder layers to the high-level semantic feature abstraction modules that the model architecture deep-dives within a cross-language framework into relationships of emotion features. The encoded output is compressed into fixed-dimensional emotion vectors by the unified mapping module that maintains parity/consistency of vector scale during language bridging between the first source language and the target language to build continuous transition trajectories within emotion dimensionality through language text.

2.4.1 Transfer Training Path and Interlingual Feature Calibration

A source language weight-based strategy is used for the transfer training process. In the first stage, a sentiment corpus in a high-resource language is used to pre-train the sentiment classification layer of the model, leaving the main network backbone frozen. Once a target language corpus is input into the model, only the top-level projection and classification layer parameters will be updated by gradients. This allows the model to create an initial mapping in the target language based on known sentiment labels. In the label-scarce stage

of training, pseudo-label training will take place. This batch of low-resource examples has known sentiment bias values, based on which the sentiment prediction can be computed, and finally assigned a bias value. During parameter updates, a confidence block is triggered that sets a minimum threshold for assigned bias, thereby updating only those pseudo labels with at least a 50% uncertainty of being a given sentiment bias. This ensures that the model will converge in an orderly fashion with little supervision. A parasocial-imitation mechanism reduces biased variation in task performance that could result from any differences in text style. This SEMANTIC balance penalty term will ensure that the feature vectors of target language samples are statistically similar to their counterparts in the source language corpus. During the periods when target language features are being adjusted, the training weight of the target language gradually increases so as to decrease the model's bias for the high-resource language in the later stages of training. During the training iterations, the distance between the sentiment decision boundaries of different languages is monitored. If the offset exceeds a preset threshold, a resampling strategy is triggered to weight the input of target language samples, concentrating the training focus on regions with significant distribution differences, thereby strengthening the model's ability to discriminate between cross-language sentiments.

2.4.2 Adversarial Training Mechanism and Emotional Consistency Constraints

To mitigate the interference of language features on emotion prediction, a Gradient Reversal Layer (GRL) structure is applied between the encoder and the language discriminator. This allows the model to maintain a normal flow during forward propagation and reverse the gradient direction during backpropagation. The language discriminator attempts to distinguish the language type of the input text, while the encoder suppresses language discrimination information under the influence of the inverse gradient, making the output features transferable to other languages. The emotion classifier and the language discriminator are updated synchronously, and the model training objective maintains a convergence trend in the emotion direction while weakening the control of language labels over spatial distribution. The emotion consistency constraint applies a dual-path inference mechanism on unlabeled target language samples, inputting the same text twice and adding random noise perturbations to each, imposing a consistency penalty on its emotion output to make the network output tend to a steady state [35–36]. The emotion consistency loss is defined in equation 4:

$$L_{cons} = \| f\,(x + \epsilon_1) - f\,(x + \epsilon_2)\,\|_2 \tag{4}$$

where f is the emotion prediction function,ϵ_1 and $\epsilon_2$are the perturbation term. This constraint maintains a stable trajectory for emotion prediction during the transfer process, preventing label direction drift in the model. Upon training convergence, the language discrimination accuracy and emotion classification accuracy are calculated separately. If the language discriminator continues to improve while the emotion accuracy decreases, the inverse gradient strength is adjusted to refocus the network on the emotion direction, thus maintaining a balance between cross-language transfer and robustness enhancement[37–38].

2.5 Emotional Hierarchy Classification and Fine-Grained Recognition Strategy

In this section, a structure is developed that consists of two layers to distinguish emotions, separating the emotion expression into a layer for polarity identification and a layer for fine-grained classification to deal with differences in both intensity and emotions across languages. In the preprocessing phase, the original text is transformed into a single, unified token sequence that is then fed to a pretrained BERT model, and an emotion semantic matrix is developed, allowing the model to iteratively encode the implicit tone characteristics and emotional hints within the sentences as context-based parameters. Once the polarity identification phase begins, the tenure representation layer of the encoder is frozen so the model can make the most of establishing the direction of the emotion gradient at the outset, and only then work out the differences in intensity of emotion classification. This output vector sequence is then passed to a CRF (Conditional Random Field) layer in order to develop dependencies concerning label transfer so that when emotion transitional words or negation structures exist, the model will produce stable outputs, limiting the negative effects of local word segment misclassifications on echoing the broader sentiment during judgement cycles. Multiple candidate emotion labels are formatted hierarchically and, as the network recursively selects objects to identify the closest polarity label, it categorizes candidates a priori to give the network a directed directionality for the next (final) step, which consists of fine-grained classification. This ensures that the emotion path is a continuous mapping trajectory in the decision space, rather than an independent segmented structure, thereby reducing drift in samples that have ambiguous boundaries.

