Sentiment Analysis of Audio, Text, Emojis, and Ratings

G. Sushrut Reddy
Department of Information Technology
Chaitanya Bharati Institute of Technology
Hyderabad, India
goodellysushrut@gmail.com

Hari Charan
Department of Information Technology
Chaitanya Bharati Institute of Technology
Hyderabad, India
jatothharicharan@gmail.com

K. Chintu
Department of Information Technology
Chaitanya Bharati Institute of Technology
Hyderabad, India
kodavathchintu@gmail.com

Dhanush
Department of Information Technology
Chaitanya Bharati Institute of Technology
Hyderabad, India
dhanushsampathi958@gmail.com

***Abstract*—User sentiment is frequently conveyed in multiple forms like text, audio transcriptions, emojis, and ratings. Conventional sentiment analysis generally only uses text, losing valuable emotional information from other forms. To fill this void, we introduce a multimodal sentiment analysis system that combines text, audio transcriptions, and user ratings into one framework. This system translates audio and non-text data into text so that text-based machine learning models can be used for sentiment classification. Our solution utilizes an LSTM-based sentiment classification model that classifies reviews as Positive, Neutral, and Negative sentiments. We also use a text generation model, google/flan-t5-large, to produce marketing recommendations based on sentiments identified, giving businesses concrete feedback to enhance product descriptions and features.**

The outcomes emphasize the strengths of multimodal sentiment analysis, showing stronger sentiment comprehension and marketing optimization. By processing reviews from various forms of input, we gain richer sentiment knowledge and more effective product placement. The most important contributions of this paper are (1) building a multimodal pipeline for sentiment analysis, (2) using an LSTM model to perform sentiment classification, and (3) coupling sentiment analysis with marketing suggestion generation. This study highlights the power of multimodal sentiment analysis in enabling businesses to gain richer customer feedback insights, drive better decisions, and optimize customer engagement strategies.

***Index Terms*—Sentiment Analysis, Multimodal Fusion, Audio to Text, Emoji Conversion, Text Processing, Ratings-Normalization, Machine Learning**

I. Introduction

In today's digital era, user-generated content dominates online platforms. From product reviews to social media exchanges, individuals express opinions and emotions through various means. These sentiments can help businesses and organizations gauge customer satisfaction, market trends, or public opinion. Accurate sentiment analysis is key to enhancing user experience and enabling data-driven decisions.

Sentiment analysis originated in Natural Language Processing (NLP) and machine learning, initially developed for text data. It has since been applied to analyze customer feedback and public opinion. However, with the rise of online communication, people now provide feedback using emojis, voice messages, and star ratings, presenting both challenges and opportunities for advanced multimodal analysis systems.

A. Background of the Problem

In the past, sentiment analysis has changed from simple keyword detection methods to deep learning approaches. The original systems had only to contend with structured text, often lacking on nuances such as sarcasm or sentiment. Modes of communication have changed with messaging and social networking like WhatsApp and Instagram, providing images, audio recordings, and emojis, making sentiment richer. The majority of current systems are nonetheless text-centric, neglecting critical emotional indicators in other media.

B. Motivation for the Study

The inspiration behind this work stems from a desire to develop a sentiment analysis system that captures how individuals interact online. Basic text-based sentiment models do not accurately capture sentiment when presented with multimodal inputs. By closing the gap between common NLP models and the interactive nature of online communication through interpretation of sound, emojis, and star ratings into meaningful textual content, we intend to enhance accuracy and relevance in sentiment identification.

C. Issues and Gaps in Existing Work

While recent work has delved into multimodal sentiment analysis, there are significant gaps left to be filled. The majority of models analyze only a single or restricted set of modalities—text, audio, or visual inputs by way of emojis—separate from one another. Some models integrate text and audio or text and emojis, yet no general framework is available that smoothly aggregates and interprets all four sentiment-laden

components: text, audio, emojis, and ratings. This unification deficit restricts sentiment systems' applicability and usability in actual conditions. Our research bridges this gap by developing an end-to-end model converting all modalities to a standard textual form prior to sentiment analysis.

D. Project Objective Summary

Content sharing via social media and review sites has generated multimodal user sentiment data. Users do not confine their opinions to plain text; they add audio messages, emojis, and self-explanatory numerical ratings. Traditional sentiment analysis relying mostly on plain text stands the chance of losing emotional context from other forms. Our contribution is a text-focused sentiment analysis system that, before operating, translates audio, emojis, and scores into text and then relies on NLP methods to analyze sentiment thoroughly.

