National College of Ireland

Sentiment Analysis Using Text and Facial Emotions

MSc Research Project
Artificial Intelligence

Nouman Ali
Student ID: x23239221

School of Computing
National College of Ireland

Supervisor: Anh Duong Trinh

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Nouman Ali
Student ID:	x23239221
Programme:	MSc Artificial Intelligence
Year:	2024
Module:	MSc Research Project
Supervisor:	Anh Duong Trinh
Submission Due Date:	12/12/2024
Project Title:	Sentiment Analysis Using Text and Facial Emotions
Word Count:	9210
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Nouman Ali
Date:	12th December 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	yes
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	yes
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	yes

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis Using Text and Facial Emotions

Nouman Ali
x23239221

Abstract

Sentiment analysis, an important area in natural language processing and computer vision, is focused on the problem of understanding emotions in textual and visual data. Specifically, this report focuses on the multimodal sentiment analysis (MSA) approach based on text and face data for sentiment classification. The global sentiment analysis market is currently at 3.3 billion dollars and expected to grow at 14.8% through 2027 emphasizing its need in fields. This study explores traditional and advanced models: DTs, RFs, RNNs, BiLSTMs have been employed for text analysis using TF-IDF and CNNs with pretrained ResNet50V2 for FER. Findings show that Random Forest yielded the highest accuracy of 71.1% and ResNet50V2 yielded the best prediction accuracy of 60.13% in text classification and facial sentiment detection, respectively. Consequently, the study points at the possibility of enhancing sentiment prediction through the integration of modalities. Further studies need to be conducted to improve results interpretability, as well as expand fusion methods, generative models, and real-time systems.

1 Introduction

Opinion mining or sentiment analysis is an essential subfield of natural language processing and, to some extent, computer vision that focuses on the comprehension and classification of feelings a user expresses through multiple channels. Initially, sentiment analysis was deprived of text data, but recent research has broadened the spectrum of the data types: audio, video, and physiological signals (Poria et al., 2017). From these, text and facial data are powerful modalities because they capture the rich aspects of human emotion including word of mouth and non-verbal signals.

According to the research conducted by Market Research Future, the sentiment analysis market was valued at 3.3 billion in the year 2020, and the study has displayed that the market can grow at a CAGR of 14.8% from 2021 to 2027 across various sectors, including marketing, health care and education. Just the simple analysis of text-based sentiment has seen improvements with the introduction of feature extraction techniques such as TF-IDF (Salton & Buckley, 1988) and other embedding techniques such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and BERT (Devlin et al., 2019). These approaches have promoted an improved appreciation of semantic and contextual interpretations of textual data.

However, complete dependence on textual data has four issues. For example, sarcasm, irony, and other implicit expressions, which may be involved in text messages, are incomprehensible to pure text analysis (Cambria et al., 2017). Facial expression analysis, on the other hand, allows registering patients' affective states by micro facial expressions,

postures, and even gaze direction. As for the facial sentiment analysis, LeCun et al. (1998) and He et al. (2016) have shown that the CNNs-based models can get very high accuracy rates when identifying facial expressions in varying datasets.

MSA addresses the aforementioned issues of unimodal approaches by adopting these modalities. The literature review reveals that approaches based on several data streams perform better than their single-data stream counterparts (Baltrušaitis et al., 2019). For instance, in cases where the direction of textual sentiment cannot be determined clearly, facial expressions may be of great help, and the reverse. This combination has fostered advancements in applications from customer feedback and feedback assessment to Mental Health diagnosis (Morency et al., 2011).

1.1 Motivation

However, even such progress in artificial intelligence and deep learning the sentiment analysis systems are pretty often inaccurate because people's emotions are tricky to predict. Though text-based interventions are helpful, it is difficult to pick up on the tenor of the mood expressed in the claims such as through frowns or low tone voice. On the other hand, the purely image-based models may lose the essence of emotional content by not perceiving the textual content that usually accompanies the emotions.

These gaps can be eliminated by the incorporation of text and facial data in sentiment analysis. For instance, while a customer may write a review about a product and filme while at the same time sharing his facial expressions, there will be better insights on customer satisfaction. However, recent approaches in dealing with factors in multimodal sentiment analysis have been sequential and limited by fundamental issues in feature combination, computation time, and possibly understandable results.

This work also seeks to add to this growing field by proposing a sound multimodal sentiment analysis system. When it comes to facial sentiment analysis, CNN and ResNet50V2 are used, as the models have been successfully implemented in the prior image classification tasks. Text sentiment analysis in the study uses both TF-IDF feature extraction and the decision tree, random forest, RNN and BiLSTM for classification. It also proposes the main idea of using all these models separately but also attempts to combine them to understand how they can work collectively for accurate sentiment prediction.

1.2 Research Question

The following key research questions guide this study:

· How do deep learning models (CNN, ResNet50V2) and machine learning classifiers (Decision Tree, Random Forest, RNN, BiLSTM) perform in multimodal sentiment analysis, integrating facial sentiment detection and text-based sentiment classification using TF-IDF features?

1.3 Research Objectives

The objectives of this study are as follows:

- To design and develop a multimodal sentiment analysis framework that integrates textual and facial data.

- To evaluate the performance of CNN and ResNet50V2 models for facial sentiment analysis.
- To compare the efficacy of decision trees, random forests, RNNs, and BiLSTM classifiers for text sentiment analysis using TF-IDF features.

The present research follows a systematic approach toward multiloaded sentiment analysis. For facial sentiment analysis, resulting preprocessed image data is fed to two distinct models, CNN and ResNet50V2, which have been fine-tuned for sentiment analysis. Text data are preprocessed, and features extracted from them are transformed with the help of the TF-IDF technique. The classification models used here are Decision tree, Random Forest, Recurrent Neural Network, and Bidirectional Long-short Term Memory.

Multimodal sentiment analysis can be implemented in customer feedback systems, social network monitoring, and mental health evaluation systems. In this study, textual and facial data for sentiment classification are used to create an environment that will require higher accuracy and reliability in sentiment classification. The conclusions of the work could be helpful for business people and investigators who attempt to analyse human emotions more broadly.

The remaining report is structured as follows: The related work section discusses the literature review in the field of sentiment analysis for both the text and face images followed by a methodology chapter detailing the steps conducted in the study which is followed by the design chapter putting forward the architecture of the system in detail which is followed by the implementation chapter detailing the steps undertaken during the implementation. This is followed by the results and discussion chapter detailing the results obtained for the implementation of the study. This is followed by the conclusion and future work chapter discussing the findings of the study and providing a detailed step wise guide for the future prospectus of the study.

2 Related Work

Multimodal sentiment analysis has recently received lots of attention as researchers try to close the gap between the textual and the visual data for better sentiment classification. Many prior studies that analyzed text-only or image media fail to grasp the richness of sentiment expressed in actual social media outlets, with more and more people gaining the capability of posting both text and pictures and presenting much more nuanced messages. The several types of mixed-methods approach for combining two resources explored in these studies entail interests in integrating image and textual data for construct such as sentiment analysis attend mechanisms, deep learning frameworks, and feature fusion. This literature review critically duplicates and contrasts these approaches in addressing their advantages, shortcomings, and puts forward to the domain.