2.5.1 Hierarchical Label Inference and Emotional Boundary Calibration

In the polarity classification phase, a linear discriminant head is used to compress the output of BERT into three directions of emotion. The flow of gradients is regularized with cross-entropy loss to ensure stability. Ratios of weights are tailored to discourage over-learning of neutral text so that the decision boundary is compactly distributed. After making the polarity decision, the process moves to a fine-grained multi-label discriminant matrix that produces independent logits predicting each emotion type. This is achieved in a way that allows for semantic patterns (i.e., irony, anger, joy, and annoyance) to form separate branches, distinguishing their similarities through encoding. In the training, significant additional hard example mining is done on difficult-to-distinguish labels (i.e., therefore likely to require a hard example mining process, including irony, satire, or implicit attack). This forces the network to invest more gradients in a semantically ambiguous region and limits the possibility of ironic text being tagged neutral or negative. When

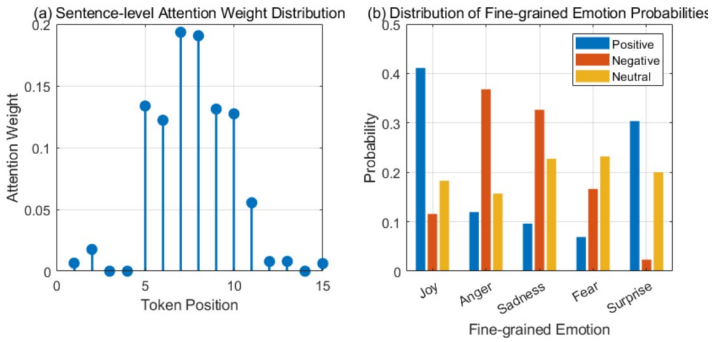


Figure 4 Visual analysis of the multi-level emotion recognition model. Figure 4(a) Distribution of sentence-level attention weights. Figure (b) Fine-grained emotion probability distribution.

inferring labels, if the output appears as a duplicate multi-class crossover, the label sequence is reordered based on polarity prior features to preserve an output classification that follows semantic commonsense logic. During the training process, similarity constraints are applied to the model, and the label similarity matrix is continuously monitored during the prior stage. If the label bias is too large, the emotion boundary calibration module is activated to smooth the logits, maintaining a reasonable interval between adjacent emotion categories and preventing the model from making random predictions due to noisy samples.

2.5.2 Attention-weighting Mechanism and Enhancement of Complex Emotional Representation

To handle cross-linguistic irony and implicit intent, a dual-path attention matrix is applied, creating an interactive weight distribution between text content and contextual tone. Attention weights are assigned based on word position, negation marks, modal particles, and emotion-triggered phrases. During backpropagation, the model tends to maintain the gradient of local weight regions, making strongly emotive words the core driving force for the output. A local emotion aggregation function is set up for this structure, weighting and summing the word segmentation vectors to form a sentence-level emotion representation, calculated as shown in equation (5):

$$s = \sum_{i=1}^{n} a_i h_i \tag{5}$$

where, h_i represents the token feature vector, and a_i represents the attention allocation coefficient. The outputs is input to a multi-label classifier to complete fine-grained inference. The ironic corpus includes an additional contrast consistency objective during training that forces the model to output similar emotional directions in both the original and perturbed text, weakening the influence of language differences on ironic semantic judgment. If, during inference, there appears to be an emotional conflict trend, the syntactically correct pathway is initiated, and the model increases the weight of negative phrases and tone reversal words to stabilize the model's judgment in the presence of polysemous expressions. Ultimately, this ensures that the cross-language complex emotion encoding is converging in a similar direction, that semantic deviation is minimized, and that emotional continuity and intensity are congruent.

The multi-level emotion recognition model for text understanding and emotion classification is shown in Figure 4. In 4(a), the horizontal axis shows the position of the token, and the vertical axis shows the attention weight. The results indicate that the model gives the greatest attention weight to the tokens near positions 7 and 8. This shows that words in the middle of a text may be more likely to contain emotional information, while words at the beginning or end may contribute less. Hence, this analysis shows that the model can focus on one of the key emotional components of the sentences. Figure 4(b) shows the model's emotion distribution trends for positive, negative, and neutral texts. The horizontal axis represents fine-grained emotion categories, and the vertical axis represents probability. Positive texts are primarily focused on the categories "Joy" and "Surprise," while negative texts are focused on the categories of "Anger" and "Sadness." This demonstrates how well the model can identify the emotional categories and its ability to make fine-grained predictions.

