WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Django-Based Framework for Multimodal Sentiment Analysis Using Text, Audio, and Image Inputs: Design, Implementation, and Applications

Prof. Prabhu Kichadi[1], Pranita Purushottam Kulkarni2, Shreya Niwas Patil[3], Samruddhi Khot[4], Pratiksha Bhatale[5],

[1] Asst. Professor, Dept of Computer Science & Engineering, VSM's Somashekhar R. Kothiwale Institute of Technology, Nipani, Karnataka, India, 591237.
[2] Dept of Computer Science & Engineering, VSM's Somashekhar R. Kothiwale Institute of Technology, Nipani, Karnataka, India, 591237
[3] Dept of Computer Science & Engineering, VSM's Somashekhar R. Kothiwale Institute of Technology, Nipani, Karnataka, India, 591237
[4] Dept of Computer Science & Engineering, VSM's Somashekhar R. Kothiwale Institute of Technology, Nipani, Karnataka, India, 591237
[5] Dept of Computer Science & Engineering, VSM's Somashekhar R. Kothiwale Institute of Technology, Nipani, Karnataka, India, 591237

A B S T R A C T :

an advanced multimodal sentiment analysis system capable of interpreting human emotions across various data types, including text, audio, images, and document formats. The system integrates several modern technologies—Django for the user interface, SQLite for backend data handling, TensorFlow for deep learning implementation, NLTK for linguistic processing, Libros for audio feature extraction, and OpenCV for image-based emotion recognition. Unlike traditional sentiment analysis models that rely solely on text input, this framework utilizes specialized models for each modality to provide more nuanced and context-aware emotional assessments.

Text analysis leverages state-of-the-art natural language processing methods, including transformer architectures and contextual embeddings, to enhance sentiment classification accuracy. Audio data is analyzed using prosodic and spectral features to capture emotional tone, while images are evaluated through convolutional neural networks to detect facial expressions. For document-based inputs, the system combines optical character recognition (OCR) with sentiment scoring algorithms to extract and assess emotional content from textual data.

The application features an interactive web dashboard that visualizes real-time sentiment outputs for each modality, offering comprehensive emotional insights. It also supports the generation of detailed reports and downloadable summaries for practical use in fields such as customer service, marketing analysis, psychological evaluation, and digital content monitoring. The modular and adaptable design ensures high performance across diverse data environments and makes the system suitable for both research and commercial deployment.

Keywords: Multimodal Sentiment Analysis, Emotion Recognition, Deep Learning, Natural Language Processing (NLP),Audio Sentiment Detection, Image-Based Emotion Detection, Optical Character Recognition (OCR), Django Framework, SQLite Database, Real-Time Sentiment Dashboard, TensorFlow, Libros OpenCV, Transformer Models, Convolutional Neural Networks (CNNs)

1. Introduction

Understanding human emotions has become a crucial aspect of modern digital interactions, especially with the rise of social platforms and multimedia communications. While traditional sentiment analysis systems have predominantly focused on text data—such as tweets, reviews, or blog posts—the growing prevalence of audio, image, and video content demands more comprehensive tools. Addressing this need, this project aims to design a unified platform capable of analyzing sentiments across multiple media types using Django for the web interface and SQLite for lightweight, effective data storage.

Sentiment analysis, a branch of Natural Language Processing (NLP), is evolving to encompass affective computing—a field concerned with recognizing and responding to human emotions. Multimodal sentiment analysis combines various communication formats, including text, speech, facial expressions in images, and even visual documents, to form a more holistic understanding of sentiment. This expansion is vital for deriving deeper emotional insights, which can be particularly valuable in applications like customer feedback systems, psychological health assessments, and intelligent virtual assistants.

The proposed system uses an array of modern tools to handle different data modalities effectively. Text inputs are processed using NLP libraries such as NLTK and deep learning models like transformers for contextual sentiment understanding. Audio inputs are analyzed using Librosa to extract features like pitch and tone, offering insights into emotional states. For image analysis, OpenCV and convolutional neural networks (CNNs) are employed to detect facial cues indicative of specific emotions. Document sentiment is extracted using OCR techniques, followed by text-based sentiment scoring. All these modalities are integrated into a Django-powered interface, with data managed by SQLite to maintain efficiency and portability.