2.1 Foundations of Multimodal Sentiment Analysis: Hybrid Models and Sequential Processing

The work on developing superior sentiment analysis, which is based on several modalities, had to start with previous models on extant textual and visual sentiment analysis. Kumar and Garg (2019) developed a preliminary mixture of SentiBank for image evaluation and SentiStrength for textual sentiment analysis with a correct rate of 91.32% from the

Twitter information. In multi-modal OCR, for text extraction this model uses OCR, indicating that text boundary from image is a critical area of focus. Nevertheless, due to its higher accuracy, the simple model of feature fusion did not capture deeper inter-modality interactions, which became the topic of subsequent models aimed at inter-modal dynamics.

Expanding on hybrid methodologies, Tembhurne, & Diwan, (2020) took the next step using Recurrent neural networks (RNNs): LSTM and GRU absorbing timeframe dependence of sequential data. For the multimodal data which are visual, textual and audio, the need to underscore that RNNs are appropriate for sentiment analysis in social media since there is temporal information present. However, their proposed technique was successfully solved to high accuracy and outperformed the baselines while being fairly effective in handling multimodal interaction, even though that approach could not scale up due to the time complexity of large social media collections (Tembhurne & Diwan, 2020).

2.2 Advanced Attention Mechanisms and Modality Selection Techniques

While developing multimodal sentiment analysis, attention mechanisms become crucial to enhancing the sentiment-information from the source modality. In light of this, Gan et al. (2024) introduce the Multimodal Fusion Network (MFN) with multi-head self-attention mechanisms for accuracy gains of no less than 0.11%, 0.13%, and 0.38% on the Twitter, Flickr, and Getty data sets. Due to the fact that visual and textual feature extraction for this model was done separately, it reduced inter-modal noise, creating a paradigm for modality attention; helpful when dealing with noisy data from social media ((Gan et al., 2024).

In addition to this, Yadav & Vishwakarma (2021) proposed the Deep Multi-Level Attentive Network (DMLANet), which incorporates spatial & channel attention mechanism to model heterogeneous inter-modality correlation, outperforming all other models on MVSA-Single, MVSA-Multiple and Flickr datasets tests. Beside, the improvement of accuracy and F1 score of their model's performance in our experiment confirmed the advantage of their model for managing large-scale data obtained from different sources, while the high computational cost become a limitation for real-time analysis. The multi-level attention enabled DMLANet to attend sentimentally important area in both images and text which is quite advanced level compared to previous fusion models like (Yadav & Vishwakarma, 2021).

To address these issues, Al-Tameemi et al. (2023), proposed the DMVAN, where the various levels fine-tune attention mechanisms that attained spectacular 99.8% accuracy on Binary_Getty, 96.9% on Twitter, and 96.2% on the EMO-G dataset. The multi-head attention across the different levels of documents enriched sentiment classification, proving its generalization when dealing with highly structured data. However, as evidenced by Al-Tameemi et al. (2023), high dependence on well-labeled data raised the question of transporting it into the unstructured and noisy portion of the social media data where the concept of document structure is incomprehensible.

2.3 Graph Neural Networks, Cross-Modal Correlations, and Feature Fusion

The subsequent work in the framework of multimodal sentiment analysis incorporated Graph Neural Networks (GNNs) to capture dependencies in multimodal data. Yang et al. (2021) proposed the Multi-channel Graph Neural Network with Sentiment-awareness known as MGNNS that incorporated global dataset attributes concurrent with the multi-head approach, which yielded high success rates for three substantial social media data sets. Ablation study showed that the MGNNS model has ¿20% higher accuracy than baseline models given the fact that GNNs handle the inter-modal dependencies and are particularly helpful when there are strong text-image relations in the datasets. However, this use of GNNs imposed a degree of computational cost which was detrimental for real time domains when implemented on less robust hardware (Yang et al., 2021).

To improve the inter-modality alignment, ITIN was proposed by Zhu et al. (2023) to generate an accurate correspondence between regions of images with related textual stores, and achieve relatively high accuracy rates across numerous social media datasets. The cross-modal alignment of the across-the-board affective regions of the ITIN means that affective region-word relationships were better captured and the models based on concatenation of words were a lot worse. Thus, the successes achieved within the framework of the model indicate the effectiveness of using regional alignment in multimodal sentiment analysis, which, however, still depends on datasets containing distinguishable image-text pairs (Zhu et al., 2023).

Featuring fusion, an equally novel strategy that does not involve GNN-based procedures, was introduced in Lopes et al. (2021) and Kusal et al. (2024). With the help of the same CNN methods as Lopes et al., used for image processing with an additional AutoML selection method, 95.19% accuracy was achieved on the B-T4SA dataset. It is much more flexible and tunable than traditional ways of model selection because it applies AutoML but it is a "black-box approach" (Lopes et al., 2021). Kusal et al. (2024) used a ResNet50 model to handle visuals and T5 model for text where an attention-based feature fusion process captures the positive/ negative sentiment variations of the Instagram dataset at the cost of explainability. The integration of both visual and textual features of the emotions, for example, extremely positive, negative, and neutral really helped in the fine details of flagging the sentiments given it was not so good for model interpretability as it needed (Kusal et al., 2024).

2.4 Interpretability, and Semantic Correlation

The concern with the interpretability of models has emerged as a critical issue in multimodal sentiment analysis since the field's pioneers aim to guarantee that AI delivers comprehensible explanations for its conclusions. Al-Tameemi et al. (2023) added the LIME interpretability framework into DMVAN to help users to track back model decisions to enhance the suitability of DMVAN in applications where transparency is necessary. When LIME integration, It claimed the trade-off of accuracy and explainability because complex DMVAN attained the results of 99.80% along with the hurdle of maintaining the transparent decision making of the.

To enhance the interpretability of the proposed model, Zhang et al. (2023) extended the discussion by introducing metaphor sentiment analysis using synchronised faces and text. As a result of compiling a metaphor-specific dataset with corresponding visual

prompts, this paper has shown that combining facial expression with metaphorical text also provided better accuracy in single-modality models for complex emotional considerations. Nonetheless, their model has the constraint that it only focused on different metaphor-based datasets which reduced the potential of machine-aided analysis of social media containing less context-based metaphorical language (Zhang et al., 2023).

Biswas et al. (2022) gave another definition to interpretability while announcing the inclusion of object attributes within images, augmenting classification accuracy through the utilisation of SVM, CNN and BiLSTM models. By doing so, the researchers found that their object-specific sentiment cutes greatly improved classification thus improving on the datasets that contain structurally based features of objects such as product images or emotive objects. Although useful, the process of applying VISDOM must be limited to more general sentiment contexts, such as social media, which are characterized by brief acceptations rather than a focus on object-specific visual representation (Biswas et al., 2022).

Cross-modal semantic correlation integration was also initially embarked upon by Ke Zhang et al. (2021) whom developed a Cross-Modal Semantic Content Correlation (SCC) model. Since image regions and text semantics were represented by CNN and GloVe embeddings, SCC reached high accuracy on social media datasets mainly when there are impressive text-image correspondence. However, SCC is said to degrade in situations where the text and images have low similarity, which limits its function that is based on text-image relationships (Ke Zhang et al., 2021).