The architecture of the multilingual emotion recognition system is presented in Fig. 5. The system takes in three types of input data: multilingual text, emojis/images, and dialectal slang. The input data are processed in parallel through three modules: cross-lingual semantic alignment, multimodal fusion, and dialect adaptation. The outputs of each module are subsequently combined and enhanced by cross-lingual transfer and adversarial enhancement. Lastly, a hierarchical emotion classifier generates both polarity and fine-grained emotion labels. The entire framework has an end-to-end structure and relies on semantic space alignment, attention mechanisms, and adversarial enhancement to account for important factors associated with diverse emotion expression in multilingual contexts, multimodal information fusion, and informal and dialectal language.

3. MULTILINGUAL EMOTION RECOGNITION PERFORMANCE VERIFICATION SYSTEM

The evaluation experiments presented in this research are founded on a multilingual emotion corpus, comprising English, Chinese, Spanish, Arabic, and other languages,

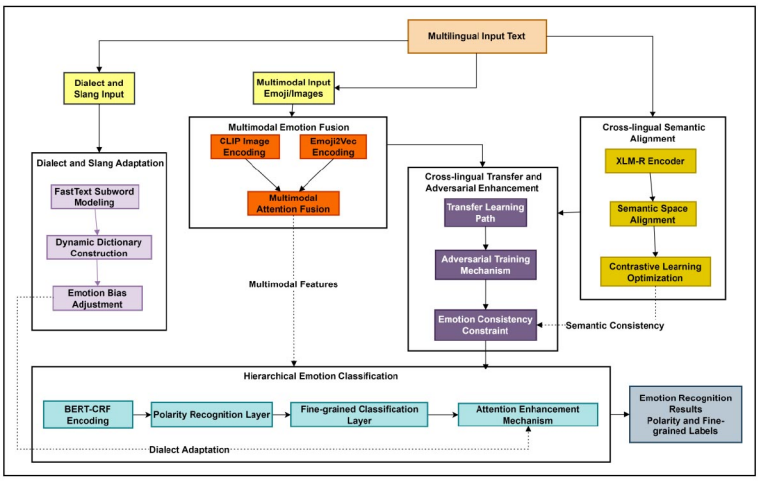


Figure 5 Overall architecture of the multilingual emotion recognition system.

as well as low-resource language datasets. The source data used to establish the corpora include public domain open-source multilingual emotion corpora, publicly available social media corpora, cross-platform real user comments collected in dynamic platforms, and manually annotated emotion text (with anonymized metadata), all of which have been deduplicated, anonymized, and filtered from other noise that would lead to poor data quality. Samples from each language have also been divided into separate layers by polarity and percentages of fine-grained emotions. Additionally, to investigate the model's ability to adapt to a real-world internet language environment, dialect and slang extension sets, emoji, and multi-modal supplemental sets were constructed. The preprocessing of the dataset ensures that, to avoid accidental data cleaning affecting the original expression, there is consistency in the encoding format used, the specification of unified symbols, and colloquial features are preserved used during the process. Based on this corpus of data, the experimental procedures were designed to take into account language independence, and include cross-language transfer, stress testing challenging examples, multi-modal intervention, and dialect and internet language scenarios. Next, it verified and analyzed the performance of the models with respect to five aspects: accuracy, transfer stability, complex emotion recognition ability, multi-modal consistency, and dialect robustness.

3.1 Performance Validation of Cross-Language Emotion Classification

In this phase, a multilingual test is established, dividing the test batches by language. The model output is compared with the real labels one by one, the number of correct predictions is counted, and the language switching status during each inference round is recorded to avoid sample interference caused by mixed batches. The testing process starts with the English corpus, followed by Chinese, Spanish, Arabic, and low-resource languages. Each language is scored independently, and the overall mean and inter-language dispersion are calculated. When data is sent to the system, the word order is kept intact, and the maximum sequence length is fixed. Excessively long samples are truncated, and short texts are used to maintain input consistency. To evaluate the model's adaptability in real-world scenarios, temperature adjustment and soft label correction are not used during the testing phase; only the highest probability label of the inferred output is used to record the results. After the test, the distribution of each category is checked, and groups with low sample sizes are retested to ensure sufficient test sample coverage for both sentiment polarity and fine-grained labels. If a language shows an excessive bias towards neutral sentiment, high-sensitivity samples in that language are added for retesting to determine whether the model exhibits a stable discrimination imbalance with weakly-expressed languages.

In Figure 6, the horizontal axis represents language category and emotion category, and the vertical axis represents recognition performance indicators. Figure 6(a) uses Accuracy and F1 Score as the vertical axis indicators to compare the overall emotion recognition performance of five languages: English, Chinese, Spanish, Arabic, and Russian. The results show that English performs best (accuracy approximately 89.2%, F1 approximately 87.8%), followed by Spanish and Chinese, while Arabic and Russian are slightly lower. This is related to the size of the training corpus and the differences in expressions of emotion. Figure 6(b) uses the emotion category as the horizontal axis to show the accuracy differences of English, Chinese, and Arabic in the five emotion categories. It can be observed that English performs better in Joy and Surprise emotions, while Chinese and Arabic show a slight decline in performance for negative emotions such as Sadness and Fear. The data shows the stability of the model during cross-language transfer and the impact of language differences on the fine-grained emotion recognition effect.