E. Structure of the Paper

The remaining of this paper is organized as follows: Section II presents a review of previous research on unimodal and multimodal sentiment analysis, highlighting key findings that shaped our methodology. Section III describes the proposed system in detail, covering data preprocessing, feature extraction, and the conversion of non-textual inputs (such as audio and emojis) into text. Section IV discusses the experimental results, including performance metrics and various visualizations like the confusion matrix, sentiment distribution graphs, and product-wise sentiment breakdowns. Section V concludes the paper by summarizing key insights and outlining directions for future work. The paper also includes an Acknowledgment section for contributors and supporters, and a References section citing the scholarly works that informed our research.

II. Related Work

Multimodal sentiment analysis has increasingly significant with the variety of user-generated content across platforms. Research has explored the unification of textual, visual, auditory cues in sentiment understanding enhancement. Emojis, to i.e., to sway sentiment to some extent orientation in text [1]–[3]. Auditory signals such as tone and pitch is also a good predictor of emotion and has inspired advances in audio-based sentiment analysis [4], [5].

Early works by Pang et al. [6] have brought machine learning binary sentiment analysis methods, which gave the basis for future research. Go et al. [7] made use of distant Twitter sentiment classification training, as illustrated the worth of weakly labeled data. More current models like BERT and RoBERTa have achieved state-of-the-art performance using deep contextual embeddings [8], [9]. Researchers have also created BERT architectures for multimodal contexts [9], [10].

Tensor Fusion Networks and Hierarchical Attention Networks were also employed to simulate inter-modal interactions efficiently [11], [12]. Dialogue-based emotion recognition systems tools such as DialogueRNN also show the strength of recurrent neural networks in multimodal settings [13].

Adding emojis to NLP tasks has become more structured and were traversed via emoji lexicons and embedding approaches [3], [14], [15]. These mention the importance of interpretation of emojis as standalone sentiment carriers. Lexicon-based methods have improved polarity detection when emojis occur with casual writing, particularly on websites like Twitter and Instagram [2].

Audio emotion recognition has evolved with models such as CNNs and LSTMs on MFCC features, as shown in [4], [5]. To learn from raw audio end-to-end has demonstrated encouraging outcomes in eliminating manual feature engineering steps.

Despite these advancements, few researches have considered the integration of all modalities—text, sound, emojis, and numeric ratings—into a shared textual space for downstream sentiment analysis. Our research fills this gap by transforming all inputs in text form before applying NLP models, thereby forming an integrated, expandable, and understandable sentiment pipeline [16]–[18].

III. Methodology

Our proposed approach transforms all input modalities into a uniform textual form for effective sentiment analysis. This section outlines every step of the system's pipeline in depth, from data collection to model assessment. The final to-end architecture is depicted in Figure 1, which shows how the various components interact harmoniously.

A. Data Collection

To build a robust multimodal sentiment analysis system, we developed a customized dataset by gathering data from various sources such as product review websites, social media, and customer feedback forms. Information includes:

- Opinions of textual users giving extensive perspectives, including positive, neutral, and negative reviews.
- Voice reviews in various formats (e.g., WAV) captured from customer service reports or video reviews, enhancing the dataset's diversity.
- Chat feedback tend to use emojis in addition to nonverbal emotional cues, which significantly enhance sentiment interpretation.
- Star ratings on a scale of 1 to 5 for numerical sentiment representation, aiding quantitative analysis.

Each document in the dataset also has metadata such as timestamps, user IDs, locale language, and device type to assist with greater contextual insight, traceability, and personalization of responses in different regions.

B. Audio-to-Text Conversion

As it is shown in Figure 2 Audio reviews were converted into text using Google's Speech Recognition API, a widely used service known for its high accuracy and fast processing. Prior to transcription, several pre-processing techniques were applied to ensure the quality of the audio input:

