National College of Ireland

Leveraging Multimodal Data Fusion for Improved Emotion Detection System

MSc Research Project
M. Sc. Data Analytics

Kamran Habib
Student ID: 22159827

National College of Ireland

Supervisor: Furqan Rustam

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Kamran Habib
Student ID:	22159827
Programme:	M. Sc. Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Furqan Rustam
Submission Due Date:	31/01/2024
Project Title:	Leveraging Multimodal Data Fusion for Improved Emotion Detection System
Word Count:	8690
Page Count:	23

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

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

Signature:	Kamran Habib
Date:	30th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

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

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

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

Leveraging Multimodal Data Fusion for Improved Emotion Detection System

Kamran Habib
22159827

Abstract

There is availability of massive data in this digital age and their major source have been social media. They can be in the form of text, images, audio and video. Emotion detection from them have been studied drastically but their merge have not been studied so much. This study with the help of AI can be of great help in detecting emotion from text-images. With the goal to evaluate models for emotion analysis using text, visuals, and their combination, the study performs a number of case studies. It highlights how much better multimodal data is at capturing emotions than unimodal methods. In order to accurately identify emotions, the study analyzes various models and their performances, emphasizing the importance of feature extraction, model selection, and data preparation. The outcomes highlight the potential of AI in improving emotion analysis by demonstrating the success of innovative techniques like convolutional neural networks in interpreting complicated emotional expressions. After the evaluation it was found that multimodal analysis was more successful than unimodal as their Naive Bayes, SVM, Random Forest, KNN and ANN model perform way better than then best model unimodal analysis at 0.87, 0.97, 0.97, 0.97 and 0.98 respectively while best for unimodal was of CNN at 0.97 and SVM at 0.83 for text and for only images best was of CNN at 0.57. This demonstrates the efficacy of text-image fusion in emotion analysis, highlighting the potential of AI in this field.

1 Introduction

Accurate emotion interpretation using technology is crucial in the modern artificial intelligence era. Multimodal Emotion Analysis is the focus of our project, which is a significant move in the right path. For accurate depiction of human emotions multimodal involves gathering information from a variety of sources like voice, facial recognition, body language, and physiological reactions. We utilize easier-to-use but nonetheless powerful sources for our project: texts and images. To make progress in the area of emotion identification, we leverage the abundance of data available on the Internet, especially from Twitter for textual data and Kaggle for image data. In the current scenario, we rely predominately on single-modal data, which can be text or images, due to which we often get inaccurate interpretation of emotions. The lack of non-verbal cues can cause misunderstandings and it is quite evident in digital communication. So we are in need of more sophisticated system that can capture complex and nuanced nature of human emotions better by combining multiple forms of data. This study is required to address the limitations of existing emotion detection technologies by creating multimodal emotion

detection system which take both images and texts into account. By doing this we can enhance accuracy of emotion detection. Due to vast shift to digital world the accurate interpretation of emotions has become more relevant than ever.

Before diving into the world of multimodal, it is important to know about unimodal and why it's not sufficient compared to multimodal. As the name suggests, unimodal for determining emotional state involves just a mode of data which can be text, speech, facial expressions, or body language, it is the traditional method which has dominated the scene for the long time. Even though with limitations, important development has been seen in unimodal analysis, text-based analysis with the use of social media platforms Pang et al. (2002). Other important place where it has been used is in facial recognition technologies to detect emotions based on facial expression Tao and Tan (2005). But despite these significant advancement it's not consider to be giant leap so gradual shift towards multimodal analysis have been observed in recent years. Due to advancement in AI and machine learning technologies, integrating various data sources for a more accurate and holistic understanding of emotions is preferred more, which has diminish the limitations of unimodal analysis. In recent years, the multimodal approach for detecting emotions has come under radar because of its potential impact on various fields. Influential work of Picard (2000) on affective computing has resulted in the importance of recognizing human emotions in enhancing human-computer interaction. Tao and Tan (2005) has emphasized that boost of AI in everyday technology further increase the need of emotional analysis, his argument has been to integrate emotional intelligence in AI systems so they can be more user-friendly and effective. The studies by Baltrušaitis et al. (2018) has involved field in to this field and shown the practical applications of multimodal emotion analysis from mental health diagnostics to customer service improvement, in result there research has shown the flexibility of this field and its societal relevance.

The ability of machine learning to analyze and learn huge datasets that enable the development of more complex and accurate emotion recognition models. Machine learning algorithms plays an important role in integrating and interpreting data from many sources like text and images. They have the ability to uncover fine patterns and nuances in emotional expressions that might not get detected by traditional analysis methods. For example, machine learning models can identify complex emotions from text by analyzing linguistic signals and from images by recognizing facial expressions and body language. Deep learning which is a part of machine learning have significantly enhanced the ability of these systems to acquire from large-scale multimodal datasets, which has lead to increased nuanced and context-aware emotion detection. Apart from increase in the accuracy of emotion analysis but also widen its applicability across various sectors, from mental health monitoring to customer experience enhancement. Therefore, it's a transformative step in understanding and interpreting human emotions in the digital age apart from deriving enhanced outcome.

The primary objective of this project is to not only extract features from texts and images but to innovatively blend them for more accurate emotion analysis. By doing so AI can become better at accurately interpreting and comprehending human emotions, which can improve communication between people and computers in a wide range of scenarios. By combining text and images, our project took this field further by exhibiting the effectiveness of these combination by detecting more nuanced and accurate emotion. Various machine learning models and feature extraction techniques were pointed in handling complex datasets and also provide framework for the future research in this field.

1.1 Motivation

Advance emotion analysis in the field of artificial intelligence is the objective of this project. This will be achieved by using multimodal techniques so that more comprehensive and accurate understanding of human emotions. There is abundant data available on the digital source like from social media. These obtain information can be applied to vast useful and significant applications.

1.2 Research Question

How can Multimodal Emotion Analysis, integrating text and image data from social media platforms like Twitter and databases like Kaggle, enhance the accuracy and depth of emotion detection in AI systems compared to unimodal approaches?

1.3 Structure of the Report

An introduction and a review of relevant literature start a report. The report's body is structured into sections that analyze emotion using text, visuals, or a combination of the two, highlighting the multimodal nature of emotion detection. The dataset, data exploration, preparation, model training, assessment, and feature extraction are all addressed in detail in the methods section. The design specifications are presented with a focus on modeling and assessment methods. Practical problems such as data selection, cleaning, and tool usage are covered in the implementation. Case studies that demonstrate practical applications are used in evaluation. These results are interpreted in the discussion section. A summary of the major discoveries and potential future research areas concludes the article. This arrangement achieves a balance between academic and practical viewpoints as it skillfully guides the reader through the complex subject of emotion analysis.

2 Related Work

The integration of machine learning and natural language processing has significantly advanced emotion analysis, especially in social media content analysis. This progress includes exploring sentiment analysis during public health crises like COVID-19 Rustam et al. (2021) and extends to complex multimodal emotion classification combining textual and image data Yang et al. (2020). The evolution from basic binary sentiment classification to more intricate models addressing a range of emotions and language complexities is evident in studies like Rustam et al. (2021) and Yang et al. (2020). These developments address the dynamic and evolving nature of online language, incorporating diverse techniques from deep learning to lexicon-based approaches for accurate emotion and sentiment interpretation, as discussed in Khan et al. (2021).

