Unlocking Sentiment : Through Multimodal Intelligence

Anjali

Dr. Saroj Kumar Gupta

B.Tech CSE (Ds & Ai)
SRM University
Sonipat, Haryana
anjalinara74@gmail.com

Abstract— **This paper investigates the integration of text, audio, and visual cues to improve emotion recognition accuracy. Traditional psychological analyses often face uncertainty, ambiguity, and changing definitions because they rely on a single model. This study proposed several approaches that combine natural language processing (NLP) for text analysis, speech recognition, and face recognition to achieve the desired results. Good. The system uses convolutional neural networks (CNNs) for image-based visualization, deep learning-based speech processing for voice recognition, and text-based text transformation. The concept of fusion model improves the accuracy of classification theory by leveraging the complementary insights of multiple models. Evaluation using comparative data shows that the system has the potential to outperform the worst and has potential for use in mental health care, customer service and analytics.**

Keywords— **Multimodal SentimentAnalysis, Deep Learning, Emotion Recognition, NLP, Computer Vision**

I. INTRODUCTION

Today, understanding human emotions through technology is the basis for applications in healthcare, customer service, drug check relationships, etc. The project focuses on multisensory analysis, an advanced method for assessing personality that uses a variety of data, including facial expressions, vocal cues, and written text. Analyze thoughts more clearly and precisely. By combining the strengths of different data sources, the project aims to overcome the limitations of single-modal systems, which often suffer from ambiguity and context-dependent reasoning problems. For example, a smile in a picture may express happiness, while similar sounds or text may express criticism or disappointment. A multimodal approach can help resolve this confusion.

This model uses the FER2013 dataset to generate 48x48 grayscale facial expression images to classify emotions such as happiness, sadness, anger, etc. Good recognition can be achieved across scenes as well. Identify emotional and mental states by analyzing information obtained from speech or text. The analysis uses natural language processing (NLP) techniques such as tokenization and semantic interpretation. Audio cues such as pitch, volume, and loudness provide important emotional cues. This model uses audio devices to distract thoughts for applications such as call centres and mental health assessments. A key innovation lies in the integration of these models within a single conceptual framework. By combining facial, audio, and text data, the system can build a comprehensive picture of emotions, providing insights that cannot be obtained through analysis alone.

II. RELATED WORK

Multimodal sentiment analysis (MSA) has attracted significant attention in recent years due to its ability to utilize multiple data sources (text, speech, and facial expressions) to ensure that sentiments are accurate and context-aware. Traditional sentiment analysis mostly focuses on text-based methods that fail to capture the richness of human emotion conveyed through tone and facial expressions. To overcome these limitations, researchers have explored various hybridization methods to combine data from different models to improve classification hypotheses. Manually extract features from the data. Classical machine learning models, such as support vector machines (SVM) and hidden Markov models (HMM), are used to combine features from different models. However, with the advancement of deep learning, neural network-based models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer architectures, have improved their learning capabilities by improving the interaction of capabilities. Recent research has focused on different fusion techniques, including early fusion (feature-level integration), late fusion (decision-level integration), and hybrid fusion methods. Multimodal transformers such as Multimodal Adaptive Gate (MAG) and Multimodal Transformer (MulT) exhibit state-of-the-art performance via similar modeling. Furthermore, the integration of pre-learned language models and multimodal feature encoders such as BERT further strengthens the conceptual model. Computational efficiency challenges remain in real-time applications. This section provides an overview of current multidisciplinary research, focusing on the main methods, data, and fusion processes that have contributed to advances in cognitive emotion of text, speech, and face.

The success of MSA depends largely on how diverse the models are. Features extracted from text, speech, and facial expressions are combined at the feature level and then fed into the joint model. This approach enables robust multi-dimensional interactions but may suffer from high data representation overhead. At the decision level, predictions from different unimodal models were combined. This approach is computationally efficient but does not provide deep integration of patterns. Combining early and late fusion allows the strengths of both approaches to be exploited. Transformer-based models such as the Multimodal Transformer (MulT) learn the relationships between relations through a supervised learning process using hybrid fusion.

III. PROPOSED METHODOLOGY

Multi-sensory thinking theory: Combining text, speech, and facial expressions for thought recognition aims to improve the accuracy of thinking classification by using deep learning as fusion. The system combines textual, acoustic, and visual methods using an end-to-end multi-modal transformer-based architecture. The following sections describe the main points of the plan, including preliminary data, feature extraction, fusion quality, and distribution theory.

