Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 15, No. 3, 2026

Study on Sentiment Analysis Methods for Uyghur Videos

Jiazhi Wang, Jiarong Zhang *, Haijiao Guan, Wenxiu He

School of Information Engineering, Tarim University, Aral, Xinjiang, 843300, China
* **Corresponding author**: Jiarong Zhang

Abstract: As a representative minority language in China, Uyghur has important application value in cross-regional communication, trade exchanges, and multilingual information services. Compared with high-resource languages, research on sentiment analysis for Uyghur videos remains limited, especially in terms of systematically constructed multimodal sentiment annotation datasets and highly adaptable analytical models. To address this issue, this paper investigates the task of sentiment analysis for Uyghur videos. A Uyghur video sentiment dataset containing three modalities—text, audio, and video—was constructed. Based on the ToxVidLM framework, the original audio and text encoders were replaced: Whisper and H-RoBERTa were substituted with XLS-R-uyghur-cv and CINO, respectively, which are better suited to Uyghur-language scenarios, while VideoMAE was retained as the video encoder, thereby forming an improved model. Experimental results show that the improved model outperforms the original framework on three sentiment-related tasks, namely emotional tension, emotion intensity, and sentiment polarity. This indicates that, in low-resource language scenarios, the combination of dataset construction and encoder adaptation can effectively improve the performance of sentiment analysis for Uyghur videos.

Keywords: Video Sentiment Analysis; Low-resource Language; Multimodality.

1. Introduction

With the advancement of technologies such as natural language processing, speech understanding, and computer vision, sentiment analysis has gradually evolved from single-text analysis to multimodal analysis integrating text, speech, and video information [1]. Compared with unimodal approaches, multimodal sentiment analysis can simultaneously utilize cues such as semantic content, vocal prosody, and facial or action expressions, and is therefore closer to the process of emotional expression in real communication. In recent years, the widespread application of pretrained models has further improved the representation capabilities of text, speech, and video features, providing new technical support for multimodal sentiment analysis research.

However, existing studies mainly focus on high-resource languages such as English and Chinese, while relatively little attention has been paid to low-resource languages such as Uyghur[2]. The main challenges faced by sentiment analysis for Uyghur videos lie, on the one hand, in the lack of relatively complete multimodal annotated datasets, and on the other hand, in the insufficient adaptability of existing general-purpose models, most of which are trained on high-resource languages, to Uyghur text structures, speech prosody, and video expression scenarios[3]. Existing studies have shown that sentiment recognition for low-resource languages can achieve certain improvements through methods such as cross-lingual transfer and multilingual pretraining[4–5]. However, in video scenarios, how to fully exploit the complementary information of text, audio, and video remains a question worthy of further study.

2. Research Status

In general, sentiment recognition research has undergone a gradual expansion from unimodal to multimodal settings and from high-resource to low-resource languages. In the area of text sentiment analysis, the research focus has shifted from traditional feature-engineering methods to pretrained language models represented by Transformer, with continuous enhancement in text semantic representation capability[6]. The development of generative models has further expanded the application space of sentiment analysis methods[7]. In speech emotion recognition, self-supervised learning has significantly improved speech feature representation. Models such as HuBERT and WavLM can more effectively extract emotion-related information such as intonation, rhythm, and pauses[8–9], while emotion2vec further strengthens modeling capability for emotional representation[10]. In video understanding, methods such as TimeSformer and VideoMAE have promoted the development of visual sentiment analysis from static facial expression recognition to dynamic video understanding [11–12].

On this basis, multimodal sentiment analysis has gradually become an important research direction. Related studies are no longer limited to simple feature concatenation, but place greater emphasis on cross-modal interaction, fusion modeling, and joint optimization[13–14]. At the same time, task design has gradually extended from a single sentiment polarity judgment to multidimensional joint characterization involving emotion categories and sentiment intensity, thereby providing a finer-grained analytical framework for describing complex emotional states[15–16].

