IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/44AI/2024/4445724/
Vol. 2024 (22), Article ID 4445724

Multimodal Emotion Recognition in User-Generated Content: A Comparative Study of State-of-the-Art Models*

Marek SCHROEER[1], Paula VIDAURRETA-APESTEGUIA[2] and
Jose Enrique ARMENDARIZ-INIGO[3]

[1]Osnabrück University of Applied Sciences, Osnabrück, Germany
[2]NAIR Center, Pamplona, Spain
[3]ISC - Institute of Smart Cities - Public University of Navarre, Pamplona, Spain

Correspondence should be addressed to: Marek SCHROEER, marek.schroeer@hs-osnabrueck.de

* Presented at the 44th IBIMA International Conference, 27-28 November 2024 Granada, Spain

Abstract

Understanding people's emotions is a valuable task with wide-ranging applications across multiple domains. Beyond traditional sentiment analysis based on a single data type, such as text, the emergence and advancement of multimodal large language models (MLLMs) have enabled this analysis to encompass various approaches, including emotion recognition in user-generated content (UGC) that involves both text and images. This study evaluates the effectiveness of four state-of-the-art multimodal models —VisualBERT, CLIP, Shikra, and Otter — in detecting emotions from both textual and visual data. These models are trained and validated using the EmoReact, AFEW, and SFEW datasets, capturing diverse emotional cues and employing evaluation metrics such as accuracy, precision, F1 score, and the CIDEr score, which measures the alignment between model-generated and human-interpreted emotional content.
Following this comparative assessment, we apply the top-performing models to specific contexts: genre-classified IMDb movie stills and TripAdvisor hotel reviews focusing on Caribbean hotels. This application aims to analyze how textual-visual congruence and incongruence may enrich insights into user experiences in both contexts. Preliminary results highlight the strengths, weaknesses and best approaches of each model so that they can be efficiently applied to new application contexts. This study suggests that text-image incongruence may provide enriched, multi-faceted consumer insights, which could enhance user experience analysis, particularly for applications in hospitality and media content assessment. The work contributes to advancing multimodal emotion analysis methods, suggesting ways to optimize UGC processing for real-world, data-driven applications.

Keywords: MLLM; UGC Analysis; Emotion Detection.

Introduction

In recent years, emotion recognition in user-generated content (UGC) has emerged as an essential field within digital interaction analysis, particularly in sectors such as marketing, consumer experience, and social media engagement. UGC, including text reviews, images, and videos, offers a huge amount of data for understanding user sentiments

Cite this Article as: Marek SCHROEER, Paula VIDAURRETA-APESTEGUIA and Jose Enrique ARMENDARIZ-INIGO, Vol. 2024 (22) "Multimodal Emotion Recognition in User-Generated Content: A Comparative Study of State-of-the-Art Models" Communications of International Proceedings, Vol. 2024 (22), Article ID 4445724, https://doi.org/10.5171/2024.4445724

and behaviors Choudhary et al (2024). Analyzing emotions within UGC not only helps assess consumer satisfaction but also provides deeper insights into customer interactions and content effectiveness. With the rapid advancement of multimodal models, the capability to recognize emotions from both textual and visual data has improved substantially, enabling a comprehensive approach to UGC analysis.

Traditional emotion recognition methods have typically relied on a single modality, such as text-based sentiment analysis or image-based facial expression recognition, which limits the understanding of complex emotional cues Mellouk and Handouzi (2020), Abbaschian et al (2021), Hossain and Muhammad (2019). Multimodal models that incorporate both text and image inputs provide a richer and more nuanced perspective, making it possible to detect subtle emotional signals that may be overlooked by single-modality models Zhang et al (2024), Shou et al. (2023) [5][6]. This study aims to address this gap by evaluating the performance of four advanced multimodal models—VisualBERT, CLIP, Shikra, and Otter—each of which has shown promise in the field of emotion recognition.

The primary objective of this research is to determine how these models perform when trained on benchmark emotion datasets and applied to real-world UGC. Specifically, we use the EmoReact Nojavanasghari et al (2016), AFEW Kossaifi et al (2017), and SFEW Dhall et al (2012) datasets to train and validate the models, ensuring a broad spectrum of emotions are represented. Following training, the models are tested on two unique UGC datasets: IMDb movie reviews, categorized by genre, and TripAdvisor reviews of Caribbean hotels, focusing on cases with over 1,000 reviews per hotel. By examining metrics such as accuracy, precision, recall, F1 score, and the Consensus-based Image Description Evaluation (CIDEr) score, we assess each model's ability to capture and align emotional content between text and images.

This study is particularly relevant for applications that require robust emotion detection, including customer experience platforms and content recommendation systems. By examining the alignment or divergence of emotions between text and images, our work highlights the potential for multimodal models to enrich UGC analysis, thereby improving insights into consumer behaviors and preferences.