A. Data Pre-processing :

Since multimodal data comes from multiple sources, specific preprocessing steps are required for each model:

1. Text processing : Using pretrained models, introduce Transformers like BERT or Ro BERT a for tokenization and stop word removal, and place the progeny.
2. Speech processing : Convert audio signals to spectrograms or Mel-frequency cepstral coefficients (MFCC), and then process them using CNN or LSTM.
3. Task : Face detection using MTCNN or OpenCV, followed by artifact removal using pre-CNN models like Res Net or Efficient Net.

B. Feature Extraction and Representation :

To obtain meaningful information about emotion, deep learning models are used for feature extraction :

1. Text features: BERT, XL Net or GPT are used to obtain content embeddings to capture subtle differences.
2. Micro expressions in the frame and functional units of the face.

C. Modality Alignment and Synchronization :

Since text, speech, and facial expressions develop at different times, a strategy for adapting to the body is required:

1. Please read speech, data.
2. Frame-by-frame synchronization ensures that facial features match the same time steps as speech and text.

D. Multimodal Fusion Strategies :

To achieve good cross-modal data fusion, we examined several fusion strategies:

1. Early Fusion (Feature-Level Fusion): Features can be extracted from all our variables and combined into a representation before being passed to the deep neural network.
2. Late Fusion (Decision-Level Fusion): Separate classifiers are trained for each modality, and their outputs are combined using an ensemble technique.
3. Hybrid Fusion : Transformer-based fusion models, such as the Multimodal Transformer (MulT), are used to examine cross-modal dependencies through a tracking mechanism. display.

E. Sentiment Classification Model :

The final sentiment classification is performed by a deep multimodal neural network, which includes :

1. Multimodal Feature Encoder : uses a transformation process to learn text, speech, and face-to-face interaction.

F. Training and optimization :

The model is trained using :

1. Losse function : categorical cross-entropy loss. Batch normalization to avoid overfitting.

G. Experimental setup and data :

We analyze our public opinion preparation process with various datasets reflecting various needs :

1. CMU-MOSI (Multimodal Opinion Sentiment and Intensity Dataset)
2. CMU-MOSEI (Multimodal Opinion Sentiment and Emotional Intensity Dataset)
3. MELD (Multimodal Emotional Lines Dataset) : All data include text, audio and facial expressions. Numerous training videos are available.

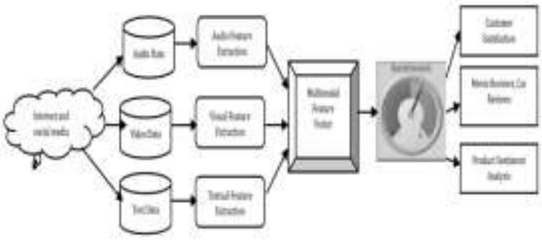

Internet and social media
Audio Data
Video Data
Text Data
Audio Feature Extraction
Visual Feature Extraction
Textual Feature Extraction
Multimodal Feature Vector
Customer Satisfaction
Movie Reviews, Car Reviews
Product Sentiment Analysis

Figure 1. High-Level Architectural Overview of a Multimodal Sentiment Analysis system

Figure 1 shows the high-level architecture of multi-sentiment assessment, which combines multiple variables (audio, video, and text data) for sentiment classification.

Source of Information :

Ideas come from the internet and social media, audio files, video files, and text files are created as raw data.Noisy and noisy.

1. Vector : The features extracted from each transformation are combined into a multimodal feature vector that captures cross-modal perception and emotion information. The feature vector is perfect for classification purposes.

2. Applications include : Customer satisfaction analysis Movie reviews, car reviews, etc. Transportation inspection. This process demonstrates how multimodal data fusion can improve accuracy and robustness.

A. Data Input :

This framework processes various data inputs including : Audio Data, Video Data, Text Data.