Despite its potential, multimodal sentiment analysis is fraught with challenges. Each media type has unique data properties, noise levels, and interpretation complexities. Emotional expressions are often subtle, culturally influenced, or context-dependent, making accurate classification difficult. Additionally, integrating data from multiple sources raises questions about timing, synchronization, and fusion strategy. Common techniques to address these issues include early fusion (merging raw features), late fusion (combining model outputs), and hybrid fusion strategies that blend both approaches.

As this field matures, future developments will likely focus on building more robust and adaptive systems that can generalize across datasets and scenarios. Innovations may include incorporating additional input types such as physiological signals (e.g., heart rate), expanding domain-specific applications, and enhancing real-time performance. Practical applications are vast—ranging from enhancing customer experience through emotion-aware chatbots to supporting mental health diagnostics via speech and expression analysis. Multimodal sentiment analysis also holds promise in education, entertainment, and marketing, where emotional feedback plays a critical role in user engagement.

In conclusion, developing a multi-modal sentiment analysis system represents a significant step toward more emotionally intelligent computing. By integrating diverse data types and leveraging advanced machine learning and web technologies, such systems can provide deeper, more accurate insights into human emotions. This approach not only enhances technological interaction but also opens new avenues in various domains, making it an essential area of research and development in the coming years.

Nomenclature

NLP (Natural Language Processing): A field of artificial intelligence focused on the interaction between computers and human languages, particularly in understanding and processing text and speech.
Multimodal Sentiment Analysis: The process of detecting emotions or sentiments from multiple data types, such as text, audio, images, and documents, to form a comprehensive emotional understanding.
Affective Computing: A branch of computing that deals with the design of systems capable of recognizing, interpreting, and responding to human emotions.
Django: A high-level Python web framework used for building secure and maintainable web applications.
SQLite: A lightweight, embedded SQL database engine that provides local data storage in a simple and efficient manner.
Transformers: Deep learning models used for understanding the context of words in a sentence, enabling advanced natural language understanding.
NLTK (Natural Language Toolkit): A Python library used for working with human language data for text processing and analysis.
Librosa: A Python package for analysing and extracting features from audio signals, commonly used in music and speech analysis.
OpenCV: An open-source computer vision library used for processing and analysing visual data such as images and videos.
CNN (Convolutional Neural Network): A class of deep learning algorithms widely used for image recognition and classification tasks.
OCR (Optical Character Recognition): Technology that converts different types of documents, such as scanned paper documents or images, into editable and searchable text.
Fusion Strategies: Techniques used to combine data or model outputs from multiple modalities

2. SURVEY AND CONCLUSIONS FROM LITERATURE

The field of sentiment analysis has matured from analyzing plain text to understanding complex emotional cues across audio, images, and video. This evolution supports a more holistic interpretation of human emotion, forming the basis for affective computing systems in real-world applications.

Sentiment analysis has evolved significantly, moving beyond text-based interpretation to include audio and visual data, enabling systems to understand human emotions more holistically. A review of recent literature reveals the breadth and depth of techniques applied in this rapidly expanding field.

Early work on sentiment analysis was primarily focused on textual data, leveraging natural language processing (NLP) techniques. Models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers have demonstrated high accuracy in understanding context and emotion in written content. However, challenges like sarcasm, mixed sentiments, and the computational intensity of deep models remain persistent. Future research aims to produce more lightweight and efficient models, especially for deployment on mobile platforms.

In parallel, sentiment analysis in audio data has gained momentum. Techniques focusing on acoustic features—like pitch, energy, and tone—have shown promise in detecting emotions from speech. CNN and LSTM-based architectures have been particularly effective with sequential data like audio. However, these models often face limitations in training time and in handling multilingual contexts. Expanding to cross-language and real-time recognition remains an active area of development.

Image-based sentiment analysis, particularly through facial expression recognition, has seen notable progress with CNNs applied to static images. These systems can detect nuanced emotional cues but may falter with non-standard or obscured facial expressions. To overcome this, researchers are moving toward dynamic analysis in video streams for better emotional tracking.
Several studies have focused on specific data types—such as long-form documents or user reviews—and have explored domain-specific challenges. Long documents pose difficulties in maintaining context over extended content, while customer reviews often contain sarcasm or mixed opinions that traditional models struggle to capture.