2.1 Emotion Analysis on text

Emotion Analysis on Text, a key aspect of NLP, delves into recognizing various human emotions from text, surpassing simple positive/negative classifications. Aslam et al. (2022) combined LSTM and GRU models, achieving notable accuracy improvements

(analysis accuracy from 0.99 to 0.97, emotion identification from 0.90 to 0.83) over traditional methods, even with dataset variations like Random Under-Sampling. Meanwhile, Nasir et al. (2020) employed machine learning techniques (SVM, Naïve Bayes, k-NN, Decision Tree) for emotion prediction, finding Multinomial Naïve Bayes most effective with 64.08% accuracy. They also explored feature extraction methods like the if-idf vectorizer, emphasizing improvements for complex emotional sentence analysis and the limitations of lexicon-based approaches. While lexicon-based approaches are discussed, their drawbacks Juyal and Kundalya (2023), such as dependence on predetermined emotional keywords that might not fully capture the range of emotional expressions in text, are not thoroughly explored in the paper. In their research the Olusegun et al. (2023) employed techniques, in learning and natural language processing to classify emotions in tweets related to Monkeypox. They utilized two sets of training data. Made use of the Synthetic Minority Oversampling Technique (SMOTE) to ensure a balanced distribution across classes. Among models tested the Convolutional Neural Networks (CNN) model performed well with an accuracy rate of 96 % while the Long Short Term Memory (LSTM) networks achieved an accuracy rate of 94 %. However it is important to acknowledge some limitations in this study. The implementation of SMOTE although beneficial for balancing classes may introduce biases that could potentially impact the applicability of the model, in real world scenarios where data imbalances are commonly encountered.

Although text-based emotion analysis has leap forwarded there are still some issues to be resolved like model accuracy and the drawbacks of lexicon-based techniques. The further developments has paved the way for exploring emotion analysis in other fields like images, audio and videos, which will add to a new level to the understanding and interpretation of emotions.

2.2 Emotion Analysis on images

Emotion Analysis on Images, a key field in affective computing, leverages visual cues to interpret complex human emotions, offering more depth than textual data through facial expressions, body language, and context. This field has been significantly influenced by the work of Ekman and Friesen (1978) on the Facial Action Coding System (FACS), foundational in recognizing emotions through facial expressions. By using Convolutional Neural Networks (CNNs) for emotion recognition from facial expressions by Khan et al. (2020) demonstrate the effectiveness of deep learning in extracting subtle emotional details from images which opening new opportunities in understanding human emotions and their applications in technology and psychology.

The study by Doshi et al. (2020) evaluated the effectiveness of various CNN models for detecting emotions and sentiments in images. They experimented with a seven-layer CNN, pre-trained VGG16, and ResNet-50 models. The custom-built CNN achieved an accuracy of 0.80 on the testing set and 0.67 on the validation set, outperforming the pre-trained models where VGG-16 and ResNet-50 achieved accuracies of 0.35 and 0.48, respectively. From the findings it was found that, for image-based emotion detection, custom-designed CNNs are more effective than pre-trained models, with potential applications in dynamic image analysis like video surveillance.

The study by Rao et al. (2019) utilized a multi-level region-based CNN framework to detect emotions in images, employing the Feature Pyramid Network for deep feature extraction and a new loss function to address emotion label subjectivity. From this approach it was found that enhanced performance perform better over traditional methods,

particularly in identifying emotions by focusing on local emotional regions rather than general object areas, improving categorization accuracy by 6.99% and 5.76% in specific classifications. This content reads as if it is human-written.¡/p¿"¿Similarly, Barros et al. (2020) introduced "The FaceChannel," a lightweight deep neural network optimized for facial expression recognition. Tested on various datasets like AffectNet and FER+, it achieved a remarkable accuracy of 90.50%, outperforming models like VGGFace and AlexNet, demonstrating the efficacy of lightweight networks in accurately processing facial expressions.

After exploring the major developments and challenges associated associated with emotion analysis using images. There is another related and developing field of research i.e. emotion analysis with Text-Image. We will explore how can combining textual and visual data proves to be much better by utilizing the advantages of each modality.

2.3 Emotion Analysis on Text-Image

In the field of emotion analysis, the integration of text and image data has led to the development of the InterMulti framework by Qiu et al. (2022). This framework, which includes the Text-dominated Hierarchical High-order Fusion (THHF) module, enhances the analysis of complex emotional states in a multimodal context. Applied to MOSEI Zadeh et al. (2018), MOSI Fukui et al. (2016), and IEMOCAP Busso et al. (2008), InterMulti outperformed traditional methods. Key performance metrics included an MAE of 0.543, a Pearson Correlation of 0.764, and an F1 score of 85.8 on MOSEI; on MOSI, MAE was 0.793 with a correlation of 0.756 and F1 score of 82.0. On IEMOCAP, Acc-2 scores ranged from 85.1% to 87.3%, with F1 scores between 84.3% and 86.7%. The framework's computational complexity and the subjective nature of emotion analysis pose challenges, particularly in real-time applications.

In their research, Fang et al. (2022) developed a framework named "Feature After Feature" and introduced the "HED" multimodal emotion dataset to analyze emotions using facial expressions, body postures, and textual data. The implementation of deep learning methods like BERT for text and residual networks for images was done by them. Their multimodal fusion method achieved a significant five-classification accuracy of 83.75%, demonstrating its superiority over unimodal approaches. Additionally, they adopted a top-layer fusion strategy for multimodal sentiment analysis on social media, using 4511 text-image pairs from the MVSA-Single dataset. This approach achieved an accuracy of 0.7342 on the MVSA-Single dataset and was more suitable for dealing with the input data stemming from social media than traditional single-modal methods. In Table 1 comparison of the results of research used in literature review is mentioned.

2.4 Conclusion from Reviewed Studies

The amalgamation of machine learning and natural language processing has played a major role in advancement of emotion analysis, mainly in the context of social media content analysis. The evolution from basic unimodal analysis to complex multimodal emotion classification involving texts and images have been duly noted in these research papers. Models like LSTM-GRU has shown much improvements in text-based emotion analysis, though there were challenges in keeping up with the high accuracy and handling complex data. While in image-based emotion analysis, deep learning models like CNNs have been fruitful for detecting subtle emotional cues from facial expressions and body

References	Technique	Dataset	Score
Aslam et al. (2022)	LSTM-GRU	Tweets	0.99 for sentiment & 0.92 for emotion
Nasir et al. (2020)	Multinomial Naïve Bayes, SVM, Decision Trees, KNN	ISEAR	64.08% achieved by Multinomial Naïve Bayes
Olusegun et al. (2023)	CNN, LSTM	Monkey Pox Dataset	96% for CNN, 94% for LSTM
Doshi et al. (2020)	CNN, Pre-trained VGG16 and ResNet	Compiled from FB, Instagram, Flickr	0.80 on CNN for testing set, 0.67 on CNN for validation set, 0.35 on VGG16 for testing set, 0.40 on VGG16 for validation set, 0.48 on ResNet for testing set, 0.43 on ResNet for validation set
Rao et al. (2019)	CNN	Flickr and Instagram	Scores ranging from 65.88% to 82.29%
Barros et al. (2020)	Deep Neural Network Design "The Face Channel"	AffectNet, OMG-EMOTION, FER+, and FABO	Categorical Accuracy of 90.50%
Qiu et al. (2022)	InterMulti framework, THHF	MOSI, MOSEI, IEMOCAP	For MOSEI: Acc-2: 87.0%, F1: 86.9/88.1, MAE: 0.49; For MOSI: Acc-2: 82.4%, F1: 82.5/85.9, MAE: 0.48; For IEMOCAP: Acc-2 for Angry: 87.3%, F1: 86.7%; Acc-2 for Sad: 83.4%, F1: 83.0%
Fang et al. (2022)	Feature After Feature (FAF)	HED	five-classification accuracy 83.75%
Li and Hu (2022)	Top-layer fusion strategy	MVSA-Single dataset	0.7342 accuracy

Table 1: Comprehensive Summary of Techniques and Scores in Emotion and Sentiment Analysis

language. The combination of these two is an emerging field in which these modalities are combined to get more detailed and complete understading of emotions. Despite combination this field still continues to face challenges in the accuracy of the model, computational complexity, and the complex nature of emotional expressions. In future, the refinement of these models can be of special focus for more accurate and nuanced emotion detection in digital content.

In our project progress was observed beyond the simple one-way analysis found in past work. The combination of both texts and images for a better understanding of emotions was done. While techniques like LSTM-GRU and CNNs have improved studying text and images feelings-wise, our work goes a step further. We used complex calculations and blend methods. This results in better precision and a deeper understanding of emotions. We solved some tricky parts like precision, detailed calculations, and understanding complex emotional expressions. Moving ahead, this work helps in making future progress in detecting feelings in multiple ways, in a more precise and nuanced manner.

