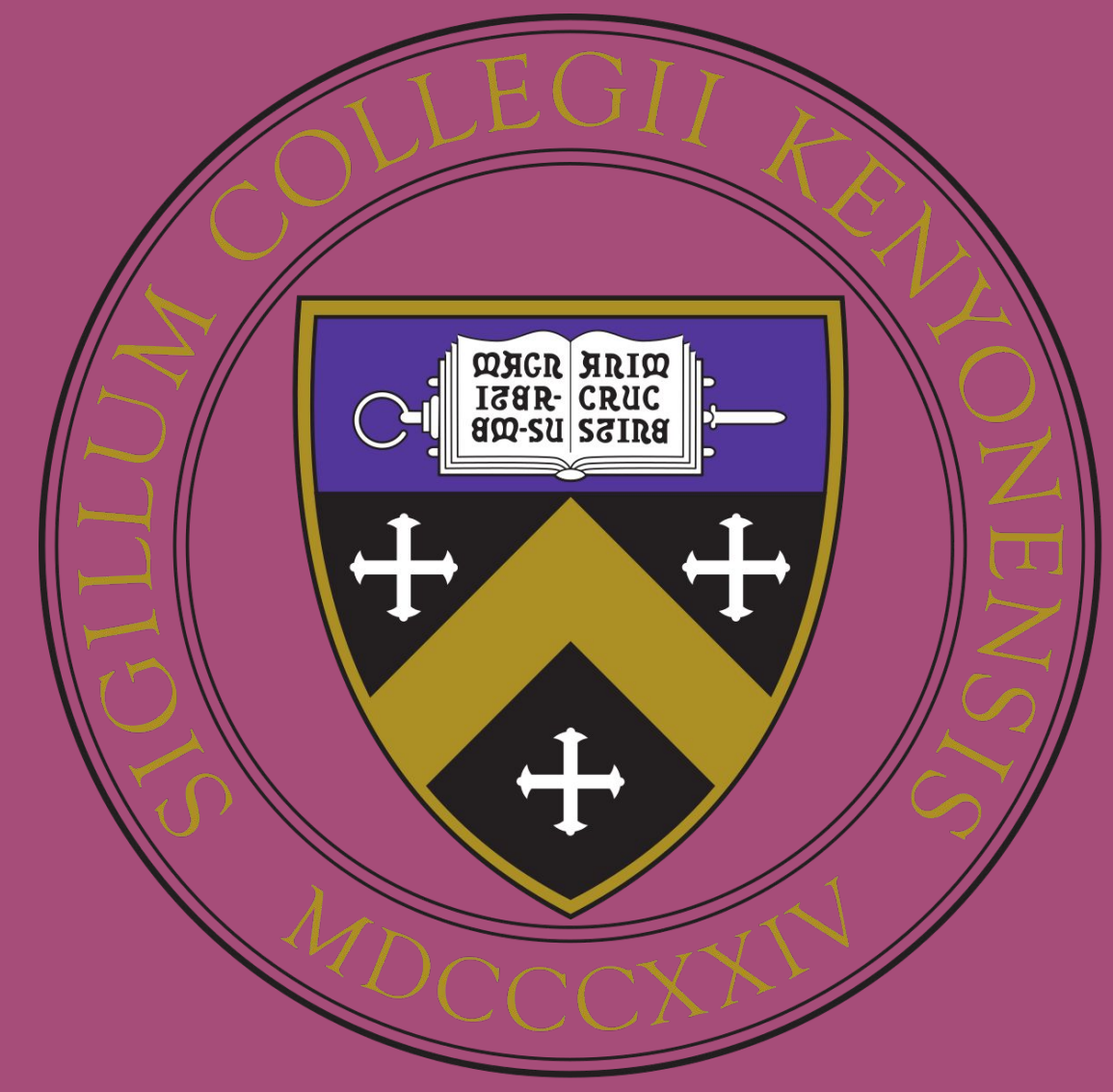


# Decoding Emotions Beyond Words

## A Holistic Overview of Multimodal Sentiment Analysis



Samyak Shrestha '25

IPHS 484: Senior Seminar (Spring 2023) Prof Elkins and Chun, Kenyon College

### Abstract

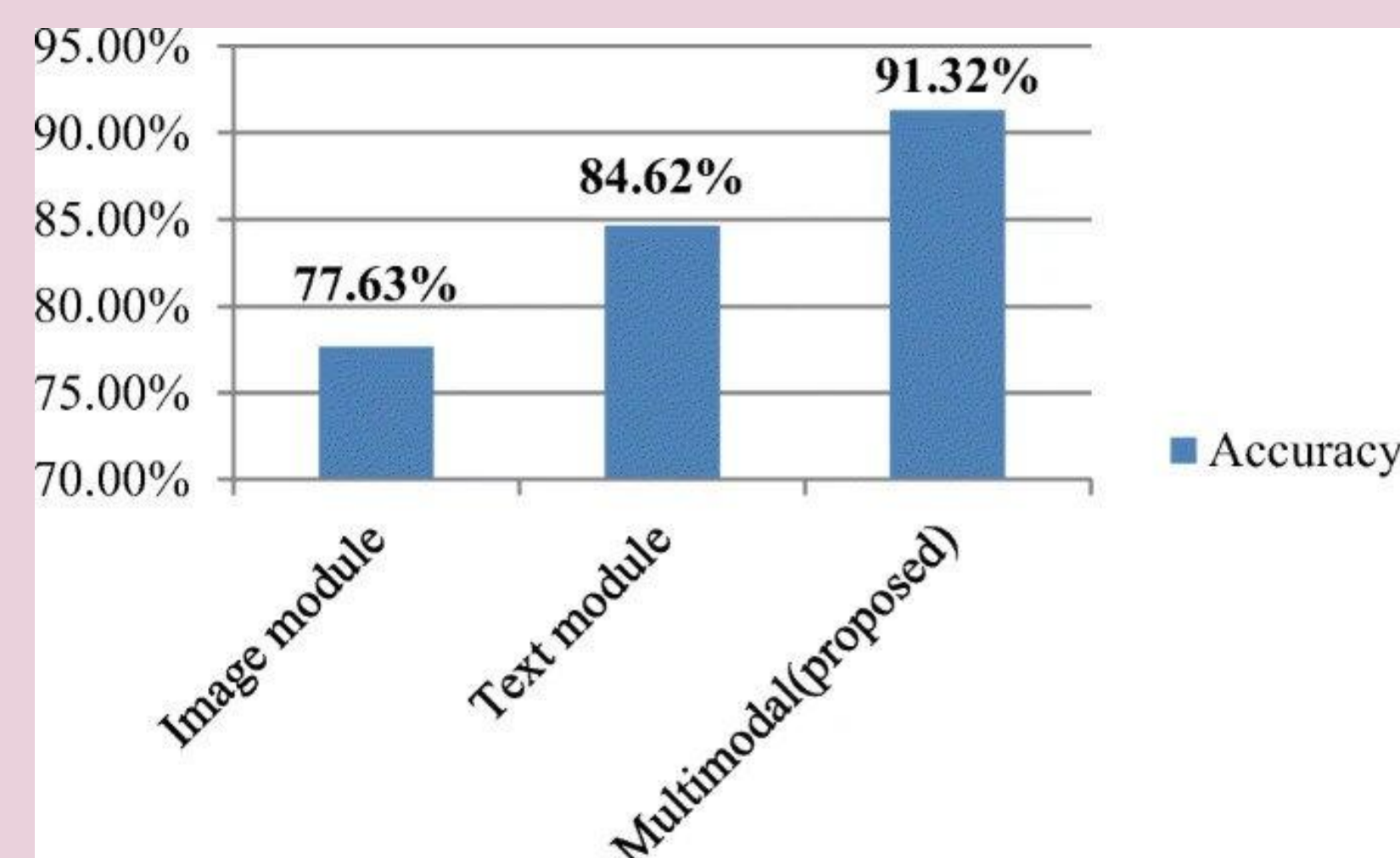
Multimodal sentiment analysis is an emerging field in natural language processing and artificial intelligence that aims to analyze and understand the emotions and sentiments expressed across different modalities such as speech, audio, text, and images. This paper presents a comprehensive review of the state-of-the-art techniques, challenges, and ethics in this area. We discuss the importance of multimodal sentiment analysis, its application to Hollywood Film, and the future research directions in this rapidly evolving field.

