Indian Institute of Information Technology Kottayam

MULTIMODAL SENTIMENT ANALYSIS

By

2020BCS0069 CHRISTO SOJAN
2020BCS0096 ANGATI BALA MURALI
2020BCS0184 KUMAR AMITANSHU
2020BCS0024 SYAM SIVADAS

Guided By,

DR. MANU MADHAVAN

INTRODUCTION

- Multimodal Sentiment Analysis (MSA) is a subfield of computational sentiment analysis that leverages information from multiple modalities, such as text, audio, image, and video, to infer and evaluate the sentiment or emotional state of an individual or group.
- The human communication process is inherently multimodal. When we express sentiments, we not only use words but also tone, pitch, facial expressions, and body language. MSA taps into this holistic approach, aiming to capture sentiment more accurately by combining information from various modalities.
- By integrating diverse data sources, MSA helps in overcoming the limitations and ambiguities present in individual modalities, leading to enhanced performance and robustness in sentiment prediction tasks.
- The key components of MSA includes Text Analysis, Audio Analysis, Visual Analysis, Fusion Techniques, and Machine Learning Models.

SCOPE

- Social Media Monitoring - Enhanced understanding, emotion detection
- Customer Feedback Analysis - Deeper insights, personalized, responses etc.
- Content Recommendation - Accuracy, emotionally intelligent recommendations, contextual recommendations, feedback loops, ad personalization
- Healthcare - Patient experience and satisfaction, telemedicine and virtual health, pharmacy services
- Human - Computer Interaction-Emotion Recognition in User Interface, Gesture based control, VR and AR
- Education - Student engagement, Personalized learning, Feedback
- Entertainment - Content Creation, Recommendation, Live entertainment
- Product Development - Idea generation, Market expansion, customer feedback, product campaign ideas

CHALLENGES

- **Data Fusion:** Integrating information from different modalities is a complex task. Deciding how to combine textual, visual, and auditory information to arrive at a holistic sentiment analysis is a significant challenge
- **Heterogeneity:** Data from different modalities may have different data formats, structures, and scales. For example, text is sequential, while images are spatial, and audio is temporal.
- **Feature Extraction:** Extracting meaningful features from different modalities can be difficult.
- **Data Sparsity:** Depending on the context, certain modalities may be missing or sparse. For example, in a text message with an image, the text may contain sentiment, but the image may not always be informative.
- **Privacy and Ethical Concerns:** Handling multimodal data may raise privacy concerns, as it can reveal sensitive information about individuals.
- **Domain and Culture Dependency:** Sentiment analysis is highly dependent on the context and culture.

LITERATURE REVIEW

Title	**Multimodal Sentiment Analysis: Addressing Key Issues and Setting up the Baselines[1]**
Authors	Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh and Amir Hussain
YOP	2019
Journal	55th Annual Meeting of the Association for Computational Linguistics (ACL) and the IEEE International Conference on Data Mining (ICDM)
Modalities	Text, Audio, Visual
Datasets	MOUD, MOSI, IEMOCAP
Models used	CNN, 3D-CNN, OPENSMILE, LSTM, SVM
Summary	The research paper explores multimodal sentiment analysis using deep-learning architectures for sentiment classification. It considers text, audio, and visual modalities for understanding emotions in videos. **The authors use CNN, 3D-CNN, openSMILE, and bc-LSTM models. The bc-LSTM fusion method outperforms SVM in accuracy, while audio and text modalities play crucial roles.** The paper emphasizes the need for context, different modalities, and generalizability of multimodal sentiment classifiers. Future work should focus on extracting semantics from visual features and incorporating contextual dependency learning.

Results Obtained	**The results show that the bc-LSTM fusion method consistently outperforms the SVM fusion method across all experiments**. **The audio modality performs better than the visual modality in both the MOSI and IEMOCAP datasets.** The text modality plays a crucial role in both emotion recognition and sentiment analysis, with its unimodal performance being substantially better than the other modalities. **The fusion of modalities shows more impact in emotion recognition than in sentiment analysis.** The paper also discusses the performance of the models in speaker-exclusive experiments, where the models have to classify emotions and sentiments from speakers they have never seen before.
Current Status	There are still major issues that remain mostly unaddressed in this field, such as the **consideration of context in classification, the effect of speaker-inclusive and speaker-exclusive scenarios, the impact of each modality across datasets, and the generalizability of multimodal sentiment classifiers**. The document serves as a benchmark for future research in MSA and highlights the need for further exploration and improvement in these areas.