2.5 Conclusion

The previous literature on multimodal sentiment analysis displays a pattern of transition from the early integration of modality fusion approaches based on the early hybrid models to the next level of sophisticated integration, (ii) the attention-driven integration of multiple modalities, (iii) and the graph-aided integration of multiple modalities. As suggested in some of the foundational models like Kumar and Garg (2019) and Tembhurne and Diwan (2020), using visual data provided an essential feature; however, it was integrated very simply and hence could not capture much sentiment. The approach of attention mechanisms as presented by Gan et al. (2024) and Yadav and Vishwakarma (2021) boosted the success of sentiment analysis through the selective enhancement of focus to the sentiment-rich components of input data.

Recent advancements of generative neural networks and feature fusion have taken the field even further with Zhu et al. (2023) and Lopes et al. (2021) enlightened about inter-modal alignment and automated model selection techniques. When it comes to interpretability however, there still is a problem because of the attention and graph structures employed in DMVAN and ITIN. Several studies such Zhang et al. (2023) and Biswas et al. (2022) proposed to improve interpretability of multimodal systems through metaphor analysis and object attribute integration are stimulating to give a clue for future research in this direction.

3 Methodology

In this chapter, the techniques used for performing multimodal sentiment analysis using textual and visual data are described. It emphasizes algorithms that have been developed to perform concept-based extraction of sentiment from text and emotion from images.

These ideas will be employed in both modalities to improve the precision of the sentiment classification task.

3.1 Introduction to Multimodal Sentiment Analysis

Multimodal sentiment analysis in the domain of this thesis is a process of analyzing text and images, or information from various data modalities in an effort to establish a sentiment or the emotional content. This approach is driven by the understanding that verbal communication undertaken by individuals is generally non-verbal, but in part, involves the use of words or images. It is thus possible to gain a better understanding of the overt emotions by analysing text contents and captions with corresponding images (Poria et al., 2017; Baltrušaitis et al., 2019).

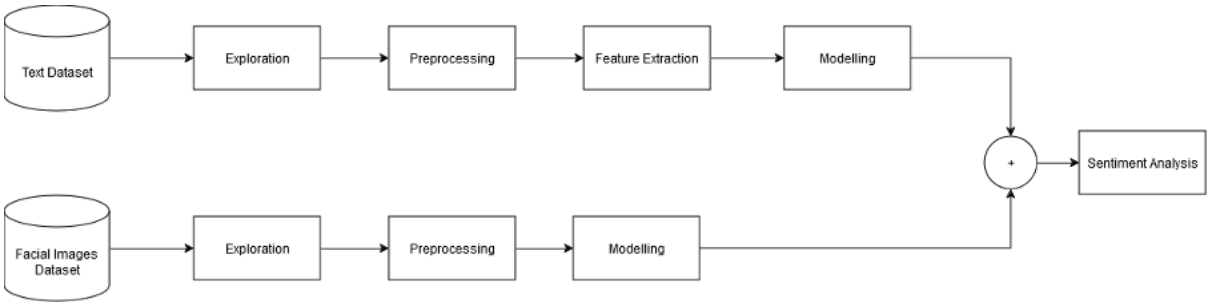

Figure 1: Enter Caption

3.2 Text-Based Sentiment Analysis

3.2.1 Data Preparation and Preprocessing

The textual data consists of social media posts, which often contain informal language, slang, and irregular grammar. Preprocessing steps are essential to clean and normalize the text data for effective analysis. Key preprocessing techniques include:

- **Text Cleaning**: Removal of punctuation, special characters, numbers, and excessive whitespace to reduce noise in the data.
- **Normalization**: Converting text to lowercase and expanding contractions (e.g., "can't" to "cannot") for standardization.
- **Stopword Removal**: Eliminating common words that do not contribute significantly to the sentiment (e.g., "and," "the," "is") using predefined stopword lists (Bird et al., 2009).
- **Stemming and Lemmatization**: Reducing words to their root forms to minimize variations (Porter, 2001). In this context, the Snowball Stemmer is utilized due to its effectiveness and efficiency.

3.2.2 Feature Extraction

Preprocessing through feature extraction makes the textual data compatible with the machine learning models converting textual data into the numbers. The Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique is employed to convert text into feature vectors (Ramos, 2003). TF-IDF provides an advantage of knowing the importance of a particular word in the document while contrasting with the entire collection of documents.

3.2.3 Machine Learning Models

Decision Tree Classifier A Decision Tree Classifier is used to model the text data's sentiment (Breiman et al., 1984). This model splits the data based on feature values, creating a tree-like structure that represents decisions and their possible consequences. It is chosen for its interpretability and ability to handle both numerical and categorical data.

Random Forest Classifier To improve upon the Decision Tree, a Random Forest Classifier is implemented (Breiman, 2001). This ensemble learning method constructs multiple decision trees and merges their outcomes to produce more accurate and stable predictions. Random Forests reduce the risk of overfitting associated with individual decision trees.

3.2.4 Neural Network Models

Recurrent Neural Networks (RNN) Recurrent Neural Networks are employed to capture sequential patterns in the text data (Goodfellow et al., 2016). RNNs maintain a hidden state that can capture information about previous inputs, making them suitable for text data where the order of words matters.

Bidirectional Long Short-Term Memory (Bi-LSTM) Bi-LSTM networks, an extension of RNNs, are utilized to capture dependencies in both forward and backward directions (Hochreiter & Schmidhuber, 1997). This is particularly useful in sentiment analysis, as the context provided by preceding and succeeding words can influence the sentiment of a phrase.

Deep Neural Networks Deep Neural Networks with multiple hidden layers are explored to capture complex patterns in the data (LeCun et al., 2015). These models use non-linear activation functions like ReLU (Rectified Linear Unit) to learn intricate relationships between features and target variables.

3.2.5 Model Evaluation

The models are evaluated using metrics such as accuracy, precision, recall, and F1-score to assess their performance in sentiment classification. Cross-validation techniques are employed to ensure the robustness of the models.

3.3 Image-Based Emotion Recognition

3.3.1 Data Preparation and Preprocessing

The visual data comprises images categorized into seven emotion classes: angry, disgust, fear, happy, neutral, sad, and surprise. Preprocessing steps for the images include:

- **Resizing**: Standardizing images to a consistent size to ensure uniform input to the models.
- **Color Mode Conversion**: Converting images to RGB to align with the expected input format of pre-trained models.
- **Normalization**: Scaling pixel values to a range of [0, 1] to facilitate faster convergence during training.
- **Data Augmentation**: Applying transformations such as rotation, flipping, zooming, and shifting to increase the diversity of the training data and reduce overfitting (Shorten & Khoshgoftaar, 2019).

3.3.2 Modelling

Convolutional Neural Networks (CNN) Convolutional Neural Networks are used due to their proficiency in image recognition tasks (Krizhevsky et al., 2012). The CNN architecture includes:

- **Convolutional Layers**: Extracting features from images using filters that detect patterns such as edges and textures.
- **Pooling Layers**: Reducing the spatial dimensions of the feature maps to decrease computational load and capture dominant features.
- **Fully Connected Layers**: Performing high-level reasoning and classification based on the features extracted.

Transfer Learning with ResNet50V2 To leverage pre-trained knowledge, the ResNet50V2 model is used via transfer learning (He et al., 2016). ResNet50V2 is a deep CNN trained on the ImageNet dataset, capable of extracting rich feature representations from images. By fine-tuning this model on the emotion recognition dataset, the model can achieve high accuracy with less training time and data.