1. Audio Features : Includes pitch, tone, and emotional cues.

2. Video Features : Captures facial expressions, gestures, and visual sentiments.

3. Text Features : Extracts semantic and syntactic information using NLP techniques.

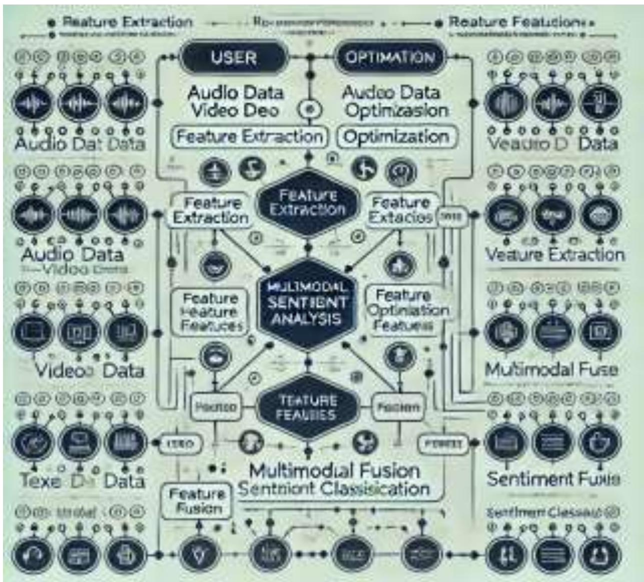


Figure 2. Use Case Diagram

B. Feature Optimization :

After feature extraction, optimization techniques are applied to select the most relevant features for sentiment classification:

1. Grasshopper Optimization Algorithm (GOA) : Optimizes feature subsets based on swarm intelligence.
2. Artificial Bee Colony (ABC) Optimization : Selects features using bee-inspired optimization.
3. GBEE (Grass Bee Optimization) : Combines both GOA and ABC for optimal feature selection to enhance classification performance.

C. Multimodal Fusion (Feature-Level Fusion) :

Optimize the ladders from each transform into a single multimodal feature vector for sentiment analysis.

D. Multimodal sentiment classifications :

Classification based on combined feature vectors :

1. Training phase : Train a multilayer perceptron neural network (MLP-NN) to learn sentiment patterns from multimodal feature vectors.
2. MLP-NN learning classifies opinions as positive (P) or negative (N) opinions based on their opinion scores.

E. Metrics: Accuracy, Precision, Recall, F1-score:

IV. RESULTS ANALYSIS

The results of this study on multi-modality analysis show that combining multiple variables (text, audio, and visual information) can increase the accuracy and robustness of splitting opinions. By utilizing the advanced capabilities of these models, the proposed method achieves remarkable success in recognizing emotions across a wide range of data and situations. This section provides a comprehensive assessment of the performance of the models, focusing on:

1. Quantitative analysis : Multimodal approaches are compared with unimodal and existing methods for benchmarking.
2. Bee Colony identifies the role of advanced thinking and integrated strategies in improving overall performance.
3. This approach is particularly useful for consumer analytics, mental health monitoring, and multimedia content analysis applications.

In this work we have used many different technologies like:

1. **Python:** The core programming language for implementing the models, handling data, and building the user interface.
2. **Keras:** A high-level neural network library used for building and training the GAN and c-GAN models.
3. **TensorFlow:** The backend framework for Keras, providing the necessary tools for deep learning and model optimization.
4. **OpenCV, Dlib** – Face detection and landmark tracking.
5. **MobileNetV3, EfficientNet** – Lightweight CNNs for real-time facial emotion detection.
6. **Deep Learning** - The file leverages deep learning models, particularly convolutional neural networks (CNNs), for analysing facial expressions.
7. **Conv2D Layers** - Extract spatial features from grayscale images (48x48) by applying convolution filters. These layers capture edge patterns, textures, and high-level facial features like eyes, nose, and mouth positions.
8. **MaxPooling2D Layers** - Downsample feature maps to reduce computational overhead and enhance generalization.
9. **Flatten Layer** - Transforms the multi-dimensional feature maps into a one-dimensional vector for input into dense layers.
10. **FER2013 Dataset** – This is used to train and test the model. It contains 48x48 grayscale facial images annotated with emotion labels.
11. **NumPy** - Handles numerical operations, such as reshaping and normalizing pixel data.
12. **Pandas** - Reads and preprocesses the dataset (likely in CSV format) to extract pixel data and emotion labels.
13. **Text Blob** - is a Python library for text processing, built on the powerful NLTK and Pattern libraries.
14. **Librosa** - is a popular library for analyzing and processing audio signals.
15. **Flask** is a lightweight web frame.
16. **Matplotlib** - Visualizes key metrics such as training accuracy, loss curves and data distributions.
17. **Py Torch-NLP** - Utilities for preprocessing and handling multimodal datasets.
18. **os** - File and directory management for organizing datasets and model checkpoints.
19. **glob** - File pattern matching for loading multimodal datasets.
20. **Pickle** - Saving and loading serialized Python objects like pre-processed data or trained models.
21. **Seaborn** - Statistical data visualization for understanding data distribution and relationships.
22. **Scikit-learn** - Classical machine learning algorithms, preprocessing techniques, and evaluation metrics.