Recent advancements advocate for a multimodal approach, integrating text, audio, and visual cues to enhance sentiment accuracy. Fusion strategies—early, late, and hybrid—are employed to combine data from multiple modalities. While these systems show superior performance, they demand higher computational resources. Thus, the future lies in optimizing these frameworks for real-time applications like emotion-aware virtual assistants, intelligent feedback systems, and interactive media platforms.

In conclusion, the integration of multimodal sentiment analysis marks a significant milestone in affective computing. By combining insights from text, audio, and visual inputs, these systems can achieve a deeper, more nuanced understanding of human emotions. Continued research aims to improve efficiency, cross-domain adaptability, and real-time performance, solidifying sentiment analysis as a cornerstone of human-computer interaction.

Sl. No	Title	Authors	Highlights
1	*A Survey on Sentiment Analysis in Text, Audio, and Visual Modalities*	Liu, B., & Zhang, L.	Overview of methods across multiple modalities.
2	*Deep Learning Approaches to Text Sentiment Analysis*	Socher, R., & Manning, C. D.	Use of CNNs, RNNs, and Transformers for text data.
3	*Audio-Based Emotion Recognition: Techniques and Applications*	Schuller, B., & Zhang, Y.	Acoustic feature extraction for audio-based emotion detection.
4	*Facial Expression Recognition for Emotion Detection in Images*	Zhao, G., & Pietikäinen, M.	CNN-based emotion recognition from images.
5	*Sentiment Analysis of User Reviews Using NLP Techniques*	Pang, B., & Lee, L.	NLP for analyzing customer reviews.
6	*Multi-Modal Sentiment Analysis: Combining Text, Audio, and Video*	Zadeh, A., & Poria, S.	Framework combining multiple data types.
7	*Challenges and Techniques in Document-Based Sentiment Analysis*	Wang, S., & Li, J.	Focus on long-form text challenges.
8	*Transformers in Sentiment Analysis: A Survey*	Devlin, J., & Vaswani, A.	Role of BERT and GPT in contextual emotion detection.
9	*Speech Emotion Recognition Using CNN and LSTM Networks*	Haq, S., & Jackson, P.	Use of sequential deep learning for audio.
10	*Real-Time Sentiment Analysis of Social Media Posts*	Pak, A., & Paroubek, P.	Sentiment detection in tweets and short messages.

3. GAPS IDENTIFIED

Research Gaps and Future Opportunities in Multimodal Sentiment Analysis

Multimodal sentiment analysis continues to evolve as researchers aim to build systems that can understand emotions through text, audio, and visual inputs. Despite notable progress, several critical challenges remain that hinder the full realization of emotionally intelligent systems.

A major issue lies in fusion strategies—current methods struggle to effectively align and integrate inputs from various modalities. Developing adaptive or dynamic fusion approaches could significantly enhance sentiment detection accuracy. In text-based sentiment analysis, many models perform well on short messages but face difficulties with longer documents and nuanced expressions like sarcasm and irony. Advancing models to capture deeper contextual understanding remains a priority.

Audio sentiment analysis also faces limitations in handling linguistic and cultural diversity. Building culturally sensitive models capable of cross-language emotion recognition can help generalize findings. Similarly, image-based emotion detection largely focuses on facial features, often overlooking body language or scene-based emotional cues. Expanding visual analysis beyond facial expressions would offer richer insights.

The computational demands of advanced deep learning models, especially for real-time or mobile applications, restrict their deployment. Research into lighter, optimized architectures is crucial. Another challenge is handling low-quality or noisy data—whether it's distorted audio, blurred images, or informal text. Robust models that tolerate such imperfections are needed for real-world scenarios.

Furthermore, most sentiment systems are designed for English and struggle with other languages due to lack of multilingual support and cultural nuance. There's a need for cross-lingual models that adapt effectively across global audiences. In audio and video sequences, tracking emotion dynamically over time is underdeveloped. Creating models that detect shifting emotions in long content like speeches or vlogs could significantly improve analysis accuracy.

Domain generalization remains a concern—models trained in one field (like social media) often fail in others (like healthcare or finance). Domain-adaptive learning and unsupervised techniques could address this. There's also a need for better classification of mixed and neutral emotions, as current models often oversimplify complex sentiments.