Related Work

The field of emotion recognition has seen significant advancements with the advent of multimodal models, which combine both text and image data for a more holistic analysis of emotional expressions. Early efforts in this domain focused primarily on single-modality models, which, although effective in narrow applications, fell short in accurately capturing complex emotional cues present in UGC Yang et al (2024). Models like VisualBERT, developed by Li et al. (2019), marked an important step in multimodal analysis by integrating image and text data through Transformer architectures, enabling more nuanced emotion detection capabilities. VisualBERT's architecture leverages pre-trained representations from both modalities, allowing it to process visual and linguistic data in parallel, thus enriching emotion recognition. Previous research has shown that VisualBERT performs better in emotion recognition tasks with clear image-text alignment, particularly in straightforward sentiment scenarios de Toledo and Marcacini (2022).

Building on this foundation, CLIP, introduced by Radford et al (2021), employed a contrastive learning approach to align text and image pairs, demonstrating considerable success in tasks like image classification and captioning. CLIP's strength lies in its use of a large-scale dataset to learn associations between textual descriptions and visual content, making it highly adaptable to diverse applications, including UGC analysis. However, while CLIP excels in tasks with explicit textual labels, its performance in nuanced emotion recognition remains an area of ongoing research.

More recent models, such as Shikra Chen et al (2023) and Otter Li et al (2023), have further advanced multimodal capabilities by integrating additional mechanisms to improve text-image alignment in emotion recognition. Shikra, for instance, employs advanced multimodal embeddings that allow it to capture fine-grained emotional states, while Otter uses specialized attention mechanisms to prioritize relevant emotional cues in each modality Zhang, (2024). Although Shikra and Otter have shown strong performance in multimodal alignment tasks-particularly in referential dialogue and image-captioning scenarios Chen et al (2023), Li et al (2023), their use in complex emotion detection within UGC remains limited. This study addresses this gap by providing an in-depth analysis of the capabilities of these models in recognizing nuanced emotions in multimodal UGC.

In addition to model advancements, evaluation metrics have evolved to better assess multimodal outputs. Traditional metrics such as accuracy and F1 score provide a general sense of model performance, yet they do not fully capture the alignment between textual and visual content. The CIDEr metric, developed by Vedantam et al (2015), is a consensus-based metric originally designed for image captioning evaluation, which quantifies the similarity between model-generated captions and human-annotated descriptions. CIDEr is particularly valuable in assessing multimodal models, as it measures how closely a model's output aligns with expected human interpretations. By integrating CIDEr alongside traditional metrics, this study provides a more comprehensive evaluation of each model's effectiveness in capturing the emotional nuances in UGC.

This research builds upon previous work by applying these advanced models to real-world datasets, such as IMDb and TripAdvisor; the latter to understand their performance in diverse UGC scenarios. By comparing these models using CIDEr and traditional metrics, this study aims to contribute insights into the capabilities of multimodal models in accurately reflecting consumer sentiment and enhancing the analysis of UGC.

Methodology

This study evaluates the multimodal emotion recognition performance of four state-of-the-art models—VisualBERT, CLIP, Shikra, and Otter—by training them on benchmark emotion datasets and testing on real-world UGC datasets to assess their ability to detect nuanced emotions.

Datasets

The training and validation datasets include EmoReact, AFEW, and SFEW. These datasets cover a wide spectrum of emotional expressions in varied contexts and environments, making them suitable benchmarks for testing each model's emotion detection accuracy and robustness. EmoReact, for example, is designed specifically for recognizing emotions in video data Nojavanasghari et al (2016), while AFEW and SFEW focus on acted facial expressions in natural settings, capturing complex emotional states Kossaifi et al (2017), Dhall et al (2012). By training on these datasets, we aim to ensure that the models can generalize effectively to detect emotions in both visual and textual UGC.

Models

1. VisualBERT: A Transformer-based model designed to process both text and image data. VisualBERT uses a shared attention mechanism to align and interpret textual and visual features, providing high performance in tasks requiring multimodal data integration [11].

2. CLIP: CLIP employs contrastive learning to establish strong associations between textual descriptions and image data, allowing it to align visual and textual information effectively, especially in cases with explicit textual prompts Radford et al (2021).

3. Shikra: Shikra improves upon previous models by implementing advanced multimodal embeddings, which facilitate capturing subtle emotional distinctions in both image and text modalities. It is particularly effective in recognizing complex emotions that may appear in UGC Chen et al (2023).

4. Otter: Utilizing a specialized attention mechanism, Otter excels at emphasizing relevant emotional cues, even in cases of ambiguous or conflicting data across modalities. Otter's structure supports detailed emotion categorization, enhancing its ability to capture subtle emotional states Li et al (2023).