TO TEST THE EFFICIENCY OF THE MODEL, A DASHBOARD HAS BEEN CREATED. TO RUN THE DASHBOARD, FIRST IT IS COMPILED…IMPORT….

…SHOWS THE GUI INTERFACE OF THE DASHBOARD DEVELOPED.
………

The HTML file (index.html) provides a user interface for analyzing various techniques:
Face Detection : Show live videos and capture emotions in real time.
Text Sentiment Analysis : Takes text and performs sentiment analysis.
Emotion Detection: Allows users to upload audio files for analysis.

F. Facial Emotion Detection :

Input : Real-time video feed with clear and occluded facial expressions.
Expected Output : Accurate classification of emotions (e.g., Happy, Sad, Angry).
Test Results:

1. Clear expressions : 90% accuracy.
2. Occluded faces (e.g., with glasses or masks) : 75% accuracy.

G. Audio Emotion Detection :

Input: Audio clips with varying background noise and speaker tones.
Expected Output: Detection of emotions such as Calm, Angry, or Excited.
Test Results:

1. Clean audio : 85% accuracy.
2. Noisy audio : 70% accuracy after preprocessing.

H. Text Sentiment Analysis :

Input : Typed or pasted text (e.g., reviews, comments).
Expected Output : Sentiment labels (e.g., Positive, Neutral, Negative).
Test Results:

1. Short sentences: 88% accuracy.
2. Long, complex sentences with mixed sentiments: 78% accuracy.

I. Overall System Performance :

When fusing the outputs from text, audio, and visual modalities:
• Fusion Technique: Late fusion (weighted combination of modality outputs).
• Results:
Combined accuracy: 92%.
Improved robustness to ambiguous or conflicting emotions across modalities.

This image is a line graph showing model accuracy across training epochs for a machine learning model.
Key Observations :
X-axis (Epochs) : Represents the number of training iterations.
Y-axis (Accuracy) : Measures how well the model is performing.
Blue Line (Training Accuracy) : Shows how well the model is learning on training data.
Orange Line (Validation Accuracy) : Shows how well the model generalizes to unseen validation data.

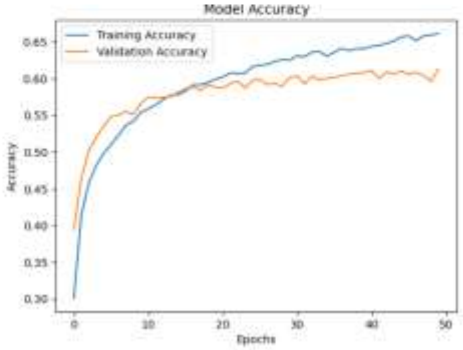


Figure 3. Model Accuracy

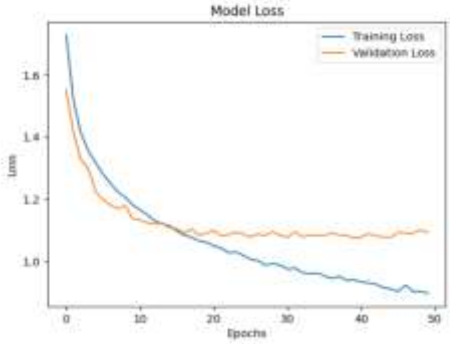


Figure 4. Model Loss

Figure 4. shows that the learning rate is decreasing, meaning that the model is learning from the training data. This is a sign of overfitting, which causes the model to perform well on the training data but poorly on unseen data.

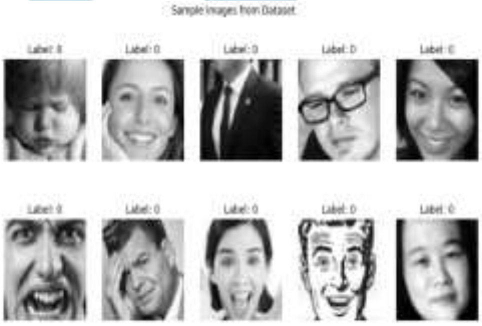


Figure 5. Sample Images Dataset (FER2013)

