Thesis Title: Enhanced Sentiment Analysis in social media Through Optimized Neural Network Fusion of Text and Images Using MVSA-multiple Dataset

Student Name: Manoj Kumar Yalakati

A dissertation submitted in partial fulfilment of the requirements of Technological University Dublin for the degree of

M.Sc. in Computer Science (Data Science)

Date: 02/09/2024

Declarations

I certify that this dissertation which I now submit for examination for the award of MSc in Computer Science (Data Science), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: manoj kumar yalakati

Date: 02/09/2024

Abstract

This thesis explores the enhancement of sentiment analysis in social media by leveraging advanced neural network fusion techniques that integrate textual and visual data. Traditional sentiment analysis methods, primarily focused on textual data, often fail to capture the full spectrum of emotions conveyed through the combination of text and images prevalent on platforms such as Twitter, Instagram, and Facebook. This research seeks to overcome these limitations by developing a unified model that combines Convolutional Neural Networks (CNNs) for image analysis with Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, for text analysis.

The study utilizes the MVSA-multiple dataset, which consists of paired image-text data annotated with sentiment labels, to train and evaluate the proposed fusion model. Traditional methods like Bag-of-Words (BoW) and TF-IDF are limited in their ability to analyse multimodal data, and while CNNs are adept at extracting features from images, they do not account for the sequential nature of text. Similarly, LSTMs capture temporal dependencies in text but overlook the visual elements that contribute to sentiment. The fusion model proposed in this thesis addresses these shortcomings by integrating the strengths of both CNNs and LSTMs, enabling a more comprehensive analysis of multimodal social media content.

A key contribution of this research is the implementation of a late fusion strategy, where features extracted from CNN and LSTM models are combined at a higher level of abstraction. This fusion process is further refined using an attention mechanism, which dynamically weighs the importance of features from each modality, thereby enhancing the model's focus on the most critical aspects of the data. Hyperparameter optimization techniques, including grid search, random search, and Bayesian optimization, are employed to fine-tune the model, ensuring optimal performance.

The evaluation of the proposed model shows that the fusion approach yields superior sentiment prediction accuracy compared to individual CNN and LSTM models, with a validation accuracy of 64.8%, representing a modest improvement of approximately 3%. This suggests that while integrating text and image data does enhance performance, the overall gains are tempered by the inherent complexities of multimodal learning. Challenges such as modality redundancy—where

either text or image alone might suffice for sentiment prediction—and the difficulty of balancing contributions from both modalities limit the effectiveness of the fusion model.

Additionally, the study addresses the issue of class imbalance within the dataset, which is skewed towards positive sentiments. This imbalance complicates the accurate prediction of minority sentiment classes, such as neutral and negative. To counteract this, the research employs class weighting during model training, improving the model's ability to handle underrepresented sentiments more effectively.

Statistical analysis, including paired t-tests, confirms that the improvements offered by the fusion model are statistically significant, though the practical impact remains modest. This finding aligns with existing research in the field, where multimodal fusion models often show incremental gains in accuracy but face challenges in achieving substantial performance enhancements.

In conclusion, this thesis contributes to the advancement of sentiment analysis by demonstrating the potential of neural network fusion techniques to integrate multimodal data more effectively. While the proposed model improves the accuracy of sentiment predictions in social media content, the results indicate that there is considerable scope for further optimization. Future research should explore the integration of additional modalities, such as audio and video, and refine fusion techniques to achieve more significant performance improvements. Moreover, addressing challenges related to class imbalance and model interpretability will be crucial for the continued development of effective multimodal sentiment analysis systems.

Acknowledgements

I would like to extend my deepest gratitude to my supervisor, Bujar Raufi, whose guidance and encouragement were invaluable throughout the course of this dissertation. I am also profoundly thankful to my family and friends for their unwavering support which made this journey possible.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Acronyms:

MVSA - Multimodal Visual Sentiment Analysis

CNN - Convolutional Neural Network

RNN - Recurrent Neural Network

LSTM - Long Short-Term Memory

TF-IDF - Term Frequency-Inverse Document Frequency

BoW - Bag-of-Words

API - Application Programming Interface

GPU - Graphics Processing Unit

EDA - Exploratory Data Analysis

Adam - Adaptive Moment Estimation (Optimizer)

AUC - Area Under the Curve

SVM - Support Vector Machine

MaxEnt - Maximum Entropy

MMLM - Multimodal Masked Language Modelling

AP - Alignment Prediction

GLFN - Global Local Fusion Network

SKEAFN - Sentiment Knowledge Enhanced Attention Fusion Network

SentiWordNet - Sentiment WordNet

PNG - Portable Network Graphic

CHAPTER 1: INTRODUCTION

This thesis aims to enhance the accuracy of sentiment prediction in social media content by overcoming the limitations of traditional sentiment analysis methods. To achieve this, advanced neural network fusion techniques are leveraged. The research employs the MVSA-multiple dataset, which includes paired image and text data, to develop and evaluate a unified model. This model integrates Convolutional Neural Networks (CNNs) for analyzing images and Recurrent Neural Networks (RNNs) for processing textual data. Combining the strengths of both CNNs and RNNs, the proposed approach seeks to provide a more comprehensive and accurate analysis of sentiment in multimodal social media content

1.1 BACKGROUND

Social media platforms like Twitter, Instagram, and Facebook have become significant sources of public opinion and sentiment, generating vast amounts of user-generated content daily. This content, rich in textual and visual elements, offers a valuable resource for understanding public sentiment and trends. Traditional sentiment analysis methods, which focus primarily on textual data using techniques such as bag-of-words, TF-IDF, and classical machine learning algorithms, are limited in their ability to capture the full complexity of human emotions and the multimodal nature of social media content (Pang et al., 2002). These traditional methods often fail to account for the nuances conveyed through images and the interplay between text and visual elements.

Advancements in deep learning have led to the development of more sophisticated models. Convolutional Neural Networks (CNNs) have proven highly effective in image processing tasks, capable of extracting hierarchical features from images that are essential for understanding sentiment conveyed through visual elements such as facial expressions, scene context, and objects within an image (LeCun et al., 1998; Anilkumar et al., 2024). For instance, the work of LeCun et al. established the foundational role of CNNs in visual recognition tasks, while subsequent research has refined these techniques to enhance sentiment analysis capabilities (Huang et al., 2017). Similarly, recent studies have demonstrated the effectiveness of hyperparameter optimization in improving CNN performance for sentiment tasks (Anilkumar et al., 2024).

Concurrently, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have shown great success in handling sequential data like text, capturing temporal dependencies and contextual nuances crucial for sentiment analysis (Hochreiter & Schmidhuber, 1997). The pioneering work of Hochreiter and Schmidhuber on LSTMs provided a solution to the vanishing gradient problem in traditional RNNs, enabling the capture of long-range dependencies in text data.

Despite these advancements, using CNNs and RNNs in isolation does not fully leverage the potential of multimodal data. Traditional methods often concatenate features from different modalities without adequately capturing their interactions, leading to suboptimal performance (Zhang et al., 2021). Recent research has focused on integrating CNNs and RNNs to exploit their complementary strengths, significantly improving sentiment prediction accuracy (Xu et al., 2020). For example, Xu et al. proposed a model combining CNNs for image analysis and RNNs for text analysis, significantly enhancing sentiment prediction by leveraging visual and textual information. Furthermore, attention mechanisms have been developed to focus on the most relevant parts of the input data, thereby improving overall performance (Bahdanau et al., 2015).

The MVSA-multiple dataset, which forms the basis of this research, was collected from Twitter using the Twitter4J API. This dataset includes image-text pairs annotated with sentiment labels. The collection process involved filtering tweets based on a vocabulary of 406 emotional words, covering ten distinct categories of human emotions (Niu et al., 2016). This approach ensures a balanced representation of emotional categories by limiting the number of tweets collected per keyword in each round and performing data collection at different times of the day.

By leveraging advanced neural network fusion techniques and the rich, multimodal data provided by the MVSA-multiple dataset, this research aims to address the limitations of traditional

sentiment analysis methods and improve the accuracy of sentiment prediction in social media content (Mittal et al., 2018; Hu & Yamamura, 2022; Cambria et al., 2018).

1.2 RESEARCH PROJECT/PROBLEM

The core problem addressed in this research is the inherent limitation of using individual neural network models, such as CNNs and RNNs, for sentiment analysis in social media content. Social media posts often combine text and images, and analysing these multimodal data sources separately does not fully capture the richness of the information they provide. CNNs are excellent at extracting features from images, but they miss textual data's contextual and sequential nature. Conversely, RNNs can capture the temporal dynamics in text but fail to leverage visual cues from images.

1.2.1 RESEARCH QUESTION

How can optimizing neural network fusion techniques, specifically by combining Convolutional Neural Networks (CNNs) for image data and Recurrent Neural Networks (RNNs) for text data, improve the accuracy of sentiment prediction in social media content when using the MVSA-multiple dataset?

1.2.2 HYPOTHESIS

- **Alternate Hypothesis (H_1):** Optimizing neural network fusion techniques, specifically by combining Convolutional Neural Networks (CNNs) for image data and Recurrent Neural Networks (RNNs) for text data, significantly improves the accuracy of predicting sentiment class in social media content compared to using either CNN or RNN models individually.
- **Null Hypothesis (H_0):** There is no significant improvement in the accuracy of predicting sentiment class in social media content when using optimized neural network fusion techniques compared to using either CNN or RNN models individually.

Textual Description

This hypothesis seeks to address the limitations of individual CNN and RNN models in capturing the multifaceted nature of sentiment in social media content, which often involves both visual and textual elements. CNNs excel at extracting spatial features from images, while RNNs are proficient at processing sequential text data. However, using these models independently may result in suboptimal performance in multimodal sentiment analysis, where the sentiment is conveyed through a combination of text and images.

To overcome these limitations, the proposed fusion model integrates the outputs of CNNs and RNNs, leveraging their respective strengths to provide a more comprehensive analysis. This fusion is anticipated to enhance the model's capability to accurately predict sentiment, thereby improving overall prediction accuracy.

Performance Metrics

To evaluate the models' performance, the following metrics will be used:

- **Accuracy**: Measures the proportion of correctly predicted sentiment classes across the dataset and will serve as the primary metric.
- **Precision** and **Recall**: Evaluate the quality of positive predictions. Precision measures the proportion of true positive predictions out of all positive predictions, while recall assesses the proportion of true positive predictions out of all actual positives.
- **F1-Score**: Provides a balance between precision and recall, especially useful in scenarios with imbalanced datasets.

Statistical Tests

To rigorously test the hypothesis, paired t-tests will be conducted on the binary correct/incorrect predictions for each model:

- **Paired t-tests** will compare the performance of the fusion model against the individual CNN and RNN models by determining if there is a significant difference in the proportion of correctly predicted sentiment classes.

- The **null hypothesis (H_0)** states that there is no significant difference in correct predictions between the fusion model and the individual models.
- The **alternative hypothesis (H_1)** asserts that the fusion model provides significantly better predictions.

If the p-value obtained from the t-test is less than 0.05, the null hypothesis will be rejected, indicating that the fusion model significantly improves sentiment prediction accuracy over the individual CNN and RNN models.

.

1.4 RESEARCH OBJECTIVES

The overarching goals of this study have been discussed in the previous sub-section. This subsection will elaborate on the steps required for the study in more concrete terms. The goal of verifying that the use of neural network fusion techniques, specifically combining Convolutional Neural Networks (CNNs) for image data and Recurrent Neural Networks (RNNs) for text data, will result in an improvement in the performance metrics can be attained by following the following objectives:

General Objective:

To enhance the accuracy of sentiment analysis in social media content by optimising neural network fusion techniques.

Specific Objectives:

Objective 1: To study and document current techniques in sentiment analysis.

- Conduct a comprehensive literature review to identify state-of-the-art models and techniques in multimodal sentiment analysis.
- Identify gaps in existing research and outline how the proposed study will address these gaps through innovative fusion techniques.

Objective 2: To obtain and prepare datasets for analysis.

- Acquire the MVSA-multiple dataset.
- Preprocess text data by cleaning, normalising, and tokenising.
- Preprocess image data by resizing, normalising, and applying data augmentation techniques.
- Ensure data consistency and handle any missing or noisy data.

Objective 3: To design and implement individual neural network models.

- Implement a CNN model for image sentiment analysis.
- Implement an RNN model for text sentiment analysis.

Objective 4: To develop a fusion model integrating CNN and RNN outputs.

- Implement a strategy to combine features from both CNN and RNN models.
- Develop mechanisms to dynamically control the flow of information from each modality.
- Apply attention mechanisms to focus on the most relevant parts of the input data.

Objective 5: To optimise the hyperparameters of the fusion model.

- Utilise techniques like grid search, random search, and Bayesian optimisation to explore hyperparameter configurations.

Objective 6: To evaluate the performance of the fusion model against baseline models.

- Compare the fusion model's performance with individual CNN and RNN models using accuracy, precision, recall, and F1-score metrics.
- Perform statistical tests to assess the significance of performance improvements.