A shortage of benchmark multimodal datasets hinders progress in standardized evaluation. Creating synchronized and annotated datasets across text, audio, and visual inputs would accelerate research. Analyzing long-form documents, too, requires more granular, hierarchical sentiment aggregation methods to manage sentiment shifts.

Real-time systems face latency issues, making them less effective in applications like live monitoring or customer service. Efficient models and edge-computing strategies could enable real-time performance. Integration with human-computer interaction (HCI) is also underdeveloped; adaptive systems that interpret sentiment during live conversations could revolutionize interaction design.

Lastly, there are ethical concerns, including bias and vulnerability to adversarial inputs. Ensuring fairness, transparency, and security in sentiment models is vital for their responsible use in sensitive domains.

4. Objectives

The primary aim of this project is to develop a **multi-modal sentiment analysis system** that can process and interpret emotions from various data types including text, audio, image, and DOCX files. Traditional sentiment analysis systems often rely on a single modality, typically text, which limits their ability to fully capture the emotional state of a user. By combining inputs from multiple sources, the system will be able to provide a more accurate and holistic understanding of sentiment, recognizing that human emotions are expressed in more than just words.

Each media format contributes a unique layer of emotional context. For example, **text data** can reflect sentiment through word choice and syntax, while **audio data** adds depth by capturing tone, pitch, and intonation. **Images**, especially facial expressions, often communicate emotions non-verbally, and **documents like DOCX files** offer more structured and comprehensive content for sentiment assessment. Integrating all these elements allows the system to better identify subtle emotional cues and contextual nuances that single-modality systems may overlook.

To support this integration, the system will incorporate **dedicated preprocessing and analytical pipelines** for each media type. Text will be processed using natural language processing (NLP) techniques, possibly supported by transformer models for contextual understanding. Audio files will undergo feature extraction using tools like Librosa to analyze vocal emotions, while image inputs will be analyzed using computer vision models, including convolutional neural networks (CNNs), to detect facial or visual sentiment indicators. For DOCX documents, optical character recognition (OCR) may be combined with text-based sentiment scoring methods to extract and evaluate sentiment from long-form content.

A significant part of the project involves **fusing the output from different modalities**. Sentiment scores derived from each media type will be combined using suitable fusion techniques—whether through early fusion (merging raw features), late fusion (aggregating individual model outputs), or hybrid strategies. This fusion will help generate a final sentiment decision that takes all input cues into account, offering a more balanced and comprehensive analysis.

Another essential component of the project is the development of a **user-friendly web interface**. Built using Django for backend logic, and optionally supported by front-end frameworks like Bootstrap or Vue.js, the interface will enable users to upload files in various formats and instantly view the sentiment results. Accessibility and ease of use are prioritized to ensure that both technical and non-technical users can benefit from the system without requiring deep expertise in AI or programming.

To manage and persist user inputs, outputs, and history, **SQLite** will be used as the backend database. Chosen for its lightweight and embedded nature, SQLite is particularly suitable for prototyping and small to medium-scale applications. It allows users to revisit previous analyses, generate reports, and manage historical sentiment records efficiently. Django models will be used to represent data structures and streamline database operations through Django's ORM system.

A critical feature of the system is its ability to **display meaningful insights** based on the analysis. Users will not only receive an overall sentiment score but will also be able to see how each modality contributed to the final result. Detailed breakdowns by media type, along with supporting visualizations

such as graphs and pie charts, will enhance interpretability. Libraries such as Chart.js or D3.js will be used to create interactive and visually appealing displays.

Overall, this project aims to create an integrated and intelligent sentiment analysis solution that combines the power of multiple data formats, robust machine learning models, and intuitive user interaction. By enabling richer emotional understanding through multi-modal fusion, this system could be effectively used in applications ranging from customer feedback systems and virtual assistants to psychological assessments and content moderation.

5. Software Requirement Specifications

Sl. No.	Component	Specification	Description
1	Processor	Intel i5 or higher	Provides sufficient speed for data processing and running ML models.
2	RAM	8 GB or more	Ensures smooth multitasking and efficient memory usage.
3	Storage	20 GB or more	Required for storing datasets, model files, and project assets.
4	GPU (Optional)	NVIDIA GPU (Recommended)	Enhances model training performance, especially for deep learning tasks.