The inferior performance of low-resource languages (lower accuracy of Arabic) arises mainly from the lack of training data available and culturally-based differences in expression, resulting in inadequate semantic alignment. Resource-abundant languages (e.g., English) perform better

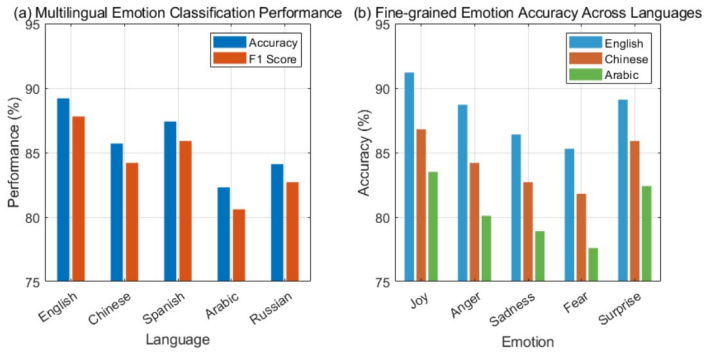


Figure 6 Performance comparison of multilingual fine-grained emotion recognition. Figure 6(a). Overall performance comparison of emotion recognition models in different languages. Figure 6(b). Distribution of fine-grained emotion classification accuracy in different languages.

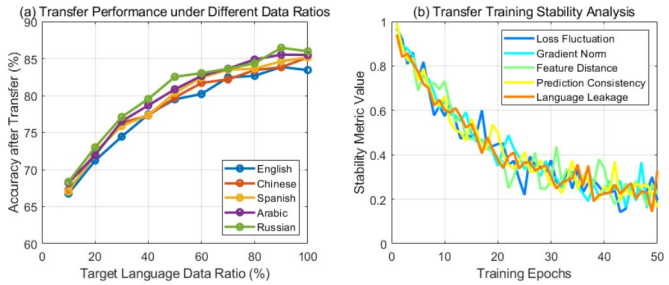


Figure 7 Cross-language transfer stability assessment Figure 7(a). Transfer performance under different data proportions Figure 7(b). Transfer training stability analysis.

due to an adequate sample of training data, affirming that current cross-language transfer mechanisms are still quite sensitive to data distribution and, therefore, semantic enhancement strategies applied to low-resource languages need to be enhanced.

3.2 Cross-language Transfer Stability Assessment

This phase sets up a primary language training and target language testing process, recording the target language performance changes with training epochs to determine if the model experiences sentiment prediction slippage during transfer learning. Multiple sets of target language data proportions are prepared and injected into the training process at four levels: 10%, 30%, 60%, and 100%. The performance trajectory of the model from the low-resource stage to the fully trained stage is observed, with a focus on identifying the performance inflection point and gradient stability. The transfer test group includes two modes: "freeze the underlying encoder" and "do not freeze", to determine the impact of the retention of multilingual basic representations on sentiment transfer, thus ensuring the reliability of the final strategy source. After each transfer training session, adversarial examples and highly ironic examples in the target language are extracted to check that the model output is consistent with the direction before transfer. The language classification probability output by the language discrimination module is also statistically analyzed to see whether there is a decrease in the language feature leakage. If the leakage is not mitigated, it is recorded as a failed transfer. Finally, a robustness conclusion for transfer learning is given based on the stability and fluctuation level of the performance curves at each stage.

Figure 7 illustrates the model's performance and training stability in the cross-language emotion transfer task. In Figure 7 (a), the X-axis represents the percentage of target language data (%), and the Y-axis represents the accuracy after transfer (%). Different colored curves correspond to the transfer effects for English, Chinese, Spanish, Arabic, and Russian. It can be observed that as the percentage of data increases, the transfer accuracy for each language gradually increases, but there are slight differences. In Figure 7(b), the X-axis represents the number of training epochs, and the Y-axis represents the stability index values comprising loss fluctuation, gradient norm, feature distance, prediction consistency, and language leakage. The curves of each index indicate that the model's stability begins to improve as training progresses. Compared with the performance of the model early in training, the indices show a rapid decrease, while prediction consistency and feature distance begin to stabilize, which suggests that the model is able to maintain some training stability during cross-language transfer. The rate at which different indices converge also shows that the model can potentially adapt to some of the noise, uncertainty, or differences introduced by language.

Transfer performance was slightly improved with an increasing data ratio, but very gradually converged under low-resource conditions, particularly for cross-language performance, due to difficulties in balancing language invariance with effectively discriminating features for the emotion representation in the early stages or iteration of adversarial training. The improvement in gradient stability and

Table 1 Results of the Complex Emotions and Irony Recognition Ability Test.