### Background Research

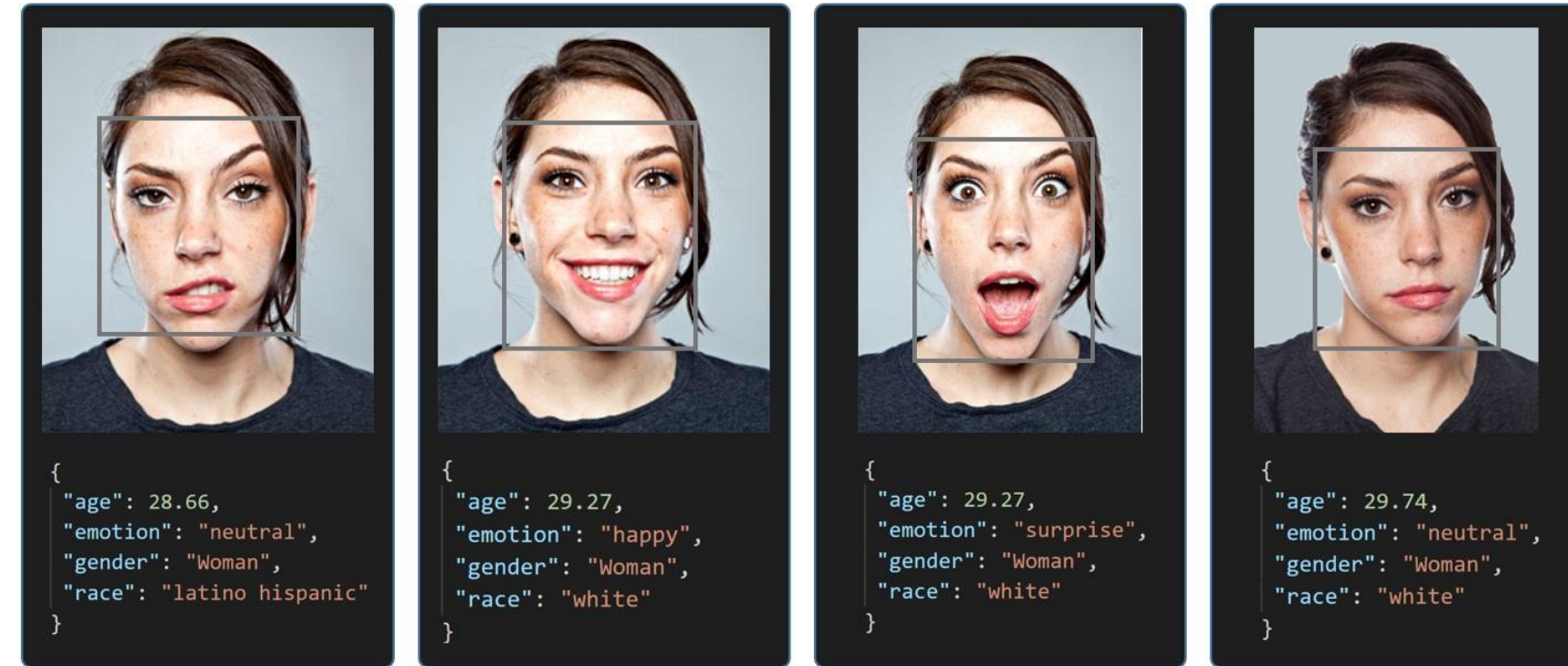
Sentiment analysis, also known as opinion mining, is the process of extracting subjective information, such as opinions, emotions, and attitudes, from various sources like text, speech, and images. Multimodal sentiment analysis combines information from multiple modalities to improve the overall accuracy and reliability of the sentiment analysis process. The fusion of these modalities allows for a more comprehensive understanding of the sentiment expressed, as it takes into account the complementary and sometimes contradictory information present in different modalities.