Sl. No.	Software/Tool	Purpose
1	Django	Backend web framework for building scalable and structured web apps.
2	Python	Programming language used for scripting, logic, and model integration.
3	SQLite	Lightweight database to manage user inputs and store results.
4	TensorFlow	Library used for building and training machine learning models.
5	OpenCV	Used for processing and analyzing visual (image/video) content.
6	NLTK	Library for text preprocessing and sentiment analysis in NLP.
7	Speech Recognition	Enables conversion of speech to text for audio input processing.
8	HTML, CSS, JS, Bootstrap	Frontend technologies to design a responsive and user-friendly interface.

6. System design Implementation

6.1 Architecture figure

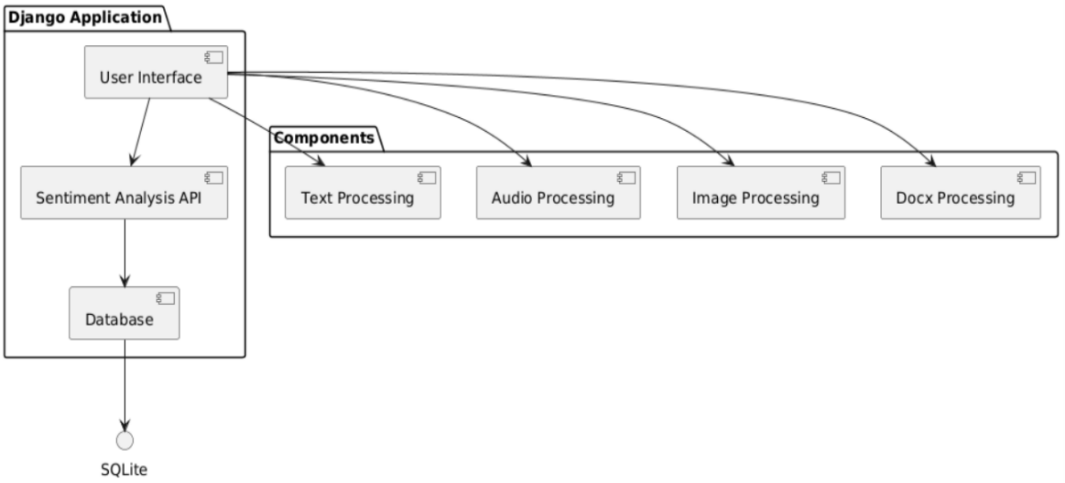


Fig 6.1 Block Diagram

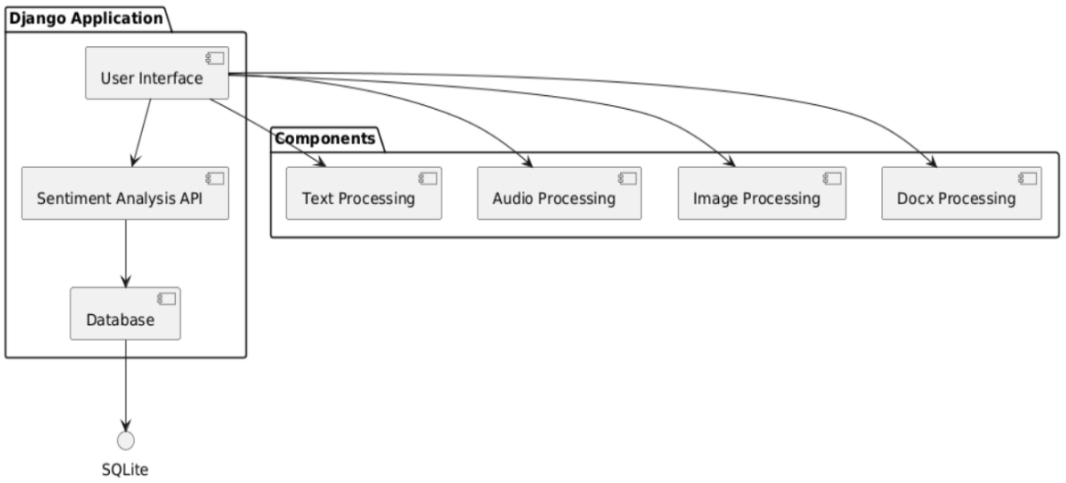
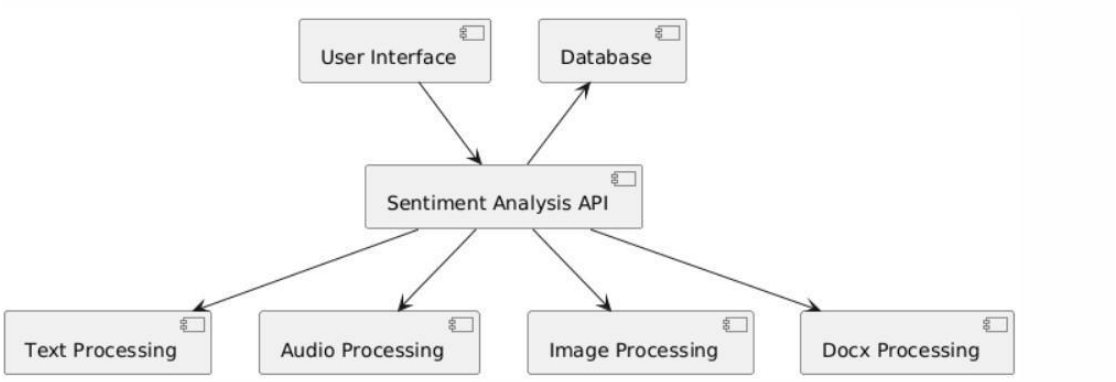
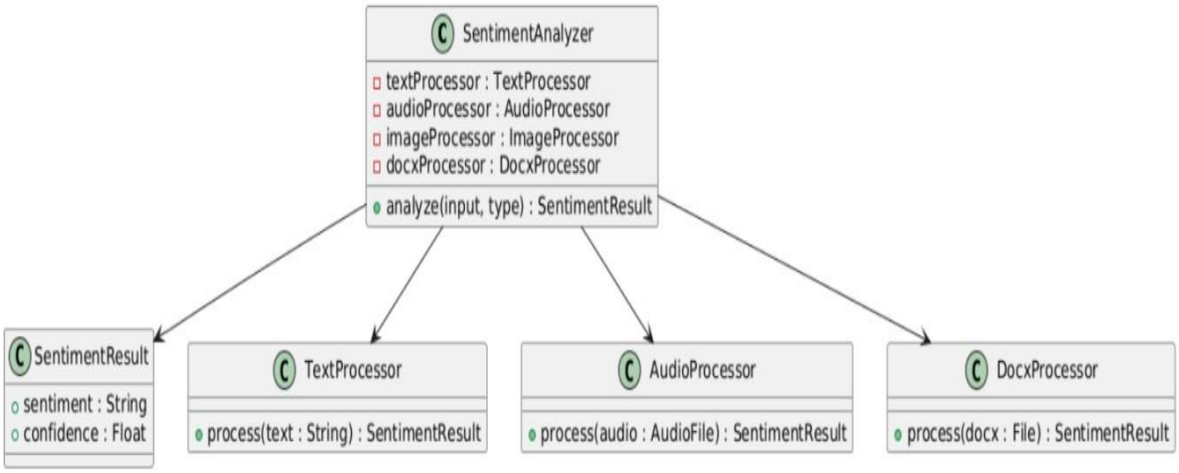