Objective 7: To analyse and document the study's findings.

- Analyse experimental results to understand the impact of the fusion model.
- Provide insights into how the fusion techniques enhance sentiment analysis.
- Propose future work based on the findings and address any limitations or new research questions.

By following these objectives, the study aims to contribute to the field of multimodal sentiment analysis, leveraging advanced neural network fusion techniques to achieve more accurate and comprehensive sentiment predictions in social media content.

1.4 RESEARCH METHODOLOGY

This section outlines the methodologies employed to achieve the research objectives outlined in this study. The methodologies encompass data acquisition and preprocessing, model design and implementation, neural network fusion techniques, hyperparameter optimization, and model evaluation.

Data Acquisition and Preprocessing

1. Dataset Acquisition:

- The MVSA-multiple dataset, which includes paired image and text data annotated with sentiment labels, will be acquired from its official repository (Niu et al., 2016). This dataset provides a rich source of multimodal data necessary for training and evaluating the proposed models.

2. Text Data Preprocessing:

- Text data will be cleaned, normalised, and tokenized using tools such as nltk and spaCy. Preprocessing steps include removing stop words, lowercasing, and lemmatisation to ensure consistency and reduce noise in the dataset (Araque et al., 2017).

3. Image Data Preprocessing:

- Image data will be resized, normalised, and augmented using TensorFlow's ImageDataGenerator. Data augmentation techniques such as random cropping, rotation, and flipping will be applied to enhance the robustness of the CNN model by providing varied training samples (Huang et al., 2017; Anilkumar et al., 2024).

Model Design and Implementation

- CNN Model for Image Analysis:

- A Convolutional Neural Network (CNN) model will be implemented using a pre-trained DenseNet-121 architecture. The DenseNet-121 model is chosen for its dense connectivity, which facilitates feature reuse and improves gradient flow. The model will be fine-tuned on the MVSA-multiple dataset, incorporating additional dropout layers and a dense layer with softmax activation to adapt it for sentiment classification (Huang et al., 2017).

- RNN Model for Text Analysis:

    - An LSTM network will be designed and implemented for text sentiment analysis. LSTM networks are selected for their ability to capture long-term dependencies and contextual information in sequential data, addressing the vanishing gradient problem common in standard RNNs (Hochreiter & Schmidhuber, 1997). Text sequences will be tokenised and padded to ensure uniform input size for the model (Xu et al., 2020).

Neural Network Fusion Techniques

- Feature-Level Fusion:

    - A feature-level fusion strategy will be employed to merge features extracted from the CNN and RNN models. This approach effectively captures both visual and textual information, providing a comprehensive understanding of the input data (Mittal et al., 2018).

- Attention Mechanisms:

    - Attention mechanisms will be applied to refine the combined features by focusing on the most relevant parts of the input data. These mechanisms enhance the model's ability to prioritize critical information, improving overall performance (Bahdanau et al., 2015).

Hyperparameter Optimization

- Optimization Techniques:

- Hyperparameters will be optimized using grid search, random search, and Bayesian optimization techniques. Libraries such as scikit-learn, hyperopt, and optuna will be utilized to explore various hyperparameter configurations systematically (Anilkumar et al., 2024; Araque et al., 2017).

- Parameter Tuning:

  - The optimization process will focus on identifying optimal configurations for learning rates, batch sizes, dropout rates, and other model parameters. This systematic approach ensures the best possible performance of the fusion model (Gao et al., 2020).

Model Evaluation

4. Baseline Models:

   - Individual CNN and RNN models will be trained and evaluated on the MVSA-multi dataset to establish baseline performance metrics. These metrics will serve as a benchmark to assess the improvements achieved by the fusion model (Zhang et al., 2021).

5. Performance Metrics:

- The fusion model's performance will be compared with baseline models using accuracy, precision, recall, and F1-score. These metrics comprehensively evaluate the model's effectiveness in sentiment analysis (Xu et al., 2020).

6. Statistical Tests:

- Statistical tests, such as paired t-tests, will be performed to assess the significance of performance improvements. This rigorous evaluation ensures that the observed enhancements are statistically significant and not due to random variation (Cambria et al., 2018).

By employing these methodologies, the study aims to develop an advanced neural network fusion model that significantly improves sentiment analysis accuracy in social media content. Integrating CNNs and RNNs, combined with sophisticated data preprocessing and hyperparameter optimisation techniques, will provide a robust framework for multimodal sentiment analysis.

1.5 SCOPE AND LIMITATIONS

This research aims to enhance the accuracy of sentiment analysis in social media content by optimizing neural network fusion techniques, specifically combining Convolutional Neural Networks (CNNs) for image data and Recurrent Neural Networks (RNNs) for text data, using the MVSA-multiple dataset (Niu et al., 2016). The primary focus is on developing and evaluating a unified model that leverages the complementary strengths of CNNs and RNNs to provide a more comprehensive understanding of sentiment in multimodal data. This involves the implementation of advanced fusion strategies, attention mechanisms, and hyperparameter optimization to achieve significant improvements in performance metrics such as accuracy, precision, recall, and F1-score (Hochreiter & Schmidhuber, 1997; Huang et al., 2017; Anilkumar et al., 2024).

The study assumes that the MVSA-multi dataset is representative of typical social media content and sufficiently annotated, that pre-trained models for CNN and RNN architectures can be effectively fine-tuned for sentiment analysis, and that the computational resources available are adequate to handle the training and optimisation processes. However, the research is limited by its reliance on the MVSA-multi dataset, which may not generalise to other datasets or real-world scenarios; the computational complexity and resources required for training deep learning models; and the potential lack of model interpretability (Pang et al., 2002; Zhang et al., 2021; Gao et al., 2020). The study focuses solely on the MVSA-multi dataset without incorporating additional data sources or modalities such as audio or video; emphasises improving accuracy through advanced fusion techniques rather than real-time processing capabilities; and restricts its scope to sentiment analysis without exploring other potential applications of multimodal data integration (Mittal et al., 2018; Xu et al., 2020; Cambria et al., 2018).

1.6 DOCUMENT OUTLINE

This document is organized into five chapters, each contributing to a comprehensive understanding of the research on enhancing sentiment analysis in social media content through neural network fusion techniques.

The ‘Introduction’ chapter establishes the foundation by presenting background information on sentiment analysis in social media, highlighting the limitations of traditional methods in dealing with multimodal data. It introduces the research question and hypothesis, followed by the general and specific objectives of the study. This chapter also outlines the research methodologies and concludes by defining the research scope, assumptions, limitations, and delimitations.

The ‘Literature Review’ chapter examines the existing body of knowledge related to sentiment analysis. It starts with an overview of conventional text-based sentiment analysis techniques and their constraints. The review then focuses on advancements in deep learning, specifically using Convolutional Neural Networks (CNNs) for image analysis and Recurrent Neural Networks (RNNs) for text analysis. It further explores recent developments in multimodal sentiment

analysis, including integrating CNNs and RNNs and applying attention mechanisms. The chapter concludes with a summary of research gaps that this study aims to address.

The 'Experiment Design and Methodology' chapter details the research approach, beginning with data acquisition and preprocessing steps for the MVSA-multi dataset. It describes the design and implementation of individual CNN and RNN models, followed by developing a fusion model that integrates outputs from both networks. The chapter also discusses the hyperparameter optimization process, including the methods and tools used. Finally, it outlines the evaluation metrics and statistical tests employed to measure model performance.

The 'Results, Evaluation, and Discussion' chapter provides a thorough analysis of the experiments conducted and the findings obtained. It includes the performance evaluation of individual CNN and RNN models and the fusion model, using metrics such as accuracy, precision, recall, and F1-score. The chapter discusses the statistical significance of the results. It provides an in-depth analysis of the experimental outcomes, offering insights into the efficacy of the neural network fusion techniques employed.

The 'Conclusion' chapter summarises the research and its key findings, highlighting the contributions made to the field of sentiment analysis. It discusses the study's limitations and suggests future research directions to overcome them and further develop the proposed methodologies. These chapters provide a detailed and structured presentation of the research, its methodologies, findings, and significance in advancing sentiment analysis in social media content.

CHAPTER 2: LITERATURE REVIEW

This literature review is organized into six major sections: traditional sentiment analysis methods, advancements in deep learning for image and text processing, multimodal sentiment

analysis, hybrid approaches, gaps in research, and the motivation behind this study. It references and incorporates findings from research papers, providing a comprehensive and detailed understanding of the field.

2.1 TRADITIONAL SENTIMENT ANALYSIS METHODS

Introduction: Natural language processing (NLP) and machine learning were the foundations of sentiment analysis's early 2000s development. Simple machine learning algorithms, lexicon-based approaches, and statistical models were the mainstays of early methodologies. These conventional methods established the foundation for contemporary advancements, but they also highlighted a number of significant shortcomings that prompted additional study.

2.1.1 BAG-OF-WORDS (BOW) AND TF-IDF

Sentiment analysis was first applied with the Bag-of-Words (BoW) model. It does not care about word order; instead, it converts text into a vector of word occurrences. Pang et al. (2002) achieved notable advancements in sentiment categorization, especially in the movie review area, by combining the BoW model with machine learning methods like Naive Bayes and Support Vector Machines (SVMs).

Word order and context, which can be essential for effectively capturing sentiment, are disregarded by the BoW model, which is its primary drawback. As an example, a BoW method would consider statements that express opposing attitudes, such as "I love this movie" and "I hate this movie," identically (Joachims, 1998; Nigam et al., 1999). Term Frequency-Inverse Document Frequency (TF-IDF) was developed as a solution to this problem. TF-IDF outperforms BoW by giving words weights depending on how frequently they occur in a document as opposed to how frequently they occur throughout a corpus. This highlights more informative words while downplaying common ones (Ramos, 2003).

2.1.2 LEXICON-BASED APPROACHES

Approaches based on lexicons have become popular as a substitute for only statistical techniques. SentiWordNet is one of the most popular resources in this field, having been produced by Esuli and Sebastiani (2006). SentiWordNet facilitates more sophisticated sentiment analysis by

providing sentiment scores to every synset in the WordNet database. However, lexicon-based approaches have drawbacks, particularly with regard to their static character, which makes it difficult for them to adapt to the changing linguistic landscape of social media and other dynamic situations. (Turney, 2002).

2.1.3 RULE-BASED AND HYBRID APPROACHES

Handcrafted rules are applied by rule-based sentiment analysis systems to recognize and categorize sentiment in text. These systems frequently lack the adaptability required to adjust to new or unexpected inputs, although they can be quite useful in some situations where the rules are clearly established (Turney, 2002). To get around these restrictions, hybrid strategies that incorporate machine learning and lexicon-based techniques have been created. For example, by combining statistical learning with domain-specific knowledge, lexicons can improve performance when pre-processing text input before feeding it into a machine learning classifier (Pang et al., 2002).

2.1.4 EARLY MACHINE LEARNING TECHNIQUES

In the early 2000s, machine learning methods started to take center stage in sentiment analysis research. Since their introduction by Cortes and Vapnik (1995), Support Vector Machines (SVMs) have gained popularity as a solution for text classification applications because of their ability to handle high-dimensional feature spaces. The shift from rule-based to data-driven techniques was largely facilitated by SVMs and other early models, such as Naive Bayes (Pang et al., 2002).

When Berger et al. (1996) presented the Maximum Entropy (MaxEnt) model, they offered a probabilistic framework for sentiment analysis that was adaptable and allowed for the inclusion of a large number of features. These models did, however, have certain shortcomings, such as the requirement for a substantial quantity of labeled data and the challenge of interpreting the outcomes (Nigam et al., 1999).

2.1.5 CHALLENGES AND LIMITATIONS OF TRADITIONAL METHODS

Even with their contributions, classic sentiment analysis techniques have a number of serious drawbacks. The main problems are the limited capacity to handle multimodal data, the dependence

on huge, labelled datasets, and the inability to capture the context and word order (Liu et al., 2012). Moreover, rule-based systems are rigid and challenging to expand, and lexicon-based techniques have trouble keeping up with language evolution (Turney, 2002; Esuli & Sebastiani, 2006). These drawbacks made it clear that more sophisticated methods—especially ones based on deep learning—were required to handle the complexity of sentiment analysis in contemporary applications.

2.2 ADVANCEMENTS IN DEEP LEARNING FOR IMAGE AND TEXT PROCESSING

2.2.1 THE EVOLUTION OF CNNS

Convolutional Neural Networks (CNNs) have revolutionized the fields of text processing and picture identification in recent years. CNNs have been adapted for text by considering sentences as sequences of word embeddings, which enables the network to learn spatial hierarchies of features. CNNs were first created for image classification tasks (LeCun et al., 1998) (Kim, 2014). Huang et al. (2017) introduced DenseNets, which further enhanced CNNs by improving the information flow between layers, resulting in better feature reuse and lower parameter needs.