Overall, existing achievements have provided a solid methodological foundation for multimodal sentiment analysis, but most related studies are based on high-resource languages and mature public datasets. For sentiment analysis of Uyghur videos, what truly needs to be addressed is not merely model transfer, but also data construction, modality representation, and language-specific adaptation. Therefore, conducting research on multimodal sentiment analysis for Uyghur videos under a unified task framework still has certain research significance

3. Method

3.1. Construction of a Multimodal Sentiment Dataset for Uyghur Videos

To support subsequent model training and experimental analysis, this paper constructs a multimodal sentiment dataset for the task of sentiment analysis on Uyghur videos through three stages: data collection and cleaning, data preprocessing, and sentiment annotation. The final dataset contains three modalities—text, audio, and video—and provides a unified data foundation for subsequent model training, result comparison, and experimental analysis.

3.1.1. Data Collection and Cleaning

The raw data mainly come from publicly accessible Uyghur video resources, covering a variety of scenarios such as trade, communication, conversation, news, entertainment, education, and social interaction. This was intended to ensure, as much as possible, diversity in topic types, expression styles, and speaking styles, while also taking into account accent differences and expression habits across different regions. Since public videos vary greatly in terms of clarity, completeness, and language purity, the original samples could not be directly used for subsequent modeling. Therefore, before preprocessing, unified cleaning was required. During the cleaning process, the focus was placed on removing duplicate uploads, corrupted files, audio-video asynchrony, samples with severe language mixing, and samples with weak emotional information, while retaining video clips that were complete in content, emotionally distinguishable, and suitable for joint analysis across three modalities. Through the above cleaning process, the samples were initially standardized in terms of content integrity, modality consistency, and emotional distinguishability, thus providing a relatively stable data basis for subsequent three-modal preprocessing and sentiment annotation.

3.1.2. Data Preprocessing

After sample screening and cleaning, this paper further performs unified preprocessing on the three modalities of text, audio, and video in order to reduce differences in clarity, subtitle quality, background noise, and temporal boundaries among videos from different sources, and to ensure consistency in subsequent sentiment annotation and model input. For the video modality, natural segments that could relatively completely express a single emotional content were used as the basic units for segmentation and organization. While preserving major expressions, actions, and temporal variation information, efforts were made to avoid semantic interruption or emotional fragmentation caused by mechanical segmentation. The text modality was mainly obtained from original subtitles and speech transcription results, and was manually proofread in combination with video context to correct obvious transcription deviations and semantic errors. At the same time, noise information such as invalid symbols and redundant punctuation was removed so that the text content could be as consistent as possible with the actual expression. The audio modality was extracted from the original videos and segmented according to boundaries consistent with the video modality. Through format unification and basic denoising processing, emotion-related cues such as intonation, speech rate, pauses, and stress were preserved as much as possible. After the above processing, the three modalities achieved good consistency in sample granularity, temporal boundaries, and input form, thereby providing a reliable data basis for subsequent manual annotation, structured organization, and multimodal modeling.

3.1.3. Data Annotation and Dataset Construction

Considering that emotional expression in Uyghur videos is not merely reflected as a simple positive or negative tendency, but is often accompanied by changes in emotional state and differences in expressive strength, this paper did not adopt a single sentiment polarity label. Instead, by integrating information from text, speech, and video modalities, it constructed a multitask annotation system consisting of emotional tension, emotion intensity, and sentiment polarity. Among them, sentiment polarity is used to characterize the overall emotional direction of a video; emotion intensity reflects the explicit degree of emotional expression; and emotional tension describes the sense of tension, emotional fluctuation, and state variation during the expression process. These three tasks complement the characterization of video sentiment from the three aspects of direction, degree, and state, thereby better fitting the actual characteristics of emotional expression in Uyghur videos.

