**Literature Review: Recent Advances in Multimodal Sentiment Analysis**

**Abstract**

Sentiment Analysis (SA) is a crucial research area in Natural Language Processing (NLP) and Computer Vision, especially in analyzing and integrating multimodal data such as text, speech, and images. The introduction of deep learning has significantly improved performance in multimodal sentiment analysis, leading to the emergence of various models and optimization techniques. This paper reviews recent methods in multimodal sentiment analysis, including deep learning-based fusion models, sentiment classification techniques optimized with attention mechanisms, and integration strategies for different modalities (e.g., text, emojis, speech, and visual data). By comparing the core ideas, algorithms, and experimental results of various methods, we summarize the advantages and limitations of current research and explore future directions.

**1. Sentiment Analysis Using a Hybrid GoogleNet-ResNet Model for Text and Emojis**

**Methodology**

This study introduces an innovative deep learning hybrid model that integrates text and emoji data to enhance sentiment analysis accuracy. The main techniques include:

* **Text Representation**: Uses **Word2Vec** and **Emoji2Vec** for feature extraction, with **Continuous Bag of Words (CBOW)** and **Skip-gram (SG)** for text vectorization.
* **Deep Learning Architecture**: Employs a **hybrid GoogleNet and ResNet** convolutional neural network (CNN) with **Depthwise Separable Convolution** to extract features and improve classification accuracy.
* **Optimization Method**: Uses the **Osprey optimization algorithm** for hyperparameter tuning to accelerate convergence and enhance classification performance.

**Experimental Results**

Tested on the **Flipkart product review dataset**, this method achieved **98.3% accuracy, 96% precision, 97.6% recall, and 96.82% F1-score**, outperforming traditional unimodal sentiment analysis approaches.

**2. End-to-End Multimodal Sentiment Analysis Based on SAE (Sentiment Analysis Engine)**

**Methodology**

This study proposes an **end-to-end multimodal sentiment analysis framework** that integrates video, audio, and textual information using the following techniques:

* **Speech Feature Extraction**: Uses **DeepSpeech** for speech-to-text conversion and **LSTM (Long Short-Term Memory)** for sentiment feature extraction.
* **Visual Feature Extraction**: Utilizes **ResNet-50** to capture temporal features from video frames and applies a **Multi-scale Efficient Channel Space Attention (MECS) mechanism** to recognize micro-expression features.
* **Multimodal Fusion**: Implements **self-attention mechanisms** for deep integration of different modalities, generating text descriptions and performing sentiment analysis.

**Experimental Results**

Tested on **NExT-QA and CMU-MOSI datasets**, this method achieved **F1 scores of 91.14 (Class 2-7 classification) and 86.5 (IEMOCAP testing)**. The study also highlights the **importance of speech features**, as removing the speech module reduced overall accuracy by 5%.

**3. UMSA: A Multimodal Sentiment Analysis Framework for Urdu**

**Methodology**

Focusing on **low-resource languages**, this study introduces **Urdu Multimodal Sentiment Analysis (UMSA)**, featuring:

* **Dataset Construction**: Creates the **Urdu Sentiments Dataset (USD)** containing Urdu-language video reviews.
* **Model Fusion**: Uses **early feature fusion** and **ensemble models** for text, audio, and video modalities.
* **Modality-Specific Models**:
  + **Audio**: Employs **LSTM and Random Forest** for sentiment classification.
  + **Text**: Uses **Logistic Regression and BERT (Bidirectional Encoder Representations from Transformers)** for sentiment analysis.
  + **Video Frames**: Utilizes **CNN and Random Forest** for emotion recognition.

**Experimental Results**

This method achieved **over 80% accuracy** on the Urdu dataset and demonstrated effectiveness in real-world applications through case studies.

**4. Conflict-Aware Multimodal Sentiment Analysis (MCAN)**

**Methodology**

This study introduces a **Multi-level Conflict-Aware Network (MCAN)** that focuses on **modeling conflicts between different modalities**, incorporating:

* **Alignment and Conflict Separation**: Extracts aligned and conflicting features at the unimodal and bimodal levels.
* **Conflict Modeling Branch**: Implements **Discrepancy Constraint** at feature and output layers to avoid reliance on unstable generated labels.

**Experimental Results**

Tested on the **CMU-MOSI and CMU-MOSEI datasets**, this approach demonstrated superior performance in multimodal sentiment analysis.

**5. Multimodal Fusion Network (MFN) Based on Attention Mechanisms**

**Methodology**

This study introduces a **Multi-head Self-Attention Mechanism** for **visual-text sentiment analysis**, utilizing:

* **Denoising Mechanism**: Applies **neural networks and attention mechanisms** to filter out noisy information across modalities.
* **Fine-grained Feature Learning**: Extracts **local region features** from multimodal data, using varying neuron counts in hidden layers to model heterogeneous information.

**Experimental Results**

Tested on **Twitter, Flickr, and Getty Image datasets**, this method outperformed 11 existing approaches, achieving accuracy improvements of **0.11%, 0.13%, and 0.38%, respectively**.

**6. Multimodal Sentiment Analysis with Word-level Fusion and Reinforcement Learning (GME-LSTM(A))**

**Methodology**

This study proposes a **word-level multimodal fusion approach**, featuring:

* **Gated Multimodal Embedding (GME)**: **Reduces interference from noisy modalities**.
* **LSTM with Temporal Attention**: Integrates word-level fusion across input modalities and applies attention to crucial time steps.

**Experimental Results**

Tested on the **CMU-MOSI dataset**, this method achieved **state-of-the-art performance** in sentiment classification and regression tasks. The study also emphasized the importance of the **Temporal Attention Layer** in sentiment prediction.

**7. Datasets and Experimental Evaluation**

Commonly used datasets in MSA research include:

* **CMU-MOSI**: Contains video, audio, and text data, widely used for sentiment recognition tasks.
* **CMU-MOSEI**: A large-scale multi-modal sentiment analysis dataset with extensive emotion annotations.
* **IEMOCAP**: Used for speech and video emotion analysis, providing rich facial expression and voice data.

Performance evaluation metrics typically include:

* **Accuracy**
* **F1-score**
* **Precision and Recall**

**8. Conclusions and Future Research Directions**

This paper reviews recent advancements in multimodal sentiment analysis, covering fusion strategies for text, speech, and visual data and deep learning model optimizations. Overall, multimodal approaches outperform unimodal methods in sentiment classification, particularly in handling complex emotional information. However, several challenges remain:

1. **Dataset Diversity and Scalability**: Existing datasets primarily focus on English, while low-resource languages (e.g., Urdu) require further expansion.
2. **Modeling Inter-modal Conflicts**: Although some studies address inter-modal conflicts, distinguishing useful information from noise remains an open problem.
3. **Computational Efficiency**: Multimodal fusion methods tend to have high computational costs. Improving inference speed while maintaining accuracy is a key research direction.

Future research should explore **Large Language Models (LLMs)**, **Self-supervised Learning**, and **lightweight deep learning architectures** to further advance multimodal sentiment analysis.