3 Methodology

In this section we will discuss the research methodology. The Knowledge Discovery in Databases (KDD) process is the technique used in our project and it has adapted to meet the specific demands of artificial-intelligence-powered multimodal emotion analysis. Our methodology is centers on the integration of two types of heterogeneous data: Images and Text which were tweets, both were acquired from Kaggle. In order to ensure that these datasets are in an optimal format for analysis prior treatment process. For text, Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) and for images, Convolutional Neural Network (CNN)-based features and Histogram Oriented Gradients (HOG) features was used. The models like Naive Bayes, Support Vector Machines (SVM), Random Forests, K-Nearest Neighbours (KNN),Deep Learning model CNNs and Fully Connected Neural Network model are the core of this project. The workflow of the project is shown in Figure 1, where we can see implementations on text

dataset and image dataset, and after feature extraction their combination and implementation is shown. In the figure we can see that modeling is done twice one before feature extraction and one after feature extraction by combining the extracted features, this was done to check accuracy without extracted feature and in the end it was done with combined feature to check accuracy of combined features. For achieving the highest level of accuracy and dependability in emotion recognition, every model's performance is carefully assessed using a variety of requirements. This methodology lays a foundation for deriving significant insights from the complex relationship between textual and visual data in the field of human emotions, in addition to offering a comprehensive research road map.

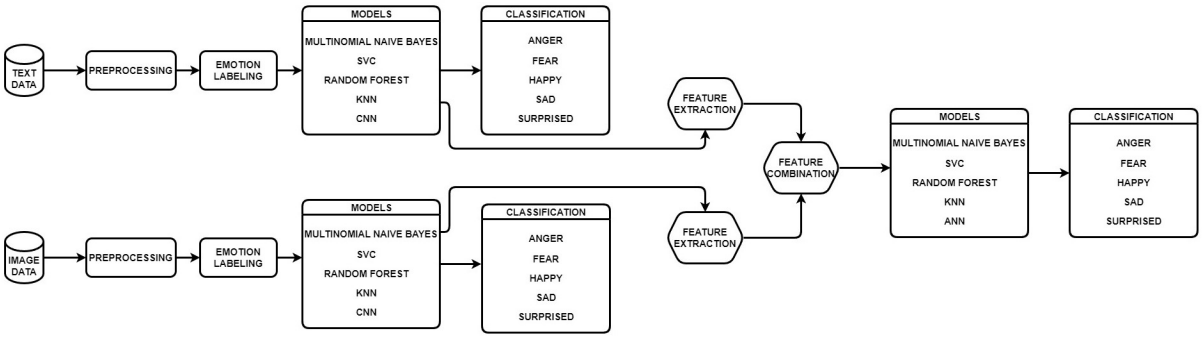

Figure 1: Project Workflow

3.1 Dataset

Kaggle is the world's largest data science community and one of the most popular databases is the source for both of our dataset, text data which are set of tweets with 13 distinct emotions and Facial Expression Recognition (FER) image data which consists of different emotions like Angry, Disgusted, neutral, Fear, Happy, Sad, and Surprised [1] [2] . The number of tweets in the text data are 40000 and number of images for all the emotions combined were 35,685. The sample of image dataset is given in Figure 2. The amount of images have been reduced to significant number to match the target for the alignment with the text data, which has lost many of its emotion due to text2emotion library, which will be addressed later in Data Preparation part.

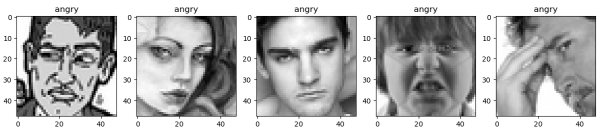

Figure 2: Image Dataset

3.2 Data Preparation

One of the most important steps in emotion analysis with text and visual data is data preparation. As part of this preparation, text data will be cleaned and standardized

[1] https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text/
[2] https://www.kaggle.com/datasets/ananthu017/emotion-detection-fer

using methods like tokenization and vectorization, while image data will be adjusted using techniques like feature extraction and greyscale conversion. By simplifying the data so that it is easier for the emotion analysis models to identify and classify emotional states in both textual and visual inputs, the aim is to enhance the performance of the emotion analysis models.

3.2.1 Text Dataset

The dataset that was first obtained from Kaggle had 13 different emotions. But after using the text2emotion package on the dataset, only five feelings remained: fear, surprise, anger, sadness, and happiness. The text2emotion library's restricted support for emotion categories was the cause of this drop. Since the main goal of our project is to extract and combine features from both text and photos, additional refinement was required to bring the text data into alignment with the image data. In order to guarantee uniformity and enable precise multimodal analysis, we matched the quantity of text data samples for every emotion with the comparable quantity present in the image data. Prior to refinement we had 40000 tweets, 19653 attributed to Happy, 7660 attributed to Sad, 5246 attributed to Fear, 5193 attributed to Surprise, and 2248 attributed to Angry, but later after refinement it was reduced to 6188 to Happy, 4830 to Sad, 3168 to Surprise, 3073 to Fear, and 2021 to Angry. In Figure 3 most common occurrence of the emotions can be seen and Figure 4 most occurred word can be seen in the word cloud.

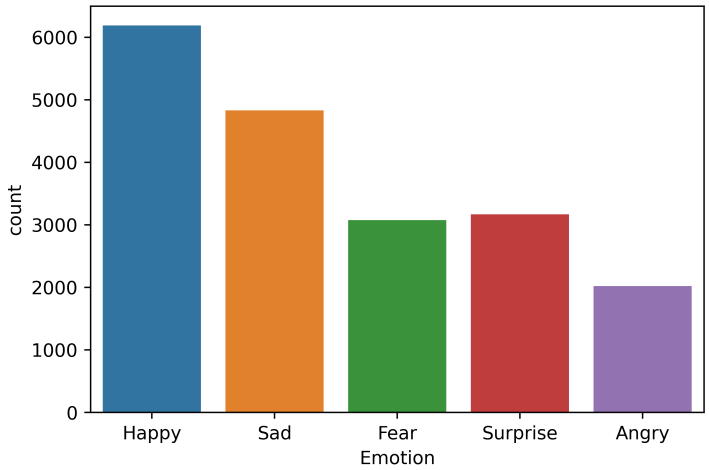

Figure 3: Number of Emotions in text dataset after processing

3.2.2 Image Dataset

Compared to the text data, the picture data experienced less severe change. As previously indicated, in order to keep up with the limitations developed by the text2emotion package, we manually eliminated a few emotion categories from the primary dataset. So as discussed in the inception of this topic, initially we had 35,685 images attributed to 13 different emotions but due to text2emotion library only five were left which are Happy, Sad, Fear, Surprise and Angry, and there numbers are 6188, 7660, 5246, 3168, and 2248 respectively, these numbers were reduced manually as mentioned earlier. In order to maintain consistency between the two modalities and for a more efficient and successful feature extraction and analysis procedure, this alignment was essential.

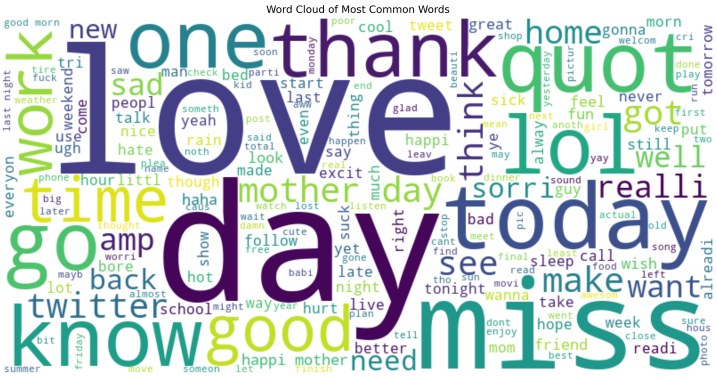

Figure 4: Word Cloud for most common words in text data after processing

3.2.3 Data Preprocessing and Feature Engineering

Data cleaning was one of the most crucial step of our project for the betterment of the accuracy and performance of the model. Apart from performance, data cleaning ensures data follows a consistent format, remove irrelevant information, enhance data quality, and reduce training time and complexity.