Future Scope	● **Semantics Extraction:** Enhancing sentiment analysis by extracting deeper meanings from visual cues. ● **Cross-Modal Feature Exploration:** Understanding interactions between different modalities to refine fusion techniques. ● **Contextual Dependency:** Incorporating learning methods to grasp the relations between segments, enhancing sentiment classification. ● **Benchmark Datasets:** Establishing standardized datasets to ensure model comparisons are fair and push advancements. ● **Generalizability:** Evaluating model performance across datasets to ensure their applicability in diverse contexts. Exploring these facets can drive MSA advancements and widen its domain applications.

Title	**Multi-Level Attention Map Network for Multimodal Sentiment Analysis[2]**
Authors	Xiaojun Xue, Chunxia Zhang, Zhendong Niu and Xindong Wu
YOP	2023
Journal	IEEE Transactions on Knowledge and Data Engineering journal.
Modalities	Texts and Images from user generated content
Datasets	MVSA-Single, MVSA-Multi, and Multi-ZOL.
Models used	Multi-Level Attention Map Network (MAMN), which consists of three modules: multi-granularity feature extraction, multi-level attention map generation, and attention map fusion.
Summary	This research paper presents new way to approach (MSA) as **“Multi-Level Attention Map Network”** (MAMN). The paper addresses the **challenges of noise reduction, feature extraction and correlation capture in MSA tasks**. The model utilizes techniques such as **Attention mechanisms, Gated mechanisms, and Multi-task learning.**

Results Obtained	The experimental results show that the proposed **MAMN model outperforms methods in terms of accuracy and effectiveness for both document-based and aspect-based MSA tasks**. The model achieves significant improvements in sentiment classification performance on the evaluated datasets.
Current Status	They contributed to the field “**Multimodal Sentiment Analysis**” via **novel model that addresses the challenges of noise reduction, feature extraction, and correlation capture**.However, the current status of MSA is an ongoing research area, and there is still room for further advancements and improvements in the field.
Future Scope	Paper suggest exploring **more advanced fusion methods, i**nvestigating the impact of different modalities on sentiment analysis, and exploring other application areas for MSA such as **product marketing** and **public opinion monitoring.**

Title	**Multimodal Sentiment Analysis via RNN variants[3]**
Authors	Ayush Agarwal, Ashima Yadav and Dinesh Kumar Vishwakarma
YOP	2019
Journal	IEEE Xplore
Modalities	Text, Audio and Video
Datasets	CMU-MOSI Dataset
Models used	Four variants of Recurrent Neural Networks (RNNs) for sentiment analysis (models are based on LSTM and GRU architectures) : - GRNN - LRNN - GLRNN - UGRNN
Summary	They conducted **experiments on the CMU-MOSI dataset and show that their approach achieves better sentiment classification accuracy** than existing methods on individual modalities and also after fusing the modalities using attention networks.

Results obtained	The research paper reports the accuracy achieved by the different RNN variants on the CMU-MOSI dataset. **GRNN performs best for text, GRNN and GLRNN perform best for audio, and UGRNN performs best for video**. After fusing the modalities, LRNN and GLRNN achieve the best results for multimodal sentiment analysis, with an accuracy of 78.05%.
Current Status	They contributed to the field "**Multimodal Sentiment Analysis**" via **proposing novel RNN variants and evaluating their performance on the CMU-MOSI dataset**. It demonstrates the effectiveness of using multimodal data for sentiment classification.
Future Scope	- Credibility analysis - Used for medical purposes such as detection of autism in a child - Used for predicting the sentiments, emotions, and genre of a movie by its trailer.