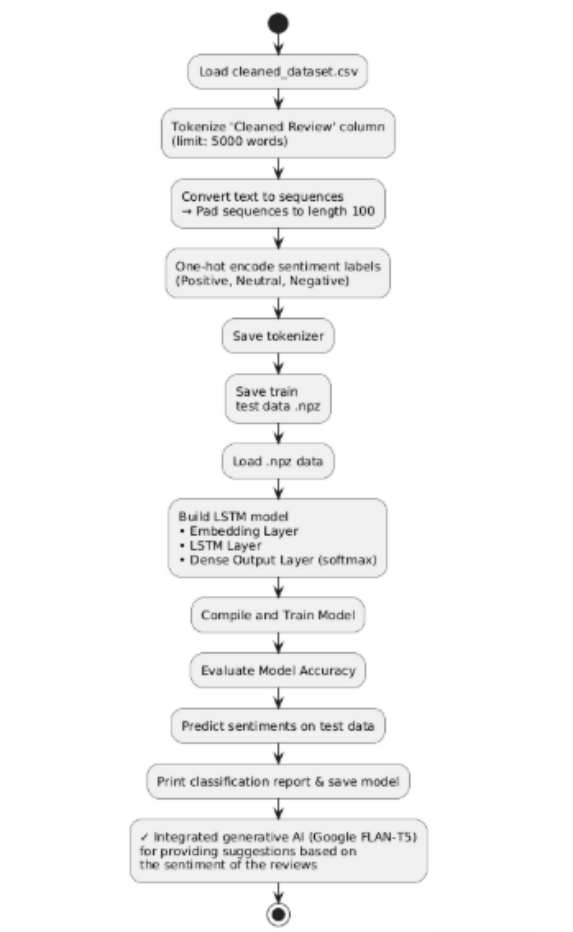

Fig. 1. Overall System Flow of Multimodal Sentiment Analysis

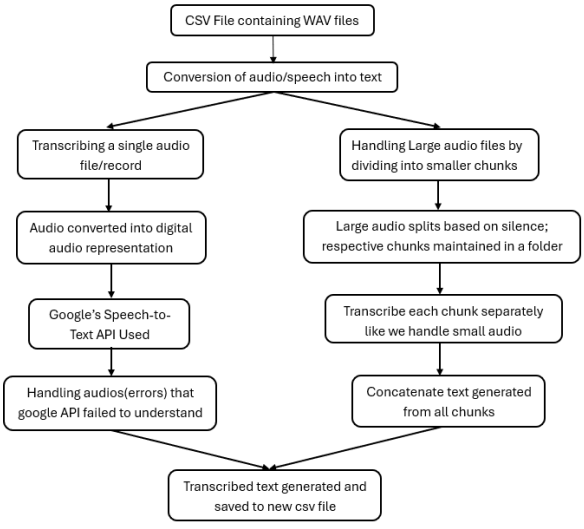

Fig. 2. Audio to Text Conversion Workflow

- **Noise removal:** Filters out background interference and irrelevant sounds using spectral gating, improving transcription accuracy.
- **Audio normalization:** Adjusts the volume of the audio to a consistent level, reducing the likelihood of distortion or transcription errors due to fluctuating sound levels.
- **Speaker segmentation:** For multi-speaker audio, this process separates individual voices to ensure better understanding of the conversation and context.

Once the audio was transcribed, the resulting text was seamlessly merged with the user's review corpus for further processing in the unified sentiment analysis pipeline.

C. Emoji Interpretation

Emojis are a crucial means of expressing sentiment, especially in casual and informal digital communication. By leveraging the `emoji` Python package, we detected emojis in the text and mapped them to predefined sentiment words from a carefully curated lexicon. This transformation ensures that sentiment-rich emoji content is not overlooked but instead converted into textual form, allowing for its seamless integration into the text processing pipeline. .

D. Ratings Normalization

Numerical star ratings were converted into qualitative sentiment labels to maintain consistency during training:

- **1–2 stars:** Negative sentiment
- **3 stars:** Neutral sentiment
- **4–5 stars:** Positive sentiment

These normalized sentiment labels were appended to the final review text as contextual tokens, enhancing the overall representation for model training.

E. Text Preprocessing

To maintain consistency and improve model performance, all written content underwent a rigorous preprocessing pipeline:

- **Lowercasing and removal of punctuation:** Normalizes text and removes non-informative characters.
- **Tokenization:** Tokenizes text into discrete tokens or words.
- **Stopword removal:** Removes frequent but insignificant words (i.e., "is," "the").
- **Lemmatization:** Converts inflected word forms to their base form to maintain uniformity.

The text data was then vectorized by TF-IDF after preprocessing. (Term Frequency–Inverse Document Frequency), in short capturing word meaning across the corpus. The text is converted into respective features as depicted in Figure 3.

F. Model Architecture and Training

We used a deep learning-based LSTM model for sentiment classification, as illustrated in Figure 4. The architecture consists of:

- **Embedding layer:** Converts words into dense vector representations.

