Multimedia Sentiment Analysis

Sajan Arora,
Khoury College of Computer Sciences,
Northeastern University,
Boston, MA, USA,
arora.saj@northeastern.edu

Abstract

This paper introduces a comprehensive system for sentiment analysis that incorporates multiple data types: text, images, and audio. Traditional sentiment analysis techniques primarily focus on text, which limits their effectiveness in situations where multimodal data is common. In this project, we developed separate models for each data type—text, image, and audio—and their outputs were combined to produce a unified sentiment score. The results shows that this approach greatly enhances the accuracy and reliability of sentiment detection.

Introduction

Sentiment analysis is an important tool for understanding public opinion, emotions, and attitudes expressed in different types of data. Traditional methods have mainly focused on text, often missing the emotional context found in images and audio. This project aims to overcome these limitations by developing a system that can analyze sentiment across text, images, and audio. By using models designed specifically for each type of data and combining their results into a unified sentiment score, this approach aims to give a more accurate and detailed understanding of sentiment.

Related Work

Our project on multimodal sentiment analysis is based on existing research. It focuses on combining different types of data, like text, images, and audio. The study by Lai et al. (2021) called ”Multimodal Sentiment Analysis: A Survey” gives a detailed overview of the datasets, methods, and uses in this field. It shows how recent advances in deep learning have made sentiment analysis more accurate and reliable. The authors explain that using different types of data together has greatly improved the performance of these systems.

Another important contribution to this field is the work by Alam, Ryu, and Lee (2020) titled ”Sentiment Analysis using a Deep Ensemble Learning Model.” This study looks at how deep ensemble learning models can improve the accuracy and efficiency of sentiment analysis. The authors show that by combining different machine learning techniques, ensemble models can capture the complex patterns in sentiment data more effectively, leading to better performance across various datasets. Their research highlights the potential of ensemble learning to overcome the limitations of single models, making it a key technique for advancing sentiment analysis.

Similarly, the work by Choube and Soleymani (2020) titled ”Punchline Detection using Context-Aware Hierarchical Multimodal Fusion” introduces a new neural architecture that combines text, audio, and visual data to detect humor, particularly punchlines. The paper presents a hierarchical fusion approach using Gated Recurrent Units (GRUs) to model context, achieving top results on the UR-FUNNY database. This method is especially relevant to our project as it shows the effectiveness of multimodal fusion in understanding humor, which often involves complex interactions between different types of data.

These studies emphasize the importance of integrating multiple data types and advanced machine learning techniques in sentiment analysis. Our work builds on these foundations by developing a comprehensive system that combines sentiment analysis from text, images, and audio, using the strengths of deep learning and multimodal fusion to create a unified sentiment score.

Data Collection and Preprocessing

Text Data

The Sentiment140 dataset, which contains 1.6 million tweets labeled as either positive or negative, was selected for text sentiment analysis. For computational efficiency and to ensure balanced data representation, a subset of 20,000 tweets—10,000 positive and 10,000 negative—was used. https://www.kaggle.com/datasets/kazanova/sentiment140

Preprocessing Steps

Noise Removal

We removed URLs, user mentions, hashtags, and special characters to clean the text. This step was crucial in eliminating non-informative elements that could distort the sentiment analysis.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