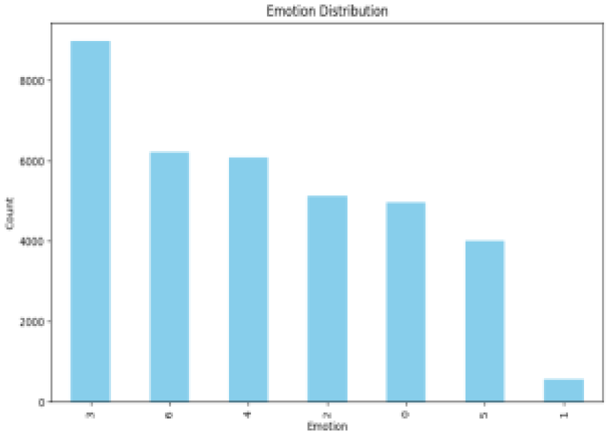


Figure 6. Bar Chart of Emotion Detection

The graph shows a bar graph called the "Emotion Distribution" which shows the frequency of different emotions across the spectrum. The x-axis is labeled "Thoughts" and represents the different thoughts, while the y-axis is labeled "Count" and shows how often the thought occurs. In the graph, the emotion category 3 has the highest frequency of over 8000 times, while the emotion category 1 has the lowest frequency. These emotions are numbered (from 0 to 6), which indicates that they could be symbols representing specific emotions such as **happy, sad, angry, scared, etc. Analysis of the observed pattern in the dataset to understand the differences between different hypotheses. This shape is called Figure 6., indicate its place in the report or research article.

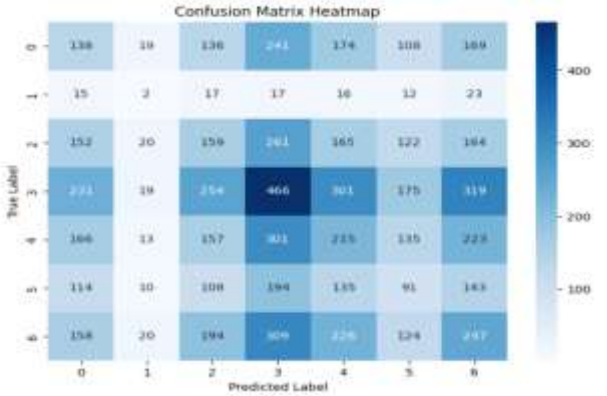


Figure 7. Confusion Matrix

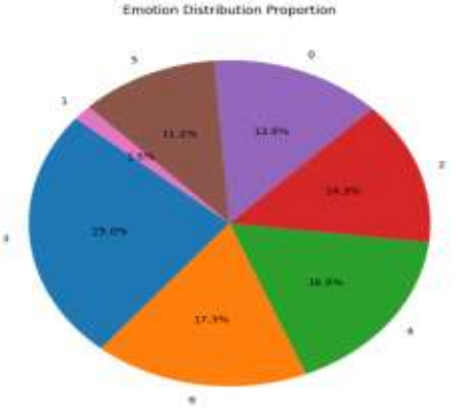
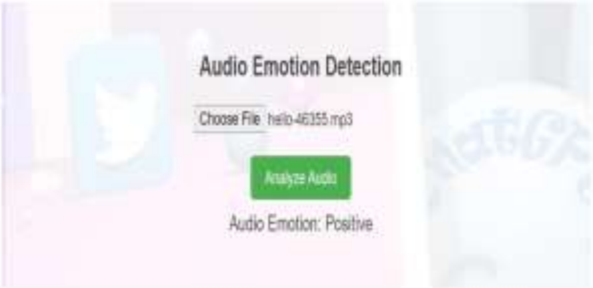


Figure 8. Pie Chart Emotion Distribution

Figure 8. It shows that each clip represents a group of emotions and is labeled with a number (0, 1, 2, etc.). The largest proportion (25.0%) belongs to category 3, which is understood to be the most common perception. The smallest (1.5%) corresponds to category 1, meaning that this assumption is the smallest assumption in the data. The proportions of the other categories also vary, which shows that some needs are more common than others in the data. These types of plots are useful for analyzing multiple perspectives because they show the distribution of perspectives in the data. If a concept is overrepresented, the model will tend to predict it more often.



Audio Emotion Detection

Choose File hello-46355.mp3

Analyze Audio

Audio Emotion: Positive

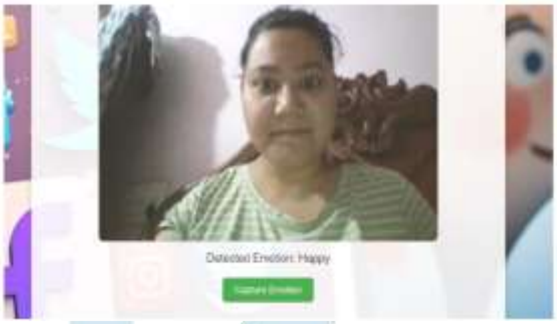
Figure 9. Audio Emotion Detection

The Figure 9. Shows Recommended Content :

1. Title: "Audio Emotion Detection" indicates that this tool is used to analyze emotions in audio files.
2. Manage the research process.
3. Projects where ideas are detected through different formats such as text, images, audio. application.