- **Feature Extraction**: The pre-trained layers act as a fixed feature extractor.
- **Fine-Tuning**: Adjusting the weights of the top layers to tailor the model to the specific emotion recognition task.

Model Enhancements To improve model performance, techniques such as:

- **Batch Normalization**: Normalizing the inputs of each layer to stabilize and accelerate training (Ioffe & Szegedy, 2015).
- **Dropout Regularization**: Randomly dropping units during training to prevent overfitting (Srivastava et al., 2014).

are employed in the CNN architectures.

3.3.3 Model Evaluation

Similar to text-based models, image-based models are evaluated using accuracy and loss metrics. Confusion matrices are also analyzed to understand the model's performance across different emotion classes.

3.4 Conclusion

Using both machine learning paradigm for textual data and deep learning approach for image data this working hypothesis presupposes that human sentiment expression is a manifold phenomenon. Specifically, the new models Bi-LSTM for text prediction and ResNet50V2 model for image classification from transfer learning enhances the accuracy of the sentiment analysis system. Decision-level fusion of these modalities improves the system's performance in capturing sophisticated emotional information inherent in multichannel data.

4 Design Specification

While establishing the development of the multimodal sentiment analysis models for this study's analysis, it has been crucial to balance the use of machine learning methods and Deep Learning techniques. The approach involves two primary modalities: textual sentiment analysis and image-based emotion recognition. This section also describes how the models were designed and implemented starting with the methods used for designing, the architectural style, the framework used and how the models were implemented concerning all the requirements.

Figure 2 below depicts the system architecture of the study.

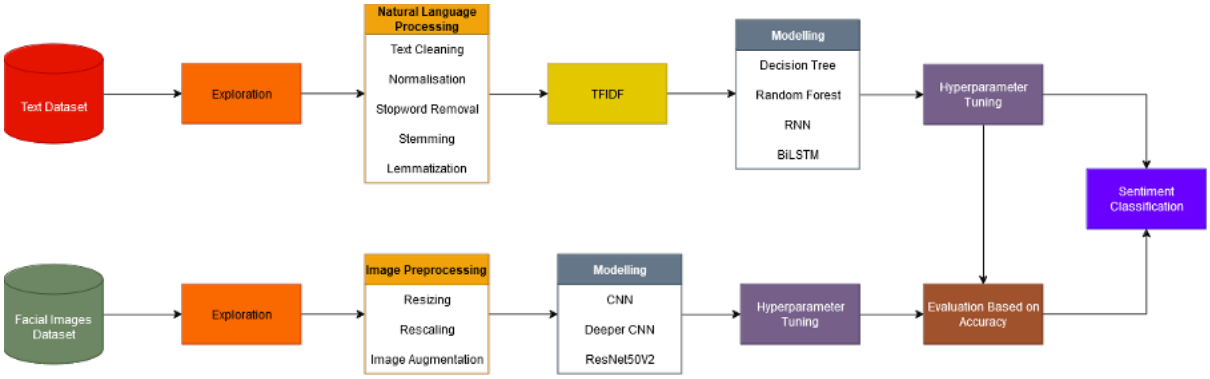

Figure 2: System Architecture

4.1 Text-Based Sentiment Analysis Models

The text-based sentiment analysis is mainly concerned with the task of sorting text data based on sentiment (positive, neutral, and negative). Data pre-processing, feature engineering, and integration of various supervised and unsupervised, traditional and deep learning techniques is part of the design.

4.1.1 Model Architecture

Multiple models were designed and implemented to classify sentiments based on the extracted features.

Various models were created and tested for the purpose of classifying sentiments through the identified features. The traditional approaches employed for classification included DecisionTreeClassifier from the Scikit-Learn package and RandomForestClassifier.Traditional and novel approaches also included TensorFlow with Keras as the neural networks approach.

Decision Tree Classifier Decision Tree Classifier works on a simple concept of dividing the dataset into subsets of features by making a decision based on those features of the data set. We decided on various hyperparameters where the criterion used was Gini impurity measure and the maximum depth used was set to 100 and 10000 respectively to avoid overfitting. To make the experiment reproducible, a random state of 123 was set while performing each state.

Random Forest Classifier Thus, the Random Forest Classifier in the form of the set of decision trees was applied in the analysis, which would allow maximizing prediction and minimizing overfitting. We set the number of trees that made up the ensemble to be 100, 200, 300, 400 and 500. As in the Decision Tree, both 'gini' and 'entropy' splitting criteria, as well as different maximum depths, were tried. One kept a random state of 123 in order to keep it in the same format.

Recurrent Neural Network (RNN) To model sequential data dependencies an RNN was developed as a simple form of recurrent neural network. The architecture of the machine comprised an input layer integral to TF-IDF feature vectors, the dense hidden layer comprises 32 neurons with "tanh" activation. The output layer was built with three neurons with 'softmax' activation function to yield probability in regard to each sentiment class. In an attempt to modify the depth of the network, we implemented a number of layers and tested different activation function including the 'tanh' and the 'relu'.

Bi-LSTM Model To tackle the temporal information, Bi-LSTM networks were used for contextual information from both prior and future token in the text sequences. The architecture was comprised an input layer modified to be compatible with LSTM input dimensions, a Bidirectional LSTM layer with units ranging from 64 to 512 and dense layers with unit 128, 64 using 'relu' activation. The output layer was also applying 'softmax' in order to classify the input into sentiment classes. To reduce overfitting we used techniques like dropout layers etc.

4.1.2 Implementation Details

The given text-based models were written in Python along with Pandas, NumPy, NLTK, Scikit Lern, TensorFlow, and Keras. This dataset was then divided into the training and testing data through a 85-15 ratio in order to assess the effectiveness of the model. ,cross-validation and grid search were employed for hyperparameters tuning to the models parameters. Evaluation criteria were based on the classification performance; accuracy, precision, recall, as well as F1 scores.

4.1.3 Prerequisites

Adequate computational resources were required, ideally with GPU support for training deep learning models efficiently. All necessary Python libraries and frameworks were installed, and the data was preprocessed and cleaned with features and labels ready for model ingestion.

4.2 Image-Based Emotion Recognition Models

The image-based emotion recognition component focuses on classifying images into one of seven emotion categories. The design involves data preprocessing, model architecture, and implementation of convolutional neural networks (CNNs) and transfer learning techniques.

4.2.1 Model Architectures

Two primary types of models were designed for image-based emotion recognition: CNNs and models with transfer learning of ResNet50V2.

Convolutional Neural Networks CNNs were used for fully automated and dynamic identification of spatial pyramids of features in images. First, it could use convolution layers and this layers had different number of filters like 64 filters with 3×3 kernel to extract the features Second, the network could contains pooling layers to minimize the size of the output data to reduce complexity. We used batch normalization following the convolutional layers to enhance and accelerate training and dropout after pooling and dense layers to minimize overfitting. Two dense layers were incorporated to perform highly cognitive reasoning and classification data, and there was the output layer with seven neurons under the 'softmax' activation function.

To test the depth of the network we created stack more convolutional and dense layers, activation function for hidden layers was chosen to be 'relu'. TensorFlow and Keras were used to create the models.