Sample Type	Sample Size	Correct Predictions	Agreement with Human Labels (%)	Attention Focus Coverage (%)	Concentrated Misclassification	Negative Sarcasm Bias (%)
Sarcasm	150	128	85.3	72.5	Implicit Attack	12
Irony	120	96	80	70.1	Emotion Masking	15.8
Implicit Attack	100	82	82	68.4	Sarcasm	10.5
Emotion Masking	130	104	80	66.2	Irony	13.2
Overall	500	410	82	69.3	-	12.9

prediction consistency at higher data ratios suggests that the adversarial mechanism is working to effectively limit interference from the language feature; however, the use of extensions to optimize the generalization ability into low-data regions still requires some additional consideration for further optimization.

3.3 Detection of Complex Emotions and Irony

This phase involves constructing test sets of satirical, ironic, implicitly offensive, and emotionally masked texts, comparing the model's output with manually annotated text. The test sets are sourced from multilingual social media platforms, retaining user emoticons, line breaks, and colloquialisms, and not removing repeated punctuation to ensure authentic representation of tone and conditions. During inference, output smoothing and post-processing logic are disabled, directly recording the emotion tag score and the highest tag category. For each sample, the model's attention weight position is recorded in order to check whether the focus is concentrated on interjections, contrasting words, and negation structures, rather than semantically irrelevant positions. After model inference, the results are verified by three independent annotators, and controversial samples undergo a second review to ensure alignment and labeling are unbiased. Difficult examples, such as reverse emotional statements and cross-linguistic puns, are recorded during evaluation, and their hit rates are calculated separately. If the model shows a strong bias trend in the negative satirical category, targeted difficult examples are applied to test the model's output consistency under stress. The final output measures the model's discrimination stability and misclassification concentration positions in satirical and mixed emotion scenarios.

Table 1 demonstrates the model's performance on tasks involving complex emotion and ironic text recognition on a test set of four sample types—irony, satire, implicit attack, and emotion masking—along with the overall performance. The model correctly predicted 128 ironic text samples, or 85.3% of the total samples, with an attention coverage of 72.5%. This demonstrates that the model appears to perform robustly when focused attention is consistently applied to key tone words and contrast structure when determining ironic tone. The model correctly predicted 96 satirical text samples, accurately predicting 80% of the samples with attention coverage of 70.1%, evidenced by consistently lower coverage than ironic text. This suggests that the model retains some bias for complex sentence structure when facing emotion masking. There were 82% correct predictions in the implicit attack sample type, with a negative irony bias, once again extremely low at only a 10.5% recognition in potential attack intentions, indicating strong stability when handling implicit attack. The performance in the emotion masking sample was: 104 correct predictions, accurately including 80% of the samples, with an attention coverage of 66.2% indicating some recognition misjudgments, specifically through the satirical category when maintaining multi-layered tone modifications.

In summary, the model accurately classified 410 samples out of 500 (82% overall accuracy), with an attention coverage of 69.3% and a negative irony bias of 12.9%. This indicates that the model has a strong performance in terms of differentiating complex emotions and irony. However, more work is necessary to improve the attention distribution in order to minimize misclassification clusters in the presence of competing tones. In addition, the attention coverage indicates that the model's implicit tonal sensitivity is still compromised, and contextual modeling must be improved to counteract local dependency influences.

3.4 Consistency Detection of Multimodal Sentiment Analysis

This stage of the study examines the reliability of emotional output under varying combinations of input that consist of text, emoji, and imagery. Three varying input types are constructed: text only, text + emoji, and text + image. Each text has a matching emoji and a related image label, which is input into the system, while output differences are recorded through inputs. During testing, the text portion remains completely steady while visual and symbolic information is added incrementally, observing if there is a reasonable directional shift in output emotion. If the text is overtly emotional, and the emoji and image do not provide further information, then the output of the model should maintain directional stability. If there is visual content that provides emotional supplementation, the magnitude of emotional shift is noted to see whether the model accurately interprets the multimodal signals. The return on output of emoji and images is noted in order to determine whether the multimodal information creates possible residual emotional interference. During testing, when multimodal information isinput into the model, auxiliary decoding strategies were pushed aside so as to eliminate the model's ability to follow a bias path under multimodal

Table 2 Results of the Multimodal Emotion Analysis Consistency Test.

Input Combination	Sample Size	Output Direction Consistency (%)	Average Emotion Shift	False Trigger Count	Output Recovery Consistency (%)
Text Only	200	92.5	0	3	100
Text + Emoji	200	90.8	0.12	5	97
Text + Image	200	91.2	0.18	6	95
Text + Emoji + Image	200	89.7	0.2	7	94

Table 3 Results of robustness tests on dialects and slang.