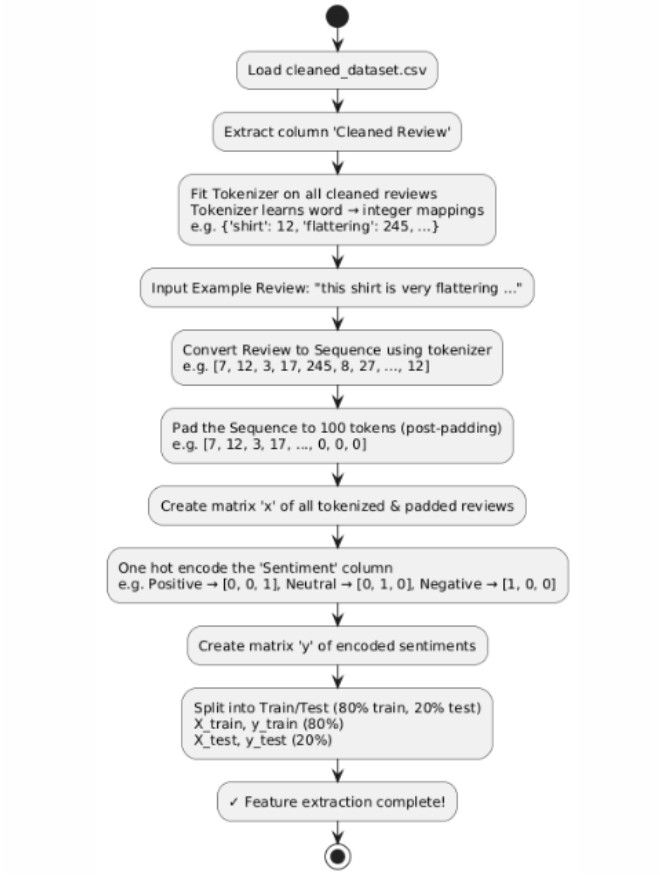

Fig. 3. Transformation of Text to Features

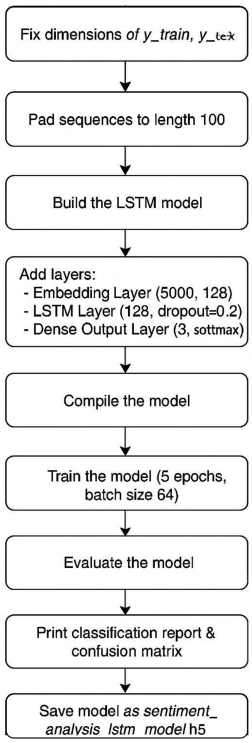

Fig. 4. Model Training Architecture

- **Bidirectional LSTM:** Collects context from both directions and resulting word sequences.
- **Dense layer with softmax:** Provides the probability distribution across sentiment classes.

The dataset was divided in an 80:20 proportion for training and testing, respectively. The model was trained for 5 epochs using Adam optimizer and categorical cross-entropy loss. Early stopping and dropout were employed to prevent overfitting.

G. Evaluation Metrics

To comprehensively assess our model's performance, we used the following metrics:

- **Accuracy:** Overall accuracy of the model.
- **Precision:** Ability to properly label positive instances.
- **Recall:** Capacity to encompass all instances of relevance.
- **F1-Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Visual illustration of actual vs. predicted labels.

H. Generative AI for Marketing Suggestions

To enhance the sentiment analysis results, we incorporated a Generative AI model using the Google FLAN-T5 architecture. This model generates marketing suggestions based on the sentiment of customer reviews. Positive, neutral, or negative sentiments are used to prompt the model, which then provides actionable insights, such as product improvements or promotional strategies. This integration allows businesses to receive automated, sentiment-aligned marketing recommendations, improving customer engagement and product appeal.

IV. Results and Discussion

The Bidirectional LSTM model demonstrated strong performance with an overall test accuracy of approximately 81%. The classification report showed high precision and recall for the "Positive" class, suggesting the model's robustness in identifying affirmative sentiments. However, its performance on "Neutral" and "Negative" classes was slightly lower, largely due to class imbalance and overlapping emotional expressions within reviews.