Before formal annotation, this paper first referred to relevant task settings in existing multimodal sentiment analysis research and, based on the characteristics of Uyghur video corpora in linguistic expression, speech prosody, and visual presentation, formulated unified annotation guidelines. The category definitions, judgment criteria, and easily confused cases for each task were clearly specified. On this basis, a portion of representative samples was selected for pilot annotation. Problems encountered during pilot annotation, such as unclear category boundaries and large discrepancies in sample judgments, were analyzed in a concentrated manner, and the annotation rules were further revised and refined. In the formal annotation stage, this paper adopted a combination of independent dual annotation and disagreement review. Samples with obvious annotation disagreements were re-examined in order to improve the consistency and reliability of the annotation results. After annotation was completed, the samples were further organized in a structured manner, with the correspondence among text, audio, and video modalities, sample IDs, and task labels uniformly recorded. The training set, validation set, and test set were then split by original video units so as to avoid potential information leakage caused by similar segments from the same source being distributed across different subsets. The final dataset can simultaneously support the three tasks of emotional tension, emotion intensity, and sentiment polarity, thus providing a unified data basis for subsequent model training, result comparison, and experimental analysis.

3.2. Encoder Replacement

Based on the ToxVidLM multimodal modeling framework [17], this paper improves the sentiment analysis model for Uyghur videos. Considering that this task involves multiple sources of information, including textual semantics, speech prosody, and dynamic facial expressions in videos, this paper retains the original multimodal joint modeling idea of the framework and takes text, audio, and video as unified inputs. Joint prediction of emotional tension, emotion intensity, and sentiment polarity is achieved through modality feature extraction, representation mapping, and fusion modeling. In the specific implementation, the focus of improvement is placed on the adaptation of modal encoders. For the text

modality, CINO is used as the encoder to generate contextual semantic representations of the input text. For the audio modality, XLS-R-uyghur-cv is used as the speech encoder to extract prosodic features such as intonation, rhythm, stress, and pauses in speech,hereinafter referred to as XLS-R-uy. For the video modality, VideoMAE is adopted to encode the spatiotemporal features of video clips and extract facial changes, local actions, and dynamic visual information from continuous frame sequences,hereinafter referred to as VM. After the independent encoding of the three modalities is completed, the features of each modality are further mapped into a unified representation space and jointly modeled using the original multimodal fusion method, thus constructing an improved model for the task of sentiment analysis on Uyghur videos.

4. Experimental Results and Analysis

To verify the effectiveness of the encoder adaptation method, this paper conducts comparative experiments under the same task settings and a unified fusion framework, comparing the original encoder configuration with the adapted encoder configuration. Accuracy and F1 are used as evaluation metrics. Accuracy and F1 are presented in Table 1 and Table 2, respectively.

Table 1. Accuracy Results

Model	Emotional Tension	Emotion Intensity	Sentiment Polarity
Whisper+VM+H-RoBERTa	0.6085	0.5136	0.6282
XLS-R-uy+VM+CINO	0.7113	0.6069	0.6596

Table 2. F1-score Results

Model	Emotional Tension	Emotion Intensity	Sentiment Polarity
Whisper+VM+H-RoBERTa	0.6132	0.5095	0.5988
XLS-R-uy+VM+CINO	0.7125	0.6054	0.6529

The experimental results indicate that, after encoder adaptation, the model performance improves on all three tasks: emotional tension, emotion intensity, and sentiment polarity. Specifically, under the original encoder configuration, the Accuracy and F1 scores are 0.6085 and 0.6132 for the emotional tension task, 0.5136 and 0.5095 for the emotion intensity task, and 0.6282 and 0.5988 for the sentiment polarity task. Under the adapted encoder configuration, the Accuracy and F1 scores for the three tasks improve to 0.7113 and 0.7125, 0.6069 and 0.6054, and 0.6596 and 0.6529, respectively. Overall, the improved model outperforms the original configuration on all three tasks, with more obvious gains in the emotional tension and emotion intensity tasks, indicating that encoder adaptation plays a positive role in modeling emotional features in Uyghur videos.

Further analysis shows that although the general-purpose text encoder and speech encoder in the original configuration possess strong general representation capabilities, they still have certain limitations in adapting to text structural features and speech prosodic features in low-resource Uyghur scenarios. After adopting text and speech encoders that are more targeted to Uyghur-language scenarios, the model can more effectively extract semantic and structural information from text, as well as emotional prosodic cues from speech, thereby improving multimodal sentiment recognition performance. This suggests that, in the task of sentiment analysis for Uyghur videos, performance improvement depends not only on the multimodal fusion framework itself, but is also closely related to the degree of adaptation of the underlying encoders to the specific language scenario.