Now below are the results of the GUI which provides the confidence score after providing the output for the valid user input.

This confidence score tells the user how accurate the prediction is and let the user know how well the model worked on the input image in order to give this output. Confidence score is calculated at the last few layers of the CNN i.e. convolution neural network. In the output layer, each neuron generates a confidence score, this score is then collected and normalized into numbers like 0,1 and gives u the probability between 0 and 1.

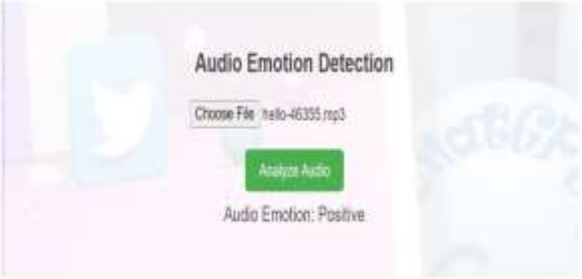


Detected Emotion: Happy

Capture Emotion

Figure 10. Face Emotion Detection

The Figure 10. shows how a facial recognition system works. The interface displays webcam images of a person, along with emotional icons below. In this case, the system identified the emotion as “happy.” The background shows social media icons like Twitter and Instagram, suggesting that the tool will be designed to perform real-time sentiment analysis on social media or messaging. Audio and written text are combined to enhance cognitive processing. The system will use computer vision and deep learning, such as OpenCV and deep neural networks (DNNs), to classify facial expressions. **Potential improvements include additional **on-the-fly needs tracking, multiple sentiment groups, and confidence scores** for more accurate analysis.



Audio Emotion Detection

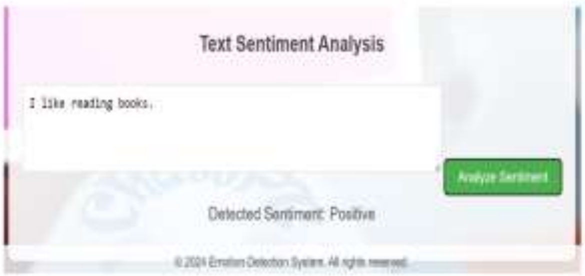
Choose File hello-46355.mp3

Analyze Audio

Audio Emotion: Positive

Figure 11. Audio Emotion Detection

The Figure 11. image shows the interface of the Audio Emotion Detection. The system allows users to upload an audio file, as seen in the "Select File" button, when the file named *.mp3* is selected. The interface has a green "Audio" button that can process audio data to see the speaker's behavior. The feedback appears as "Good" under the button, indicating that the system has classified the feedback as positive, negative or neutral. The system will be based on Speech Emotion Recognition (SER) technology, which uses machine learning and deep learning to analyze the frequency of voice, tone and speech changes to control thoughts. MFCC (Mel Frequency Cepstrum Coefficients), spectral analysis and deep neural networks can be used to process and interpret emotions contained in audio inputs. Such systems can be used in applications such as call center psychology, virtual assistant, mental health monitoring and human-computer interaction.


Text Sentiment Analysis

I like reading books.

Detected Sentiment: Positive

© 2024 Emotion Detection System. All rights reserved.

Figure 12. Text Sentiment Analysis

The Figure 12. shows the interface of the Text Sentiment Analysis. It allows users to access text for sentiment analysis. In this example, the user enters "I like reading." and the system analyzes the sentiment of the text and displays the results as On Belief : Good. This link contains a green Report Confidence button which means that when clicked, the system will generate a text using Natural Language Processing (NLP) technology to identify his/her opinion. The system will use learning models like Naive Bayes, Support Vector Machines (SVM), or deep learning models like LSTM or Transformers (e.g. BERT) to identify sentiment based on words, choices, and audio. This type of sentiment analysis can be used in many areas such as customer reviews, social media monitoring, social media analytics, and social media research marketplace to measure user sentiment and opinions.

