Sentimental Analysis on Video

By

Akshay Mohan

Submitted to
The University of Roehampton

In partial fulfilment of the requirements
for the degree of

MASTER OF SCIENCE IN DATA SCIENCE

Abstract

In the current digital era, user-generated content on websites like YouTube has developed into a rich source of information as well as opinions. Particularly, YouTube review videos offer insightful information on various products, services, and topics. Sentiment analysis, the process of identifying the emotional tone of the text, may help consumers and organizations gain these insights. Traditional sentiment analysis techniques, primarily designed for text, struggle to capture the full spectrum of sentiments in YouTube review videos, where emotions are conveyed not only through text but also through visual cues.

This project addresses this challenge by introducing a multimodal sentiment analysis framework tailored for YouTube review videos. The main goals include text-to-video conversion, sentiment analysis of both text and images, and merging the outcomes to produce a whole sentiment analysis visualisation. To accomplish these objectives, the project makes use of deep learning models, computer vision techniques, and natural language processing (NLP) techniques.

This project's main objective of conducting sentiment analysis on YouTube review videos was accomplished successfully. Deep learning techniques, video processing, and Natural Language Processing (NLP) were combined to define attitudes such as positive, negative, or neutral. It also has the potential for further developments and future studies.

Declaration

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

Signed

Akshay Mohan

Date: 06/09/2023

Acknowledgements

I would like to convey my sincere gratitude to everyone who helped this project be completed successfully. I want to start by sincerely thanking my supervisor, **Changjiang He**, for his important advice, constant guidance, and insightful comments throughout this project.

My friends and colleagues greatly helped achieve the project's success by supporting and sharing their knowledge, and I am grateful to them. I also thank the **University of Roehampton** for providing the facilities and resources that were required. Especially the faculties who supported and taught the basics of Data Science.

I am grateful, for the support and kindness of my family throughout this project. Their understanding and constant encouragement have been a source of motivation for me.

Finally, I want to extend my thanks to all the writers, researchers, and contributors whose articles and works provided invaluable data sources for this project. I appreciate everyone involved for supporting and motivating me through this entire project.

Table of Contents

1. Introduction 6
Research Question or Problem Statement 6
Aims 7
Objectives 7
Legal, Social, Ethical and Professional Considerations 8
Background 9
Report overview 10
2. Literature and Technology Review 12
Literature Review 12
Technology Review 14
3. Methodology 16
Project Management 18
4. Implementation 20
5. Result 31
Evaluation 31
Related Work 33
6. Conclusion 35
Reflection 35
Future Work 35
7. References 37
8. Appendices 38

1.Introduction

In the age of online platforms, online contents have a significant role as a source of information and viewpoints. Being one of the largest platforms for sharing videos, YouTube hosts a large collection of review videos where users share their thoughts, reviews and experiences about various products, services, and subjects. Content creators, brands, and consumers can produce valuable insights by analysing the sentiments expressed in these videos.

The primary focus of the traditional sentimental analysis is on textual data from product reviews and social media posts. However, in YouTube review videos, sentiment is not only conveyed through text but also through visual cues. Hence, there is a need to develop advanced sentiment analysis techniques that integrate both textual and visual information to capture the complete sentiment conveyed in the videos.

This project aims to propose a multimodal sentiment analysis framework for YouTube review videos. The objective of the project is to combine the sentimental analysis results of textual transcript data and visual data. This project aims to enable a more comprehensive understanding of the sentiment expressed in YouTube review videos, thereby providing valuable insights for various stakeholders.

Research Question or Problem Statement

YouTube has a huge number of review videos. Both customers and businesses are finding it tough to locate the ones which are beneficial. Consumers want to find videos a good way to assist them make informed decisions, however it could be time-consuming to look at a bunch of videos and try and parent out what the reviewer is announcing. Businesses need to understand what their customers think of their merchandise, but it can be hard to track down all the video reviews and analyse them.

Sentiment analysis can help in solving this problem as it is a way of understanding the emotional tone of a data. By analysing the data from video reviews, we can figure out whether the reviewer or the speaker is expressing positive, negative, or neutral emotions. This information can be used to help consumers find helpful videos and to help businesses understand what their customers think of their products.