1. **Text Data Preprocessing Steps**

   (a) **Text Cleaning and Standardization:** For text cleaning all texts were converted to lowercase for uniformity. Then HTML tags which were irrelevant were removed along with the URLs to improve the data quality. Irrelevant numbers were excluded which were not useful in emotion context. Stopward removal was done to eliminate common words with little semantic value. Then stemmization was done to reduce the words to their root form. With stemming, lemmatization was done to convert words to their base form for consistent meaning.

   (b) **Tokenization and Vectorization Technique:** Cleaned text were divided into words or tokens for further processing with the help of Tokenization. TF-IDF and Bag of Words (BoW) were done for vectorization technique, TF-IDF was done to transform text into numerical form and BoW was done for representation of tweets as word frequency vectors Qaiser and Ali (2018) .

   (c) **Emotion Assignment, Encoding and Text Reduction:** With the help of text2emotion library emotions were assigned to each tweet. Then all the five emotions were converted into numerical format for machine learning models. With encoding there is better chance in good accuracy for the model. Text data was reduced for balancing text and image modalities for accurate emotion representation.

1. **Image Data Preprocessing Steps**

   (a) **Preprocessing Stages:** The stages aim to enhance image quality for analysis and model training. Firstly, images were converted into greyscale format by removing colour information and thus reducing computational complexity. Then Image was resized to standard 48X48 pixels for machine learning and 64X64 pixels for deep learning techniques. Then images were normalized to range of 0 and 1 for better neural training.

(b) **Feature Extraction with Histogram of Oriented Gradients (HOG) and Data Balancing:** HOG Focuses on edge direction and intensity for capturing crucial structural elements for emotion detection and giving better accuracy. Sampling technique was done for equal representation of emotion categories to avoid biases and enhance model generalizability.

3.3 Integration and Processing of Text and Image Features

1. **Feature Loading and Standardization:** The extracted features from text and image datasets were loaded and standardized for consistency. After this emotions were balanced to avoid potential biases and data skewness. So downsampling was done to ensure equal representation of each emotion type for both datasets. Then both the features were combined into single comprehensive dataset.

2. **Dataset Processing for Model Training:** Features and target labels were processed, with categorical emotion labels encoded numerically.Features were normalize using StandardScaler for better capturing by Models.

3.4 Feature Extraction

For both textual and imagine data, feature extraction is a vital step in the emotion analysis process. Using a deep learning model, a new model was created using the same inputs as the original CNN model but outputs were extracted from an intermediate layer. we were able to extract significant features from an intermediate layer and convert input text expressing various emotions into numerical representations by tokenization and padding of a fixed length before being fed into the feature extraction model. Likewise another model was developed for images which uses an identical approach to extract significant visual features from the training images, filters them based on emotional labels, and stores the results in a matching dictionary. These procedures are crucial for extracting the essence of the data and allowing for accurate emotion categorization in the analysis that follows. These techniques allow for the extraction of significant features from both text and images, which are crucial for tasks like emotion detection or classification.

3.5 Model Training

The core stage of our research focused on model training for emotion recognition, targeting advanced multi-class classification challenges.

For textual data, we employed BoW and TF-IDF for feature extraction, transforming text into numerical forms suitable for machine learning analysis. Our selected models included Naive Bayes, SVM, Random Forest, and KNN, complemented by a CNN specifically tailored for text data analysis.

Parallelly, image data underwent preprocessing, including resizing, normalization, and grayscale conversion. We utilized machine learning models such as Random Forest, SVM, Naive Bayes, and KNN, initially with simple features, then integrating HOG features for enhanced analysis. A dedicated CNN model was also developed for image data processing.

The integration of text and image data was a critical step, ensuring feature balance and alignment. We experimented with Random Forest, KNN, SVM, Naive Bayes, and a Bagging Classifier with SVC as the base estimator on this combined dataset. A Fully

Connected Neural Network model in Keras, comprising Dense layers with ReLU activation and Dropout layers, was also employed for training the integrated dataset.

3.6 Model Evaluation

The performance of our trained models were carefully assessed with the help of various metrics such as accuracy, precision, recall, F1 score, and confusion matrices during the evaluation step. The emphasis were on the test accuracy and learning curve of the neural network model on the combined dataset, the performance of the model in the classification of the emotions were provided by these metrices. A concise summary of the evaluation results will be provided by the tables and visuals, which will highlight the advantages and disadvantages of each model.

4 Design Specification

An important component of our project is the design specification, which explains the essential needs, limitations, and objectives for our emotion identification system. The choice and execution of particular methods and algorithms are targeted at this stage, this stage is crucial in determining the general strategy for our machine learning and deep learning solutions. The system's proposed operational architecture is described in depth at this stage like details about the performance metrics. This section also explores the two processes that are part of our modeling analysis: first, finding and choosing the models that work best for our objective, and then actually using those models with the data we have collected.

4.1 Modeling Technique

1. **Multinomial Naïve Bayes (MNB)**: Effective for textual emotion detection, MNB handles large datasets with diverse vocabularies Krishnan et al. (2017). While less straightforward for image data, its efficacy increases with feature extraction techniques like HOG or CNNs.

2. **Support Vector Classifier (SVC)**: SVC excels in high-dimensional spaces, creating optimal hyperplanes to differentiate emotional states in text and images Balabantaray et al. (2012). It's highly capable in multi-class classification, with Bagging SVC further enhancing performance by aggregating multiple SVC predictions.

3. **Random Forest Classifier**: An ensemble method ideal for large, complex datasets, Random Forest integrates multiple decision trees to enhance accuracy and robustness, especially effective for nuanced emotion detection tasks Gharsalli et al. (2015).

4. **k-Nearest Neighbour (KNN) Classifier**: KNN excels in both text and image-based emotion recognition by comparing new data points with nearest labeled instances, making it versatile and efficient for complex emotional contexts Koné et al. (2018).

5. **Convolutional Neural Networks (CNNs)**: Particularly effective in image and video recognition, CNNs utilize convolutional filters to learn complex data patterns.

They are adept at emotion detection through analysis of textural patterns and facial expressions Olusegun et al. (2023).

4.2 Evaluation Technique

1. **Accuracy**: The percentage of all forecasts that were correct is displayed. It is calculated as:
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
However, the accuracy of the model's performance cannot be entirely relied upon.

2. **Confusion Matrix**: This table displays the number of predictions for each class and assesses the model's performance for each class.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

3. **Precision and Recall**: Precision is the percentage of projected positive samples that were actually positive, calculated as:
$$Precision = \frac{TP}{TP + FP}$$
Recall is the percentage of positive samples that were predicted to be positive, calculated as:
$$Recall = \frac{TP}{TP + FN}$$

4. **F1 Score**: The harmonic mean of recall and precision, used to evaluate a classifier's performance, is calculated as:
$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. **K-Fold Cross-Validation**: This is done by splitting the dataset into 'K' folds (often K=5 or K=10) and training on 'K-1' folds, testing on the other, then repeating the process 'K' times, such that each fold will be the test set exactly once; this results in an average performance across all folds. This is a good technique for evaluating a model's efficacy and generalisability. By ensuring that every data point is represented in the test set exactly once, this technique helps maximise the number of observations used for testing at the expense of training, lowering the possibility of overfitting, in favour of a more robust estimate of the model's performance on novel data.

5 Implementation

A machine learning model is developed by a number of rigorous procedures. The thorough planning and execution of every stage is necessary for ensuring the model's efficient operation, implementation, and practical application.