Dialect/Slang Category	Sample Size	Baseline Accuracy (%)	Extended Mechanism Accuracy (%)	Embedding Concentration Change (%)	Fine-tuning Convergence Epochs	Error Word Ratio (%)
High-resource Dialect	150	83.5	90.2	12.1	3	7
Medium-resource Dialect	120	78.2	85.4	10.3	4	9.2
Low-resource Dialect	100	70.1	78	8.7	5	12.5
Internet Slang	130	75.4	82.3	9.5	3	10
Spoken Abbreviations	110	80.3	87	11	4	8.3

inputs, while simultaneously observing and documenting atypical false triggering examples during testing, such as negative associations when no emotional symbols are in the text, or mappings were incorrectly put in slang meanings.

Table 2 depicts the performance of the model in terms of its ability to detect the consistency of multimodal emotion parsing in order to compare the consistency and reliability of emotion output when parsing a combination of input modalities. When only the text is input, the output direction consistency is highest at 92.5%, the average emotion shift is 0 (indicating the model was consistent in detection), and output recovery consistency is 100% (the model is accurately capturing the emotion information of pure text with no influence externally). When the emoji is added as an input modality, the output direction consistency decreases slightly to 90.8%, the average emotion shift increases to 0.12, the number of false triggers increases to 5, and the output recovery consistency is 97%. This slight shift may indicate that the model is experiencing a subtle fluctuation when processing the symbolic input. In the case of adding images, the output direction consistency is 91.2%, the emotion average shift is 0.18, there are 6 false triggers, and the output recovery is 95%. In this case, visual input enhances emotion capture, but it can also apply a small amount of bias in processing the information.

When combined with text, emoji, and image inputs, consistency dropped to 89.7%, with a maximum emotion shift of 0.2 and 7 false triggers. Output recovery consistency was 94%, indicating that multimodal information provided more emotional cues. However, the model still maintained high overall stability under high interference conditions. This result suggests that the model has a strong ability to integrate emotional signals under multimodal inputs, but complex information increases the risk of output shift and false triggers.

3.5 Robustness Test of Dialects and Slang

In this phase. multilingual dialect corpora and user-generated slang samples are chosen, which follow the level of colloquialism/level of dialect, and added to the model. It keeps the spellings, surviving words, abbreviations, and local cultural symbols without any grammatical changes to create the test set, allowing the model to deal directly with authentic, non-standard text rather than regularized text. The dynamic dictionary expansion element associated with the CWR is not enabled during overall vocabulary inference. The baseline performance is measured the first time for the items in the test set, and is retested when the expanded model with the new vocabulary is deployed. Since they are mainly concerned with the magnitude of changes in the word vectors and/or movement towards a concentration, using items identified by the model to assess changes in vocabulary wording, with failed recognition samples if response time allows follow-up tasks. These incorrect samples are added to a short-term fine-tuning loop, and the convergence speed is observed to determine the model's ability to absorb new words. A repeated injection mechanism is implemented for low-resource dialect samples to determine whether the model's accuracy quickly improves after a small amount of exposure. If it still remains in the confusion range, it is considered to have insufficient dialect absorption capacity. Finally, an evaluation conclusion is drawn based on the convergence time, parameter variation magnitude, and error word trends of the two rounds of testing.

Table 3 shows the model's performance in dialect and slang robustness tests. By comparing the baseline recognition accuracy with the accuracy under the dynamic expansion mechanism, the model's adaptability to non-standard text is evaluated. For high-resource dialects, the baseline accuracy is 83.5%, which improves to 90.2% after enabling the expansion mechanism, with an increase of 12.1% in the concentration of the embedding vector distribution. Fine-tuning convergence requires only 3 rounds, and the error word ratio is 7%, indicating the model's strong and rapid adaptability to resource-rich dialects. In regard to low-resource dialects, baseline accuracy starts at just 70.1% while accuracy after dictionary expansion improved to 78%, and embedding concentration increased from 8.7%. Obtaining fine-tuning convergence required 5 rounds, and the error word ratio was as high as 12.5% indicating continued difficulty in recognizing slang due to the lack of data?. In the case

of internet slang performance, baseline accuracy was 75.4% before increasing to 82.3% after dictionary expansion, with embedding concentration improved from 9.5%. Obtaining fine-tuning convergence required only 3 rounds while the error word ratio was 10%, demonstrating improvement in the model's ability to adapt to online language.

Overall, the combination of applying dynamic dictionary expansion and fine-tuning mechanisms effectively improves the model's accuracy in recognizing dialects, slang, and colloquial abbreviation texts while also promoting an increased concentration in word vector distribution, contributing to strong robustness and adaptability to non-standard expressions. The improved accuracy of dialect is attributed to the rapid dictionary dynamicity to adapt to non-standard expressions. These results demonstrate the potential for modality fusion and dynamic dictionary to help address the issues of relying on a single implementation of a modality or relying solely on standard language, while continuing to improve the component-specific area and subsequent performance in recognition accuracy. However, the noise-suppressing mechanisms require further development for optimal performance.