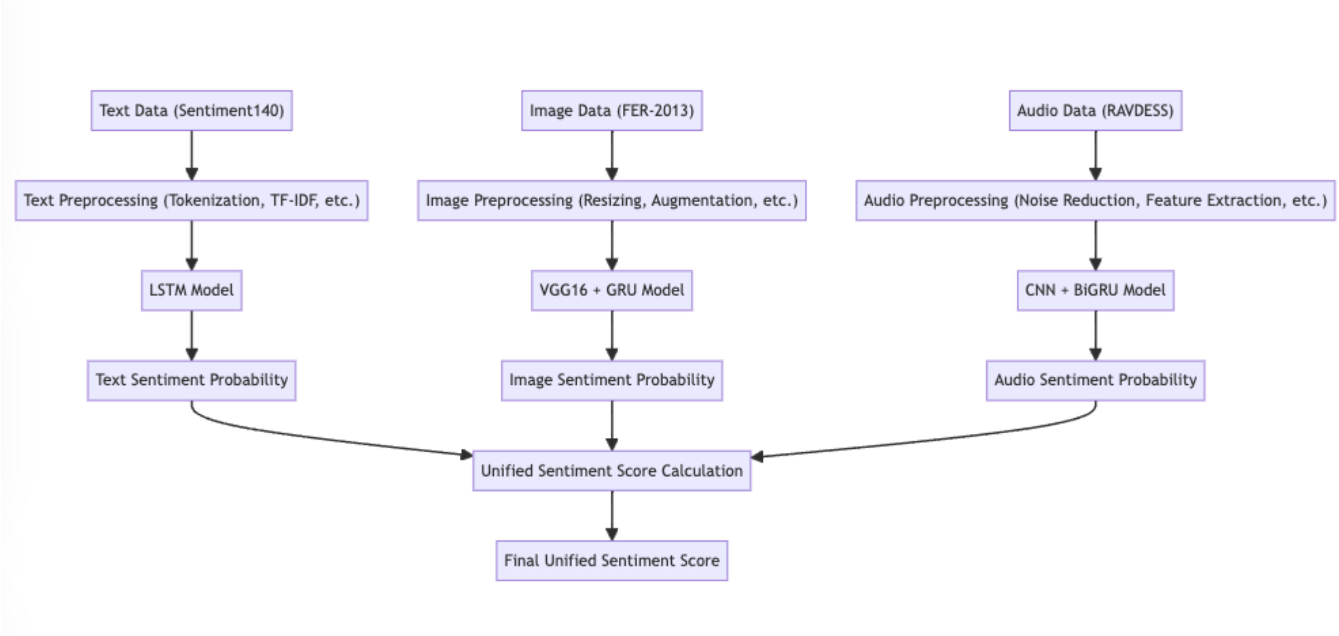
Text Data (Sentiment140)
Image Data (FER-2013)
Audio Data (RAVDESS)
Text Preprocessing (Tokenization, TF-IDF, etc.)
Image Preprocessing (Resizing, Augmentation, etc.)
Audio Preprocessing (Noise Reduction, Feature Extraction, etc.)
LSTM Model
VGG16 + GRU Model
CNN + BiGRU Model
Text Sentiment Probability
Image Sentiment Probability
Audio Sentiment Probability
Unified Sentiment Score Calculation
Final Unified Sentiment Score

Figure 1: Workflow for Multimodal Sentiment Analysis

Tokenization

The cleaned text was tokenized using the Natural Language Toolkit (NLTK), splitting it into individual words (tokens). This step was essential in preparing the text for further analysis.

Stop Words Removal

We filtered out commonly used words (e.g., “the,” “and,” “is”) using NLTK’s stop words list. This helps the model focus on the more meaningful content of the text.

Stemming and Lemmatization

Words were reduced to their base or root forms through stemming and lemmatization. This step helps reduce the complexity of the text and improves the model’s ability to generalize.

Feature Extraction

The processed text was transformed into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. This highlights the importance of words within the text, allowing the model to focus on the most informative terms.

Image Data Preprocessing

We used the FER-2013 dataset, containing 35,887 grayscale facial images labeled with one of seven emotions, for image sentiment analysis. To maintain balanced representation, we selected a subset of 24,000 images for training. The following preprocessing steps were applied: You can find the FER-2013 dataset on Kaggle here: FER-2013 Dataset on Kaggle.

Resizing

All images were resized to 48x48 pixels to standardize the input dimensions for the model. This ensures consistency across the dataset and compatibility with the model’s architecture.

Normalization

Pixel values were normalized to a range between 0 and 1. This step helps improve the model’s learning efficiency by ensuring uniformity in pixel intensity.