Transfer Learning with ResNet50V2 In order to fine-tune the pre-existing knowledge, the ResNet50V2 was adopted through transfer learning. ResNet50V2 is a deep Convolutional Neural Network that is trained on the ImageNet data set that is able to produce deep feature extraction from images. If this model was further trained on the emotion recognition dataset, it would get a high accuracy with lesser time and less data.

The specific architecture used ResNet50V2 network without the final classification layers to create a feature extractor architecture. Additional top layers incorporated were flatten layer to convert the output of the base model to a 1D vector, dense layers with units say 64 neurons 'relu' activation function to help in feature enhancement, and an output layer with 7 neurons and 'softmax ' activation function. We chose whether to freeze base model layers or let the deeper layers to be fine tuned, Moreover, we adjusted regularization rates such as dropout rates and batch normalization to improve the models' stability.

4.2.2 Implementation Details

The image-based models were implemented using Python with TensorFlow and Keras libraries. Data generators were utilized for efficient loading and preprocessing of image data during training. Training parameters included a batch size of 64 and epochs ranging from 30 to 50, monitoring for convergence. Model performance was assessed using accuracy and loss on the validation dataset.

4.2.3 Prerequisites

High-performance GPU support was recommended due to the computational demands of training deep CNNs and transfer learning models. All necessary libraries were installed, and images were organized into directories by class for compatibility with Keras data generators.

5 Implementation

The idea of the proposed multimodal sentiment analysis involves creating and evaluating machine learning and deep learning models for text based sentiment analysis and image based emotion recognition. To ensure modularity, scalability, and performance optimization, the implementation is divided into two main components: Text Based models for sentiment analysis and Image based models for emotion recognition. The present chapter describes the implementation strategies, the adopted model architectures, the applied hyperparameters, and the utilized tools and technologies.

5.1 Text-Based Sentiment Analysis Models

Preprocessing the data was very useful in sanitizing and offering the textual data in a suitable form to feed the classifiers. Our data cleaning process began with the elimination of null data and duplicate data to make the set free from ambiguity. They used the script and changed the format, all punctuation and special characters were deleted using Regex to decrease noise level. Some words were spelled out to make them standard; for instance, contractions like cannot were spelt as can not. Pure English stopwords were removed using the NLTK word list by only retaining meaningful words. Word preprocessing included stemming to reduce words to their stem using the Snowball Stemmer in order to keep the number of features as low as possible. Further data preprocessing was done to convert this data into numerical form and included performing word count, count of upper case characters and count of special characters on the text entries or text strings.

The text data was cleaned and the cleaned text data was transformed into feature vectors using Term Frequency-Inverse Document Frequency (TF-IDF). The maximum number of features was limited to 10,000 to capture the most important terms only but not much more than this because of large computation and space complexity.

5.1.1 Model Implementation

Several machine learning and deep learning models were implemented to classify sentiments based on the extracted features. The models and their hyperparameter settings are summarized in Table below.

Model	Hyperparameters	Tunning Range
Decision Tree (DT)	criterion='gini'	criterion='gini', 'entropy'
	$\max_d epth = 100$	$\max_d epth = (10 to 10,000)$
	$\text{random}_s tate = 123$	
Random Forest (RF)	$\text{n}_e stimators = 100$	$\text{n}_e stimators = (100 to 500)$
	criterion='gini'	criterion='gini', 'entropy'
	$\max_d epth = 100$	$\max_d epth = (10 to 10,000)$
	$\text{random}_s tate = 123$	
Recurrent Neural Network (RNN)	layers=[32], activation='tanh'	layers=[32 to 512], activation='tanh', 'relu'
	optimizer='adam'	optimizer='adam', 'sgd'
	epochs=5, $\text{batch}_s ize = 32$	epochs=(5 to 50)
Bidirectional LSTM	$\text{LSTM}_u nits = 64$	$\text{LSTM}_u nits = (64 to 512)$
	layers=[128, 64], activation='relu'	layers=[128 to 512]
	optimizer='adam'	optimizer='adam', 'rmsprop'
	epochs=5, $\text{batch}_s ize = 32$	epochs=(5 to 50)

Table 1: Hyperparameter Settings for Text-Based Sentiment Analysis Models

5.1.2 Model Implementation

Decision Tree Classifier: To model the Decision tree, we employed Scikit-Learn's DecisionTreeClassifier. Hyperparameters were set by; Criteria='gini'-The impurity measure to use Maximum depth=100 to avoid over fitting of the algorithm. To make the results reproducible, a random state of 123 was set. This model was developed with the respect of TF-IDF feature vectors and its efficiency is quantitatively expressed with the help of accuracy.

Random Forest Classifier Random Forest Classifier was adopted using Scikit-Learn's RandomForestClassifier. The hyperparameters established were number of estimators= 100, criterion= 'gini', maximum depth= 100. The random state was set to 123 for reproducibility purposes. The ensemble model using DT was developed to be generalized and to minimize over training as opposed to the single tree.

RNN The models of TensorFlow and Keras were used to construct the RNN. The architecture were an input layer that accepts 725 dimensional TF-IDF vectors, followed by a dense hidden layer with 32 neurons that were activated by 'tanh', and the output layer with three neurons for sentiment classes that were activated by 'softmax'. Other hyperparameters used were 'adam', 'categorical_crossentropy', 5, and 32 for epochs and batch size respectively. This model was trained and checked, accuracy and loss with all epochs in the tab.

Bidirectional LSTM The Bi-LSTM was designed and developed using TensorFlow and Keras. The architecture comprised an input layer reformatted to meet LSTM dimensions, a Bidirectional LSTM layer with 64 units, dense layers with 128 and 64 neurons with the activation function 'relu', and the output layer with three neurons having activation function 'softmax'. Hyperparameters chosen were the 'adam' optimizer, chose the 'categorical_crossentropy' loss function, set the number of iterations as 5 and the batch size to be 32. As such, the model aimed to incorporate past and future context within the textual data extracted from the documents.

5.2 Image-Based Emotion Recognition Models

The image-based component focuses on classifying images into one of seven emotion categories. The implementation includes data preprocessing, model training using CNNs, and transfer learning techniques.

Multiple models were developed and their hyperparameter settings are summarized in Table below

Model	Hyperparameters	Tuning Range
Convolutional Neural Network (CNN)	filters=[64], kernel$_size = 3$	filters=[32 to 128], kernel$_size = (2to5)$
	activation='relu'	activation='relu'
	dropout$_rate = 0.25$	dropout$_rate = (0.2to0.5)$
	optimizer='adam', epochs=50	epochs=(30 to 100)
	batch$_size = 64$	
CNN with Deeper Layers	filters=[64], layers=[256, 128, 64]	layers=[32 to 512]
	activation='relu'	
	dropout$_rate = 0.25$	dropout$_rate = (0.2to0.5)$
	optimizer='adam', epochs=50	
	batch$_size = 64$	
ResNet50V2 Transfer Learning	base$_model = ResNet50V2$	
	include$_top = False$	
	custom$_layers = [Flatten, Dense(64)]$	Dense units=[32 to 256]
	activation='relu', dropout$_rate = 0.5$	dropout$_rate = (0.2to0.5)$
	optimizer='adam', epochs=30	epochs=(20 to 50)
	batch$_size = 64$	

Table 2: Hyperparameter Settings for Image-Based Emotion Recognition Models

5.2.1 Model Implementation

CNN In this literature review, we deployed the CNN with the TensorFlow and Keras. In the architecture setup it comprised of a convolution layer that had a filter of 64 with a kernel size of 3X3 using an activation function of 'relu', furthermore it included batch normalization. To decrease spatial dimensions, a MaxPooling layer with 2 as the size was incorporated deep into the model after convolutional layers. We also included a dropout layer which had a dropout rate of 0.25 to minimize over-fitting. The features were flattened and for regression a dense output layer of 7 neurons were used with 'softmax' activation to enable classification. While designing convolutional neural network, hyperparameters used were 'adam' as optimizer, 'categorical_crossentropy' as a loss, no epoch was set to 50 and the batch size was set at 64. The model used augmented data to make it perform better because of increased generalization.