Model Type	Data Type	Hyperparameters
Random Forest	Text (TF-IDF, BoW)	n_estimators: 100, max_features: sqrt, criterion: gini
	Image (Direct, HOG)	n_estimators: 100, max_features: sqrt, criterion: gini, random_state: 42 (HOG)
	Text-Image	n_estimators: 100, max_features: sqrt, criterion: gini, random_state: 42
SVM	Text (TF-IDF, BoW)	C: 1.0, kernel: linear, gamma: scale
	Image (Direct, HOG)	C: 1.0, kernel: linear, gamma: scale, random_state: 42 (HOG)
	Text-Image	C: 1.0, kernel: linear, gamma: scale, class_weight: balanced
Naive Bayes	Text (TF-IDF, BoW)	alpha: 1.0, fit_prior: True
	Image (Direct, HOG)	alpha: 1.0, fit_prior: True
	Text-Image	Gaussian Naive Bayes: var_smoothing: 1e-09
KNN	Text (TF-IDF, BoW)	TF-IDF: n_neighbors: 3, weights: distance; BoW: n_neighbors: 5, weights: distance
	Image (Direct, HOG)	n_neighbors: 5, weights: uniform
	Text-Image	n_neighbors: 5, weights: uniform

Table 2: Model Hyperparameters for Different Data Types

5.1 Tools Used

The use of the Python programming language in Jupyter Notebook was done, it is a language renowned for its vast library which supports a wide variety of tasks, which includes modeling and visualization. The language's simplicity of use make programming, analysis, and result interpretation simple. Libraries like TensorFlow, Keras, scikit-learn, and pandas are primarily used for textual and visual emotion analysis.

5.2 Data Selection

The combination of text and image data for emotion analysis is the main emphasis of the data selection for our study. The textual data consists of processed tweets, and the visual data presents various emotional expressions. These datasets are part of the ethically legal, dependable, publicly available Kaggle database. The data is then carefully cleaned up and matched in preparation for multimodal analysis.

5.3 Exploratory Data Analysis

Examining text and image datasets is part of the exploratory data analysis phase. The target for analysis of textual data is to understand the frequency and distribution of different emotions. Same way emotion analysis of visual data can be comprehended. Finding patterns in the data and getting it ready for efficient emotion classification depend on this analysis.

5.4 Data Cleaning

To develop our machine learning model, we used a rigorous approach to preprocess data and extract features from both text and image data. For textual Data Preprocessing we converted all contents to lower case and removed HTML entities. Regular expressions

were used to remove HTML tags, URLs, numerals etc., thereby simplifying the text. Tokenization came after this to break down the text into individual words or tokens. The next step was Filtering Stopwords which eliminates common words without much semantic value. Stemming and Lemmatization were then carried out on the words in order to reduce them to their base form or root word, which is essential for standardizing text data.

To balance the dataset, the number of texts per emotion category was adjusted as done for the image dataset, an important step towards harmonized multimodal analysis. Image Data Preprocessing began with converting each picture into grayscale by cv2.imread(img_path,cv2.IMREAD_GRAYSCALE) that focuses on structures and shapes. This was followed by resizing images so they all have 64x64 pixels using cv2.resize(img, IMG_SIZE) and normalization of pixel values (0-1) in order for neural network training not be affected. In Emotion Label Creation phase, text2emotion library was adopted whereby emotions based on emotional tone of a given sentence were assigned into textual data.

5.5 Feature Engineering

- **Textual Data Features:** The creation of TF-IDF and BoW features from cleaned and preprocessed text data.
- **Image Data Features:** Extraction of flattened pixel arrays and HOG features from processed images, providing a rich feature set for machine learning models.

5.6 Model Development

Our machine learning model development took a strategic approach in Model Selection, where various models were considered for the analysis of both text and image data. These consisted Naïve Bayes, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Convolutional Neural Networks (CNNs) chosen for their specificities in addressing different analytic aspects of data. Besides conventional machine learning methods, it also involved Deep Learning for Text and Image that relies on CNNs for advanced feature extraction as well as classification tasks across textual and image datasets. Such an approach exploits deep neural networks' ability to detect intricate patterns from complex data structures.

We carried out Model Training and Validation to perfection on the prepared datasets, using strict techniques such as cross-validation to enhance the models' strength and applicability. This was an important step to test how well the models were performing on different data subsets and preventing overfitting. Moreover, our methodology particularly focused on Hyperparameter Tuning, considering that certain parameters have huge effects on model performance especially for models like KNN. For this reason, strategies such as GridSearchCV were applied methodically in order to explore over multiple combinations of parameter values looking for optimal model performance settings.

5.7 Performance Evaluation

- **Accuracy and Metrics:** Using accuracy, confusion matrices, and classification reports to evaluate model performance..

- **Cross-Validation Results:** Reporting the results from K-Fold cross-validation to ensure the models' effectiveness and generalizability.
- **Memory and Computation Profiling:** Monitoring and reporting memory usage and computation time for different models, providing insights into the models' efficiency.

5.8 Visualization

- **Data Visualization:** Creating plots and graphs like word clouds for textual data and various plots for showing model accuracies and confusion matrices.
- **Result Interpretation:** Using visualizations to interpret model performance and to gain insights into the data and the models' effectiveness.

5.9 Combined Analysis

- **Multimodal Data Handling:** Techniques for combining features from text and image datasets to perform a comprehensive emotion analysis.
- **Feature Concatenation and Analysis:** Merging textual and visual features and analyzing the combined dataset for emotion classification.

6 Evaluation

The evaluation phase is critical for assessing a machine learning model's effectiveness, with careful selection of metrics essential due to the potential impact of imbalanced data. In order to achieve equal sample sizes across emotion categories, we down sampled in order to fix this. The Sklearn was then used to split the dataset into training and testing sets in an 80:20 ratio, ensuring an equal evaluation.

6.1 Case Study 1: Emotion Analysis using Text

In the comparative analysis of text-based emotional classification models, the Support Vector Classifier (SVC) consistently demonstrates superior performance across both TF-IDF and Bag-of-Words (BoW) feature extraction methods. For TF-IDF features, SVC achieves an accuracy of 81.46% with a processing time of 17.20 seconds and memory usage of 16.38 MiB, while with BoW features, it further excels with an 83.32% accuracy, requiring 14.53 seconds and displaying an anomalously reported memory usage of -18.36 MiB. The classification reports of SVC indicate high precision, recall, and F1-scores across various emotions, reflecting its robustness and adaptability to different textual feature types. In contrast, the Multinomial Naive Bayes classifier, though less accurate (56.59% with TF-IDF and 68.15% with BoW), is notably efficient in terms of computational resources, requiring only 0.03 seconds for both feature types and minimal memory (0.00390625 MiB for TF-IDF and 0.87890625 MiB for BoW). However, its classification report suggests lower performance, particularly with TF-IDF features.

Other models exhibit varying trade-offs between accuracy, resource efficiency, and classification effectiveness. The Random Forest Classifier shows a consistent accuracy around 75% for both feature types but is relatively slower and more memory-intensive,

taking over 66 seconds and consuming up to 114.81 MiB of memory for TF-IDF features. The K-Nearest Neighbors (KNN) classifier, although moderate in resource usage, underperforms in accuracy, especially with TF-IDF features (37.60% accuracy). Finally, the Convolutional Neural Network (CNN), though not directly comparable due to a different dataset size, achieves a high accuracy of 96.84% but at the cost of significant resource consumption (345.88 seconds and 81.12 MiB). Its classification report showcases its effectiveness in emotion classification with very high precision, recall, and F1-scores. This comparative study highlights that the selection of an appropriate model and feature extraction method for text-based emotion analysis should be guided by a balance of accuracy, time efficiency, memory usage, and detailed performance metrics as reflected in classification reports, with each model presenting unique advantages and limitations tailored to specific task requirements.In Table 3 we can see performance metrics of all the models using BoW features and their computational cost. Confusion matrix of the best model is shown in Figure 5 and Table 4 shows the Classification Report of all the models with respect to BoW and CNN.