Data Augmentation

We applied techniques such as rotation, zoom, and horizontal flipping to increase the variety of training data. This enhances the model’s robustness by exposing it to a wider range of image variations.

Audio Data Preprocessing

For audio sentiment analysis, we used the RAVDESS dataset, which includes 7,356 emotional speech audio files. You can find the audio sentiment dataset on Kaggle here: RAVDESS Emotional Speech Audio Dataset on Kaggle.The preprocessing steps included:

Noise Reduction

We reduced background noise by using the first second of the audio as a noise reference. This was essential for improving audio clarity and feature extraction accuracy.

Feature Extraction

We extracted key features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma, spectral contrast, and tonnetz to represent the audio signals. These features are critical for accurate emotion recognition.

Data Augmentation

To diversify the audio data, we applied augmentation techniques like pitch shifting and time stretching. This increases the model's robustness by exposing it to a broader range of audio conditions.

Model Development

Text Sentiment Analysis

For text sentiment analysis, a Long Short-Term Memory (LSTM) model was developed. LSTM networks are a type of recurrent neural network (RNN) that are particularly effective for sequence prediction tasks, making them ideal for analyzing text data.

Model Architecture The LSTM model architecture comprised the following layers:

- **Embedding Layer**: This layer transforms words into dense vectors of fixed size that offers a continuous and meaningful representation of words in the text
- **LSTM Layers**: Two LSTM layers were utilized to capture the temporal dependencies within the text. The first LSTM layer returns sequences that feed into a second LSTM layer for further processing.
- **Dropout Layers**: Dropout layers were applied after each LSTM layer to reduce overfitting by randomly setting a portion of the input units to zero during training.
- **Dense Layer**: The final dense layer used a sigmoid activation function for binary classification, distinguishing between positive and negative sentiment.

Model Training and Results The model was trained using binary cross-entropy as the loss function and the Adam optimizer. Early stopping was employed to halt training when the validation loss ceased to improve, preventing overfitting. The final test accuracy was 70.78%, with balanced precision, recall, and F1-scores across both positive and negative sentiment classes. The training and validation accuracy plots, indicate steady improvement, with validation accuracy stabilizing around 71

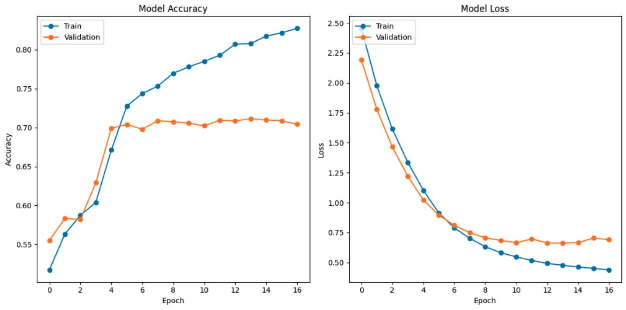

Figure 2: LSTM model

Image Sentiment Analysis

For image sentiment analysis, a hybrid model combining Convolutional Neural Network (CNN) with Gated Recurrent Unit (GRU) layers was used. The CNN was utilized for feature extraction from the images, while the GRU layers were used to capture sequential dependencies within the extracted features.

Model Architecture The hybrid model architecture included the following components:

- **VGG16 Base**: We used the VGG16 model, pretrained on ImageNet, as the base for feature extraction. The convolutional layers were fine-tuned to better suit the FER-2013 dataset.
- **Time Distributed Layer**: This layer was applied to the output from the CNN that had been transformed into a one-dimensional format that enables the model to effectively manage sequential data
- **GRU Layer**: A GRU layer was included to model temporal rela- tionships in the sequential data which helps the model to capture the sequence of features extracted by the CNN
- **Dense Layers**: The model includes dense layers with ReLU activation for classification, with the final layer using softmax activation for multi- class classification

Model Training and Results The model was trained using categorical cross-entropy as the loss function and the Adam optimizer. Early stopping and learning rate reduction on plateau were employed to optimize training. The final test accuracy was 61.12%. The confusion matrix reveals that "happy" is the most accurately predicted emotion, with notable confusion between similar emotions such as "fear" and "sad." ROC curves indicated AUC values ranging from 0.81 to 0.95, demonstrating strong overall performance.