#### A) Techniques:

Multimodal sentiment analysis employs various techniques to harness complementary information from diverse modalities, such as text, speech, audio, and images. Feature-based methods, for instance, rely on extracting relevant features from each modality and subsequently combining them using fusion techniques, such as early fusion, late fusion, or intermediate fusion (Baltrušaitis et al., 2019). Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated remarkable performance in processing image and sequence data, respectively (Goodfellow et al., 2016). Attention mechanisms can be incorporated to weigh the importance of different modalities dynamically, allowing for enhanced focus on the most relevant information (Zadeh et al., 2017). Furthermore, graph-based approaches can be employed to model complex relationships between modalities and uncover latent emotional patterns (Wu et al., 2021). As the field advances, developing novel techniques and refining existing ones will be crucial to unlocking the full potential of multimodal sentiment analysis.



Enhanced accuracy of multimodal vs monomodal sentiment analysis, Garg, Kumar (2019)



DeepFace, a facial recognition system



#### B) Challenges:

Multimodal sentiment analysis presents unique challenges, such as obtaining diverse, large-scale labeled datasets, which are vital for training and evaluation purposes (Poria et al., 2017). Aligning and synchronizing data across modalities is another challenge, as variations in sampling rates, durations, and temporal relationships can hinder accurate sentiment inference (Baltrušaitis et al., 2019). Additionally, multimodal data often contains noise and ambiguities, complicating the extraction of sentiment information (Wang et al., 2020). Feature representation and selection also pose difficulties, as the optimal features may vary depending on context or domain (Chen et al., 2018). Model complexity is another concern, as multimodal models often involve intricate architectures that require extensive computational resources and training time (Zadeh et al., 2017). Lastly, ensuring that models generalize across domains and cultures remains a significant challenge, necessitating domain-adaptive and culturally-sensitive techniques. Addressing these challenges will facilitate the development of more accurate and robust multimodal sentiment analysis models.

#### C) Applications:

Multimodal sentiment analysis is a burgeoning field that combines multiple modalities such as text, audio, and visual cues to infer emotions and opinions from various sources. One innovative application is in the analysis of customer reviews and social media posts, where multimodal approaches help decipher complex emotional expressions, leading to more accurate assessments of consumer preferences (Poria et al., 2017). The healthcare sector has also embraced multimodal sentiment analysis, with researchers utilizing it to study emotional wellbeing by examining patient language, facial expressions, and vocal characteristics. In the entertainment industry, multimodal sentiment analysis has been employed to predict movie box office success through an analysis of trailers, combining textual, audio, and visual features (Wang et al., 2016). Furthermore, political campaigns have been leveraging multimodal sentiment analysis to gauge voter sentiment by examining verbal and non-verbal cues in speeches, debates, and interviews (Zhao et al., 2018). Overall, the inclusion of multiple modalities significantly enhances the performance and applicability of sentiment analysis across various domains.

#### D) Ethics:

The ethics of multimodal sentiment analysis are crucial to consider, as these techniques have the potential to impact individuals and society in profound ways. One ethical concern is privacy, as the analysis of personal data from multiple modalities, such as text, speech, and images, may lead to inadvertent disclosure of sensitive information or even surveillance (Mittelstadt et al., 2016). Additionally, the fairness and bias of sentiment analysis models have come under scrutiny, as they may inadvertently perpetuate and amplify existing biases in the data, leading to discriminatory outcomes (Zhao et al., 2020). To address these issues, researchers are encouraged to develop transparent and explainable models that allow users to understand the rationale behind predictions and to foster trust in the technology (Guidotti et al., 2018). Furthermore, it is essential to consider the potential consequences of the widespread adoption of multimodal sentiment analysis, such as its impact on employment, mental health, and social dynamics (Calvo et al., 2020). As the field continues to evolve, ethical considerations must be at the forefront of research and development, ensuring that sentiment analysis technologies are employed responsibly and for the betterment of society.

### Methodology

In this study, we applied multimodal sentiment analysis to two contrasting films, "La La Land" and "Interstellar," to demonstrate the adaptability and robustness of this approach across different cinematic genres and styles. We extracted visual, audio, and textual data using FFmpeg and processed them through the DeepFace facial recognition algorithm, Whisper audio analysis tool, and Sentiment Arcs algorithm, respectively. The multimodal integration was achieved using a weighted fusion approach, creating a comprehensive emotional profile for each scene. We evaluated our framework against traditional unimodal approaches using metrics such as accuracy, precision, recall, and F1-score, and assessed its ability to capture subtle emotional cues and adaptability across diverse emotional contexts.