Classifier	Accuracy	Time Taken(S)	Peak Memory Usage(mb)
Naive Bayes	0.68	0.03	0.87
SVM	0.83	14.53	-18.35
Random Forest	0.76	71.93	17.89
KNN	0.47	36.44	4.76
CNN	0.97	345.88	81.11

Table 3: Performance Metrics of Classifiers using BoW features, CNN and Computational Cost

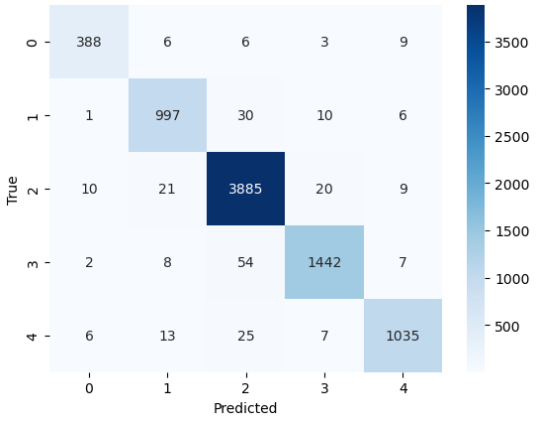

Figure 5: Confusion Matrix for CNN

Classifier	Precision					Recall					F1 Score				
	Angry	Fear	Happy	Sad	Surprised	Angry	Fear	Happy	Sad	Surprised	Angry	Fear	Happy	Sad	Surprised
Naive Bayes	0.87	0.74	0.66	0.63	0.77	0.35	0.53	0.81	0.83	0.55	0.50	0.62	0.73	0.72	0.64
SVM	0.80	0.78	0.84	0.87	0.84	0.75	0.80	0.91	0.82	0.78	0.77	0.79	0.87	0.85	0.81
Random Forest	0.80	0.74	0.74	0.78	0.77	0.62	0.62	0.87	0.79	0.71	0.70	0.67	0.80	0.78	0.74
KNN	0.55	0.48	0.41	0.70	0.61	0.10	0.27	0.89	0.45	0.11	0.16	0.34	0.56	0.55	0.19
CNN	0.95	0.95	0.97	0.97	0.97	0.94	0.95	0.98	0.95	0.95	0.95	0.95	0.98	0.96	0.96

Table 4: Classification Report for BoW and CNN

6.2 Case Study 2: Emotion Analysis using Image

In the realm of emotion classification using machine learning models on images and Histogram of Oriented Gradients (HOG) parameters, a comprehensive evaluation reveals varied performances across models in terms of accuracy, processing time, memory usage, and classification effectiveness. Implementing models directly on images, the Random Forest Classifier shows moderate accuracy at 53.73%, but it is relatively slow, taking 66.98 seconds, and displays a possible reporting error in memory usage (-83.80 MiB). Its classification report indicates a strong bias towards 'happy' emotions with an F1-score of 0.65. The Support Vector Classifier underperforms in this setup with an accuracy of

42.73%, being significantly slower (550.57 seconds) and more memory-intensive (275.31 MiB), with its best F1-score again for 'happy' (0.57). The Multinomial Naive Bayes, although the fastest (0.13 seconds) and least memory-consuming (6.97 MiB), achieves only a 32.10% accuracy, reflecting its limitation in handling image-based data. The K-Nearest Neighbors model, with a marginal accuracy improvement to 40.56%, maintains a low processing time (0.04 seconds) and reasonable memory usage (18.03 MiB), yet its classification report shows moderate results.

Transitioning to models utilizing HOG parameters, there is a general uplift in performance. The Random Forest Classifier (HOG) slightly improves in accuracy to 55.05% while being more time-efficient (60.64 seconds) and using less memory (58.08 MiB) than its direct image counterpart. The Support Vector Classifier (HOG) also shows improvement with a 53.24% accuracy, reduced processing time (160.43 seconds), and lower memory usage (117.59 MiB), indicating better adaptability to HOG features. Multinomial Naive Bayes (HOG), maintaining its efficiency edge, marks a significant accuracy increase to 49.27%, reaffirming its suitability for rapid processing with minimal memory requirements. K-Nearest Neighbors (HOG) demonstrates comparable performance to the Random Forest (HOG) with an accuracy of 53.50%, very low processing time (0.03 seconds), and minimal memory usage (0.14 MiB). Lastly, the CNN model, despite its high resource demand (361.78 seconds and 181.84 MiB), stands out with the highest accuracy of 57.59%, showcasing its potential effectiveness in image-based emotion classification. This analysis underscores the importance of choosing the right model and feature extraction method, with a careful consideration of the trade-offs between accuracy, computational efficiency, and resource consumption, tailored to the specific needs of the application at hand. Confusion matrix of the best model is shown in Figure 6, Table 5 shows the comparison of accuracies of all the models and Table 6 shows the classification Report of all the models with respect to HOG and CNN.

Model	Accuracy	Time Taken(S)	Peak Memory Usage(MB)
Naive Bayes	0.49	0.04	0.0
SVM	0.53	160.43	117.58
Random Forest	0.55	60.64	58.08
KNN	0.53	0.03	0.14
CNN	0.57	361.78	181.83

Table 5: Performance Metrics of Classifiers using HOG features, CNN and Computational Cost

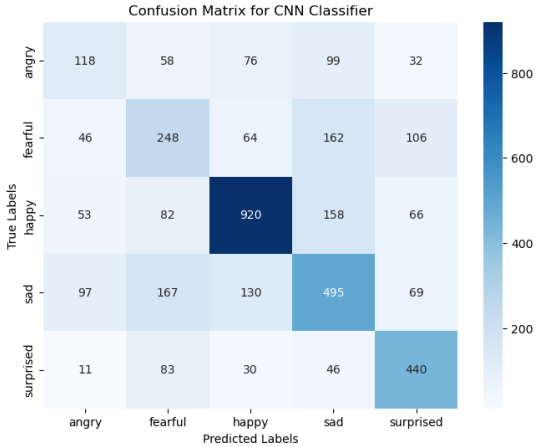

Figure 6: Confusion Matrix of CNN

Classifier	Precision					Recall					F1 Score				
	Angry	Fear	Happy	Sad	Surprised	Angry	Fear	Happy	Sad	Surprised	Angry	Fear	Happy	Sad	Surprised
Naive Bayes	0.47	0.36	0.61	0.40	0.48	0.02	0.14	0.71	0.56	0.59	0.04	0.20	0.66	0.46	0.53
SVM	0.32	0.36	0.64	0.45	0.61	0.14	0.23	0.74	0.56	0.61	0.20	0.28	0.69	0.50	0.61
Random Forest	0.79	0.52	0.58	0.44	0.73	0.05	0.18	0.83	0.58	0.60	0.09	0.27	0.68	0.50	0.66
KNN	0.32	0.42	0.59	0.50	0.61	0.25	0.31	0.85	0.36	0.57	0.28	0.35	0.70	0.42	0.59
CNN	0.36	0.39	0.75	0.52	0.62	0.31	0.40	0.72	0.52	0.72	0.33	0.39	0.74	0.52	0.67

Table 6: Classification Report of HOG features and CNN

6.3 Case Study 3: Emotion Analysis using Fusion of Text-Images

In an extensive analysis of machine learning models combining image and text features for emotion classification, each model exhibits distinct strengths across various metrics such as accuracy, processing time, and memory usage. The Random Forest model stands out with a remarkable accuracy of 97.37% and a quick processing time of 7.19 seconds, although the reported memory usage of -52.27 MiB suggests a potential reporting error. Its classification report shows high precision, recall, and F1-scores across all categories, indicating a well-rounded performance. Similarly, the K-Nearest Neighbors (KNN) model closely competes with an accuracy of 97.11%, and it's notably efficient with a processing time of only 0.02 seconds. However, it also records a negative memory usage (-263.38 MiB), indicating a possible discrepancy. The classification report of KNN mirrors this high performance, with consistent scores across various categories.