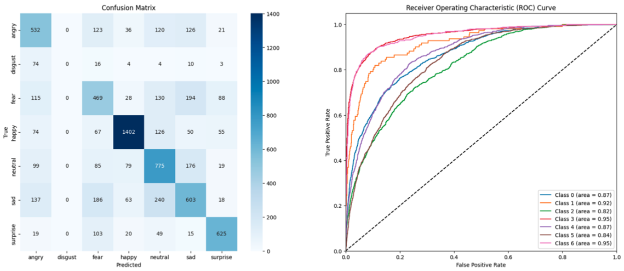

Figure 3: VGG16 + GRU Hybrid Model

Audio Sentiment Analysis

For audio sentiment analysis, we formed a hybrid model that combines Convolutional Neural Networks (CNNs) with Bidirectional GRU layers. The CNN layers handled feature extraction from the audio inputs, while the GRU layers modeled the sequential nature of the audio data.

Model Architecture The CNN + Bidirectional GRU model architecture included:

- **Conv1D Layers**: We used two 1-dimensional convolutional layers to extract features from the MFCCs and other audio features. These layers were followed by max-pooling layers to reduce dimensionality and mitigate overfitting.
- **Bidirectional GRU Layers**: Two Bidirectional GRU layers were employed to capture dependencies in both directions (past and future) within the audio sequences which enhances the model's understanding of context.
- **Dense Layers**: The output from the GRU layers was passed into dense layers and lastly we have a softmax layer for multi-class classification which differentiates between various emotions

Model Training and Results The model was trained using categorical cross-entropy as the loss function and the Adam optimizer. Early stopping and learning rate reduction on plateau were employed to optimize training. The final test accuracy was 81.94%, with ROC curves indicating strong differentiation between emotional classes.

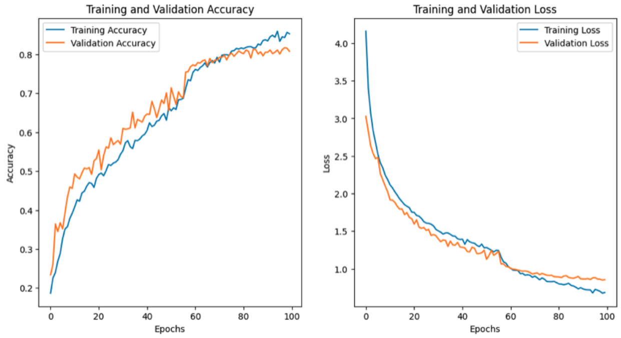

Figure 4: CNN + Bidirectional GRU model

Unified Sentiment Score Calculation

To achieve a consistent comparison across text, image, and audio modalities, unified sentiment scores were calculated by averaging the predictions from each model. This approach provided a comprehensive understanding of how sentiment is interpreted across different data types.

Unified Scores

The unified sentiment scores for text, image, and audio were as follows:

- **Unified Sentiment Scores for Text:** [0.6993828 0.6390063 0.28529093 ... 0.8638967 0.9825865 0.6690353]
- **Unified Sentiment Scores for Images:** [0.14285714 0.14285714 0.14285713 0.14285713 0.14285715 0.14285714]
- **Unified Sentiment Scores for Audio:** [0.12500001 0.125 0.12500001 0.125 0.125 0.12499999]

These scores suggest that text data tends to convey stronger sentiment signals, while visual and audio data provide complementary context.

Comparison of Unified Sentiment Scores

Once the unified sentiment scores were calculated for each modality, they were compared to analyze consistency and differences in sentiment detection.

The below bar chart shows the differences in unified sentiment scores across text, image, and audio modalities for ten samples. There are noticeable gaps between Text vs. Image and Text vs. Audio, indicating that sentiment is interpreted differently in text compared to the other two. In contrast, the differences between Image vs. Audio are smaller, suggesting that visual and auditory data are more closely aligned in sentiment interpretation.