For example, if you are looking for a new television, you could use sentiment analysis to find videos that analysed as more positive. This will save time and help you to find videos that are more likely to be helpful. Businesses can also use sentiment analysis to track customer satisfaction and opinions. By analysing video reviews, they could see what customers like and dislike about their products. This information could then be used to improve the products and make customers happier.

Sentiment analysis is a powerful approach or tool that helps human beings to make higher choices and enhance their services and products. By conducting sentiment analysis on YouTube review videos, we can help people to find helpful videos, businesses to understand their customers, and everyone to make better decisions.

Aims

1. Enhance viewer understanding: The project aims to improve viewers' understanding of sentiments and emotions expressed in videos by providing comprehensive sentiment analysis and emotion detection results.
2. Enable actionable insights: The project aims to generate actionable insights for various stakeholders, such as content creators, marketers, and social media analysts, by identifying and visualising the sentiments and emotions contained in the video.
3. Improve user experience: By developing a combined graph visualisation, the project aims to enhance the user experience of exploring sentiment and emotion data in videos. The visualisation will provide an intuitive aspect for the stakeholders.

Objectives

1. Video-to-Text Conversion: Convert YouTube review videos into text transcripts. This involves importing the YouTube API to get the transcription with the time from the video.
2. Perform Sentiment Analysis on Text: Apply sentiment analysis techniques to the transcribed text from YouTube review videos. This includes pre-processing the text, such as removing stop words, handling punctuation, and normalising the text. Then, employ sentiment analysis algorithms or machine learning models to classify the sentiment of the text into positive, negative, or neutral categories.
3. Perform Sentiment Analysis on Images: Apply sentiment analysis techniques to the extracted visual features to analyse the sentiment expressed in the video frames. This could involve training image-based sentiment analysis models or using pre-trained models to classify the images into different sentiment categories.
4. Combining Text and Image Sentiment Analysis: Integrate the results obtained from text-based sentiment analysis and image-based sentiment analysis to provide a comprehensive sentiment analysis of the YouTube review videos.
5. Provide Insights and Visualisations: Generate meaningful insights from the sentiment analysis results and present them in a user-friendly manner. Create visualisations, such as sentiment distributions, word clouds, or sentiment trends over time, to facilitate the interpretation and understanding of the sentiment analysis outcomes.
6. Demonstrate the Practical Utility: Showcase the practical utility and potential applications of the sentiment analysis on YouTube review videos. Highlight how the obtained sentiment

information can aid users in making informed decisions and assist businesses in improving their products or services based on customer feedback.

Legal, Social, Ethical and Professional Considerations

This section outlines the key legal, social, ethical, and professional considerations associated with the execution of this project. Identifying and addressing these considerations is paramount to ensuring the project's integrity and responsible conduct.

Legal Considerations:

1. Copyright and Fair Use: The undertaking includes analysing YouTube-evaluated films, which may additionally include copyrighted content. To mitigate legal issues, we are able to make certain that using the video facts complies with copyright legal guidelines and fair use guidelines. This includes acquiring the essential permissions for statistics collection and use.
2. Data Privacy: The task may additionally involve the gathering and analysis of user-generated content material. We will adhere to information privacy guidelines, which include the General Data Protection Regulation (GDPR), and ensure that any personally identifiable information (PII) is dealt with as it should be and anonymized.

Social Considerations:

1. Bias and Representation: Analysing sentiment from user-generated content introduces the ability for bias and skewed representations of sentiment. We will be cognizant of ability bias within the collected facts and take steps to deal with and acknowledge any capacity limitations in our analysis.
2. Societal Impact: The sentiments expressed in YouTube review movies ought to influence patron choices and perceptions. As such, we recognize the duty to accurately analyse and represent sentiments, as well as to communicate findings responsibly to save you any unintended consequences.