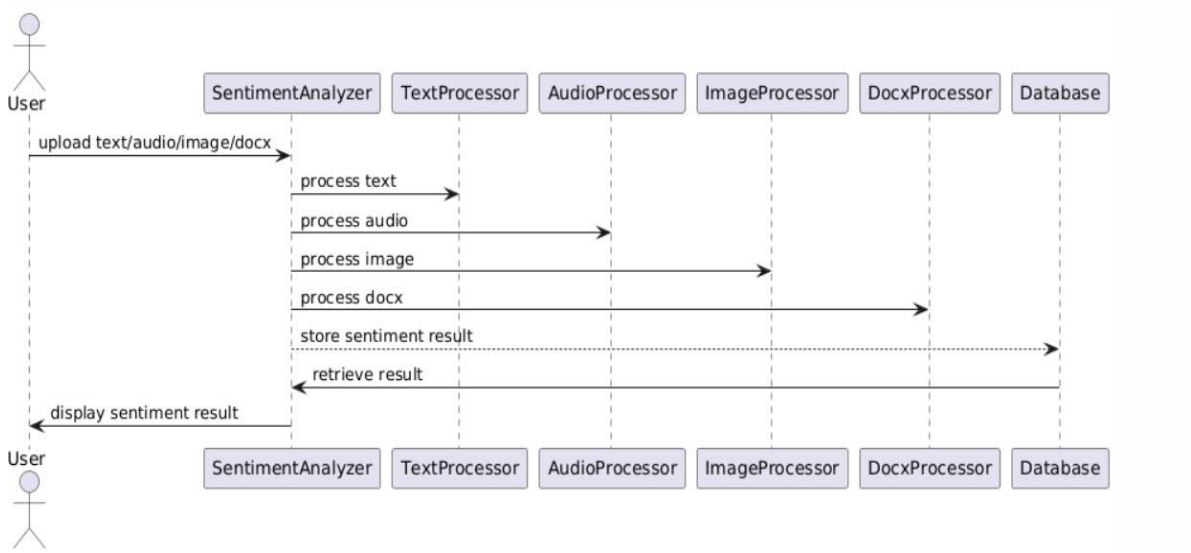
Fig 6.2 Component Diagram



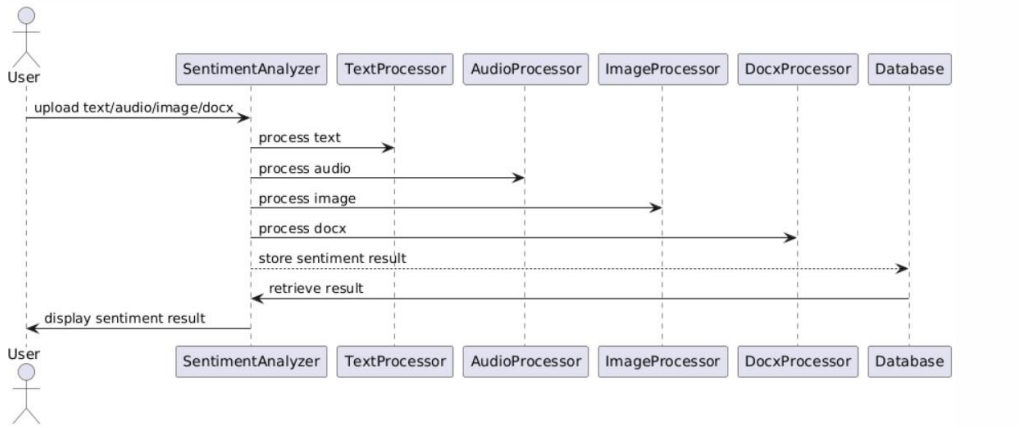
6.3 Class Diagram



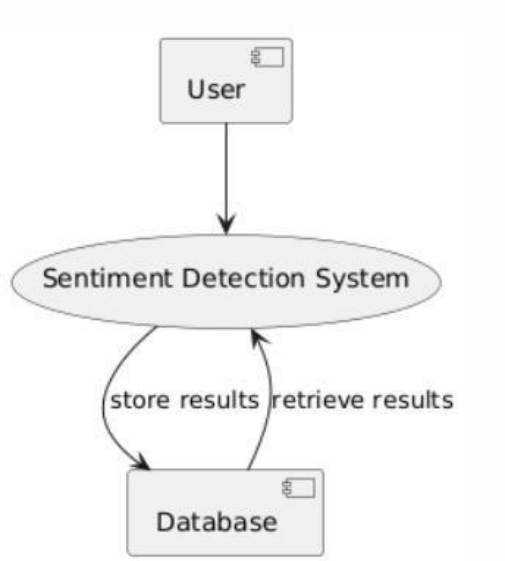
6.4 Sequence Diagram



6.5 Use case diagram



6.6 Data flow diagram



System Design Report: Multi-Modal Sentiment Analysis

1. System Overview

The Multi-Modal Sentiment Analysis System is designed to process and interpret user emotions across various input types, including text, audio, images, and documents. The system is developed using Django as the backend framework and integrates modular processing components for each media type. The goal is to deliver accurate and detailed sentiment results by combining insights from different modalities.