5. Conclusion

This paper investigates the task of sentiment analysis for Uyghur videos. To address problems such as insufficient multimodal sentiment data and limited adaptability of general-purpose encoders in low-resource language scenarios, a Uyghur video sentiment dataset containing three modalities—text, speech, and video—was constructed, along with a multitask annotation system composed of emotional tension, emotion intensity, and sentiment polarity. On this basis, adaptive improvements were made to the text and speech encoders in the original multimodal modeling framework, and combined with spatiotemporal feature representation of the video modality, a multimodal recognition method for sentiment analysis of Uyghur videos was formed. Experimental results show that, after encoder replacement, the recognition performance of the model improves on all three tasks, demonstrating that, in low-resource language video sentiment analysis, language-specific adaptation of the underlying encoders plays an important role in enhancing multimodal representation capability and overall recognition performance. To a certain extent, this study verifies the feasibility of multimodal sentiment analysis methods in Uyghur scenarios, and can also provide a reference for sentiment recognition research on videos in minority languages. Future work may further focus on expanding data scale, refining fine-grained emotion category division, and developing more effective cross-modal fusion mechanisms, so as to improve the robustness and generalization ability of the model in complex real-world scenarios.

References

[1] Wu Y, Mi Q W, Gao T H. A comprehensive review of multimodal emotion recognition: techniques, challenges, and future directions[J]. Biomimetics, 2025, 10(7): 418.

[2] Gladys A A, Vetriselvi V. Sentiment analysis on a low-resource language dataset using multimodal representation learning and cross-lingual transfer learning[J]. Applied Soft Computing, 2024, 157: 111553.

[3] Chen L, Guan S, Huang X, et al. Cross-lingual multimodal sentiment analysis for low-resource languages via language family disentanglement and rethinking transfer[C]//Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 6513-6522.

[4] Qin L, Chen Q, Zhou Y, et al. A survey of multilingual large language models[J]. Patterns, 2025, 6(1): 101118.

[5] Xu Y, Hu L, Zhao J, et al. A survey on multilingual large language models: corpora, alignment, and bias[J]. Frontiers of Computer Science, 2025, 19(11): 1911362.

[6] Areshey A, Mathkour H. Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet[J]. Expert Systems, 2024, 41(11): e13701.

[7] Krugmann J O, Hartmann J. Sentiment Analysis in the Age of Generative AI[J]. Customer Needs and Solutions, 2024, 11(1): 3.

[8] Hsu W N, Bolte B, Tsai Y H H, et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[9] Chen S, Wang C, Chen Z, et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.

[10] Ma Z, Zheng Z, Ye J, et al. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation[C]//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok: Association for Computational Linguistics, 2024: 15747-15760.

[11] Bertasius G, Wang H, Torresani L. Is Space-Time Attention All You Need for Video Understanding?[C]//Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, 2021, 139: 813-824.

[12] Tong Z, Song Y, Wang J, et al. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training[C]//Advances in Neural Information Processing Systems 35. 2022.

[13] Sun J, Han S, Ruan Y P, et al. Layer-wise Fusion with Modality Independence Modeling for Multi-modal Emotion Recognition[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 658-670.

[14] Wu Z, Gong Z, Koo J, et al. Multimodal Multi-loss Fusion Network for Sentiment Analysis[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico: Association for Computational Linguistics, 2024: 3588-3602.

[15] Firdaus M, Chauhan H, Ekbal A, et al. MEISD: A Multimodal Multi-Label Emotion, Intensity and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020: 4441-4453.

[16] Yu W, Xu H, Meng F, et al. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 3718-3727.

[17] Maity K, Poornash A S, Saha S, et al. ToxVidLM: A Multimodal Framework for Toxicity Detection in Code-Mixed Videos[C]//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024: 11130-11142.