Ethical Considerations:

1. Informed Consent: If the mission entails human contributors or sensitive statistics, we can acquire knowledgeable consent consistent with ethical pointers. For example, if manual sentiment annotation by using human annotators is required, contributors will be knowledgeable about the purpose of the annotation, and their consent could be sought.
2. Ethical Review: This mission does not involve direct human interaction or sensitive private information and, as a result, does not require a formal ethical evaluation. However, the moral considerations related to record collection, privateness, and accountable analysis might be upheld for the duration of the challenge.

Professional Considerations:

1. Professional Conduct: The assignment could be carried out in a professional and respectful manner, adhering to the standards of academic integrity, honesty, and transparency. Proper quotation of sources and an accurate representation of findings will be maintained.
2. Collaboration: Collaboration and communication with stakeholders, along with content creators and ability users of the sentiment analysis, may be prioritized. Collaboration fosters extra-comprehensive know-how of the mission's effects and promotes responsible deployment of the effects.

Background

This section provides the essential background information that is necessary to grasp the significance of the project. By offering context, references to relevant literature, and personal motivation, the goal is to clarify why this project holds importance and how it addresses existing challenges.

In today's digital landscape, YouTube stands out as a major platform for sharing videos and user-generated content. Among the diverse content available, review videos hold a significant place, where individuals express their thoughts and experiences about various products, services, and topics. These review videos have emerged as a valuable source of information and opinions for both consumers and businesses.

Traditional approaches to sentiment analysis primarily focus on textual data from sources like social media posts and reviews [1]. However, YouTube review videos are multimodal in nature, encompassing both textual and visual components. The sentiments conveyed in these videos extend beyond just text, encompassing visual cues and emotions. Consequently, there exists a gap in sentiment analysis techniques that effectively integrate both textual and visual elements to capture the complete sentiment expressed in the videos.

This project aims to bridge the gap by proposing a multimodal sentiment analysis framework specifically tailored for YouTube review videos. By combining advanced techniques for both textual and visual analysis, the objective is to enhance the accuracy and comprehensiveness of sentiment analysis within this unique context [2].

This project's relevance extends to various domains. Content creators can benefit from insights into viewer sentiments, helping them improve their content. Marketers can gain valuable information about consumer perceptions, enabling more targeted strategies. Additionally, consumers seeking informed decisions can utilise sentiment analysis to identify videos that align with their preferences.

In the business sphere, sentiment analysis aids in understanding customer feedback, thereby enabling product refinement and enhancing customer satisfaction.

Report overview

This section provides an overview of the structure of this report, outlining the upcoming sections and their respective roles in presenting the project's progression and outcomes.

1. Literature Review: In this section, we delve into the existing body of knowledge, exploring relevant studies, methodologies, and technologies related to sentiment analysis, multimodal analysis, and their applications in the context of YouTube review videos. This review forms the foundation for our project's approach and contributions.
2. Methodology: The methodology section outlines the step-by-step approach we have designed to address the research objectives. We detail the process of extracting sentiment from both text and images, including video-to-text conversion, textual sentiment analysis, image-based sentiment analysis, and the integration of these analyses.
3. Implementation: Here, we present the practical realisation of our methodology. We discuss the tools, software, and frameworks employed to develop and implement the sentiment analysis framework. This section offers insight into the technical aspects of our approach.
4. Results: This section showcases the outcomes of our sentiment analysis efforts. We present the sentiment distributions, insights derived from the data, and visualisations that aid in understanding the sentiment trends in YouTube review videos.
5. Evaluation: In this section, we assess the performance of our sentiment analysis framework. We discuss the evaluation metrics employed to measure the accuracy and effectiveness of our approach, ensuring that our analysis aligns with our research goals.
6. Related Work: Building upon the literature review, this section explores recent studies, projects, or technologies that are closely related to our research. We compare our approach with existing work, highlighting the novelty and contributions of our project.
7. Conclusion: Here, we summarise the key findings and insights drawn from our sentiment analysis of YouTube review videos. We revisit the project's objectives, discuss the implications of our results, and highlight the potential impact of our work on stakeholders.
8. Reflection: This section offers a reflective assessment of our project's journey. We discuss challenges encountered, lessons learned, and insights gained during the research and development process.
9. Future Work: We outline potential directions for future research and development in the field of multimodal sentiment analysis. This section highlights opportunities for refining and expanding our approach.
10. References: The references section lists the sources we have cited throughout the report, ensuring proper credit to prior research and studies that have influenced our work.