Title	**Fusion-Extraction Network for Multimodal Sentiment Analysis[4]**
Authors	Tao Jiang, Jiahai Wang, Zhiyue Liu and Yingbiao Ling
YOP	2020
Journal	Pacific-Asia Conference on Knowledge Discovery and Data Mining
Modalities	Visual and textual information.
Datasets	MVSA(Multiple Viewpoint Semantic Annotation)- Single and MVSA-Multiple
Models used	Fusion-Extraction Network (FENet) using fine grained attention and gated convolutional layers, BERT.
Summary	The document proposes a **Fusion-Extraction Network (FENet)** for multimodal sentiment analysis. The network utilizes an interactive information **fusion mechanism** to learn visual-specific textual representations and textual-specific visual representations. It also incorporates an **information extraction mechanism to filter redundant parts and extract valid information** from the multimodal representations. Experimental results on two public multimodal sentiment datasets show that FENet outperforms existing state-of-the-art methods.

Results	The experimental results show that **FENet outperforms existing state-of-the-art methods** on the two multimodal sentiment datasets. The model achieves **higher accuracy and F1 scores compared to baseline methods** such as SentiBank & SentiStrength, CoMN etc.
Current Status	The research paper contributes to the field of multimodal sentiment analysis by proposing a **novel model that effectively utilizes the relationship between visual and textual information**. It demonstrates improved performance compared to existing methods, indicating progress in the field.
Future Scope	The research opens up possibilities for further advancements in multimodal sentiment analysis. Future work could **explore more sophisticated fusion mechanisms, extraction techniques, and attention mechanisms** to enhance the understanding of multimodal data and improve sentiment analysis performance. Additionally, the proposed model could be **extended to other domains and datasets** to evaluate its generalizability.

Title	**Dense Fusion Network with Multimodal Residual for Sentiment Classification[5]**
Authors	Huan Deng, Peipei Kang, Zhenguo Yang, Tianyong Hao, Qing Li and Wenyin Liu
YOP	2021
Journal	IEEE Xplore
Modalities	language, acoustic speeches, and visual images
Datasets	CMU-MOSI, ICT-MMMO, YouTube, and IEMOCAP
Models used	The DFMR framework which consists of four modules-The modality-specific module, dense multimodal fusion module, multimodal residual module, and sentiment classification module.
Summary	The paper proposes the **DFMR framework** for multimodal sentiment analysis. It i**ntegrates language, acoustic speeches, and visual images using dense fusion and multimodal residual modules**. The framework achieves state-of-the-art performance on benchmark datasets and provides a promising solution for sentiment analysis using multimodal data.

Results obtained	The experimental results show that DFMR outperforms eleven state-of-the-art baselines on the benchmark datasets. It achieves **higher accuracy, F1 score, and correlation** compared to the other models.
Current Status	The research paper contributes to the field of multimodal sentiment analysis by proposing a **deep learning model that effectively integrates and analyzes multimodal data.** It demonstrates the effectiveness of the proposed DFMR model in capturing and fusing multimodal features for sentiment analysis.
Future Scope	The research paper opens up possibilities for further advancements in multimodal sentiment analysis. Future research can explore more complex fusion techniques, **investigate the impact of different modalities** on sentiment analysis, and explore the **use of other deep learning architectures for improved performance**. Additionally, the proposed model can be applied to other domains and datasets to evaluate its generalizability.

PROBLEM STATEMENT

Multimodal Sentiment Analysis has witnessed significant advancement in recent years, particularly with the development of various computational models aiming to harness the richness of multiple data modalities. However, a **comprehensive comparison of these models' efficacy**, especially in the **context of non-English languages**, remains sparse. India, with its rich linguistic diversity, offers a unique opportunity to explore the application of MSA, yet research in Indian languages remains underrepresented in the global MSA landscape. While some models have shown promise in English or other widely-researched languages, the **adaptability of these models to Indian languages** remains largely unexplored, signaling both a gap and an opportunity for refinement to capture the nuances of these languages.

OBJECTIVES

- **Comprehensive Model Analysis**
  - To evaluate each model's efficiency, accuracy, and performance metrics when applied to Indian language datasets.
- **Localization and Customization**
  - To source a rich multimodal dataset in the chosen Indian language, ensuring it captures textual, audio, and visual sentiments authentically.
  - To test and refine MSA models to ensure they handle local idioms, phrases, and expressions accurately in sentiment classification.
- **Enhancement of Existing Models**
  - To identify gaps in current MSA models when applied to Indian languages.
  - To modify and improve these models by integrating language-specific modules or features.