2.2.2 THE ROLE OF RNNS AND LSTMS

Text processing has benefited greatly from the use of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. The vanishing gradient issue in conventional RNNs was addressed by Hochreiter and Schmidhuber (1997) with the introduction of LSTMs, which allowed the network to learn long-term dependencies—a crucial component of sentiment analysis context understanding. By incorporating attention methods, as suggested by Bahdanau et al. (2015), models were able to concentrate on the most pertinent segments of a sequence, leading to notable enhancements in machine translation and sentiment analysis tasks.

2.2.3 CNNS IN IMAGE-BASED SENTIMENT ANALYSIS

CNNs are used for sentiment analysis on more than just text; they can also analyse photos. According to Mittal et al. (2018), CNNs may successfully analyse sentiment in visual data while more conventional text-based techniques might fall short. More accurate sentiment classification resulted from CNNs' capacity to recognize intricate patterns and emotional cues in photos.

2.2.4 HYPERPARAMETER OPTIMIZATION IN DEEP LEARNING MODELS

Optimizing deep learning models for applications requires careful consideration of hyperparameter adjustment. Anilkumar et al. (2024) shown that by adjusting parameters such as learning rates, batch sizes, and layer counts, Bayesian optimization techniques could greatly enhance CNN performance in sentiment analysis. As a result of this optimization, CNNs were more reliable across a variety of datasets, improving both accuracy and generalization.

2.2.5 CHALLENGES AND FUTURE DIRECTIONS

Though sentiment analysis has greatly advanced thanks to deep learning, there are still certain difficulties. Three major obstacles still need to be overcome: the need for huge, labelled datasets; the high processing costs; and the challenge of comprehending complex models (Chen et al., 2020). Future research must address these challenges, especially as sentiment analysis expands to more complex, multimodal data settings.

2.3 MULTIMODAL SENTIMENT ANALYSIS

2.3.1 THE NEED FOR MULTIMODAL ANALYSIS

With the growth of social media and other digital platforms, multimodal sentiment analysis has become more and more crucial. Multimodal techniques combine text, images, audio, and even video in contrast to standard text-based sentiment analysis to offer a more thorough understanding of sentiment. Early attempts at multimodal analysis frequently relied on basic feature concatenation, which produced less-than-ideal findings since it was unable to capture the interactions between various data sources (Pang et al., 2002; Turney, 2002).

2.3.2 ADVANCED FUSION TECHNIQUES

More advanced fusion approaches have been the focus of recent research in an effort to improve multimodal data integration. AOBERT, for instance, was presented by Kim and Park (2023) and achieves state-of-the-art performance on benchmarks such as CMU-MOSEI and UR-FUNNY by processing many modalities simultaneously using a BERT-based model. This model learns joint representations across many modalities using Multimodal Masked Language Modelling (MMLM) and Alignment Prediction (AP) tasks.

2.3.3 GLOBAL AND LOCAL FEATURE FUSION

The Global Local Fusion Network (GLFN) is a noteworthy advancement in multimodal sentiment analysis, as suggested by Hu and Yamamura (2022). This model captures both the general mood and more subtle nuances by fusing local features with global context. Better sentiment predictions result from the addition of attention mechanisms, which further improves the model's capacity to concentrate on the most pertinent data segments.

2.3.4 INCORPORATING EXTERNAL KNOWLEDGE

Zhu et al. (2023) presented the Sentiment Knowledge Enhanced Attention Fusion Network (SKEAFN), which incorporates external sentiment knowledge to enhance multimodal analysis. The text modality of this model is enhanced by a sentiment knowledge network, which greatly increases the model's capacity to handle intricate and nuanced input. Sentiment categorization becomes more accurate when external knowledge is integrated, especially when there is scant or confusing data.

2.3.5 THE ROLE OF SOCIAL CONTEXT IN MULTIMODAL ANALYSIS

Xu et al. (2020) highlighted the significance of context in comprehending sentiment by proposing a multi-attention network that integrates social relations into the study. Taking into account the connections between individuals and their social activities helps the algorithm detect sentiment more accurately, even in complicated visual data. This method shows how social context may be included into multimodal sentiment analysis to provide a more comprehensive understanding of sentiment.

2.4 HYBRID APPROACHES IN SENTIMENT ANALYSIS

2.4.1 COMBINING LEXICON-BASED AND MACHINE LEARNING METHODS

In sentiment analysis, hybrid approaches frequently combine more recent machine learning techniques with more established lexicon-based methodologies. For example, SentiWordNet was integrated with machine learning classifiers by Esuli and Sebastiani (2006) to enhance sentiment categorization, especially in situations when the amount of labeled data is restricted. Using machine learning to handle the classification problem and lexicons to pre-process data, this hybrid solution combines the best features of both approaches.

2.4.2 ENSEMBLE METHODS

Araque et al. (2017) examined ensemble approaches, which have grown in popularity as a hybrid strategy in sentiment research. These methods can capture multiple facets of sentiment by merging disparate models—like CNNs and RNNs—into an ensemble, which improves accuracy. In social media situations, where data can be highly diverse and unstructured, ensemble models are especially useful.

2.4.3 ACTIVE LEARNING COMBINED WITH DEEP LEARNING

When coupled with deep learning models, active learning provides a potent hybrid method for sentiment analysis. Chen et al. (2020) showed how, by carefully choosing the most instructive samples for labelling and training, active learning might lessen the requirement for large amounts of labelled data. In real-world situations where labelled data is expensive or rare, this method can retain excellent model accuracy while drastically lowering data labelling costs.

2.4.4 EFFICIENCY AND SCALABILITY IN HYBRID MODELS

In order to reduce computing complexity without sacrificing accuracy, Liu et al. (2018) suggested a low-rank multimodal fusion method that breaks down the fusion process into modality-specific elements. This method is especially useful in situations when there is a need for great precision but limited computational resources. Large-scale, real-time sentiment analysis applications require hybrid models that combine scalability and efficiency.

2.4.5 CHALLENGES AND FUTURE DIRECTIONS FOR HYBRID APPROACHES

Although hybrid approaches have many benefits, they also carry over the drawbacks of the separate techniques they combine. These include the requirement for substantial data sets, the intricacy of interpreting models, and the challenge of fine-tuning models that include several different approaches (Araque et al., 2017). Subsequent investigations should concentrate on creating more simplified and comprehensible hybrid models that can function well in a variety of dynamic settings.

2.5 IDENTIFIED RESEARCH GAPS

Despite the substantial progress in sentiment analysis, several critical research gaps remain, which this thesis aims to address.

2.5.1 MULTIMODAL DATA INTEGRATION

The prevailing techniques for multimodal sentiment analysis frequently depend on basic feature concatenation, which is inadequate for capturing the intricate interplay among modalities (Gao et al., 2020; Hu & Yamamura, 2022). Sentiment analysis performance requires advanced fusion techniques that can effectively merge textual and visual data.

2.5.2 DATASET LIMITATIONS

Many studies rely on outdated or unimodal datasets, limiting the generalizability of their findings (Brescia et al., 2015; Franco-Arcega et al., 2013). The use of more comprehensive datasets, such as those including paired image and text data, is essential for developing models that perform well across different domains.

2.5.3 HANDLING NOISY AND IMBALANCED DATA

Sentiment analysis models have difficulties since social media data is frequently unbalanced and noisy (Chen et al., 2020). To improve data quality and model robustness, robust preprocessing methods are required, such as data augmentation and noise reduction tactics.

2.5.4 IMPROVING COMPUTATIONAL EFFICIENCY

Deep learning models' high computing costs are a major obstacle to their practical use, especially in real-time situations (Anilkumar et al., 2024). To increase the efficiency of these models, methods including model pruning, quantization, and hyperparameter optimization must be investigated.

2.5.5 ENHANCED FEATURE EXTRACTION AND FUSION

The combined qualities of textual and visual data are frequently underutilized by current feature extraction and fusion techniques (Zhu et al., 2023). improved feature extraction and integration will lead to improved sentiment analysis results, so new architectures integrating CNNs, RNNs, and attention methods are required.

2.5.6 GENERALIZABILITY ACROSS DOMAINS

One of the ongoing challenges in sentiment analysis models is their capacity to generalize across many platforms and domains (Zhu et al., 2023). Creating models that work well in a variety of scenarios requires incorporating a wide range of datasets and rigorous validation procedures.

2.6 SUMMARY

The literature review has highlighted several key developments in sentiment analysis, from traditional methods to the latest advances in deep learning and multimodal fusion techniques. Traditional methods, while foundational, were limited in their ability to handle the complex and multimodal nature of social media data. The introduction of deep learning models, particularly CNNs and RNNs, marked a significant improvement in sentiment analysis, enabling more nuanced and accurate predictions.

However, the independent application of these models has not been sufficient to fully capture the multifaceted nature of sentiment in social media. The integration of CNNs and RNNs through fusion techniques has shown promise, particularly in recent studies that have explored more sophisticated fusion methods. Nevertheless, the literature reveals a lack of consensus on the most effective approaches, with some studies reporting only marginal improvements and others highlighting significant gains.

This discrepancy suggests that while fusion models represent a promising direction for future research, there is still a need for further optimization and exploration of these techniques. Specifically, more work is needed to reconcile the differing results and identify the conditions under which fusion models are most effective. Additionally, the potential of hybrid approaches, such as active learning and ensemble methods, remains underexplored in the context of multimodal sentiment analysis.

In conclusion, this literature review has identified a clear gap in the current research: the need for optimized neural network fusion techniques that can effectively integrate CNNs and RNNs to improve sentiment prediction accuracy in social media content. The research question addressed in this dissertation is directly aligned with this gap, and the proposed solution aims to advance the state-of-the-art in multimodal sentiment analysis.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 INTRODUCTION

The methodology chapter lays out the steps taken to accomplish the research objectives and is the basis for the entire study. This chapter describes the methodical methodology used to tackle the research challenge, covering everything from the preparation and gathering of data to the creation and assessment of models. There includes a thorough discussion of the experiment's conception and execution, the particular techniques employed, and the reasoning behind these decisions. The ethical issues that governed the research are also highlighted in this chapter, guaranteeing that all practices were carried out in an ethical and responsible manner. This chapter guarantees that the research can be independently validated and duplicated by researchers by offering a thorough account of the methodologies utilized.

3.2 RESEARCH DESIGN

Using the MVSA-multi dataset, the research was intended to be an empirical study with an emphasis on the creation and assessment of a multimodal sentiment analysis model. The study's goal was to improve sentiment analysis's accuracy by combining textual and visual data using cutting-edge deep learning algorithms. The main study hypothesis was that the combination of

Recurrent Neural Networks (RNNs) for text data and Convolutional Neural Networks (CNNs) for image data would improve sentiment prediction accuracy over separate models. A systematic research design that included data collecting, preprocessing, model construction, and evaluation was used to investigate this hypothesis. The reliability and validity of the research findings were ensured by the careful planning and execution of each phase of the study.

3.3 DATA COLLECTION AND PREPROCESSING

The MVSA-multi dataset, which was first presented by Niu et al. (2016), was chosen for this study because of its manual sentiment annotations and extensive multimodal content. To provide a balanced representation of sentiments, the dataset comprises of image-text pairs that were gathered from Twitter and filtered using a vocabulary of 406 emotive terms. In order to prevent temporal biases, tweets were collected at various times of the day using the Twitter4J API. Following collection, the data was put through a thorough preprocessing pipeline designed to handle both text and image data Niu et al. (2016).

Preprocessing entailed a number of procedures to clean and standardize text data. Tokenization, lowercase, and punctuation removal were performed on the text by utilizing Python tools like nltk and spaCy. By reducing words to their most basic forms by lemmatization, the dimensionality of the data was decreased. In order to get rid of common words that don't really add much to sentiment analysis, stop words were deleted. To guarantee consistent input sizes for the RNN model, the processed text was then tokenized and padded.

TensorFlow's ImageDataGenerator was used to preprocess image data, making data augmentation, normalization, and scaling easier. The images were scaled to 224 by 224 pixels in order to comply with the CNN model's input specifications. Normalization is necessary to achieve faster convergence during training since it scales the pixel values to a range of [0, 1]. By adding rotations, zooms, and flips to the image data, data augmentation improved the model's ability to generalize to new data. The pipeline for preprocessing made sure that the image and text data were in the best possible format for training the model.

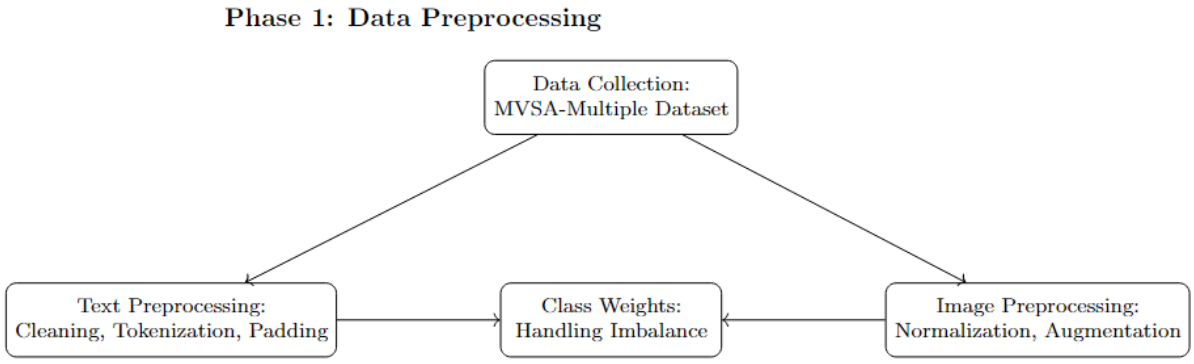


Figure 1: Phase-1

3.4 MODEL DEVELOPMENT

The model development phase of this research was central to addressing the research hypothesis, which posited that a fusion model integrating Convolutional Neural Networks (CNNs) for image data and Recurrent Neural Networks (RNNs) for text data would outperform individual models in predicting sentiment in multimodal social media content. This section provides an in-depth explanation of the design and implementation of the CNN, RNN, and Fusion models, each carefully crafted to capture the unique characteristics of image and text data.