11. Appendices: Additional information, such as code snippets, detailed visualisations, or supplementary data, can be found in the appendices.

2. Literature and Technology Review

Literature Review

In recent years, user-generated content on online platforms has gained immense prominence, with YouTube emerging as a dominant platform for sharing video content. The abundance of review videos on YouTube provides a wealth of information and opinions across a wide range of subjects. To extract meaningful insights from this content, sentiment analysis has emerged as a valuable tool that can enhance both consumer decision-making and business strategies. While traditional sentiment analysis primarily focuses on textual data, YouTube review videos present a unique challenge due to the integration of both textual and visual cues.

Sentiment analysis is the task of determining the sentiment polarity of a piece of text, such as whether it is positive, negative, or neutral. It has been widely applied in various domains, such as customer reviews, product ratings, and social media. This project involves constructing a sentiment analysis model for text by evaluating three distinct models: Simple Neural Network (SNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). The choice of model and its architectural layers contribute to the fluctuation in accuracy and loss rate.
Traditional sentiment analysis methods mainly focus on textual data. However, there has been a growing interest in multimodal sentiment analysis, which considers both textual and audio-visual information. This is because audio-visual information can provide additional cues for sentiment analysis, such as the speaker's facial expressions, voice tone and body language. There are many papers that have undergone studies about sentimental analysis on video.

The paper by Wöllmer et al. (2020) studies the problem of sentiment analysis in YouTube movie reviews. The authors propose a multimodal sentiment analysis approach that combines features from text, audio, and video data. The features from the three modalities are fused using a late fusion approach [3]. This paper is the most comprehensive, as it discusses the problem of sentiment analysis in YouTube movie reviews, which is a challenging task due to the presence of multiple modalities and audio-visual information that can provide additional cues for sentiment analysis, such as the speaker's tone of voice, facial expressions, and body language.

The paper by Ban et al. (2020) is the most technical, as it proposes a multimodal aspect-level sentiment analysis model based on Deep neural networks (DNNs). Aspect-level sentiment analysis (ALSA) is the task of identifying the sentiment polarity of each aspect of a sentence. Traditional ALSA methods mainly focus on textual data. It has been widely applied in various domains, such as customer reviews, product ratings, and social media. It has been demonstrated that deep neural networks (DNNs) are efficient for ALSA [4]. DNNs can learn complex features from both textual and visual data, and they can also capture long-range dependencies [8]. The model first extracts feature from the text and audio data using separate neural networks. The features from the two modalities

are then fused using a multi-modal attention mechanism. The fused features are then used to classify the sentiment of each aspect.

The paper by Poria et al. (2020) discusses the challenges and future perspectives of multimodal sentiment analysis. The authors identify three main challenges: data scarcity, data fusion, and domain adaptation [5].

- Data scarcity: There is a lack of large, labelled datasets for multimodal sentiment analysis. This makes it difficult to train and evaluate models.
- Data fusion: There is no single, optimal way to fuse data from different modalities. This is a complex problem that is still being actively researched.
- Domain adaptation: Multimodal sentiment analysis models often perform well on in-domain data, but they may not generalise well to out-of-domain data. This is a problem that is shared with traditional sentiment analysis.

They also discuss several promising directions for future research, such as the development of larger and more diverse datasets, the development of new fusion techniques, and the investigation of new applications for multimodal sentiment analysis.

The paper by Liu (2022) provides a comprehensive review of sentiment analysis of micro-videos, which is a relatively new area of research. The author discusses the different approaches to micro-video sentiment analysis, the challenges of this task, and the potential applications of micro-video sentiment analysis [6]. All studies agree that multimodal sentiment analysis is a promising research area with the potential to improve the accuracy and robustness of sentiment analysis. In the current time, micro-videos are most popular, such as TikTok, Instagram Reels, YouTube Shorts, and so on. Most people prefer watching videos on these kinds of platforms to save time because they can find short videos that are one minute or less.