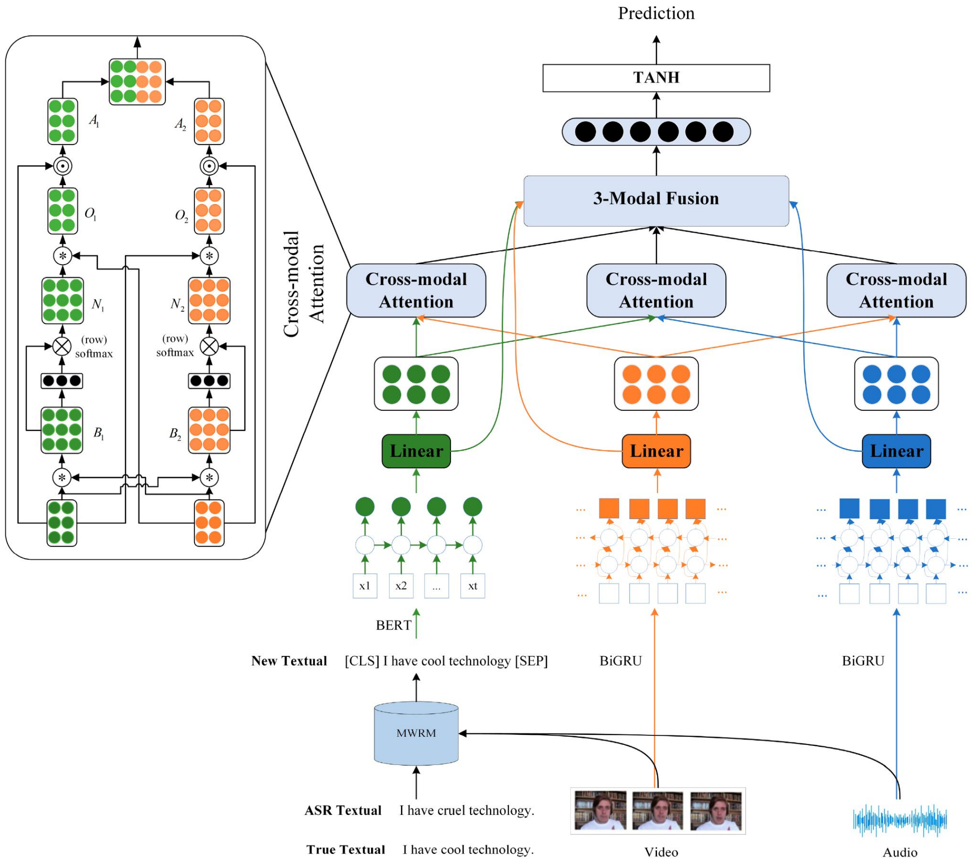
ARCHITECTURE

The Model’s Architecture is depicted in Figure

It comprises 4 main components:

- Unimodal-feature extraction
- Multimodal-word-refinement module
- Multimodal-feature-fusion module
- Prediction network

To enhance the adaptability to feature differences across different modalities and effectively model complex nonlinear interactions, we designed a cross-modal, hierarchical-fusion network that considered the present of sentiment-word-recognition errors in texts generated by ASR(Automatic Speech Recognition).