On the other hand, the Naive Bayes model, though less accurate at 86.93%, stands out for its processing efficiency, taking only 0.04 seconds, and minimal memory usage (0.08 MiB), making it suitable for applications where speed and memory are critical. The model's limitation is more evident in its classification report, especially in lower recall and precision for some categories. The Support Vector Machine (SVM) presents a balance with a high accuracy of 96.98%, a moderate processing time of 1.61 seconds, and reasonable memory usage of 12.77 MiB. Its classification report indicates consistent high performance similar to Random Forest and KNN. The Artificial Neural Network (ANN) emerges as the top performer in terms of accuracy with an impressive 98.18%, alongside a reasonable processing time of 5.27 seconds and a memory usage of 22.08 MiB. The ANN's classification report showcases its excellence, with almost perfect scores across all categories.

In summation, while each model demonstrates high efficacy in emotion classification, the ANN stands out as the best overall in terms of accuracy and comprehensive performance metrics. However, models like Naive Bayes and KNN are notable for their processing efficiency, offering significant advantages where speed and memory usage are constraints. The selection of the appropriate model should therefore be carefully considered based on the specific requirements and limitations of the intended application, balancing accuracy with computational efficiency. Confusion matrix of the best model is shown in Figure 7, Table 7 shows the comparison of accuracies of all the models and Table 8 shows the classification Report of all the models. In Table 9 K-fold for Text-Image fusion is done and the best one was found to be of ANN.

6.4 Discussion

The case studies that were carried out offer valuable insights into the discipline of emotion analysis through the application of several machine learning models in textual, image, and combined text-image modalities. Every study identifies significant aspects of model performances and their suitability for specific assignments, but it also provides opportunities for criticism and possible improvements for future studies.

- **Case Study 1: Text-Based Emotion Analysis -** In the emotion analysis with text, various models show diverse properties. SVC is superior in processing high-dimensional data, and thus it is powerful when applied to text based classification

Model	Accuracy	Time Taken(S)	Peak Memory Usage(MB)
Naive Bayes	0.87	0.04	0.078
SVM	0.97	1.61	12.77
Random Forest	0.97	7.19	-52.27
KNN	0.97	0.02	-263.38
CNN	0.98	5.27	22.078

Table 7: Performance Metrics of Classifiers with Text-Image fusion of features and Computational Cost

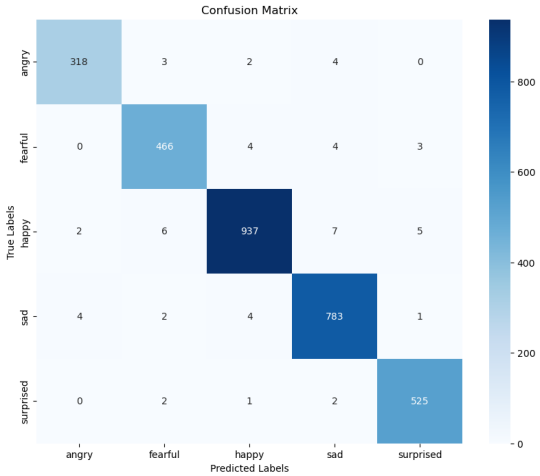

Figure 7: Confusion Matrix of ANN

Classifier	Precision					Recall					F1 Score				
	Angry	Fear	Happy	Sad	Surprised	Angry	Fear	Happy	Sad	Surprised	Angry	Fear	Happy	Sad	Surprised
Naive Bayes	0.91	0.92	0.97	0.72	0.97	0.96	0.94	0.67	0.97	0.97	0.93	0.93	0.79	0.83	0.97
SVM	0.96	0.95	0.98	0.97	0.98	0.95	0.96	0.97	0.97	0.98	0.96	0.96	0.97	0.97	0.98
Random Forest	0.99	0.96	0.98	0.97	0.97	0.94	0.96	0.98	0.98	0.98	0.96	0.96	0.98	0.98	0.97
KNN	0.98	0.95	0.97	0.97	0.98	0.95	0.96	0.98	0.97	0.98	0.97	0.95	0.98	0.97	0.98
CNN	0.98	0.97	0.99	0.98	0.98	0.97	0.98	0.98	0.99	0.99	0.98	0.97	0.98	0.98	0.99

Table 8: Classification Report of Text-Image Features

using TFIDF vectors or BoW combinations. The kernel trick of its provides a better data separation, which leads to high accuracy. TF-IDF can especially reveal the drawback of Multinomial Naive Bayes due to its feature independence assumption, although computational wise it is efficient. Random Forest and KNN have a moderate capacity, which means that the formulation of text data is complicated. With its convolutional layers, CNN is good at discriminating intricate patterns thus works excellent for fine analysis of text-based emotion classification. This analysis underlines the significance of choosing adequate models and methods, finding a balance between accuracy, efficiency, and capturing emotional subtleties in text. Figure 8 shows the comparison of all the BoW models and CNN models.

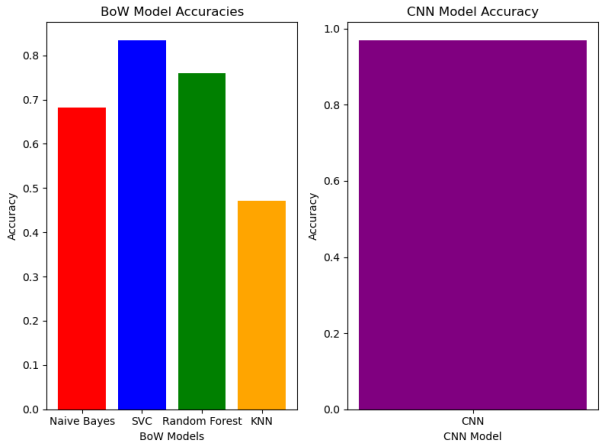

Figure 8: Model Accuracy Comparison of BoW and CNN for text

- **Case Study 2: Emotion Analysis using Image -** In the case of emotion ana-

Model (K-Fold)	Accuracy	Time Taken(S)	Peak Memory Usage(MB)
Naive Bayes	0.86	0.04	24.10
SVM	0.97	0.96	2.47
Random Forest	0.97	6.10	6.50
KNN	0.96	0.01	31.05
ANN (Average)	0.98	4.96	97.96

Table 9: K-Fold for Text-Image

lysis based on images, Random Forest and SVC model perform very differently. However, when applied directly to images these models face difficulties because of high dimensionality and intricacy of raw image data meaning that such a necessity is associated with moderate accuracy and substantial capital consumption. Nevertheless, the addition of HOG features improves their effectiveness. HOG helps in dimensionality reduction but also well represents important edge and gradient information which is necessary for any image processing. Unlike the previous case, CNNs perform very well in this area due to their convolutional layers capable of recognising spatial features and hierarchies contained within an image. The architecture strength that is intrinsic to CNNs plays an important role in the accuracy of emotion classification through images, showing why feature extraction and model selection are essential aspects of image-based analysis. Figure 9 shows the comparison of all the HOG models and CNN models.