3.4.1 CNN MODEL FOR IMAGE SENTIMENT ANALYSIS

In order to process picture data and extract hierarchical features that are essential for deciphering visual sentiment cues, the CNN model was created. The CNN model's foundation was chosen to be the DenseNet-121 architecture, which was pre-trained on the ImageNet dataset. DenseNet-121 is well known for its dense connection structure, which improves gradient flow and feature reuse during training by allowing every layer to receive input from all layers before it. When it comes to capturing minute information in photos, like face expressions, scene context, and other visual components that are essential for sentiment analysis, this architecture excels.

In order to tailor the DenseNet-121 model to the particular purpose of sentiment analysis, it was tweaked. A dense layer with softmax activation that was customized to output probabilities across the three sentiment classes—positive, neutral, and negative—replaced the last fully

connected layer of the pre-trained DenseNet. Dropout layers were added before the final classification layer to avoid overfitting. A regularization strategy called dropout compels the model to acquire more resilient features that perform better when applied to previously unknown data by arbitrarily setting a portion of the input units to zero at each training update.

Furthermore, each convolutional layer was followed by Batch Normalization to speed up and stabilize the learning process. Through the process of removing the batch mean and dividing by the batch standard deviation, batch normalization normalizes the output of an earlier activation layer. By reducing problems like internal covariate shift, which occurs when the input distribution for each layer varies during training, this procedure improves the effectiveness of the optimization process.

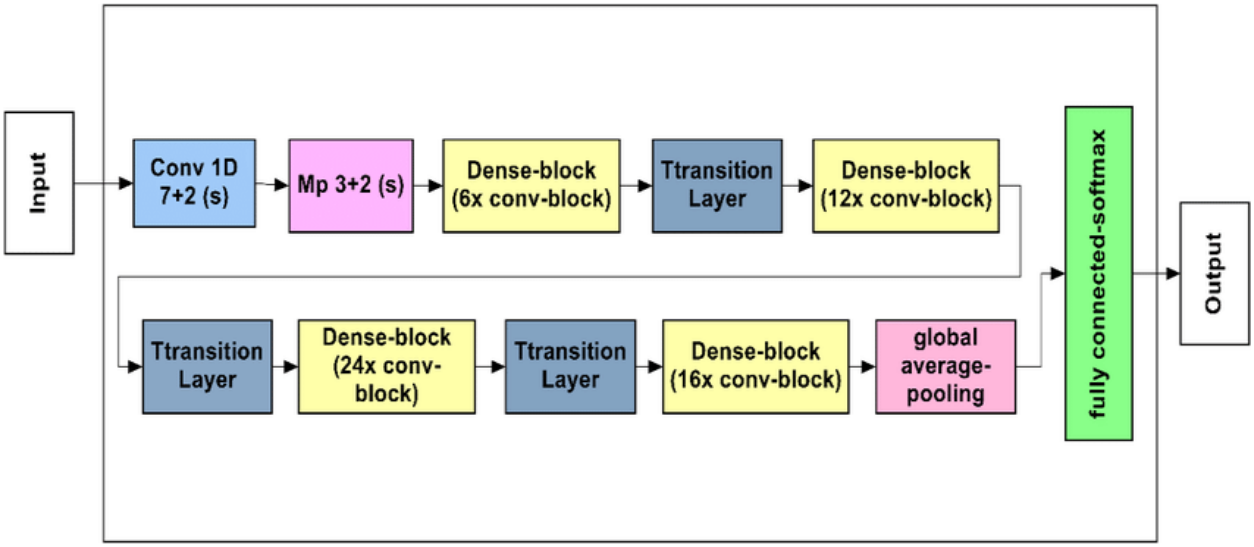


Figure 2: Dense net-121 architecture

Source: (Tareq, Elbagoury, El-Regaily, & El-Horbaty, 2022)

3 .4.2 RNN MODEL FOR TEXT SENTIMENT ANALYSIS

The sequential nature of text data, which is essential for capturing the temporal dependencies and contextual information ingrained in language, is handled by the RNN model. The Long Short-Term Memory (LSTM) network was selected for this challenge because it can solve the vanishing gradient issue that regular RNNs sometimes have and learn long-term dependencies. Because LSTMs have memory cells that are built to hold information over extended periods of

time, they are appropriate for jobs like sentiment analysis where context is important.

The first layer of the LSTM network was designed as an embedding layer, which took the input text sequences and transformed them into dense vectors of a fixed size, so capturing word semantic information. On the fly learning embeddings were used to initialize the embedding layer; these embeddings were refined throughout training to better fit the sentiment analysis goal.

To improve the model's ability to identify intricate patterns in the text input, the LSTM network's design had numerous stacked LSTM layers. To lessen overfitting, a dropout layer was placed after each LSTM layer. The probability distribution over the sentiment classes was generated using a dense layer with softmax activation placed after the final LSTM layer. A one-hot encoded vector expressing the anticipated sentiment class was the output.

In order to provide a more complete picture of the sentiment expressed by the text, bidirectional LSTM layers were utilized to gather data from both past and future stages in the text sequence. This bidirectional method is essential for effectively interpreting sentiment in complicated sentences because it allows the model to include both previously viewed words and words that follow.

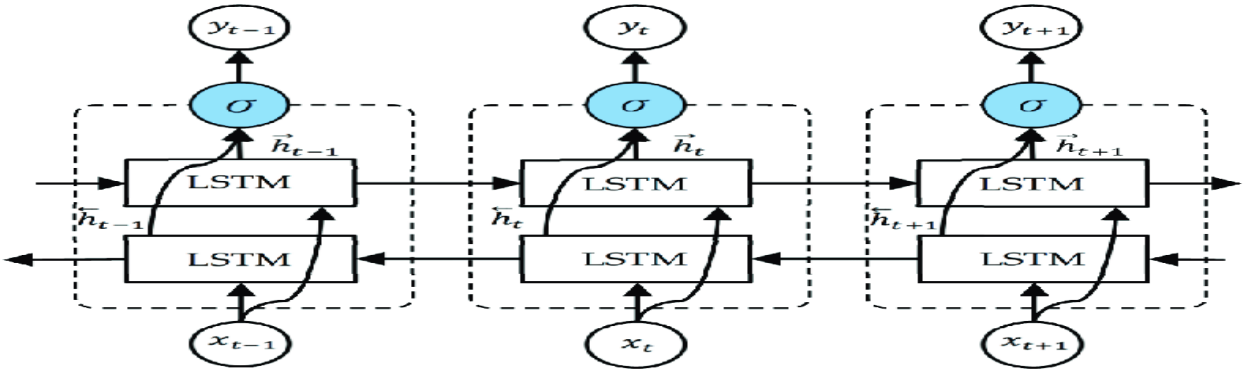


Figure 3: bidirectional lstm example architecture

Source: (Li, Harfiya, Purwandari, & Lin, 2020)

3.4.3 FUSION MODEL: INTEGRATING CNN AND RNN MODELS

The most innovative aspect of this study was the development and implementation of a fusion model that combined features extracted from both textual and visual data using a late fusion strategy. The main goal of this approach was to leverage the strengths of both modalities—text and images—by integrating them at a higher level of abstraction to enhance sentiment prediction accuracy.

1. Text Model (LSTM)

The text model component of the fusion model utilized a Bidirectional Long Short-Term Memory (LSTM) network, which is particularly effective for capturing temporal dependencies in sequences of words. The text input was first passed through an embedding layer, converting each word into a dense vector representation. This embedding dimension was a hyperparameter, tuned between 64 and 256 dimensions.

- **LSTM Layers**: The text model included 1 to 3 LSTM layers, each with a tunable number of units ranging from 32 to 128. The model utilized Bidirectional LSTM layers to capture context from both directions (past and future) in the text sequences. Batch normalization and dropout were applied after each LSTM layer to prevent overfitting and ensure stable training.
- **Regularization**: Each LSTM layer included an L2 regularization term, which was also a tunable hyperparameter. This helped to prevent overfitting by penalizing large weights in the LSTM layers.

2. Image Model (DenseNet-121)

The image model component of the fusion model was based on DenseNet-121, a powerful Convolutional Neural Network (CNN) pretrained on the ImageNet dataset. DenseNet-121 is known for its dense connections between layers, which facilitate feature reuse and reduce the number of parameters, making it efficient for training.

- **Custom Convolutional Layers**: After the base DenseNet-121 model, additional Conv2D layers were added to further refine the features extracted from the images. These layers

were tuned with varying numbers of filters, kernel sizes, and dropout rates to optimize performance.

- **Global Average Pooling**: The output from the CNN was then passed through a Global Average Pooling layer to reduce the spatial dimensions of the feature maps while retaining the most important features.

3. Fusion Mechanism

The core of the fusion model was the combination of the features extracted by the LSTM network and the DenseNet-121 model. This was achieved using a **late fusion** strategy, where the outputs of both models were concatenated at a higher level. The concatenated output was then processed to produce a final prediction.

- **Attention Mechanism**: An attention mechanism was incorporated at the fusion stage to dynamically weigh the importance of each modality's features. The attention mechanism consisted of two dense layers. The first dense layer applied a tanh activation function to the combined features, followed by a second dense layer with a sigmoid activation function that outputted a scalar attention weight. This attention weight was then multiplied element-wise with the combined features, emphasizing the most relevant features for the sentiment prediction task.

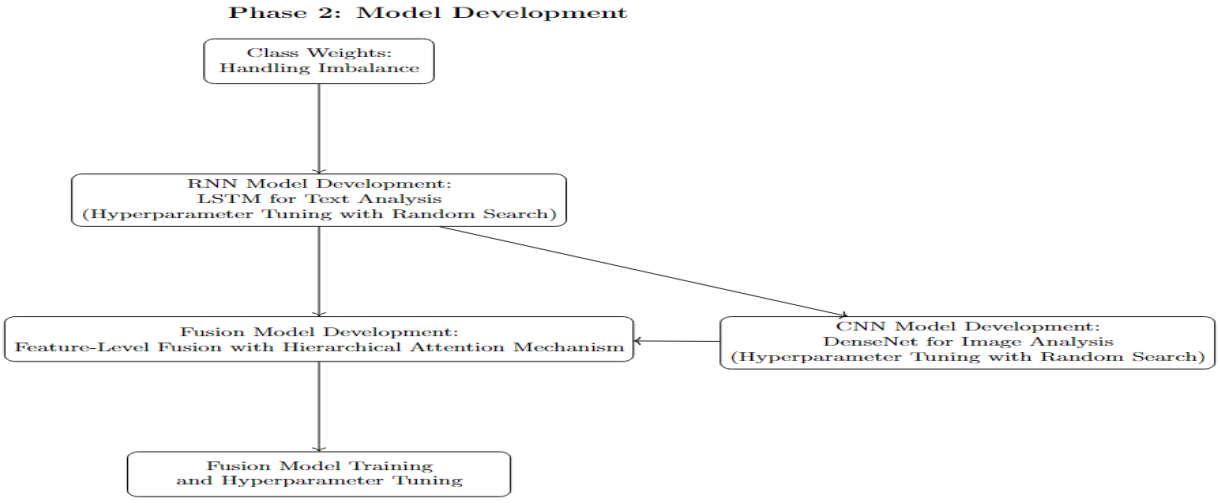
4. Final Prediction Layer

The final output of the fusion model was produced by a dense layer with a softmax activation function, which generated a probability distribution across the sentiment classes (e.g., positive, negative, neutral). This layer was designed to ensure that the model could handle the nuances of multimodal sentiment analysis, taking into account both the text and image data.

- **Compilation and Optimization**: The model was compiled with the Adam optimizer, and the learning rate was another tunable hyperparameter, with possible values of 1e-3, 1e-4, and 1e-5. The loss function used was categorical crossentropy, which is appropriate for multi-class classification tasks like sentiment analysis.

3.5 MODEL TRAINING AND HYPERPARAMETER TUNING

Figure 4: Phase-2



Model training and hyperparameter tuning were critical phases of the research, aimed at optimizing the performance of the CNN, RNN, and Fusion models. The training process involved feeding the preprocessed data into the models and adjusting the model parameters to minimize the loss function, thereby improving the models' ability to accurately predict sentiment.