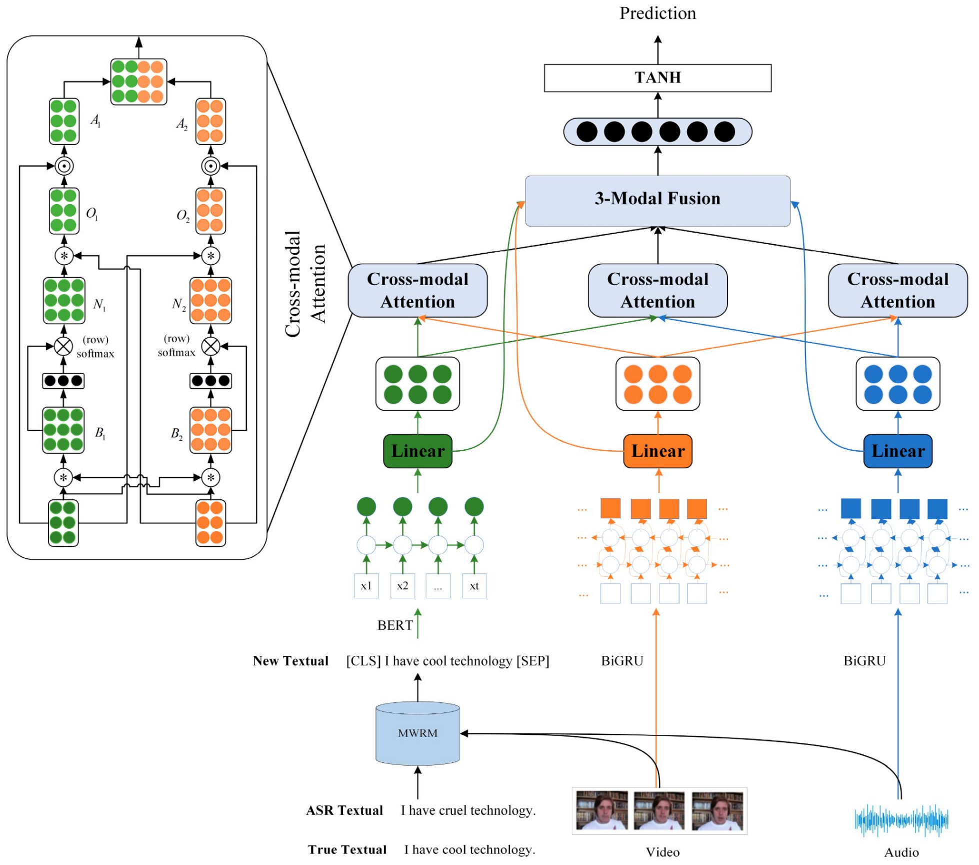
Source: https://www.mdpi.com/2079-9292/12/16/3504

- Initially, we will employ mode to extract textual feature information while another model to extract audio and visual feature information.

- the multimodal-word-refinement module dynamically supplements missing sentiment semantics by leveraging multimodal-sentiment information, resulting in the generation of new word embeddings.

- Obtained new word embeddings were then input into the multimodal-feature-fusion module, which utilized a cross-modal hierarchical-fusion network to perform feature fusion across different modalities.

- Finally, a nonlinear layer was employed to predict the final sentiment-regression labels, facilitating accurate sentiment judgment.



Source: https://www.mdpi.com/2079-9292/12/16/3504

CONCLUSION

- The overall objective across these papers is to leverage different modalities - namely text, audio, and visual data using various deep learning architectures and fusion mechanisms.
- Contextual information, especially in videos and audio, plays a crucial role in sentiment analysis. Several papers addressed the importance of capturing context using architectures like bi-directional LSTMs.
- All the research papers utilize deep learning techniques to analyze the multimodal data. Techniques such as RNNs, attention mechanisms, CNNs, and LSTMs are frequently employed.
- An evident theme across the papers is the significance of fusion mechanisms in enhancing the performance of sentiment analysis. Fusing modalities often leads to better results than analyzing them separately.
- The alignment and interaction of data from different modalities are challenging but crucial. For instance, understanding the relationship between text and images in social media content is essential for effective MSA.

REFERENCES

❑ [1] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh and A. Hussain, "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines" in *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17-25, Nov.-Dec. 2018, doi: 10.1109/MIS.2018.2882362.

❑ [2] X. Xue, C. Zhang, Z. Niu and X. Wu, "Multi-Level Attention Map Network for Multimodal Sentiment Analysis," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 5105-5118, 1 May 2023, doi: 10.1109/TKDE.2022.3155290.

❑ [3] A. Agarwal, A. Yadav and D. K. Vishwakarma, "Multimodal Sentiment Analysis via RNN variants," *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, Honolulu, HI, USA, 2019, pp. 19-23, doi: 10.1109/BCD.2019.8885108.

❑ [4] Jiang, T., Wang, J., Liu, Z., Ling, Y. (2020). Fusion-Extraction Network for Multimodal Sentiment Analysis. In: Lauw, H., Wong, RW., Ntoulas, A., Lim, EP., Ng, SK., Pan, S. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2020. Lecture Notes in Computer Science(), vol 12085. Springer, Cham.

❑ [5] H. Deng, P. Kang, Z. Yang, T. Hao, Q. Li and W. Liu, "Dense Fusion Network with Multimodal Residual for Sentiment Classification," *2021 IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China, 2021, pp. 1-6, doi: 10.1109/ICME51207.2021.9428321.