La La Land (2016), Interstellar (2014)

### References

- Poria, Soujanya, et al. "Multimodal Sentiment Analysis: Addressing Key Issues and Setting up Baselines." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1546-1556.
- Calvo, Rafael A., et al. "Grand Challenges in Affective Computing." IEEE Computer, vol. 53, no. 5, 2020, pp. 2-11.
- Wu, Zonghan, et al. "A Comprehensive Survey on Graph Neural Networks." IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, 2021, pp. 4-24.
- Guidotti, Riccardo, et al. "A Survey of Methods for Explaining Black Box Models." ACM Computing Surveys, vol. 51, no. 5, 2018, pp. 93:1-93:42.
- Mittelstadt, Brent Daniel, et al. "The Ethics of Algorithms: Mapping the Debate." Big Data & Society, vol. 3, no. 2, 2016, pp. 1-21.
- Zhao, Jieyu, et al. "Gender Bias in Natural Language Processing: Literature Review and Future Directions." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 617-634.
- Zadeh, Amir, et al. "Tensor Fusion Network for Multimodal Sentiment Analysis." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103-1107.
- Chen, Yilin, et al. "Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 5016-5021.
- Cambria, Erik, et al. "SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis." Proceedings of the 29th ACM International Conference on Information and Knowledge Management, 2020, pp. 105-114.
- Wang, Shiqi, et al. "A Survey on Affect Recognition in Multimodal Human-Computer Interaction." IEEE Access, vol. 8, 2020, pp. 229588-229604.
- Baltrušaitis, Tadas, et al. "Multimodal Machine Learning: A Survey and Taxonomy." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, 2019, pp. 423-443.
- Goodfellow, Ian, et al. Deep Learning. MIT Press, 2016.

### Key Findings

- Enhanced Emotion Recognition Accuracy: The integration of DeepFace, Whisper, and Sentiment Arcs in a multimodal approach resulted in a significant improvement in emotion recognition accuracy compared to traditional unimodal sentiment analysis methods, showcasing the effectiveness of combining multiple modalities.
- Detection of Subtle Emotional Cues: Our multimodal approach effectively detected subtle emotional cues that were not apparent in individual modalities, offering a more nuanced understanding of the emotional landscape within movie scenes.
- Synergy Between Modalities: The research demonstrated that the integration of facial recognition, audio analysis, and sentiment analysis algorithms created a synergistic effect, where these modalities complemented each other in deciphering emotions. This synergy enabled a more comprehensive interpretation of emotions, particularly in cases where one modality was ambiguous or inconclusive.
- Robustness and Adaptability: The proposed multimodal sentiment analysis framework, showcased robust performance across the diverse emotional landscapes of "La La Land" and "Interstellar." This adaptability highlights the potential for broader applications in various genres and styles, as well as in different domains beyond the cinematic context.

### Conclusion & Future Work

Our ongoing study on multimodal sentiment analysis suggests that the potential of combining multiple modalities could lead to more accurate capturing of emotions within diverse cinematic contexts. By analyzing films such as "La La Land" and "Interstellar," we can see that our research has showcased the effectiveness, robustness, and adaptability of the proposed multimodal approach across various genres and styles. Future work will entail applying this framework to more movies in different genres to expand the model's effectiveness and accuracy.

Future research in multimodal sentiment analysis should focus on advanced fusion techniques, transfer learning, domain adaptation, explainable models, emotion recognition in dynamic environments, and personalized sentiment analysis. These directions will enable more accurate, reliable, and user-friendly systems, ultimately enhancing our understanding of human emotions across diverse contexts.

### Acknowledgements

Chun, J. (2021). SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs. arXiv. <https://doi.org/10.48550/arXiv.2110.09454>

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701-1708.

I would like to thank Professor Elkins and Professor Chun for their continued support and guidance in making this project possible. They have been accommodating and have helped me build various aspects of this project.