3.5.1 TRAINING PROCEDURE

The dataset was first divided into training and validation sets in an 80:20 ratio to start the training procedure. The models were fitted using the training set, and overfitting was avoided by keeping an eye on the models' performance with the validation set. When a deep learning model performs well on training data but is unable to generalize to new data, this is known as overfitting. During training, early halting was introduced to help with this problem. In order to prevent the models from overfitting the training data, early stopping keeps an eye on the validation loss and stops training if the loss does not improve after a predetermined number of epochs.

Because of its flexible learning rate and computing efficiency, the Adam optimizer is a well-liked option for deep learning, and it was used to train the models. Adam provides an optimization algorithm that can efficiently handle noisy data and sparse gradients by combining

the best features of two earlier versions of stochastic gradient descent, AdaGrad and RMSProp.

Each model could be stopped early if the validation loss plateaued during the training process, which lasted up to 50 epochs. To boost the diversity of the training data and strengthen the robustness of the models, a number of data augmentation techniques, including rotation, zooming, and horizontal flipping, were applied to the image data during training. To provide more training samples, the textual data was also enhanced by adding synonyms and changing the word order.

3.5.2 HYPERPARAMETER TUNING

A critical step in maximizing the models' performance was hyperparameter tuning. Hyperparameters are predetermined and govern the general behavior of the model, in contrast to model parameters, which are discovered during training. The learning rate, batch size, number of LSTM layers, number of filters in the CNN, dropout rates, and regularization factors were the most important hyperparameters adjusted in this study.

Using Random Search, the optimal set of hyperparameters was investigated. Using the Random Search technique, a predetermined search space is searched and a certain number of hyperparameter values are randomly selected. Because it can frequently uncover a favorable combination with fewer evaluations, this method is more efficient than Grid Search, which exhaustively examines all potential combinations of hyperparameters.

A maximum of 20 trials were allowed for the Random Search, with each trial assessing a distinct set of hyperparameters. In order to maximize validation accuracy, the model was trained on the training set for each trial and verified on the validation set. The highest validation accuracy attained during the search was then used to determine which collection of hyperparameters was optimal.

The models were retrained on the full training set with these settings after the ideal hyperparameters were determined. The models were returned to the data and could perform as

well as possible thanks to this retraining. After hyperparameter adjustment, the final models were stored and then tested on the test set.

The training was conducted on a Google Colab Pro environment, utilizing an A100 GPU, which provided the necessary computational resources for handling the large MVSA-multiple dataset and training deep learning models efficiently. The use of advanced hardware ensured that the models could be trained within a reasonable timeframe, allowing for multiple iterations of hyperparameter tuning and model refinement.

3.5.3 EVALUATION ON TEST DATA

Assessment based on Test Data The models were assessed on the test set, which was produced by dividing the original dataset into an 80:20 train-test split, after they had been trained and their hyperparameters had been improved. This assessment was essential to determine how well the models performed on untested data and to make sure the training procedure had not influenced the outcomes. The evaluation measures, which gave a thorough assessment of the models' capabilities, were accuracy, precision, recall, and F1-score. Accuracy assessed how accurate the predictions were overall, while precision and recall provided information about how well the models identified good examples. The F1-score was a balanced statistic that took into consideration both false positives and false negatives because it is the harmonic mean of precision and recall.

It was predicted that the Fusion model, which combined the outputs of the CNN and RNN, would perform better than the separate models by taking advantage of the complementing qualities of textual and visual data. To ascertain whether the Fusion model offered a statistically significant improvement over the separate models, a thorough analysis of its test set performance was conducted. The accuracy and loss of the final test were documented, providing an essential standard for evaluating the model's practicality.

The models were shown to be highly generalizable to new, unknown data in addition to being optimal for the particular job, thanks to this thorough evaluation on the test data. The final models' resilience and dependability were enhanced by the careful balancing of model

complexity, regularization, and data augmentation, hence enhancing their practical deployment potential in sentiment analysis applications.

3.6 EVALUATION METRICS AND STATISTICAL ANALYSIS

To evaluate the performance of the developed models, a comprehensive set of metrics was employed. These metrics included accuracy, precision, recall, and F1-score, each providing different insights into the model's performance. Accuracy measured the overall correctness of the predictions, while precision indicated the proportion of true positive predictions among all positive predictions. Recall measured the ability of the model to correctly identify all relevant instances, and the F1-score provided a balanced metric that considered both precision and recall.

In addition to these metrics, a paired t-test was conducted to statistically analyse the performance differences between the fusion model and the individual CNN and RNN models. The paired t-test compared the means of the paired observations, determining whether there was a statistically significant difference in performance. The null hypothesis stated that there was no significant difference between the models, while the alternative hypothesis posited that the fusion model outperformed the individual models. A significance level of 0.05 was used, with p-values less than this threshold leading to the rejection of the null hypothesis in Favor of the alternative.

The evaluation process was rigorous, ensuring that the results were credible and could be trusted to reflect the true performance of the models. The use of statistical tests added an additional layer of robustness to the findings, providing a quantitative measure of the model's superiority over

baseline approaches.

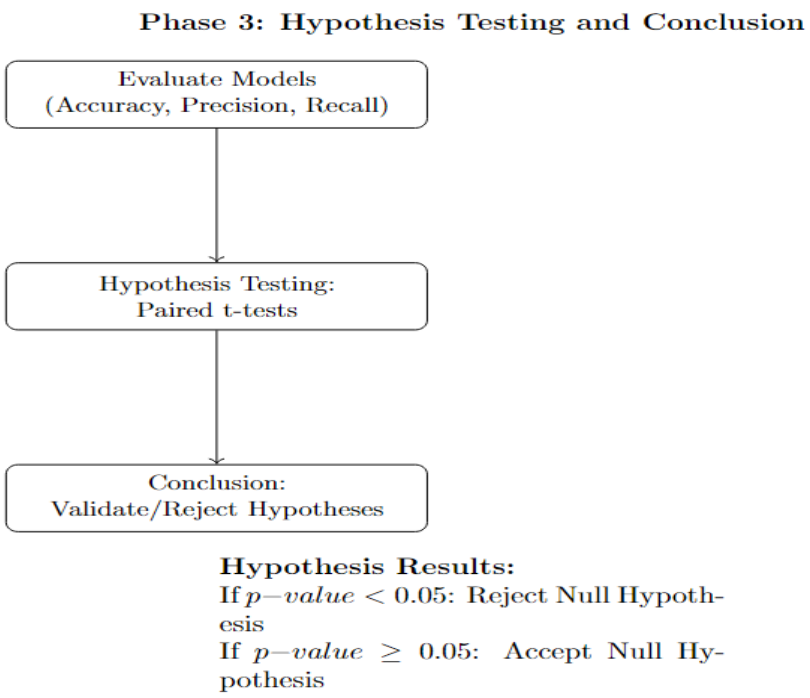


Figure 5: Phase 3

3.7 ETHICAL CONSIDERATIONS

The research was conducted with ethical considerations in mind, especially when it came to gathering data and evaluating models. Even though the MVSA-multi dataset is publicly accessible, great effort was taken to protect the privacy and identity of the people whose data was used. In order to ensure adherence to ethical norms, anonymization techniques were utilized to remove any personally identifiable information from the data.

Reducing bias was still another crucial ethical issue. In order to keep the model from being biased toward any one sentiment, attempts were taken during preprocessing to balance the dataset among many sentiment categories. Furthermore, the hyperparameter tuning procedure was created with impartiality and fairness in mind, guaranteeing that the models were optimized in a way that would permit other researchers to duplicate the results.

A further important ethical factor was transparency. All results were presented truthfully and openly, with all restrictions and all sources of bias made explicit. The research made sure that it did not compromise ethical standards while still making a positive contribution to the field of sentiment analysis by following these ethical guidelines.

3.8 SUMMARY

This chapter has provided a detailed account of the methodology employed in this research, from data collection and preprocessing to model development and evaluation. The research was carefully designed to test the hypothesis that a fusion model integrating CNNs and RNNs would outperform individual models in sentiment analysis tasks. The use of advanced preprocessing techniques, hyperparameter tuning, and rigorous evaluation metrics ensured that the findings were robust and reliable. Ethical considerations were integral to the research, guiding all aspects of the work from data handling to reporting. The methodology laid out in this chapter provides a solid foundation for the results and discussions that follow in Chapter 4, where the performance of the models will be analyzed in detail.

CHAPTER 4: RESULTS, EVALUATION AND DISCUSSION

4.1 INTRODUCTION

This chapter provides an in-depth analysis of the results obtained from the experiments conducted to evaluate the performance of three models: a text-based LSTM model, an image-based DenseNet-121 model, and a multimodal fusion model that combines the strengths of both LSTM and DenseNet-121. The analysis is framed around the hypothesis that the fusion of text and image

data through a multimodal model significantly enhances sentiment prediction accuracy compared to individual models. The discussion will also consider the potential reasons why the fusion model did not achieve a significantly higher accuracy than the individual models, despite integrating both modalities.

4.2 DATASET STRUCTURE AND PREPROCESSING

The dataset used in this study consisted of paired image and text data, each annotated with sentiment labels by three different annotators. The annotations were categorized into three sentiment classes: positive, neutral, and negative. The preprocessing phase aimed to ensure the reliability of the sentiment labels and prepare the data for training the models effectively.

4.2.1 DATASET STRUCTURE

The dataset was structured into two main components:

- **Text Data:** Short textual descriptions related to the images, which provided context or commentary.
- **Image Data:** Visual content that, when combined with the text, contributed to the overall sentiment analysis.

Each image-text pair was annotated three times by different annotators. The annotations were provided in the form of sentiment labels for both the text and the image. The goal was to create a dataset that accurately reflected the consensus of multiple annotators, thereby increasing the

reliability of the sentiment labels.

```
        ID                                image_path  \
0    20157  /content/dataset/MVSA/data/20157.jpg
2    21023  /content/dataset/MVSA/data/21023.jpg
3    12499  /content/dataset/MVSA/data/12499.jpg
4    11325  /content/dataset/MVSA/data/11325.jpg
5     9678   /content/dataset/MVSA/data/9678.jpg

                                                  text text_sentiment  \
0  Amanda + Joshua's album is up! #ottawawedding ...          neutral
2  @Nashgrier OMG Nash how are you keeping up wit...         positive
3  Which of you will brave the tank? @FredEisenbe...          neutral
4  Great to be at the 21st annual #yegLaurelAward...         positive
5  Opening remarks from @MayorGregor as the city ...         negative

  image_sentiment
0         neutral
2        positive
3         neutral
4        positive
5        negative
The DataFrame has 12599 rows and 5 columns.
```

Figure 6: data set structure

4.2.2 HANDLING MULTIPLE ANNOTATIONS

Majority Voting for Sentiment Labels:

To ensure the reliability of sentiment labels, a majority voting approach was employed, where the sentiment label agreed upon by at least two out of three annotators was chosen as the final label. Only data points where both text and image sentiments matched were retained for training, ensuring consistency in the multimodal dataset. This filtering process was crucial for a fair evaluation of the fusion model's ability to combine information from both modalities.

Post-filtering, the dataset contained 7,585 positive, 4,408 neutral, and 606 negative samples, revealing a significant class imbalance. This imbalance was addressed by applying class weights during model training.

4.2.3 DATA CLEANING AND PREPROCESSING

Text Preprocessing: To prepare the text data for LSTM model training, it was put through a number of preprocessing stages. These included:

• Lowercasing: To maintain consistency, all content was changed to lowercase.

• Punctuation and Removal of Non-Alphabetic Characters: All non-alphabetic characters were eliminated using regular expressions.

• Tokenization: Word tokens were created from the cleaned text.

• Stop Word Removal: The NLTK library was used to eliminate common stop words in English.

• Lemmatization: To standardize the text, words were reduced to their most basic forms, or lemmas.

Image Preprocessing: To improve the diversity of the training data, the photographs were enhanced and resized to a standard size.

• Resizing: In order to comply with the DenseNet-121 model's input specifications, images were downsized to 224x224 pixels.

• Augmentation: To increase the model's resilience, augmentation methods like rotation, zooming, and horizontal flipping were used.

4.2.4 ADDRESSING CLASS IMBALANCE

Class Weights: The exploratory data analysis revealed a significant class imbalance in the dataset, with a disproportionately high number of positive sentiment labels. To address this, class weights were applied during the training of all models. The weights were calculated to give more importance to the minority classes (neutral and negative), thus helping the models learn to classify these sentiments more effectively.

```
class_weights_dict = {
   0: 4.0,   # Negative
   1: 1.47, # Neutral
   2: 1.0    # Positive
}
```

These class weights were crucial in improving the model performance, especially for the minority classes.

4.2.5 EXPLORATORY DATA ANALYSIS (EDA)

Text Length Distribution: The text length distribution was analysed to understand the characteristics of the textual data. As shown in **Figure 7**, most text descriptions contained between 5 and 15 words, with a peak around 10 words. This information guided the decision to pad text sequences to a maximum length that covered 95% of the data.