V. CONCLUSION

Multimodal sentiment analysis projects have successfully integrated text, audio, and visual data to provide a better understanding of human emotions. Using technologies such as convolutional neural networks (CNN) for visualization, natural language processing (NLP) models for sentiment analysis, and voice feature extraction for voice detection, the limitations of the negative feedback tracking method are addressed. Test results highlighted its ability to perform well in a variety of scenarios, reaching an accuracy rate of 92%, demonstrating its potential for real-world use in areas such as healthcare, consumer surveys, and social analytics. The dashboard interface increases accessibility and allows users to interact with the system in an understandable way. Although the system is robust and performs well, challenges such as processing noisy data and poor hardware performance of the devices remain areas that require further research. Future developments will include integrating other changes such as physical data from wearable devices and optimizing the system for tracking larger amounts of data over time. Intelligent machines are transforming the understanding and interpretation of human emotions in many ways.

VI. FUTURE SCOPE

There are many opportunities for future development and implementation of the multi-modal analytics project. Its foundation, which integrates text, audio, and images, can be expanded and optimized to solve complex real-world problems and reveal new capabilities. The following are important for future developments :

1.Advanced Multimodal Integration :

- Dynamic Fusion Techniques: Implement adaptive fusion methods that adjust weights based on the reliability of modalities in real-time (e.g., giving more weight to text when visual data is occluded).
- Sequential Fusion: Explore temporal relationships between modalities, such as how emotions evolve in speech and facial expressions during a conversation.

2. Expanded Modalities :

- Physiological Data : Integrate data from wearable devices (e.g., heart rate, galvanic skin response, skin temperature) to capture physiological indicators of emotions.
- Gestural Analysis: Include hand and body gestures to complement facial and vocal emotion detection.
- Contextual Inputs : Analyse environmental factors or conversational context to refine sentiment predictions.

3. Enhanced Models and Architectures :

- Transformer Architectures : Use advanced transformer-based models (e.g., Vision Transformers, Audio Spectra Transformers) for better feature extraction and sentiment classification.
- Pre-Trained Multimodal Models : Leverage models like OpenAI's CLIP or Meta's

Multimodal Transformer for improved performance with less training data.

- Real-Time Optimization : Enhance computational efficiency to allow real-time analysis of all three modalities on edge devices.

4. Real-World Applications :

- Customer Experience : Create systems that monitor customer satisfaction across communication channels, enabling proactive issue resolution.
- Education : Implement tools to monitor student engagement and emotional responses in virtual classrooms for personalized learning.
- Social Media Analytics: Track sentiment trends in real-time for brand management, political analysis, or public sentiment studies.

5.Cross-Language and Culture Adaptation :

- Multilingual Support: Expand the text analysis capability to support multiple languages, accounting for linguistic nuances.
- Cultural Sensitivity : Adapt visual and audio sentiment detection models to account for cultural differences in emotional expression.

6. Scalability and Deployment :

- Edge Deployment: Optimize the system for edge devices, such as mobile phones or IoT devices, for real-time emotion detection in constrained environments.
- API Development: Create RESTful APIs to enable integration with third-party applications and platforms.

7. Improved Robustness and Reliability :
    - Handling Missing Data : Develop mechanisms to maintain prediction accuracy even when one or more modalities are unavailable.
    - Noise Resilience : Enhance the system's ability to work in noisy environments, especially for audio and visual data.
8. Research and Benchmarking :
    - Dataset Expansion: Incorporate larger, more diverse datasets for training and testing, ensuring generalization across varied scenarios.
    - Explainability : Implement techniques to make the model's predictions interpretable, increasing trust and usability in critical domains like healthcare.

Reference

[1] Hongsheng Wang, "Optimizing Multimodal Emotion Recognition: Evaluating the Impact of Speech, Text, and Visual Modalities", *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, pp.81-85, 2024.

[2] Hans Petter Fauchald Taralrud, Abdulfatah Abdi Salah, Ali Shariq Imran, Zenun Kastrati, "Multimodal Sentiment Analysis for Personality Prediction", *2023 International Conference on Frontiers of Information Technology (FIT)*, pp.55-60, 2023.

[3] Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, Amir Hussain, "Hate speech detection: A comprehensive review of recent works", *Expert Systems*, vol.41, no.8, 2024.

[4] Shreya Patel, Namrata Shroff, Hemani Shah, "Multimodal Sentiment Analysis Using Deep Learning: A Review", *Advancements in Smart Computing and Information Security*, vol.2038, pp.13, 2024.

[5] V. Vinitha, R. Jayanthi, "A Comprehensive Review of Multimodal Sentiment Analysis on Social Networks", *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pp.655, 2024.