CNN with Deeper Layers We extended the CNN by stacking extra dense layers with 256, 128 and 64 neurons with 'relu' activation function. Batch normalization and dropout were added after these layers since it was good practice for model stability, along with using it to reduce overfitting. The same settings of hyperparameters as in the first CNN model were set.

ResNet50V2 Transfer Learning In this form, we utilised the transfer learning performing a fine tune on the pre-specified ResNet50V2 model. The base model was ResNet50V2, and by using include_top=False the top classification layers were omitted. New layers were created as follows; flatten layer, one dense layer having 64 neurons with 'relu' activation function, a dropout layer with the rate of 0.5 and one output layer having 7 neurons with 'softmax' activation function. The hyperparameters used were the 'adam'

optimizer, 'categorical_crossentropy' loss function, Thirty two epochs and a batch size of 64. The features in the model were then adapted using the emotion dataset to fine tune the model.

5.3 Tools and Technologies

We utilized Python and its rich ecosystem of libraries for machine learning and deep learning.

- **Python**: The primary programming language used due to its simplicity and extensive support for machine learning and data science libraries.
- **Pandas**: Used for data manipulation and preprocessing, facilitating data cleansing, transformation, and analytics.
- **NumPy**: Provided support for numerical computations and array operations, essential for normalization, encoding, and other preprocessing steps.
- **NLTK (Natural Language Toolkit)**: Utilized for text preprocessing tasks like tokenization, stopword removal, and stemming.
- **Scikit-Learn**: Employed for machine learning models like Decision Trees and Random Forests, as well as utilities for data splitting and evaluation.
- **TensorFlow and Keras**: TensorFlow served as the backend engine for deep learning computations, providing GPU acceleration, while Keras provided a high-level API for building and training neural networks.
- **OpenCV**: Used for image processing tasks and handling image data.
- **Matplotlib and Seaborn**: Used for data visualization, plotting accuracy and loss curves, and visualizing data distributions.
- **Computational Environment**

We utilized a high-performance computing environment with GPU support (e.g., NVIDIA GPU) to handle the computational demands of training deep learning models. All libraries were installed using pip or conda package managers to ensure compatibility.

6 Evaluation

This section gives the comparison of several machine learning and deep learning models on text-based and emotion-based sentiment analysis task. The models were then examined on the test dataset in order to analyse the accuracy and loss parameters. The following section gives an overview of the various states of the configurations of the models being used as well as the results of the models.

6.1 Text Based Sentiment Analysis

Table below gives the results obtained for different models in text-based sentiment analysis

Model Configuration	Accuracy(%)	Loss
Decision Tree (Default)	65.4	N/A
Decision Tree (Max Depth=100)	68	N/A
Decision Tree (Max Depth=10000)	65.5	N/A
Random Forest (Default)	71.1	N/A
Random Forest (Max Depth=100, Criterion='Entropy')	69.6	N/A
Random Forest (Max Depth=10000)	70.7	N/A
Neural Network (Dense Layers: 32, 3)	65.3	0.9044
Neural Network (Dense Layers: 256, 320, 128, 3)	66.1	1.6838
Neural Network (Dense Layers: 512, 448, 320, 256, 128, 64, 3)	67.1	1.9264
LSTM (64 units, Dense: 32, 3)	40.3	1.0838
LSTM (256 units, Dense: 128, 64, 3)	40.3	1.0847
LSTM (512 units, Dense: 448, 320, 256, 128, 64, 3)	40.3	1.0876

Table 3: Results for Text-Based Sentiment Analysis

6.1.1 Decision Tree Classifier

1. **Default Configuration**: The Decision Tree classifier achieved an accuracy of **65.4%**. This model used default parameters and served as a baseline for further tuning.

2. **Max Depth = 100**: Increasing the tree depth to 100 improved accuracy to **68.0%**, demonstrating that a deeper tree allows for better capture of data patterns.

3. **Max Depth = 10,000**: Extending the depth to an extremely large value resulted in a marginal improvement with an accuracy of **65.5%**, suggesting diminishing returns for further increases in depth.

6.1.2 Random Forest Classifier

1. **Default Configuration**: The Random Forest classifier outperformed the Decision Tree with an accuracy of **71.1%**, highlighting the benefits of ensemble learning.

2. **Max Depth = 100, Criterion = 'Entropy'**: Restricting the depth to 100 while using the entropy criterion yielded a slightly reduced accuracy of **69.6%**, which might reflect over-pruning of trees.

3. **Max Depth = 10,000, Criterion = 'Gini'**: Increasing the depth to 10,000 resulted in an accuracy of **70.7%**, demonstrating that deeper trees provide marginal benefits when combined in an ensemble.

6.1.3 Recurrent Neural Networks (Dense Layers)

1. **Configuration 1 (32, 3)**: The simplest neural network configuration achieved an accuracy of **65.3%** with a loss of **0.9044**, indicating limited capacity to generalise due to fewer parameters.

2. **Configuration 2 (256, 320, 128, 3)**: Adding additional layers and neurons improved accuracy to **66.1%**, with a loss of **1.6838**, reflecting better learning of text features.

3. **Configuration 3 (512, 448, 320, 256, 128, 64, 3)**: Further increasing the network's complexity achieved the best accuracy among neural network configurations (**67.1%**) at a loss of **1.9264**, suggesting the ability to learn intricate patterns in the data.

6.1.4 Bidirectional Long Short-Term Memory (BiLSTM) Networks

1. **Configuration 1 (64 units, Dense: 32, 3)**: The simplest BiLSTM model yielded an accuracy of **40.3%** with a loss of **1.0838**, indicating limited effectiveness in extracting temporal patterns from text.

2. **Configuration 2 (256 units, Dense: 128, 64, 3)**: Increasing the BiLSTM units and adding more dense layers did not improve the performance, maintaining an accuracy of **40.3%** with a slightly higher loss of **1.0847**.

3. **Configuration 3 (512 units, Dense: 448, 320, 256, 128, 64, 3)**: Despite the model's increased complexity, the accuracy remained stagnant at **40.3%**, with the highest loss of **1.0876**, suggesting overfitting or inefficiencies in learning.

6.1.5 Analysis of Results

The Random Forest classifier emerged as the most effective model for text-based sentiment analysis, achieving the highest accuracy of **71.1%**. This result underscores the strength of ensemble methods in capturing complex patterns. Among neural networks, increasing the depth and complexity led to better performance, but their accuracy was generally lower than Random Forests. The LSTM models, despite being tailored for sequential data, underperformed, likely due to insufficient data preprocessing or suboptimal hyperparameter tuning for the given task.