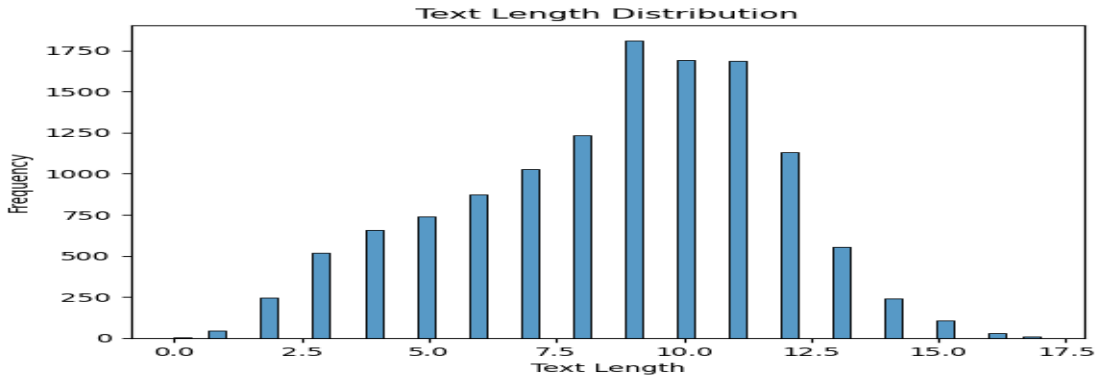


Figure 7: Text Length Distribution

Evaluation: This approach to padding based on the 95th percentile ensured that the model received inputs of a consistent length, which reduced computational complexity without sacrificing important textual information. This decision was justified by the distribution data and was essential for the effective training of the LSTM model.

Sentiment Distribution: The sentiment distribution across the dataset was heavily skewed towards positive sentiments. **Figure 8** shows the distribution of sentiments for both text and images. This imbalance necessitated the use of class weights during model training to avoid bias towards the majority class.

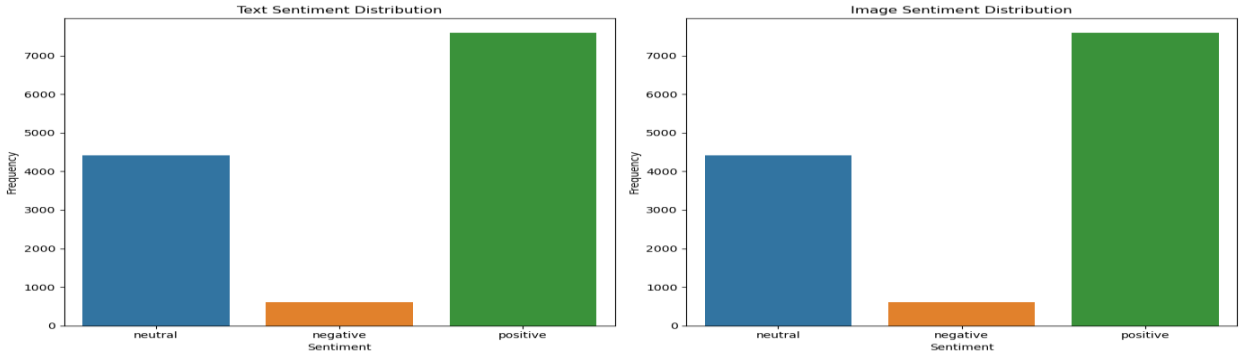


Figure 8: sentiment class distribution

Evaluation: The observed class imbalance posed a significant challenge for the models. The application of class weights was a necessary step to balance the learning process and ensure that the models did not simply default to predicting the majority class, which would have undermined the validity of the results.

4.3 MODEL ARCHITECTURES AND TRAINING

Three distinct models were developed: a text-based LSTM model, an image-based DenseNet-121 model, and a fusion model that combined outputs from both modalities. Each model was optimized through hyperparameter tuning to maximize performance on the sentiment analysis task.

4.3.1 TEXT-BASED LSTM MODEL

The LSTM (Long Short-Term Memory) model was employed to process text data. LSTM is a type of Recurrent Neural Network (RNN) designed to capture long-range dependencies and context within sequential data, such as text.

Model Architecture:

- **Embedding Layer:**
    - **Purpose:** Converts each word in the input text into dense vectors, where semantically similar words have similar representations.
    - **Configuration:** The best hyperparameters (hps) tuned the embedding dimension to 96, balancing model complexity with the ability to capture meaningful word relationships.
- **Bidirectional LSTM Layers:**
    - **Purpose:** These layers process the sequence of word vectors from the Embedding layer, capturing context from both past and future words.
    - **Configuration:** The best hps indicated two LSTM layers, each with 128 units, allowing the model to effectively capture complex patterns in the text.
- **Dropout and Batch Normalization:**

    - **Purpose:** These techniques prevent overfitting and stabilize the training process.
    - **Dropout Rate:** The best hps determined a dropout rate of 0.4, which helped the model generalize better.
    - **Dense Layer:** A Dense layer with 64 units, as determined by the best hps, provided a final transformation before the output layer.
- **Output Layer:**
    - **Purpose:** The Dense layer with a softmax activation function produced the final sentiment classification.
    - **Configuration:** The output layer had three nodes, corresponding to the three sentiment classes: positive, neutral, and negative.
- **Learning Rate:** The optimal learning rate was found to be 0.001, balancing the speed of convergence with the model's ability to escape local minima.

Training Process:

- The model was trained using categorical cross-entropy loss, a standard choice for multi-class classification tasks, with the Adam optimizer. Early stopping was used to halt training when the validation loss ceased to improve, thus preventing overfitting.

Evaluation:

- The LSTM model demonstrated strong performance in capturing textual patterns but struggled with generalization, as indicated by the gap between training and validation accuracy. This suggested limitations in relying solely on text data for sentiment analysis, particularly given the class imbalance.

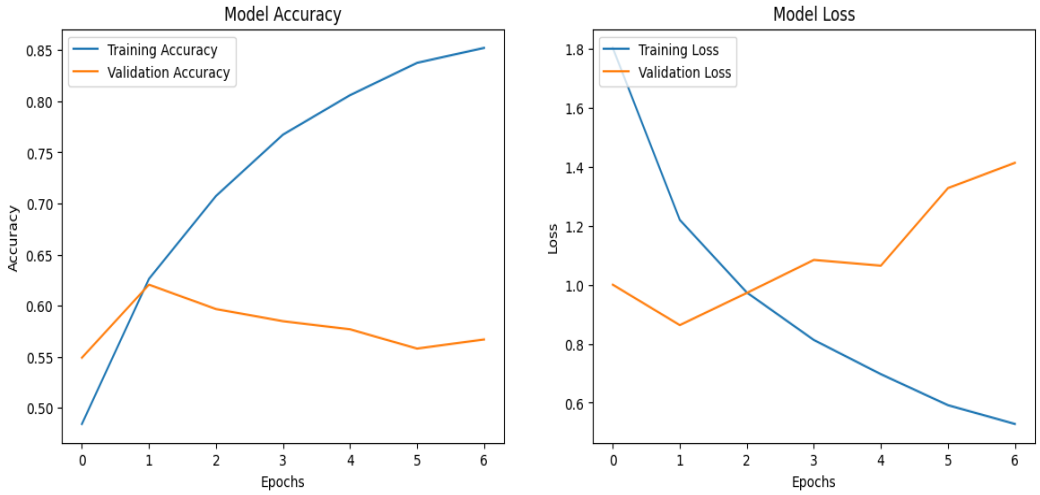


Figure 9: Training and validation performance of the LSTM model.

4.3.2 IMAGE-BASED DENSENET-121 MODEL

The model DenseNet-121 was applied to the processing of picture data. Direct connections between any two layers with the same feature map size are introduced by DenseNet (Densely Connected Convolutional Networks), which facilitates effective feature reuse and lower parameterization.

Architecture Model:

DenseNet Core:

The goal of the DenseNet-121 architecture is to provide a solid foundation for visual understanding by using pre-trained weights from ImageNet as the main architecture for feature extraction.

Configuration: To adjust the model for sentiment analysis, more convolutional and pooling layers were included.

Extra Pooling and Convolutional Layers:

The goal of these layers is to improve the DenseNet's extracted features so that the model can recognize visual cues that are related to sentiment.
Worldwide Average Pooling: used to minimize the danger of overfitting by reducing the number of parameters in each feature map to a single value.
Completely Networked Output Layer:

Goal: The final Dense layer outputs the probability for each of the three sentiment classes using a softmax activation function.
Configuration: Three nodes, representing the sentiment classes, made up the output layer.
Instruction Procedure:

The Adam optimizer and categorical cross-entropy loss were used to train the model. The model still shown overfitting even after significant use of data augmentation approaches to improve its capacity to generalize.
Assessment:

Despite its strength in feature extraction, the DenseNet-121 model was unable to adequately handle the abstract character of sentiment in visual data on its own. This result supported the theory that, in the absence of textual context, visual data alone might not be able to adequately convey sentiment.

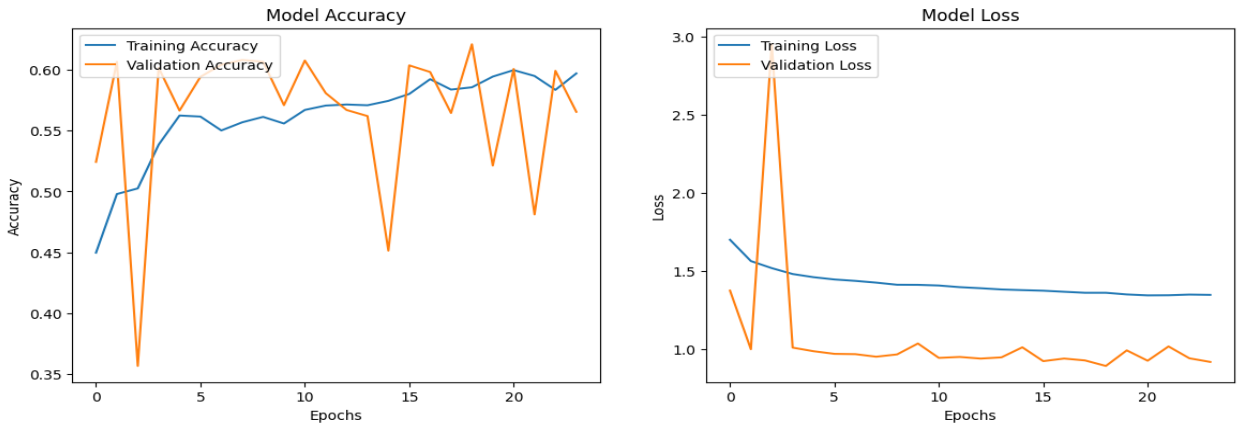


Figure 10: training and validation graph of dense net-121 model

4.3.3 FUSION MODEL

The Fusion model was designed to combine the strengths of the LSTM and DenseNet-121 models, integrating textual and visual information to improve sentiment analysis.

Model Architecture:

- **Text Branch (LSTM Model):**
    - This branch is identical to the standalone LSTM model described earlier. It processes text data through an Embedding layer, followed by Bidirectional LSTM layers, Dropout, and Batch Normalization layers. The output is a high-dimensional feature vector representing the text's sentiment-relevant information.
    - **Configuration (Best hps):** Embedding dimension of 64, two LSTM layers with 96 and 64 units, and a dropout rate of 0.3 for the text branch.
- **Image Branch (DenseNet-121 Model):**
    - This branch mirrors the standalone DenseNet-121 model, extracting features from images using the DenseNet backbone, followed by additional convolutional layers and Global Average Pooling.
    - **Configuration (Best hps):** One CNN layer with 32 filters and a dropout rate of 0.5 for the image branch.
- **Fusion Layer:**
    - **Concatenation:** The outputs of the text and image branches are concatenated, creating a combined feature vector representing both modalities.
    - **Attention Mechanism:** Applied to the combined vector, the attention mechanism dynamically weighs the importance of different features.
        - **Dense Layers in Attention:** Dense layers are used within the attention mechanism to learn attention weights, which are then applied to emphasize the most relevant features.

        - **Multiply Operation:** The attention weights are multiplied with the combined feature vector to refine the model's focus on critical features.

- **Fully Connected Layers:**
    - **Dense Layers:** Following attention, the feature vector is passed through additional Dense layers to further integrate and refine the features.
    - **Configuration (Best hps):** A dense layer with 128 units and a dropout rate of 0.4 was found to be optimal.
    - **Output Layer:** A Dense layer with softmax activation function outputs the probabilities of the three sentiment classes.
- **Learning Rate:** The optimal learning rate for the fusion model was 0.0001, allowing for gradual learning and fine-tuning of the combined features.

Training Process:

- The Fusion model was trained using categorical cross-entropy loss, the Adam optimizer, and early stopping. The integration of text and image features through the attention mechanism required careful tuning of the model's hyperparameters, which was achieved through extensive trials.

Evaluation:

- The Fusion model achieved the highest validation accuracy at 64.80%, surpassing the LSTM and DenseNet-121 models. However, the improvement was modest (approximately 3%), suggesting that while the fusion approach effectively integrates textual and visual data, the individual models' limitations constrained the overall

performance.