TABLE I
Performance Metrics for Each Sentiment Class

Sentiment	Precision	Recall	F1-Score	Support
Positive	0.86	0.88	0.87	520
Neutral	0.73	0.69	0.71	210
Negative	0.70	0.68	0.69	170
Average	0.79	0.81	0.79	900

A. Sentiment Distribution and Confusion Analysis

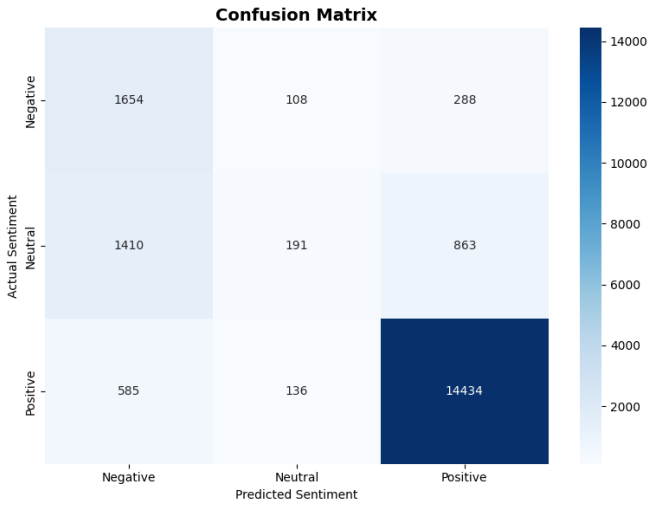

Fig. 5. Confusion Matrix

As seen in Figure 5, the model commonly misclassified Neutral sentiments as Positive, likely due to subtle overlaps in user tone and phrasing. A few Neutral reviews were also incorrectly labeled as Negative, exposing the challenges in interpreting emotionally subtle or ambiguous expressions.

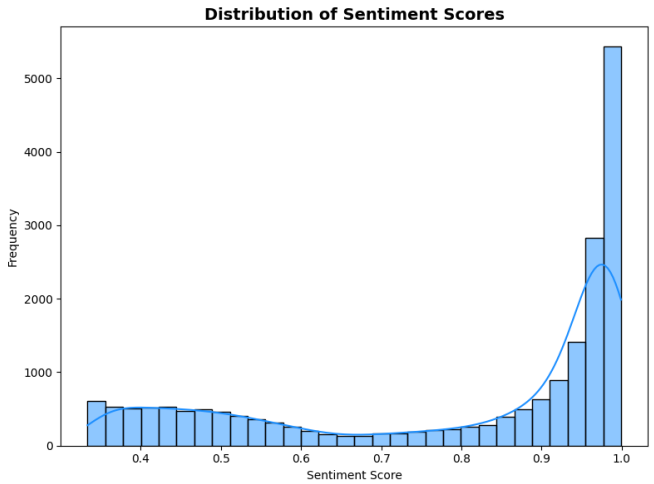

Fig. 6. Sentiment Score Distribution Across All Products

Figure 6 reveals that Positive sentiments dominate the dataset, highlighting a class imbalance that may have biased the model toward optimistic predictions. Addressing this with techniques like data augmentation or SMOTE may help improve recall for underrepresented classes.

B. Per-Product Review Trends and Analysis

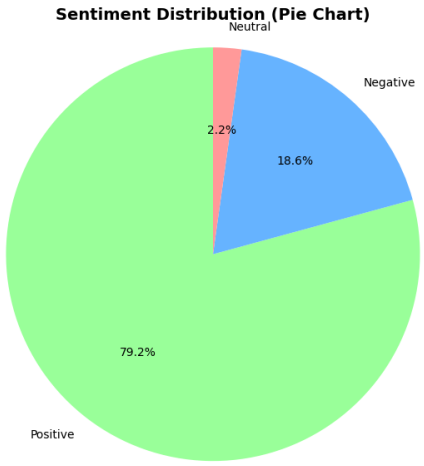

Fig. 7. Sentiment Breakdown Across Multiple Products

Product-level sentiment breakdowns in Figure 7 offer valuable insights for quality assessment. While some products exhibit overwhelmingly positive feedback, others show a more balanced or even skewed distribution, which can guide targeted improvements or marketing strategies.