These findings highlight the importance of model selection and hyperparameter optimisation in sentiment analysis tasks, with ensemble-based approaches demonstrating robustness and superior performance compared to neural networks.

6.2 Emotion-Based Sentiment Analysis

This section presents the results obtained from CNN and transfer learning models for the emotion-based sentiment analysis task. The evaluation metrics are the test accuracy and the test loss for distinct configurations of the model.

Table below enlists the results obtained for the emotion-based sentiment analysis.

Model Configuration	Test Accuracy(%)	Loss Accuracy(%)
Baseline CNN	45.02	1.4671
CNN with Batch Norm & Dropout	49.55	1.3548
Deeper CNN with Additional Dense Layers	50.69	1.2658
ResNet50V2	60.13	1.2546

Table 4: Results for CNN and Transfer Learning-based Models

6.2.1 Convolutional Neural Networks (CNNs)

1. **Baseline CNN Configuration**:
    - **Architecture**: A simple CNN with one convolutional layer (64 filters, kernel size 3x3), max-pooling, flattening, and a dense layer for classification.
    - **Performance**: Achieved a test accuracy of **45.02%** with a test loss of **1.4671**. This configuration represents a basic CNN architecture and serves as a benchmark for further enhancements.
2. **CNN with Batch Normalisation and Dropout**:
    - **Architecture**: The baseline CNN was enhanced with batch normalisation and dropout (rate 0.25) after each layer to prevent overfitting and stabilise learning.
    - **Performance**: Improved accuracy to **49.55%** and reduced test loss to **1.3548**. These enhancements resulted in better generalisation and a moderate increase in performance.
3. **Deeper CNN with Additional Dense Layers**:
    - **Architecture**: The architecture was further extended by adding more dense layers with batch normalisation and dropout. This deeper network introduced multiple layers of non-linearity.
    - **Performance**: The test accuracy increased to **50.69%**, with a further reduction in loss to **1.2658**. The deeper configuration effectively captured more complex patterns in the data.

6.2.2 Transfer Learning Using ResNet50V2

1. **ResNet50V2 Model**:
    - **Architecture**: The ResNet50V2 model, pre-trained on a large dataset, was fine-tuned for emotion-based sentiment analysis. The model was followed by a flattening layer and a dense output layer for classification.
    - **Performance**: Achieved the highest accuracy among all tested models at **60.13%**, with the lowest test loss of **1.2546**. This demonstrates the effectiveness of transfer learning, as pre-trained models are capable of extracting rich features even with limited training data.

6.2.3 Analysis of Results

Analyzing the provided results it is clearly evident that ResNet50V2 model has given better results than any other CNN based configurations. It was also established that transfer learning indeed provided a suitable technique for employing accurately identified features for the emotion-based positive/negative sentiment analysis. Out of the CNN models, modifications such as an increase in the depth of the network with addition of the batch normalisation and dropout layers yielded progressive enhancements to the accuracy as well as the loss.

However, despite moderate performance, they failed to provide results up to the mark set by the pre-trained ResNet50V2. These results support the use of transfer learning for complicated tasks such as recognition of emotions, where features are diverse and data are not always abundant.

From this analysis, one can identify the directions for further investigation – namely integrating the CNN architectures with transfer learning approaches for improved effectiveness of sentiment analysis.

6.3 Discussion

The comparison of various machine learning and deep learning algorithms on text and emotion analyses of sentiment showed different performance patterns. Analyzing the results of the performed text sentiment classification, where features extracted using the Bag of Words approach were used, it can be stated that the highest accuracy of the classifier for this task reached 71.1% with the help of the Random Forest classifier, which confirms the effectiveness of the ensemble approach to the analysis of complex patterns of textual data. Neural networks displayed a consistent improvement with greater depth and complexity, although again their results were not as promising as Random Forests. Notably, the LSTM network with the current study's sequential data achieved a mean accuracy of only 40.3% with all the configurations. Perhaps, this could be the result of inadequate optimisation of the hyperparameters, or inadequate data pre-processing with regard to sequential data.

The best configuration for emotion-based sentiment analysis was identified to be ResNet50V2; with the test accuracy of 60.13% and the least test loss of 1.2546. For that reason, the findings reaffirm the significance of transfer learning in deriving useful features, especially in comprehensi- ve and varie- gated data such as emotions. Experimenting with CNN based models, there is a progressive trend in the results in terms of the probability estimates and the loss function through the combination of such strategies as paying attention to batch normalisation, adding the effect of dropout, and incorporating deeper networks. Nevertheless, an attempt to improve the described CNN models resulted in a scenario where they could not actually perform better than the pre-trained ResNet50V2; this again confirmed the efficiency of transfer learning. These results can imply that simultaneously enhancing the transfer learning concept with the new CNN architecture could be an effective area of research for further study in the literature relating to sentiment analysis tasks, particularly for emotion detection where dataset samples are usually small.

7 Conclusion and Future Work

This work gives an extensive review on multimodal sentiment analysis for both textual data and facial expressions emotion recognition models. The results shown here highlight the achievement of the multimodal approach to address the problem of sentiment analysis. In modeling, Random Forest classifiers were the most accurate for the text-based sentiment analysis a harbinger of big gains for ensemble learning. Generally, the RNN and BiLSTM-based models are promising; however, they can be significantly fine-tuned for higher performance, so these models look for such approaches. In fact, techniques based on deep learning proved to be very effective in image-based sentiment analysis. Out of the different models tested, the ResNet50V2 transfer learning model yielded the best

result with the highest test accuracy of 60.13 % as well as the lowest test loss hence the benefits of using pretrained models to extract subtle features from limited data. Though CNN architectures were seen to progress with improvements in the configurations, none of these were found to outperform the achievements of transfer learning models.

These results underscore the possibility of using conventional approaches that include machine learning alongside deep learning to impose improved sentiment analysis involving multiple modalities. The work also discusses the drawbacks of using single channels to describe sentiment data and the utilsing of multiple channels can enhance the accuracy and interpretability of the model.

7.1 Future Work

Future studies should concentrate on the additional utilization of text and image modalities to take advantage of inter modality relations in a better manner. This involves mainly improving fusion methods, like attention mechanisms or graph-based to improve feature matching from modality to modality. Moreover, expanding to other platforms like audio or physiological signal may give another view of emotions. The greater challenge of making the fine-grained sentiment classification on multiple modalities more interpretable for such models can be pursued as a future work keeping the decision-making process more transparent. Instead of using transfer learning approaches, which are already optimized such as ResNet50V2, other domain tailored fine-tune approaches may be employed.

Another area which can be explored in the future is the creation of real time sentiment analysis systems which are prospective with large Social media datasets. Other issues, like computational cost and the ability to deal with noisy and unstructured data are going to be critical for real-life deployment. Moreover, new perspectives could be opened to using generative models, for instance, transformers, for classification tasks related to sentiment analysis, especially for sets with complex emotional patterns. These advancement then opens up wider scope for the societal implementations of multimodal sentiment analysis in various fields including but not limited to health, marketing and even in monitoring social media platforms for better and more accurate sentiment prediction systems.