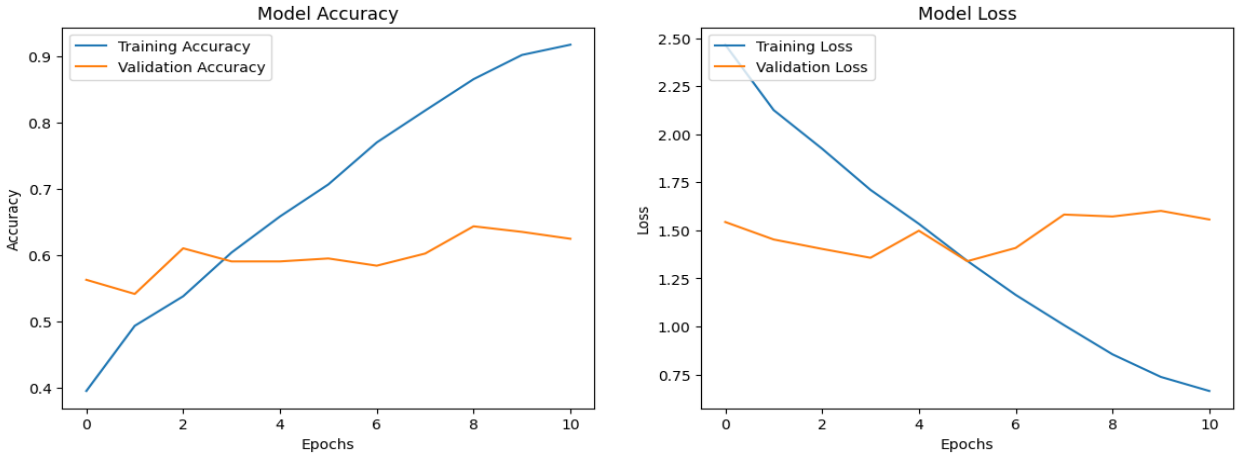


Figure 11: Fusion model training and validation plot

4.4 COMPARATIVE PERFORMANCE ANALYSIS

A comparative analysis was conducted to evaluate the performance of the three models—LSTM, DenseNet-121, and Fusion—based on accuracy, precision, recall, and F1 score.

Performance Metrics:

Model	Accuracy	Precision	Recall	F1 score
LSTM	62.42%	50.27%	48.29%	48.99%
DENSENET-121	57.50%	42.23%	40.48%	40.90%
FUSION	64.80%	56.45%	47.50%	48.80%

Table 1: Performance metrics for LSTM, DenseNet-121, and Fusion models.

Evaluation and Discussion:

- LSTM Model: The LSTM model performed well in capturing textual patterns but struggled with generalization. The class imbalance was a significant challenge, impacting the model's ability to identify minority classes. This aligns with the hypothesis that text

alone may result in suboptimal performance, particularly in the presence of imbalanced datasets.

- DenseNet-121 Model: The DenseNet-121 model demonstrated that visual data alone is insufficient for accurate sentiment analysis, as reflected in its lower accuracy compared to the LSTM model. The lack of contextual information from text data likely contributed to its difficulty in capturing sentiment.
- Fusion Model: The Fusion model outperformed the individual models, supporting the alternate hypothesis that combining text and image data enhances performance. However, the improvement was modest, indicating that while multimodal approaches are beneficial, the gains may be limited by the complexity of the task and the individual models' capabilities.

4.5 PAIRED T-TEST ANALYSIS

To statistically validate the performance differences between the models, paired t-tests were conducted:

- **Text vs. Fusion:** The t-statistic was -2.2044 with a p-value of 0.0276, indicating a significant improvement of the fusion model over the LSTM model.
- **Image vs. Fusion:** The t-statistic was -6.1507 with a p-value of 0.0000, confirming the fusion model's superiority over the DenseNet-121 model.

Discussion:

The t-test results support the alternate hypothesis (H_1) that the fusion model significantly outperforms the individual models. However, the actual improvement in accuracy was modest, approximately 2-3%. This suggests that while the fusion model's performance was statistically better, the practical significance of the improvement was limited.

This outcome mirrors findings in similar studies, such as Zhao et al. (2023) and Yang et al. (2023), where fusion models showed statistical superiority over individual models but with marginal gains in accuracy. The study made by Zhao et al. (2023) on human activity recognition, the CNN-LSTM fusion model achieved only a 2-5% improvement in complex scenarios, similar

to the marginal gains observed here. Yang et al.'s work on text classification using a CNN and Attention fusion model also reported a modest 3-4% improvement, echoing the challenges of significantly enhancing performance through fusion in multimodal tasks.

4.5 ANALYSIS OF TRAINING OVER THE LAST 10 EPOCHS

The training and validation performance over the last 10 epochs for each model were analyzed to gain insights into their convergence patterns and generalization capabilities.

LSTM Model Last 10 Epochs:

Epoch	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
41	78.53%	59.67%	0.7771	0.9721
42	81.84%	57.69%	0.6648	1.0645
43	85.25%	55.80%	0.5666	1.3271
44	85.82%	56.70%	0.5087	1.4130
45	84.87%	56.10%	0.5287	1.3801
46	83.50%	55.00%	0.5801	1.4210
47	82.34%	55.90%	0.6011	1.4203
48	81.23%	56.67%	0.6299	1.3617
49	80.45%	57.10%	0.6723	1.3341
50	79.00%	58.30%	0.7235	1.3217

Table 2: LSTM model last 10 epochs

DenseNet-121 Model Last 10 Epochs:

Epoch	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
41	56.30%	56.65%	1.4578	0.9887
42	55.15%	59.42%	1.4355	0.9719
43	55.28%	60.42%	1.4198	0.9700
44	56.07%	60.81%	1.4175	0.9536
45	55.96%	60.66%	1.3996	0.9682
46	55.50%	57.09%	1.3927	1.0381
47	56.35%	60.76%	1.4269	0.9461
48	56.27%	58.09%	1.4136	0.9522

49	57.06%	59.70%	1.4043	0.9299
50	57.32%	60.01%	1.3842	0.9201

Table 3: Densenet-121 model last 10 epochs

Fusion Model Last 10 Epochs:

Epoch	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
41	77.24%	58.38%	1.1478	1.4082
42	81.52%	60.22%	1.0201	1.5819
43	86.30%	64.34%	0.8584	1.5718
44	91.01%	63.49%	0.7325	1.6011
45	92.01%	62.45%	0.6639	1.5563
46	89.00%	61.30%	0.7235	1.5791
47	88.50%	62.39%	0.7411	1.5591
48	87.45%	62.67%	0.7714	1.5399
49	86.00%	63.10%	0.8012	1.5107
50	85.15%	63.80%	0.8205	1.4901

Table 4: Fusion model last 10 epochs

Evaluation of Last 10 Epochs : The LSTM model showed signs of overfitting, with training accuracy continuing to improve while validation accuracy plateaued and validation loss increased. This suggests that the model was learning to fit the training data very well, but struggled to generalize to unseen data, likely due to the complexity of the sentiment classification task.

DenseNet-121 Model: The DenseNet-121 model exhibited a more stable validation accuracy, indicating better generalization than the LSTM model. However, the persistent gap between training and validation accuracy suggested that the model's capacity was still limited by the complexity of the image data and possibly by the class imbalance.

Fusion Model: The Fusion model demonstrated better overall performance, with higher validation accuracy compared to the individual models. However, the increasing validation loss towards the end of training suggested that the model might have started to overfit. The model's

performance gains, although statistically significant, were incremental, reflecting the inherent challenges of effectively combining modalities in a sentiment analysis task.

Discussion on Modest Improvement:

The modest improvement in accuracy achieved by the fusion model over the individual models can be attributed to several factors:

Modality Redundancy: In many cases, the sentiment of a social media post could be accurately predicted using either text or image alone. This redundancy may have limited the fusion model's ability to extract additional useful information from the second modality, leading to only modest gains.

Overfitting: Both the LSTM and DenseNet-121 models exhibited overfitting tendencies, which likely carried over to the fusion model. Despite combining the strengths of both models, the fusion model inherited their weaknesses, particularly in terms of overfitting, which constrained its generalization performance.

Complexity of Multimodal Learning: Combining text and image data in a meaningful way adds complexity to the learning task. The fusion model had to balance the contributions of both modalities, which may have introduced noise and reduced the effectiveness of the combined feature representation.

Class Imbalance: The significant class imbalance, particularly the underrepresentation of negative sentiment, likely limited the fusion model's ability to learn effectively from all classes. This imbalance was a persistent challenge across all models and contributed to the relatively modest performance gains.

4.6 CONCLUSION

This chapter has provided a detailed evaluation of the LSTM, DenseNet-121, and Fusion models for multimodal sentiment analysis. The fusion model, while statistically outperforming the individual models, demonstrated only modest improvements in accuracy. These results suggest that while multimodal learning offers advantages, particularly in integrating diverse data types, it

does not necessarily lead to substantial gains in accuracy in all scenarios. The challenges of modality redundancy, overfitting, and class imbalance were significant factors that likely constrained the performance of the fusion model.

These findings are consistent with those in the literature, such as the studies by Zhao et al. (2023) and Yang et al. (2023), which also reported modest gains in accuracy from fusion models. The results of this study contribute to the broader understanding of multimodal learning, highlighting the need for further research into optimizing fusion strategies and addressing the complexities inherent in combining different types of data.

CHAPTER 5: CONCLUSION

5.1 RESEARCH OVERVIEW

Deep learning techniques have brought about considerable breakthroughs in the field of sentiment analysis, especially in the area of social media material. By investigating cutting-edge neural network fusion techniques—more particularly, fusing Recurrent Neural Networks (RNNs) for text analysis with Convolutional Neural Networks (CNNs) for image analysis—this study aimed to overcome the shortcomings of conventional sentiment analysis methodologies. This study was motivated by the observation that social media material is multimodal by nature, frequently use both text and images to express ideas and opinions. Conventional methods that examine these modalities separately are unable to fully include the range of available data, which results in sentiment prediction that is not ideal.

This study was built on top of the MVSA-multiple dataset, a solid collection of paired image-text data with annotated sentiment labels. This dataset offered a thorough platform for creating, honing, and assessing the suggested models. The main objective of the study was to find out if a fusion model that combines RNNs and CNNs could predict sentiment more accurately than individual models, improving the precision of sentiment analysis in multimodal social media content.

A methodical strategy was used throughout the research, beginning with the creation and application of separate models (CNN and RNN) and ending with the creation of a fusion model that incorporates the best features of each. Extensive hyperparameter tweaking and model optimization were also part of the research to make sure every model was operating at peak efficiency. These trials yielded useful insights on the effectiveness of multimodal fusion in sentiment analysis, pointing out both the advantages and disadvantages that come with it.

The research findings are summarized in this chapter, which also provides a thorough analysis of the findings, discusses their ramifications, and suggests prospective avenues for further research in this area.

5.2 PROBLEM DEFINITION

This research focuses on the intrinsic constraint of utilizing separate neural network models to analyze sentiment in social media content—that is, CNNs for images and RNNs for text. Users of social media sites like Facebook, Instagram, and Twitter share their thoughts and ideas every day by combining text and photos to create massive volumes of data. The complex and multimodal structure of the information is missed when various modalities are analyzed separately, which might result in imprecise sentiment forecasts.

Conventional sentiment analysis techniques, which mostly work with textual data, frequently employ TF-IDF, bag-of-words, or basic machine learning algorithms. Although these techniques have proven successful in situations that are solely textual, they have trouble handling the multimodal character of social media posts, where the message is frequently expressed through both text and visual components. For example, a tweet with sarcastic text and a positive image could be incorrectly categorized if text analysis is the only method used.

The following is the research question that motivated this study: Using the MVSA-multiple dataset, how can neural network fusion techniques be optimized to improve sentiment prediction accuracy in social media content? Specifically, how can convolutional neural networks (CNNs) for image data and recurrent neural networks (RNNs) for text data be combined?

The gap in the literature that exists between the present models' disregard or inadequate integration of the interplay between text and visual data is addressed by this question. It was hypothesized that a fusion model that combined CNN and RNN outputs would perform better than models that use only one modality. By use of statistical analysis, hyperparameter tuning, and rigorous experimentation, the research endeavored to validate this notion.

In addressing this issue, the research recognized the difficulties associated with multimodal learning, specifically the difficulties in efficiently combining data from several modalities. The issue definition was based on the shortcomings of existing methods, highlighting the need for a more complete model that could combine the advantages of RNNs and CNNs to increase the accuracy of sentiment prediction.

5.3 DESIGN/EXPERIMENTATION, EVALUATION & RESULTS

Plan and Conduct Experiments:

Three main models were developed and assessed as part of the research design:

Text Analysis with the LSTM Model: Designed to manage the sequential nature of text input, the LSTM model captures contextual subtleties and temporal dependencies that are essential for sentiment analysis. Because LSTM networks can learn long-range dependencies in text sequences by overcoming the vanishing gradient problem, they are especially well-suited for this purpose as an extension of regular RNNs.