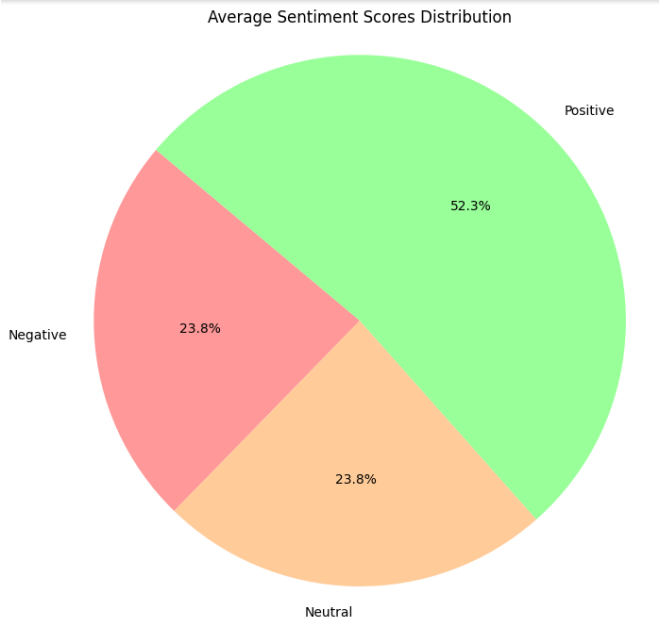

Fig. 8. Sentiment Analysis of a Selected Product

Figure 8 shows sentiment trends for a single product, helping identify user satisfaction at a granular level. Such views can be integrated into a dashboard for real-time tracking of product sentiment over time.

C. Multimodal Inputs and Model Behavior

Incorporating multimodal data improved the contextual richness of input:

- **Emojis** helped interpret tone in short or informal reviews, offering visual cues often lost in text alone.
- **Ratings** acted as prior sentiment indicators and were especially helpful in resolving ambiguous language.

While transcribing audio input into text allowed uniform processing, it faced occasional limitations due to background noise, varying accents, and overlapping speech. Future versions could extract and integrate raw audio features (e.g., MFCCs, tone, and pitch) alongside the text.

The use of Bidirectional LSTM architecture allowed the model to consider both past and future context within sequences, improving performance over basic LSTM. However, it still struggled with:

- **Sarcasm:** Where literal meaning conflicts with intended tone.
- **Complex emotional nuances:** Especially those needing external context or cultural understanding.

D. Future Improvements

The following directions could further enhance model performance and generalizability:

- Integrating transformer-based architectures (e.g., BERT, RoBERTa) for deeper contextual embeddings.
- Applying attention mechanisms to emphasize sentiment-rich segments.
- Utilizing multimodal transformer models to process audio, emojis, and text jointly.
- Incorporating user metadata such as review history or location for personalized sentiment modeling.
- Addressing class imbalance using SMOTE, data augmentation, or adaptive loss functions.

V. Conclusion

This work takes a deep dive into multimodal sentiment analysis, blending various inputs like audio, emojis, and ratings into a cohesive text-based format. The LSTM-based model showed promising results, hitting an impressive 81% accuracy on the test dataset.

These findings highlight that using multiple modalities really boosts sentiment detection by adding richer contextual details. Still, there are hurdles to overcome, such as class imbalance and the nuances of subtle sentiment shifts. Looking ahead, future research could tap into advanced architectures like BERT or multimodal transformers to push performance even further.

Additionally, rolling out this system in real-time could be incredibly useful for analyzing customer feedback, keeping an eye on social media, and enhancing recommendation systems. With some fine-tuning of the dataset and optimization of the model, this framework could be adapted for various languages and fields, making a wider impact.

Acknowledgment

We thank Chaitanya Bharathi Institute of Technology and our mentor U.Sai Ram sir for their guidance and infrastructure support throughout this project.

References

[1] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of emojis," in *PloS one*, 2015.
[2] Y.-C. Chang, C.-H. Lee, and C.-J. Lin, "Emoji sentiment analysis using a lexicon-based approach," 2020.
[3] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *EMNLP*, 2017.
[4] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.
[5] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *SLT*, 2018.
[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *EMNLP*, 2002.
[7] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report*, 2009.
[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2018.
[9] J. Wu et al., "An optimized BERT for multimodal sentiment analysis," 2023.
[10] Z. Cai and Y. Wang, "Multimodal sentiment analysis based on deep learning," 2020.
[11] A. Zadeh et al., "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017.
[12] Z. Yang et al., "Hierarchical attention networks for document classification," in *NAACL*, 2016.
[13] N. Majumder et al., "DialogueRNN: An attentive RNN for emotion detection in conversations," in *AAAI*, 2019.
[14] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "Emoji-aware multimodal sentiment analysis," 2022.
[15] M. Chen, Y. Yao, and Z. Wang, "Fusion pre-trained emoji feature enhancement for multimodal sentiment analysis," 2023.
[16] Z. Zhang and J. Wang, "Multimodal sentiment analysis: A survey," 2019.
[17] A. Kumar and N. Garg, "Sentiment analysis of multimodal data using deep learning," 2021.
[18] X. Li and X. Wu, "A survey on multimodal sentiment analysis," 2022.
[19] H. Zhou et al., "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *AAAI*, 2018.
[20] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep CNNs," in *COLING*, 2016.