References

Al-Tameemi, I.K.S., Feizi-Derakhshi, M.R., Pashazadeh, S. and Asadpour, M., 2023. Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data. *IEEE Access*, *11*, pp.91060-91081. https://doi.org/10.1109/ACCESS.2023.330771

Anundskås, L.H., Afridi, H., Tarekegn, A.N., Yamin, M.M., Ullah, M., Yamin, S. and Cheikh, F.A., 2023, June. Glove-Ing Attention: A Multi-Modal Neural Learning Approach to Image Captioning. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICASS

Atrey, P.K., Hossain, M.A., El Saddik, A. and Kankanhalli, M.S., 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6), pp.345–379. https://doi.org/10.100 010-0182-0

Baltrušaitis, T., Ahuja, C. and Morency, L.P., 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), pp.423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Bird, S., Klein, E. and Loper, E., 2009. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media.

https://thuvienso.dau.edu.vn:88/bitstream/DHKTDN/6460/1/Natural%20Language%20Processing

Biswas, S., Young, K. and Griffith, J., 2022, October. Exploring Multimodal Features for Sentiment Classification of Social Media Data. In *International Conference on Information Technology and Applications* (pp. 527-537). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-8324-7_44

Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5–32. https://doi.org/10.1023/A:1

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group. https://doi.org/10.1201/97813151394

Chen, D., Su, W., Wu, P. and Hua, B., 2023. Joint multimodal sentiment analysis based on information relevance. *Information Processing & Management*, *60*(2), p.103193. https://doi.org/10.1016/j.ipm.2022.103193

Cambria, E., Poria, S., Gelbukh, A. and Thelwall, M., 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6), pp.74-80. https://doi.org/10.1109/MIS.2017.4531228

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.248-255. https://doi.org/10.1109/CVPR.2009.5206848

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp.4171-4186. https://doi.org/10.48550/arXiv.1810.048

Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://doi.org/10.48550/arXiv.1702.08608

Fatimi, S., Sabbar, W. and Bekkhoucha, A., 2023, November. Textual-Visual multimodal sentiment analysis: A Review of Approaches and Challenges. In *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)* (pp. 1-8). IEEE. https://doi.org/10.1109/SITA60746.2023.10373735

Gan, C., Fu, X., Feng, Q., Zhu, Q., Cao, Y. and Zhu, Y., 2024. A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. *Expert Systems with Applications*, *242*, p.122731. https://doi.org/10.1016/j.eswa.2023.122731

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning*. Cambridge, MA: MIT Press. https://doi.org/10.1038/nature14539

Huang, F., Wei, K., Weng, J. and Li, Z., 2020. Attention-based modality-gated networks for image-text sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(3), pp.1-19. https://doi.org/10.1145/3388861

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778. https://doi.org/10.1109/CVPR.2016.90

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 9(8), pp.1735–1780. https://sophieeunajang.wordpress.com/wp-content/uploads/2020/10/lstm.pd

Huang, F., Zhang, X., Zhao, Z., Xu, J. and Li, Z., 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, *167*, pp.26-37. https://doi.org/10.1016/j.knosys.2019.01.019

Ioffe, S. and Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 6–11 July 2015, pp.448–456. https://doi.org/10.48550/arXiv.15

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, vol. 25, pp.1097–1105. https://doi.org/10.1145/3065386

Kumar, A. and Garg, G., 2019. Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications*, *78*, pp.24103-24119. https://doi.org/10.1007/s11042-019-7390-1

Kusal, S., Panchal, P. and Patil, S., 2024, April. Pre-Trained Networks and Feature Fusion for Enhanced Multimodal Sentiment Analysis. In *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)* (pp. 1-7). IEEE. https://doi.org/10.1109/MITADTSoCiCon60330.2024.10574938

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), pp.2278-2324. https://hal.science/hal-03926082v1

Lopes, V., Gaspar, A., Alexandre, L.A. and Cordeiro, J., 2021, July. An AutoML-based approach to multimodal image sentiment analysis. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9

Market Research Future, 2024. Sentiment Analytics Market Research Report Information By Component (Service, Professional Services, Sentiment, Support and Maintenance Services), By Organization Size (Small & Medium Enterprises (SMEs) and Large Enterprises), By Deployment (Cloud and On-Premise), By Vertical (BFSI, Retail, Transportation & Logistics, Education, Media & Entertainment, Healthcare & Life sciences, and Others), and By Region (North America, Europe, Asia-Pacific, and Rest Of The World) – Market Forecast Till 2032

Source: [Online] Available at: https://www.marketresearchfuture.com/reports/sentiment-analytics-market-4304 [Accessed 22 November 2024].

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3

Morency, L.P., Mihalcea, R. and Doshi, P., 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. *Proceedings of the 13th International Conference on Multimodal Interfaces*, pp.169-176. https://doi.org/10.1145/2070481.2070509

Pennington, J., Socher, R. and Manning, C.D., 2014. GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.1532-1543. https://aclanthology.org/D14-1162.pdf

Poria, S., Cambria, E., Bajpai, R. and Hussain, A., 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, pp.98-125. https://doi.org/10.1016/j.inffus.2017.02.003

Porter, M.F., 2001. Snowball: a language for stemming algorithms. *Snowball Publishing*. http://snowball.tartarus.org/texts/introduction.html

Ramos, J., 2003. Using TF-IDF to determine word relevance in document queries. In: *Proceedings of the First Instructional Conference on Machine Learning*, Piscataway, NJ: Rutgers University. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a

Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), pp.513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Shorten, C. and Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), pp.1–48. https://doi.org/10.1186/s40537-019-0197-0

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, pp.1929–1958. https://www.jmlr.org/papers/volume15/srivastava14a/srivastava

Tembhurne, J.V. and Diwan, T., 2021. Sentiment analysis in textual, visual and

multimodal inputs using recurrent neural networks. *Multimedia Tools and Applications*, *80*(5), pp.6871-6910. https://doi.org/10.1007/s11042-020-10037-x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, pp.5998-6008. https://user.phil.hhu.de/čwurm/wp-content/uploads/2020/01/7181-attention-is-all-you-need.pdf

Wang, H., Li, X., Ren, Z. and Yang, D., 2023. Exploring multimodal sentiment analysis via CBAM attention and double-layer BiLSTM architecture. *arXiv preprint arXiv:2303.14708*. https://doi.org/10.48550/arXiv.2303.14708

Yadav, A. and Vishwakarma, D.K., 2023. A deep multi-level attentive network for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(1), pp.1-19. https://doi.org/10.1145/3517139

Yang, X., Feng, S., Zhang, Y. and Wang, D., 2021, August. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 328-339). https://doi.org/10.18653/v1/2021.acl-long.28

Zhang, D., Zhang, M., Guo, T., Peng, C., Saikrishna, V. and Xia, F., 2021, July. In Your Face: Sentiment Analysis of Metaphor with Facial Expressive Features. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533972

Zhang, K., Geng, Y., Zhao, J., Liu, J. and Li, W., 2020. Sentiment analysis of social media via multimodal feature fusion. *Symmetry*, *12*(12), p.2010. https://doi.org/10.3390/sym12122010

Zhang, K.E., Zhu, Y., Zhang, W. and Zhu, Y., 2021. Cross-modal image sentiment analysis via deep correlation of textual semantic. *Knowledge-Based Systems*, *216*, p.106803. https://doi.org/10.1016/j.knosys.2021.106803

Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H. and Qian, J., 2022. Multimodal sentiment analysis with image-text interaction network. *IEEE transactions on multimedia*, *25*, pp.3375-3385. https://doi.org/10.1109/TMM.2022.3160060