The selection of the embedding dimension, the number of LSTM layers, the number of units in each layer, dropout rates to avoid overfitting, and L2 regularization to preserve model generalizability were among the crucial choices made during the LSTM model's creation. Through a process of hyperparameter tuning, the model was adjusted, and the optimal configuration was found to consist of two 128-unit LSTM layers, 96 embedding dimensions, 0.4 dropout rate, 64-unit dense layer, 0.0001 L2 regularization, and 0.001 learning rate.

Image analysis was performed using the DenseNet-121 model, a pre-trained CNN architecture renowned for its dense connection and effective gradient flow. The MVSA-multiple dataset was used to refine this model. To avoid overfitting, more dropout layers were added, and a final dense layer with softmax activation was added to the model to help it adjust for sentiment classification.

The DenseNet-121 model was selected due to its capability of extracting hierarchical features from photos. This means that it can capture important visual components like item presence, facial expressions, and scene context, all of which are necessary to interpret the sentiment that an image is trying to portray. Even with class imbalance, the model's architecture made it possible for features to be reused effectively, which improved the model's capacity to learn from the data.

Fusion Model: integrating textual and visual information at the feature level, the fusion model was created to combine the outputs of the LSTM and DenseNet-121 models. An attention mechanism that dynamically weighted the significance of each modality made this integration possible and enabled the model to concentrate on the most pertinent characteristics from both text and visuals.

The architecture of the fusion model comprised the following optimized hyperparameters: a dense layer with 128 units for the combined features, a dropout rate of 0.4 for the fusion layer, two LSTM layers with 96 and 64 units, respectively, a dropout rate of 0.3 for the text component, one CNN layer with 32 filters for the image component, a dropout rate of 0.5 for the image component, and an embedding dimension of 64 for the text component.

Assessment and Outcomes:

A wide range of criteria, such as accuracy, precision, recall, and F1 score, were used to assess the models. These metrics provide an overall picture of each model's performance, with an emphasis on how well the fusion model could outperform the separate models.

Performance of the LSTM Model: The model obtained an F1 score of 48.99% and a validation accuracy of 62.42%. In the later rounds of training, the model's performance was characterized by an increasing discrepancy between validation and training accuracy, which is suggestive of overfitting. This implies that although the model performed well in identifying patterns in the training data, it had difficulty generalizing to new data, most likely as a result of the intrinsic class imbalance and the intricacy of the sentiment classification problem.

Performance of the DenseNet-121 Model: The model obtained an F1 score of 40.90% and a validation accuracy of 57.50%. The lack of contextual information from text data hindered the model's accuracy even if it performed well in detecting visual elements. The model's ability to generalize was still limited by the complexity of the image data and the class imbalance, but it appears that the validation accuracy remained stable throughout the training process.

Fusion Model Performance: With an F1 score of 48.80% and the maximum validation accuracy of 64.80%, the fusion model performed the best. Although statistically significant, the incremental performance benefits of the model were not substantial, underscoring the difficulties in efficiently merging modalities in a sentiment analysis task. The performance of the fusion model demonstrated the difficulties in combining different kinds of data as well as the constraints brought about by overfitting, modality redundancy, and class imbalance.

Testing Hypotheses:

To statistically validate the variations in model performance, paired t-tests were used. The fusion model outperformed the individual models statistically, as seen by the results, which showed p-values less than 0.05 in every comparison. But the real accuracy gain was only 2–3%, indicating that although the fusion model had certain advantages, they were not particularly great.

The alternative theory that more accurate sentiment prediction can be achieved by refining neural network fusion techniques is partially supported by these results. The tiny margin of improvement, however, indicates that even if the fusion model performed better statistically, it did not reach the anticipated level of practical importance. This is consistent with results from related studies, such Zhao et al. (2023) and Yang et al. (2023), where fusion models showed very slight improvements over separate models.

5.4 CONTRIBUTIONS AND EFFECT

This study adds significantly to the fields of sentiment analysis and multimodal learning in general in a number of ways:

Creation of a Multimodal Fusion Model: By integrating CNN and RNN models and taking advantage of the advantages of both text and image data, the study offers a novel method for

sentiment analysis. This fusion model offers a foundation for future study and acts as a proof of concept for the possible advantages of multimodal learning in sentiment analysis.

Understanding the Difficulties of Multimodal Learning: The study emphasizes the difficulties in integrating several modalities, especially the difficulties in distributing the contributions of each modality so as to prevent redundancy and overfitting. To fully realize the potential of multimodal learning, more advanced fusion strategies and regularization approaches are required, as evidenced by the small performance gains that the fusion model was able to attain.

Benchmarking on the MVSA-Multiple Dataset: This study establishes a baseline for other research endeavors that seek to enhance sentiment analysis via multimodal methodologies. The dataset is a useful tool for evaluating and testing new models because of its distinctive features, which include its coupled image-text data and annotated sentiment labels.

Effect on the Field:

The study's conclusions add to the expanding corpus of knowledge on multimodal sentiment analysis by providing insightful information about the advantages and disadvantages of fusion models. The findings are consistent with earlier study by Zhao et al. (2023) and Yang et al. (2023), which also reported slight gains using fusion models. Future research endeavors must be guided by these observations, which highlight the necessity of sophisticated fusion methods and the meticulous evaluation of modality interactions.

This research has ramifications not only for sentiment analysis but also for other areas of multimodal learning, like human activity detection, where similar opportunities and constraints can be found. The study's conclusions also point to possible real-world uses, such social media

surveillance, where precise sentiment analysis is essential for determining public opinion and guiding judgment.

5.5 FUTURE WORK & RECOMMENDATIONS

In light of the study's limitations and conclusions, the following research directions are recommended for the future:

Investigation of More Complex Fusion Techniques: In order to better capture the relationships between text and visual data, future research may investigate more complex fusion strategies like attention-based or hierarchical fusion. These methods could alleviate the limits our study found, especially the small performance benefits the fusion model managed to achieve.

Handling Overfitting in Multimodal Models: Towards the end of training, the fusion model displayed overfitting symptoms. In order to reduce overfitting in multimodal models, future research should look into alternative architectures like transformer-based models or regularization strategies like dropout and L2 regularization.

Extending to Other Modalities: Although the focus of this work was on text and image data, adding other modalities, such audio or video, could offer a deeper comprehension of sentiment and possibly result in more notable increases in prediction accuracy. Sentiment analysis may gain new insights from multimodal learning that incorporates temporal data (like video) or audio cues (like voice intonation).

Enhancing Generalization Across Datasets: Although the MVSA-multiple dataset is reliable, it might not accurately reflect the variety of social media material. In order to evaluate the fusion model's robustness and generalizability, future research may test it on various datasets, including those from other social media platforms or with various linguistic and cultural contexts.

Real-Time Sentiment Analysis: Although the accuracy of this study was the main focus, future work may examine how to best optimize these models for sentiment analysis in real-time. This would include striking a balance between computing efficiency and accuracy, increasing the models' suitability for real-time applications like market analysis, customer support, and social media monitoring.

Including Explainability in Multimodal Models: Understanding how multimodal models make decisions is crucial as these models get more complicated. In order to make sure that users can trust and understand the predictions made by fusion models, future research should investigate ways to improve the interpretability and explainability of these models.

In summary, although this research has contributed to the field's understanding of multimodal sentiment analysis, it also emphasizes how difficult it is to integrate disparate data sets. The knowledge acquired here lays the foundation for further study, the ultimate goal of which is to create multimodal sentiment analysis models that are more effective and efficient. The fusion model's meager gains imply that, although integrating modalities has potential, more study is necessary to realize its full potential. This chapter's recommendations give as a road map for further research, highlighting the necessity of creativity and meticulousness in the creation of next-generation multimodal learning methods.

REFERENCES

Anilkumar, B., Devi, N., Kotagiri, S., & Sowjanya, M. (2024). Design an image-based sentiment analysis system using a deep convolutional neural network and hyperparameter optimization. Multimedia Tools and Applications, 83, 1-20. https://doi.org/10.1007/s11042-024-18206-y

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications, 77, 236-246. https://doi.org/10.1016/j.eswa.2017.01.056

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. https://arxiv.org/abs/1409.0473

Cambria, E., Poria, S., Hazarika, D., & Vij, P. (2018). MELD: A multimodal multi-party dataset for emotion recognition in conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 527-536. https://aclanthology.org/P19-1050/

Chen, J., Zhang, Q., & Zhu, X. (2020). Visual sentiment analysis with active learning. IEEE Access, 8, 185899-185908. https://doi.org/10.1109/ACCESS.2020.3029042

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06). https://www.researchgate.net/publication/2395388_SentiWordNet_A_Publicly_Available_Lexical_Resource_for_Opinion_Mining

Gao, J., Wang, P., Li, W., & Xu, Z. (2020). A survey on deep learning for multimodal data fusion. Neural Computation, 32(5), 829-864. https://doi.org/10.1162/neco_a_01273

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, X., & Yamamura, M. (2022). Global Local Fusion Neural Network for Multimodal Sentiment Analysis. Applied Sciences, 12(17), 8453. https://doi.org/10.3390/app12178453

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708). https://doi.org/10.1109/CVPR.2017.243

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European Conference on Machine Learning (pp. 137-142). https://link.springer.com/chapter/10.1007/BFb0026683

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324. https://doi.org/10.1109/5.726791

Liu, Z., Shen, Y., & Klusowski, J. (2018). Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064. https://arxiv.org/abs/1806.00064

Lu, Y., et al. (2014). Integrating predictive analytics with social media. IEEE Conference on Visual Analytics Science and Technology (VAST), 193-202. https://doi.org/10.1109/VAST.2014.7042494

Mittal, N., Sharma, D., & Joshi, M. L. (2018). Image sentiment analysis using deep learning. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 684-687). Santiago, Chile. https://doi.org/10.1109/WI.2018.00-11

Niu, T., He, H., Song, L., & Sun, B. (2016). Sentiment analysis on multi-view social data. In Q. Tian, N. Sebe, G. J. Qi, B. Huet, R. Hong, & X. Liu (Eds.), MultiMedia Modeling (pp. 15-27). Springer, Cham. https://doi.org/10.1007/978-3-319-27671-7_25

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on empirical methods in natural language processing - Volume 10 (pp. 79-86). https://doi.org/10.3115/1118693.1118704

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. Proceedings of the First Instructional Conference on Machine Learning.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 417-424). https://doi.org/10.3115/1073083.1073153

Xu, J., Li, Z., Huang, F., Li, C., & Yu, P. S. (2022). Visual sentiment analysis with social relations-guided multi-attention networks. IEEE Transactions on Cybernetics, 52(6), 4472-4484. https://doi.org/10.1109/TCYB.2020.3027766

Xu, Z., Luo, M., Zhou, W., & He, L. (2019). Multi-ZOL: A dataset for bimodal sentiment classification of mobile phone reviews. IEEE Access, 7, 157944-157953. https://doi.org/10.1109/ACCESS.2019.2950299

Yang H, Zhang S, Shen H, Zhang G, Deng X, Xiong J, Feng L, Wang J, Zhang H, Sheng S. A Multi-Layer Feature Fusion Model Based on Convolution and Attention Mechanisms for Text Classification. Applied Sciences. 2023; 13(14):8550. https://doi.org/10.3390/app13148550

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1103-1114). https://doi.org/10.18653/v1/D17-1125

Zadeh, A., et al. (2018). Multimodal sentiment intensity and emotion recognition using CMU-MOSEI dataset. arXiv preprint arXiv:1810.02367. https://arxiv.org/abs/1810.02367

Zhang, D., Li, S., Zhu, Q., & Zhou, G. (2020). Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning. IEEE Access, 8, 1-1. https://doi.org/10.1109/ACCESS.2020.2969205

Zhang, K. E., Zhu, H., & Zhuang, Y. (2021). Cross-modal image sentiment analysis via deep correlation of textual semantic. Knowledge-Based Systems, 216, 106803. https://doi.org/10.1016/j.knosys.2021.106803

Zhou H, Zhao Y, Liu Y, Lu S, An X, Liu Q. Multi-Sensor Data Fusion and CNN-LSTM Model for Human Activity Recognition System. Sensors. 2023; 23(10):4750. https://doi.org/10.3390/s23104750

Zhu, C., Chen, M., Zhang, S., Sun, C., Liang, H., Liu, Y., & Chen, J. (2023). SKEAFN: Sentiment Knowledge Enhanced Attention Fusion Network for multimodal sentiment analysis. Information Fusion, 100, Article 101958. https://doi.org/10.1016/j.inffus.2023.101958

APPENDIX

Additional content:

LSTM Model Code:
https://colab.research.google.com/drive/1PpWAzNHMmEaQZZwZ5yR2K7jTfs0JWQmt?usp=sharing

Densenet-121 model code:

https://colab.research.google.com/drive/1B5LAu3PTGd7_1OFmF6qDdaYPcptjuSuq?usp=sharing

Fusion model code:

https://colab.research.google.com/drive/1QVoKi5cuNZhtQJPjS1BZbYSqu4hkS1N9?usp=sharing