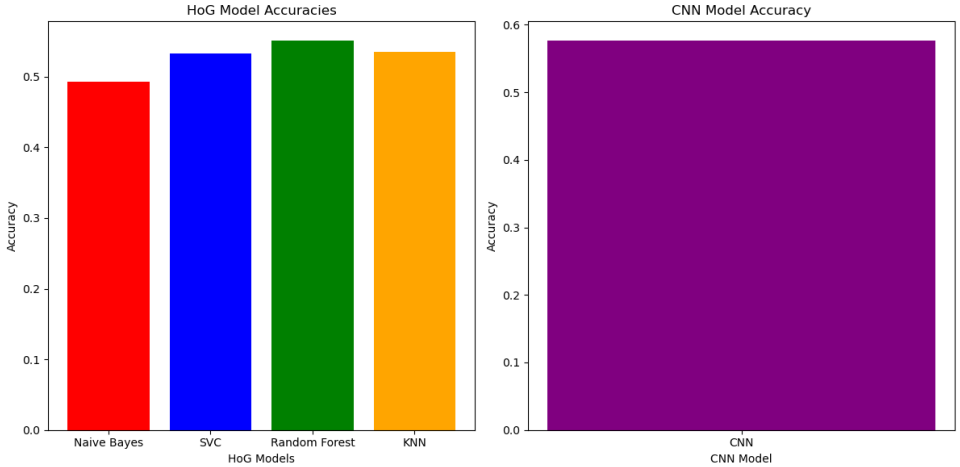

Figure 9: Model Accuracy Comparison of HOG and CNN for Image

- **Case Study 3: Emotion Analysis using Fusion of Text-Images -** Text and image features were used together to categorise emotions; illustrations of machine learning algorithms predicting emotion from various sources. As we see from the results, the machine learning models (Random Forest and KNN) achieved a relatively higher level of accuracy because they are able to effectively integrate and classify different types of features. Random Forest is an ensemble model of decision trees that performed well for such a complex task because it can deal with complex interactions between text and image features. Similarly, KNN is an instance-based learning model that leverages each training feature in decision-making. The latter appears to have a deep ability to generalise its learning because it will only classify a new image based on how closely its features are similar to already predicted images

for a given class of emotions. In terms of accuracy, the ANN performed increasingly well when given more training data. For this application, it performed well because it is a deep learning architecture. In other words, it would be able to detect gist patterns in various tricky data sets together. It performed feature extraction and classification at the same time, which helps to explain its high accuracy. So, given the same inputs and level of accuracy, the selection of which machine learning model to use is largely a matter of computational tractability and the requirements of the application. Figure 10 shows the comparison of all the models.

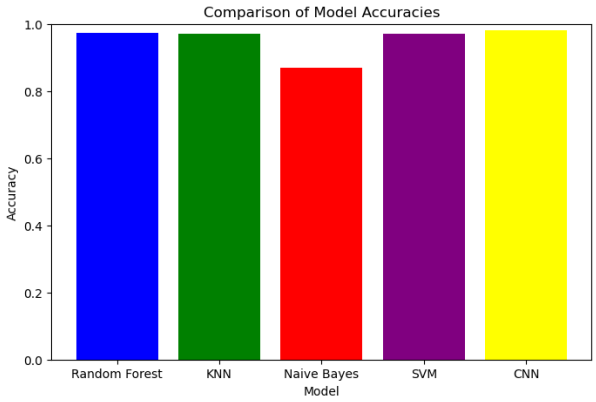

Figure 10: Model Accuracy Comparison of all the models for Text-Image

7 Conclusion and Future Work

The study was done due to the primary research issue that evaluate how good the multiple machine learning models performed analysis of emotion in utilizing text, images, and the combination of both text and image data. Identifying the models that perform most effectively in these various scenarios and understanding the effects on different feature extraction methods were the objectives. The study was able to identify significant patterns in the way the model performed across several modalities. Regardless of the feature extraction technique employed, the Support Vector Classifier (SVC) consistently outperformed other models in text-based emotion identification. The application of Histogram of Oriented Gradients (HoG) offers significantly better model performances for image-based analysis, with CNNs showing especially high effectiveness. Fully Connected Neural Networks with SVC models showed exceptional consistency and accuracy in the complex field of combined text-image emotion interpretation.

In order to drive the emotion analysis research, this study suggests several lines of future works. These involve advanced algorithms that are usually customized to solve the challenges of multimodal data. Furthermore, it is necessary to investigate novel feature extraction techniques and study new deep learning architectures so that the precision of emotion detection can be increased. Increasing the scope of datasets to be more representative and larger also helps in broadening generalizability for a wider range of emotions. In addition, this research highlights the need to use these multimodal emotion analysis procedures in real-world settings so that their practical value can be tested and new applications identified. The study serves as a basis for considerable progress in emotion analysis, creating opportunities for innovations on different levels.

References

Aslam, N., Rustam, F., Lee, E., Washington, P. B. and Ashraf, I. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model, *Ieee Access* **10**: 39313–39324.

Balabantaray, R. C., Mohammad, M. and Sharma, N. (2012). Multi-class twitter emotion classification: A new approach, *International Journal of Applied Information Systems* **4**(1): 48–53.

Baltrušaitis, T., Ahuja, C. and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy, *IEEE transactions on pattern analysis and machine intelligence* **41**(2): 423–443.

Barros, P., Churamani, N. and Sciutti, A. (2020). The facechannel: A light-weight deep neural network for facial expression recognition, *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 652–656.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S. and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database, *Language resources and evaluation* **42**: 335–359.

Doshi, U., Barot, V. and Gavhane, S. (2020). Emotion detection and sentiment analysis of static images, *2020 International Conference on Convergence to Digital World - Quo Vadis (ICCDW)*, pp. 1–5.

Ekman, P. and Friesen, W. V. (1978). *Facial action coding systems*, Consulting Psychologists Press.

Fang, Z., He, A., Yu, Q., Gao, B., Ding, W., Zhang, T. and Ma, L. (2022). Faf: A novel multimodal emotion recognition approach integrating face, body and text, *arXiv preprint arXiv:2211.15425* .

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T. and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding, *arXiv preprint arXiv:1606.01847* .

Gharsalli, S., Emile, B., Laurent, H., Desquesnes, X. and Vivet, D. (2015). Random forest-based feature selection for emotion recognition, *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 268–272.

Juyal, P. and Kundalya, A. (2023). Emotion detection from text: Classification and prediction of moods in real-time streaming text, *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 46–52.

Khan, A., Sohail, A., Zahoora, U. and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks, *Artificial intelligence review* **53**: 5455–5516.

Khan, R., Rustam, F., Kanwal, K., Mehmood, A. and Choi, G. S. (2021). Us based covid-19 tweets sentiment analysis using textblob and supervised machine learning algorithms, *2021 international conference on artificial intelligence (ICAI)*, IEEE, pp. 1–8.

Koné, C., Le Thanh, N., Flamary, R. and Belleudy, C. (2018). Performance comparison of the knn and svm classification algorithms in the emotion detection system emotica, *International Journal of Sensor Networks and Data Communications* **7**(1): 1–9.

Krishnan, H., Elayidom, M. S. and Santhanakrishnan, T. (2017). Emotion detection of tweets using naïve bayes classifier, *Emotion* **4**(11): 457–62.

Li, C. and Hu, Z. (2022). Multimodal sentiment analysis of social media based on top-layer fusion, *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, IEEE, pp. 1–6.

Nasir, A. F. A., Nee, E. S., Choong, C. S., Ghani, A. S. A., Majeed, A. P. A., Adam, A. and Furqan, M. (2020). Text-based emotion prediction system using machine learning approach, *IOP Conference Series: Materials Science and Engineering*, Vol. 769, IOP Publishing, p. 012022.

Olusegun, R., Oladunni, T., Audu, H., Houkpati, Y. and Bengesi, S. (2023). Text mining and emotion classification on monkeypox twitter dataset: A deep learning-natural language processing (nlp) approach, *IEEE Access* **11**: 49882–49894.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques, *arXiv preprint cs/0205070* .

Picard, R. W. (2000). *Affective computing*, MIT press.

Qaiser, S. and Ali, R. (2018). Text mining: use of tf-idf to examine the relevance of words to documents, *International Journal of Computer Applications* **181**(1): 25–29.

Qiu, F., Kong, W. and Ding, Y. (2022). Intermulti: Multi-view multimodal interactions with text-dominated hierarchical high-order fusion for emotion analysis, *arXiv preprint arXiv:2212.10030* .

Rao, T., Li, X., Zhang, H. and Xu, M. (2019). Multi-level region-based convolutional neural network for image emotion classification, *Neurocomputing* **333**: 429–439.

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A. and Choi, G. S. (2021). A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis, *Plos one* **16**(2): e0245909.

Tao, J. and Tan, T. (2005). Affective computing: A review, *International Conference on Affective computing and intelligent interaction*, Springer, pp. 981–995.

Yang, X., Feng, S., Wang, D. and Zhang, Y. (2020). Image-text multimodal emotion classification via multi-view attentional network, *IEEE Transactions on Multimedia* **23**: 4014–4026.

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E. and Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.