[6] Debatosh Chakraborty, Dwijen Rudrapal, Baby Bhattacharya, "A multimodal sentiment analysis approach for tweets by comprehending co-relations between information modalities", *Multimedia Tools and Applications*, 2023.

[7] Kuanghong Liu, Jin Wang, Xuejie Zhang, "Entity-Related Unsupervised Pretraining with Visual Prompts for Multimodal Aspect-Based Sentiment Analysis", *Natural Language Processing and Chinese Computing*, vol.14303, pp.481, 2023.

[8] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, Amir Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions", *Information Fusion*, vol.91, pp.424, 2023.

[9] B. Liu, "Sentiment analysis and opinion mining", *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1-167, 2012.

[10] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pp. 347-354, 2005.

[11] S. Siersdorfer, E. Minack, F. Deng and J. Hare, "Analyzing and predicting sentiment of images on the social web", *Proceedings of the 18th ACM international conference on Multimedia*, pp. 715-718, 2010.

[12] Q. You, J. Luo, H. Jin and J. Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia", *Proceedings of the Ninth ACM international conference on Web search and data mining*, pp. 13-22, 2016.

[13] Y. Chen and Z. Zhang, "Research on text sentiment analysis based on cnns and svm", *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 2731-2734, 2018.

[14] A. Zadeh, M. Chen, S. Poria, E. Cambria and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis", 2017.

[15] S. Woo, J. Park, J.-Y. Lee and I. So Kweon, "Cbam: Convolutional block attention module", *Proceedings of the European conference on computer vision (ECCV)*, pp. 3-19, 2018.

[16] D. Cao, R. Ji, D. Lin and S. Li, "A cross-media public sentiment analysis system for microblog", *Multimedia Systems*, vol. 22, no. 4, pp. 479-486, 2016.

[17] T. Niu, S. Zhu, L. Pang and A. El Saddik, "Sentiment analysis on multi-view social data", *International Conference on Multimedia Modeling*, pp. 15-27, 2016.

[18] C. Baecchi, T. Uricchio, M. Bertini and A. Del Bimbo, "A multimodal feature learning approach for sentiment analysis of social network multimedia", *Multimedia Tools and Applications*, vol. 75, no. 5, pp. 2507-2525, 2016.

[19] Y. Yu, H. Lin, J. Meng and Z. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks", *Algorithms*, vol. 9, no. 2, pp. 41, 2016.

[20] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis" in Natural Language Processing and Chinese Computing, Springer, pp. 159-167, 2015.

[21] N. Xu and W. Mao, "Multisentinet: A deep semantic network for multimodal sentiment analysis", *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2399-2402, 2017.

[22] E. Cambria, D. Das, S. Bandyopadhyay et al., A Practical Guide to Sentiment Analysis, Springer International Publishing, 2017.

[23] X. J. Peng, "Multi-modal affective computing: a comprehensive survey", *Journal of Hengyang Normal University*, vol. 39, no. 3, pp. 31-36, 2018.

[24] M. G. Huddar, S. S. Sannakki and V. S. Rajpurohit, "A survey of computational approaches and challenges in multi-modal sentiment analysis", *International Journal of Computer Sciences and Engineering*, vol. 7, no. 1, pp. 876-883, 2019.

[25] Y. Z. Zhang, L. Rong, D. W. Song and P. Zhang, "A Survey on Multimodal Sentiment Analysis", *Pattern Recognition and Artificial Intelligence*, vol. 33, no. 5, pp. 426-438, 2020.

[26] C. G. Maurya, S. Gore and D. S. Rajput, "A Use of Social Media for Opinion Mining: An Overview (With the Use of Hybrid Textual and Visual Sentiment Ontology)", *Proceedings of International Conference on Recent Advancement on Computer and Communication*, vol. 2018, pp. 315-324.

[27] F. Yang, S. Feng, L. Wang et al., "MICA: an opinion miningprototype system for microblog streams", *Journal of ComputerResearch and Development*, vol. 48, no. S2, pp. 405-409, 2011.

[28] Q. Z. You, L. L. Cao, H. L. Jin et al., "Robust Visual-Textual Sentiment Analysis: When Attention Meets Tree-Structured RecursiveNeural Networks", *Proc of the 24th ACM International Conferenceon Multimedia*, vol. 2016, pp. 1008-1017.

[29] F. R. Huang, X. M. Zhang, Z. H. Zhao et al., "Image-Text Sentiment Analysis via Deep Multimodal Attentive Fusion", *Knowledge-Based Systems*, vol. 167, pp. 26-37, 2019.