2. Basic System Flow Diagram

This diagram outlines the high-level interaction between the **User**, the **Sentiment Detection System**, and the **Database**. The user interacts directly with the system, which processes the inputs and stores or retrieves sentiment results from the database. This closed-loop ensures persistence and reusability of processed sentiment data.

3. Media-Specific Processing Architecture

This detailed architecture shows how the system handles different types of input. Upon uploading, user data is categorized and routed to the appropriate processor—**Text, Audio, Image, or Docx**. Each module conducts its specialized sentiment analysis, and results are stored for future reference. This modular design allows for scalability and efficient processing.

4. Sequence Diagram

The sequence diagram illustrates the chronological flow of operations:

- The user uploads one or more files (text/audio/image/docx).
- The Sentiment Analyzer delegates processing to the appropriate processor module.
- Each processor extracts sentiment and sends results to the database.
- The system retrieves the sentiment result and displays it back to the user.

This interaction ensures asynchronous, parallel sentiment processing.

5. Class Diagram

The class diagram defines the internal structure:

- The central class Sentiment Analyzer coordinates all modality-specific processors.
- Each processor (Text Processor, Audio Processor, etc.) implements a process() method.
- A Sentiment Result object containing sentiment type and confidence score is returned.
  This object-oriented structure ensures reusability and encapsulation of logic.

6. API-Centric Modular Flow

This architecture places the **Sentiment Analysis API** at the core. It interacts with:

- The **User Interface** for input and output.
- The **Database** for storing and retrieving data.
- The processing components for text, audio, image, and DOCX files.

This loosely-coupled design promotes easy API integration and flexibility.

7. Django Application Integration

This comprehensive diagram illustrates how the Django web application ties together all components:

- The **User Interface** allows file uploads and displays results.
- The **Sentiment Analysis API** connects to the core processing modules.
- Data is managed by **SQLite**, making it suitable for lightweight deployment.
  This full-stack integration ensures that all backend and frontend elements work in sync.

REFERENCES

[1]. "A Survey on Sentiment Analysis in Text, Audio, and Visual Modalities"
Liu, B., & Zhang, L. (2018). A survey on sentiment analysis in text, audio, and visual modalities. ACM Computing Surveys, 50(2), 1-36.

[2]. "Deep Learning Approaches to Text Sentiment Analysis"

Socher, R., & Manning, C. D. (2013). Deep learning approaches to text sentiment analysis. In Proceedings of the 2013 Conference EmpiricalMethods in Natural Language Processing (pp. 1631-1642).

[3]. "Audio-Based Emotion Recognition: Techniques and Applications"
Schuller, B., & Zhang, Y. (2017). Audio-based emotion recognition: Techniques and applications. Springer.

[4]. "Facial Expression Recognition for Emotion Detection in Images"
Zhao, G., & Pietikäinen, M. (2019). Facial expression recognition for emotion detection in images. In Emotion Recognition: A Pattern Analysis Approach (pp. 137-156). Springer.

[5]. "Sentiment Analysis of User Reviews Using NLP Techniques"
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

[6]. "Multi-Modal Sentiment Analysis: Combining Text, Audio, and Video"
Zadeh, A., & Poria, S. (2016). Multi-modal sentiment analysis: Combining text, audio, and video. IEEE Transactions on Affective Computing, 7(2), 131-144.

[7]. "Challenges and Techniques in Document-Based Sentiment Analysis"
Wang, S., & Li, J. (2019). Challenges and techniques in document-based sentiment analysis. In Proceedings of the 2019 Conference Empirical Methods in Natural Language Processing (pp. 3451-3460).
8]. "Transformers in Sentiment Analysis: A Survey"
Devlin, J., & Vaswani, A. (2019). Transformers in sentiment analysis: A survey. IEEE Transactions on Affective Computing, 10(2), 231-244.

[9]. "Speech Emotion Recognition Using CNN and LSTM Networks"
Haq, S., & Jackson, P. (2018). Speech emotion recognition using CNN and LSTM networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2469-2473).

[10]. "Real-Time Sentiment Analysis of Social Media Posts"
Pak, A., & Paroubek, P. (2010). Real-time sentiment analysis of social media posts. In Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM) (pp. 1116- 1121).