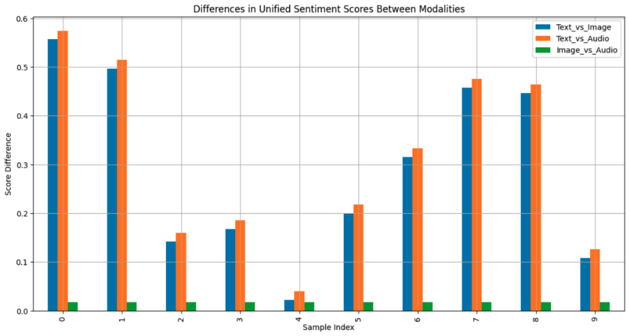

Figure 5: Comparison of Unified Sentiment Scores

Performance Evaluation

The performance of each model was evaluated using the Accuracy, Precision, Recall, F1-Score, Confusion Matrix and ROC curve:

Results

- **Text Sentiment Analysis:** The LSTM model achieved an accuracy of 70.77% on the test set. The Unified Sentiment Scores for Text is [0.6993828 0.6390063 0.28529093 ... 0.8638967 0.9825865 0.6690353]
- **Image Sentiment Analysis:** The VGG16 + GRU model achieved an accuracy of 61.12% on the test set. The Unified Sentiment Scores for Image is [0.14285714 0.14285714 0.14285713 0.14285713 0.14285715 0.14285714]
- **Audio Sentiment Analysis:** The CNN + Bidirectional GRU model achieved an accuracy of 81.94% on the test set. The Unified Sentiment Scores for Audio is [0.12500001 0.125 0.12500001 0.125 0.125 0.12499999]

Conclusion

The project proved that using a combination of text, images, and audio data is both practical and effective. This approach allowed the model to predict sentiment more accurately than when relying on just one type of data.

Future Work

- **Additional Datasets:** Experiment with more diverse datasets to validate the model's generalizability.
- **Model Improvements:** Explore advanced architectures such as transformers for text and attention mechanisms for image and audio.
- **Real-Time Implementation:** Implement the model in a real-time application for dynamic sentiment analysis.

References

Lai, S., Hu, X., Xu, H., Ren, Z., & Liu, Z. (2021). Multimodal Sentiment Analysis: A Survey. *IEEE Transactions on Affective Computing.*

Alam, M. H., Ryu, W.-J., & Lee, S. (2020). Sentiment Analysis using a Deep Ensemble Learning Model. *Journal of Information Science and Engineering.*

Choube, A., & Soleymani, M. (2020). Punchline Detection using Context-Aware Hierarchical Multimodal Fusion. *Proceedings of the 28th ACM International Conference on Multimedia.*

Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2017). A Review of Multimodal Sentiment Analysis. *Information Fusion, 37, 98-110.*

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. (2018). Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems, 33(6), 54-61.* doi:10.1109/MIS.2018.2876504

Akhtar, M. S., Ekbal, A., & Cambria, E. (2019). Multitask Learning for Multimodal Emotion Recognition and Sentiment Analysis. *Knowledge-Based Systems, 173, 43-54.* doi:10.1016/j.knosys.2019.02.034

Wang, W., Shen, J., Shao, L., & Cheng, M. M. (2019). Deep Multimodal Fusion by Leveraging Noise Signals for Audiovisual Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 29(10), 2995-3008.* doi:10.1109/TCSVT.2018.2868763

Tsai, Y. H., Ma, M. Y., Wang, L., Zadeh, A., & Morency, L. P. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 6558-6569.* doi:10.18653/v1/P19-1656

Hazarika, D., Zimmermann, R., Poria, S., & Cambria, E. (2020). Multimodal Sentiment Analysis via Hierarchical Fusion with Context Modeling. *IEEE Transactions on Affective Computing, 12(4), 1018-1029.* doi:10.1109/TAFFC.2020.2975887