4. CONCLUSION

In this paper, a comprehensive algorithm system is proposed for emotion identification in multilingual social text contexts. Using a step-wise approach, the study applied a cross-lingual semantic alignment-based mechanism, a multimodal emotion symbol fusion network, a dialect and slang adaptive expansion strategy, a cross-lingual emotion transfer and adversarial optimization structure, and an emotion hierarchical classification framework, forming an end-to-end multilingual emotion understanding framework. First, the model aligned the multilingual semantic space, then added to the emotion expression capture capability with image-guided and emoji-based accompanied emotion representations; expanded dialect word vector representations with slang and non-standard expressions; then applied the adversarial transfer and perturbation optimization model to fortify the stability of the model in cross linguistic contexts and noisy environments; then employed an emotion hierarchical classification approach to output both polarity (positive-negative) and recursively composed fine-grained emotion labels. The experimental findings demonstrated that the model can effectively and accurately initiate cross-lingual alignment, irony coding judgment, emoji and dialect modality emotion representation, and multimodal information fusion, showing an excellent capacity to maintain emotion discrimination accuracy when considering the impact on language domains on narrative and content creation of sentiment and emotion dissociation. The results provide extensive evidence to indicate that the overall meaning recognition and affective experience schematic method is appropriate for the evaluation of complex online language situations. Hence, this study provides a basis for future approaches to globalized text sentiment analysis, cross-cultural user emotion recognition, and intelligent systematic content analyses across multilingual social platforms.

FUNDING

This work was supported by the Hainan Higher Education Teaching Reform Research Project "Three Dimensions, Four Chains, Cross-border Integration: Research on the Reform of Linguistics Talent Cultivation Model Empowered by Digital Technology" (Hnjg2026ZC-105).

REFERENCES

1. Zehra W, Javed AR, Jalil Z, et al. Cross corpus multilingual speech emotion recognition using ensemble learning[J]. Complex & Intelligent Systems, 2021, 7(4): 1845–1854.
2. Zhu X, Lei Y, Li T, et al. Metts: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32(1): 1506–1518.
3. Resende Faria D, Weinberg AI, Ayrosa P P. Multimodal affective communication analysis: Fusing speech emotion and text sentiment using machine learning[J]. Applied Sciences, 2024, 14(15): 6631.
4. Chan JYL, Bea KT, Leow SMH, et al. State of the art: a review of sentiment analysis based on sequential transfer learning[J]. Artificial Intelligence Review, 2023, 56(1): 749–780.
5. Das R, Singh T D. Multimodal sentiment analysis: a survey of methods, trends, and challenges[J]. ACM Computing Surveys, 2023, 55(13s): 1–38.
6. Wang S, Shibghatullah AS, Iqbal TJ, et al. A review of multimodal-based emotion recognition techniques for cyberbullying detection in online social media platforms[J]. Neural Computing and Applications, 2024, 36(35): 21923–21956.
7. Kumar P, Pathania K, Raman B. Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data[J]. Applied Intelligence, 2023, 53(9): 10096–10113.
8. Alam MS, Mrida MSH, Rahman M A. Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI)[J]. American Journal of Scholarly Research and Innovation, 2025, 4(01): 63–100.
9. Leon M. From Lexicons to Transformers: An AI View of Sentiment Analysis[J]. Journal of Intelligent Communication, 2025, 4(2): 13–25.
10. Paul A, Nayyar A. A context-sensitive multi-tier deep learning framework for multimodal sentiment analysis[J]. Multimedia Tools and Applications, 2024, 83(18): 54249–54278.
11. Aggarwal S, Pandey A, Vishwakarma D K. Extracting cross-modal semantic incongruity with attention for multimodal sarcasm detection[J]. Applied Intelligence, 2025, 55(12): 1–22.
12. Brook G V. Multimodal Sentiment Analysis Using Transformer-Based Architectures: A Fusion of Text, Audio, and Visual Cues[J]. International Journal of Scientific Research in Computer Science and Information Technology (IJSRCSIT), 2024, 5(1): 1–7.
13. Chakhtouna A, Sekkate S, Adib A. Efficient bimodal emotion recognition system based on speech/text embeddings and ensemble learning fusion[J]. Annals of Telecommunications, 2025, 80(5): 379–399.
14. Alahmadi K, Alharbi S, Chen J, et al. Generalizing sentiment analysis: a review of progress, challenges, and emerging directions[J]. Social Network Analysis and Mining, 2025, 15(1): 1–28.