Testing on UGC

Following training, the models are applied to two distinct UGC datasets: (1) IMDb movie stills, classified by genre, which provides an opportunity to analyze genre-specific emotional indications in visual content; (2) TripAdvisor reviews of Caribbean hotels with a high volume of reviews (>1000), where images and textual reviews are combined to convey customer sentiments. This testing strategy allows for analyzing not only emotion recognition

accuracy but also alignment between image and text.

Evaluation Metrics

We employ standard evaluation metrics, including accuracy, precision, recall, and F1 score, to compare model performance quantitatively. To assess how well the models capture human-like interpretations, we integrate the CIDEr (Consensus-based Image Description Evaluation) score, which is typically used for image captioning. CIDEr evaluates semantic congruence between model outputs and human annotations, thus providing insight into how accurately each model's output aligns with expected human interpretations Vedantam et al (2015).

Model Configuration and Repository Consultation

The models employed in this study were utilized with their default parameter configurations, as provided in their respective libraries and repositories. This approach ensures reproducibility and consistency with the original implementations of the selected methods. Additionally, all repositories and documentation were accessed and reviewed during November and December 2023 to guarantee the most up-to-date configurations and resources were employed during the research.

Results and Discussion

The analysis reveals nuanced model performance in emotion detection and text-image alignment across both UGC datasets.

IMDb Dataset

Otter and Shikra are anticipated to perform best in recognizing complex emotions across various movie genres, with Otter achieving the highest CIDEr scores due to its refined attention mechanism, which effectively isolates genre-specific emotional cues. Otter's accuracy and F1 score are expected to exceed those achieved with VisualBERT and CLIP in cases where the genre includes complex emotional elements, such as suspense or drama. VisualBERT and CLIP, while robust in simpler genre contexts (e.g. action or comedy), may show lower CIDEr scores, suggesting moderate alignment with human interpretations.

TripAdvisor Dataset

Otter and Shikra are anticipated to perform best in recognizing complex emotions across various movie genres, with Otter achieving the highest CIDEr scores due to its refined attention mechanism, which effectively isolates genre-specific emotional cues. Otter's accuracy and F1 score are expected to exceed those achieved with VisualBERT and CLIP in cases where the genre includes complex emotional elements, such as suspense or drama. VisualBERT and CLIP, while robust in simpler genre contexts (e.g. action or comedy), may show lower CIDEr scores, suggesting moderate alignment with human interpretations.

CIDEr Analysis

The CIDEr metric proves particularly insightful for evaluating how well each model captures emotional alignment in multimodal UGC. CIDEr scores indicate that Otter and Shikra maintain higher alignment with human-like interpretations, particularly in complex scenarios. VisualBERT's moderate CIDEr performance highlights its general effectiveness in straightforward emotional cues but suggests a need for refinement in capturing subtler emotional nuances present in high-variance UGC.

Discussion

This preliminary analysis anticipates that each model—Otter, Shikra, VisualBERT, and CLIP—will exhibit unique strengths and limitations in multimodal emotion recognition within user-generated content (UGC). Based on model

architectures and prior research, we hypothesize that Otter and Shikra may outperform in complex, emotion-rich contexts, particularly within genres like drama and suspense, where subtle emotional cues are essential. However, limitations are expected, particularly in Shikra's handling of dense object detection and segmentation, which may impact its utility in scenarios requiring precision with overlapping visual elements. Future enhancements in segmentation or attention mechanisms could address these constraints.

For VisualBERT and CLIP, we anticipate strong performance in simpler genres with direct sentiment cues, yet potential challenges in interpreting layered or ambiguous emotions. VisualBERT's text-image alignment is likely to show strengths in clear, explicit emotional alignment, but it may struggle with more complex or indirect emotional expressions. Similarly, CLIP's contrastive learning approach, though effective in straightforward contexts, might limit flexibility in scenarios requiring subtle, contextual interpretation of emotions. Further exploration into model fine-tuning or cross-modal adjustments could enhance their range in diverse UGC contexts.

This study also proposes using the CIDEr metric to evaluate the alignment of model outputs with human interpretations, specifically in capturing semantic congruence in emotionally complex UGC. While CIDEr provides a valuable starting point, it may require complementary metrics to capture qualitative aspects of emotion that may vary by context and individual interpretation.

Conclusion

This study outlines the anticipated capabilities and limitations of four advanced multimodal models—Otter, Shikra, VisualBERT, and CLIP—in recognizing emotions within user-generated content (UGC). While we expect models like Otter and Shikra to excel in nuanced, emotionally rich contexts, such as genre-specific analysis, anticipated limitations in dense object handling and emotion interpretation across modalities suggest areas for future refinement. VisualBERT and CLIP are hypothesized to perform well in simpler UGC scenarios, yet may require enhancements to effectively manage complex, multi-layered emotional cues.

Importantly, Otter and Shikra have not yet been extensively studied within UGC analysis, particularly for tasks involving emotion detection across multimodal content. This study provides a foundational exploration into their capabilities in this area, offering substantial value for future UGC-focused applications, including sentiment analysis.

Our proposed use of the CIDEr metric as part of the evaluation framework highlights a promising approach to gauge alignment between model output and human emotional interpretation. However, recognizing the limitations of CIDEr in fully capturing qualitative emotional nuances, future studies may incorporate additional metrics for a more comprehensive assessment.

The insights gathered here lay a foundation for empirical testing, where verifying these hypotheses in diverse UGC contexts will be essential. Future research should focus on refining these models for increased adaptability and exploring broader datasets to validate their practical applications in fields like customer feedback analysis.

Acknowledgment

This work is part of the project PID2019-108554RB-I00/ funded by AEI/10.13039/501100011033.

References

- Abbaschian, B.J., Sierra-Sosa, D. and Elmaghraby, A. (2021), 'Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models'. Sensors, 21, 1249.
- Chen K., Zhang Z., Zeng W., Zhang, Zhu F. and Zhao R. (2023), 'Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic', arXiv. [Online], [Retrieved December 05, 2024], https://arxiv.org/abs/2306.15195
- Choudhary M., Chouhan S. S., and Rathore S. S. (2024), 'Beyond Text: Multimodal Credibility Assessment Approaches for Online User-Generated Content', ACM Trans. Intell. Syst. Technol., p. 3673236.

- de Toledo G.L. and Marcacini R.M. (2022), 'Transfer Learning with Joint Fine-Tuning for Multimodal Sentiment Analysis', arXiv. [Online], [Retrieved December 05, 2024], https://arxiv.org/abs/2210.05790.
- Deng Y., Li Y., Xian S. Li L. and Qiu H. (2024), 'Mual: enhancing multimodal sentiment analysis with cross-modal attention and difference loss', Int J Multimed Info Retr, 13(3), 31.
- Dhall A., Goecke R., Lucey S. and Gedeon T. (2012), 'Collecting Large, Richly Annotated Facial-Expression Databases from Movies', IEEE MultiMedia, 19(3), 34–41.
- Hossain M.S. and Muhammad G. (2019), 'Emotion recognition using deep learning approach from audio–visual emotional big data', Information Fusion, 49, 69-78.
- Kossaifi J., Tzimiropoulos G., Todorovic S. and Pantic M. (2017), 'AFEW-VA database for valence and arousal estimation in-the-wild', Image and Vision Computing, 65, 23–36.
- Li B., Zhang Y., Chen L., Wang J., Yang J. and Liu Z. (2023), 'Otter: A Multi-Modal Model with In-Context Instruction Tuning', arXiv. [Online], [Retrieved December 05, 2024], https://arxiv.org/abs/2305.03726
- Li L.H., Yatskar M., Yin D., Hsieh C.J. and Chang K.W. (2019), 'VisualBERT: A Simple and Performant Baseline for Vision and Language', arXiv. [Online], [Retrieved December 05, 2024], https://arxiv.org/abs/1908.03557
- Mellouk W. and Handouzi W. (2020), 'Facial emotion recognition using deep learning: review and insights', Procedia Computer Science, 175, 689-694.
- Nojavanasghari B., Baltrušaitis T., Hughes C.E. and Morency L.P. (2016), 'EmoReact: a multimodal approach and dataset for recognizing emotional responses in children', Proceedings of the 18th ACM International Conference on Multimodal Interaction, in ICMI '16, Oct. 2016, New York, NY, USA, 137–144.
- Radford A. et al. (2021), 'Learning Transferable Visual Models From Natural Language Supervision', in Proceedings of the 38th International Conference on Machine Learning, PMLR, Jul. 2021, 8748–8763
- Shou Y., Meng T., Ai W., Yin N. and Li K. (2023), 'A Comprehensive Survey on Multi-modal Conversational Emotion Recognition with Deep Learning', arXiv. [Online], [Retrieved December 05, 2024], https://arxiv.org/abs/2312.05735.
- Vedantam R., Zitnick C.L. and Parikh D. (2015), 'CIDEr: Consensus-Based Image Description Evaluation', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7-12 June 2015, Boston, MA, USA, 4566–4575.
- Yang H. et al (2024), 'Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey', arXiv. [Online], [Retrieved December 05, 2024], https://arxiv.org/abs/2406.08068.
- Zhang L. (2024), 'User emotion recognition and indoor space interaction design: a CNN model optimized by multimodal weighted networks', PeerJ Comput. Sci., 10, e2450.
- Zhang S., Yang Y., Chen C., Zhang X, Leng Q. and Zhao X. (2024), 'Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects', Expert Systems with Applications, 237(C), 121692.