15. Hughes E, Ward N, Bennett C, et al. Cross-Lingual Instruction Alignment in Large Language Models via Lightweight Prompt Distillation[J]. International Journal of Advance in Applied Science Research, 2025, 4(4): 9–16.
16. Pakray P, Gelbukh A, Bandyopadhyay S. Natural language processing applications for low-resource languages[J]. Natural Language Processing, 2025, 31(2): 183–197.
17. Muneer I, Saeed A, Adeel Nawab R M. Cross-Lingual English–Urdu Semantic Word Similarity Using Sentence Transformers[J]. The European Journal on Artificial Intelligence, 2025, 38(1): 21–34.
18. Sharma A, Aggarwal M. A holistic review of image-to-text conversion: techniques, evaluation metrics, multilingual captioning, storytelling and integration[J]. SN Computer Science, 2025, 6(3): 1–26.
19. Jumanazarov M. DEVELOPMENT OF DEEP LEARNING ARCHITECTURES FOR LOW-RESOURCE LANGUAGE PROCESSING: THE CASE OF UZBEK[J]. Journal of Multidisciplinary Innovation in Science and Education, 2025, 1(3): 17–31.
20. Ghasemi R, Momtazi S. How a Deep Contextualized Representation and Attention Mechanism Justifies Explainable Cross-Lingual Sentiment Analysis[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22(11): 1–15.
21. Hao LW, Liu R K. Transfer Learning Approach for Sentiment Analysis in Low-Resource Austronesian Languages Using Multilingual BERT[J]. Journal of Technology Informatics and Engineering, 2025, 4(1): 75–94.
22. Zhao S, Hong X, Yang J, et al. Toward label-efficient emotion and sentiment analysis[J]. Proceedings of the IEEE, 2023, 111(10): 1159–1197.
23. Liu G, Dong Y, Wang K, et al. A cross-lingual sentence pair interaction feature capture model based on pseudo-corpus and multilingual embedding[J]. AI communications, 2022, 35(1): 1–14.
24. Shukri IM, Hassan R, Zakaria Z. A Systematic Mapping Study of Code-mixed Embedding Technique[J]. International Journal of Innovative Computing, 2025, 15(1): 119–129.
25. Taylor S, Fauzi F. Multimodal Sentiment Analysis for the Malay Language: New Corpus using CNN-based Framework[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2025, 24(8): 1–26.
26. Askari MH, Shahzad A, Faraz A, et al. EFFECTIVE SPEECH EMOTION RECOGNITION USING R-CNN & BLSTM[J]. Kashf Journal of Multidisciplinary Research, 2025, 2(06): 293–309.
27. Gavali MP, Verma A. Ensemble of large self-supervised transformers for improving speech emotion recognition[J]. International Journal of Data Mining, Modeling and Management, 2025, 17(2): 217–244.
28. Qi F, Zhang H, Yang X, et al. A versatile multimodal learning framework for zero-shot emotion recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(7): 5728–5741.
29. Jia L, Ma T, Rong H, et al. Affective region recognition and fusion network for target-level multimodal sentiment classification[J]. IEEE transactions on emerging topics in computing, 2023, 12(3): 688–699.
30. Yang C, Wang Y, Han L, et al. Fine-grained image emotion captioning based on Generative Adversarial Networks[J]. Multimedia Tools and Applications, 2024, 83(34): 81857–81875.
31. Ma F, Li Y, Ni S, et al. Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN[J]. Applied Sciences, 2022, 12(1): 527.
32. Wu Y, Mi Q, Gao T. A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions[J]. Biomimetics, 2025, 10(7): 418.
33. Zhao S, Ren J, Zhou X. Cross-modal gated feature enhancement for multimodal emotion recognition in conversations[J]. Scientific reports, 2025, 15(1): 30004.
34. Ai Y, Zhang Y, Wang J, et al. Optimizing the interaction of service robots in elderly care institutions using multi-modal emotion recognition system based on transfer learning[J]. Discover Artificial Intelligence, 2025, 5(1): 49.
35. Wu Y, Zhang S, Li P. Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features[J]. Scientific Reports, 2025, 15(1): 8855.
36. He P, Yu J, Ge C, et al. Domain-separated bottleneck attention fusion framework for multimodal emotion recognition[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(4): 1–21.
37. W. Li. (2019). Oral English Recognition Teaching System Based on Natural Language Processing and Emotional Analysis. International Journal of Engineering Intelligent Systems, 287(3), 95–102.
38. Guo, W., Xu, Y., Shi, Z., Hu, X., Zhang, H. & Feng, L. (2026). Exploration of Enhancement Effect in Natural Language Understanding Task Based on BERT Model with Integrated Power Knowledge Graph. International Journal of Engineering Intelligent Systems, 34(1), 17–26.