The paper by Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, about ‘Pre-training of Deep Bidirectional Transformers for Language Understanding’ published by Google AI Language, introduces a ground-breaking natural language processing (NLP) model called BERT, which stands for Bidirectional Encoder Representations from Transformers [7]. This paper offers several remarkable benefits. One of its key advantages is its capacity to record bidirectional context during pre-training, which enables it to fully comprehend the semantics and connections between words in a sentence. BERT has a considerable performance and efficiency advantage due to its transfer learning capabilities, which make it highly adaptable to a variety of NLP tasks. The publication of the study also had a profound influence on the NLP community, inspiring creativity and resulting in the subsequent creation of cutting-edge models. There are certain restrictions, though. BERT needs to be adjusted, which calls for task-specific data and resources that might not always be available. The large size of the model can make deployment difficult in environments with limited

resources. Additionally, the pre-training procedure requires a significant number of computational resources due to its high computational demand [7]. Despite these flaws, BERT has made a significant contribution to language understanding and the NLP field, strengthening its position as an inspiration in the field.

In addition, the literature highlights challenges in capturing fine-grained and aspect-level sentiment, which aligns with the objectives of this project. The proposed multimodal sentiment analysis framework seeks to address these challenges by combining textual and image-based sentiment analysis, aiming to offer a comprehensive sentiment interpretation of YouTube review videos.

Overall, these papers collectively emphasise the significance of multimodal sentiment analysis for video content, including YouTube review videos. By drawing insights from these research works, I can leverage state-of-the-art techniques and methodologies to tackle the challenges of sentiment analysis in a comprehensive and efficient manner, improving the accuracy and relevance of sentiment predictions in YouTube review videos.

Technology Review

In the context of sentiment analysis on YouTube review videos, several technologies can be considered for processing both text and image data. A review of different technology options is essential to identifying the most suitable approach for achieving accurate sentiment classification in videos.

1. Natural Language Processing (NLP) Libraries and Algorithms:
   NLP libraries such as NLTK (Natural Language Toolkit) and spaCy offer a wide range of tools for text pre-processing, tokenization, and sentiment analysis. These libraries provide access to pre-trained sentiment analysis models based on traditional machine learning algorithms, such as Naive Bayes classifiers and Support Vector Machines (SVMs) [3]. While NLP libraries are reliable for text-based sentiment analysis, they may not directly handle multimodal data like videos.
2. Deep Learning Models for Text Sentiment Analysis: Deep learning models like recurrent neural networks (RNNs) and transformers (e.g., BERT) have shown exceptional performance in text-based sentiment analysis. By fine-tuning pre-trained language models for specific sentiment analysis tasks, it is possible to achieve state-of-the-art results. However, these models usually do not consider visual features and are limited to processing text data only [3].
3. Computer Vision Libraries and Techniques: For the image part of the sentiment analysis, computer vision libraries like OpenCV and deep learning frameworks like TensorFlow and PyTorch can be utilised [3]. Techniques such as object detection, facial emotion recognition, and feature extraction from video frames can be employed to extract visual features. These

visual features can then be fed into image-based sentiment analysis models to classify emotions in the images.

4. Multimodal Fusion Techniques: To effectively combine text and image-based sentiment analysis, multimodal fusion techniques can be explored. Fusion methods such as late fusion (combining the outputs of separate models) or early fusion (combining features at the input level) can be used to integrate the results from text and image analysis [5]. Additionally, methods like attention mechanisms or cross-modal embeddings can be employed to align and associate textual and visual features.

Utilising deep learning models like transformers for text sentiment analysis allows for context-aware sentiment classification and the ability to capture subtle nuances in language. Fine-tuning pre-trained models on a relevant sentiment analysis dataset can yield highly accurate sentiment predictions for the text part of the video analysis.

Computer vision libraries and deep learning frameworks offer powerful tools for extracting visual features from video frames. Techniques like facial emotion recognition can help identify emotional cues in the video content. Applying image-based sentiment analysis models trained on annotated datasets enables the classification of emotions expressed in the images.

Integrating the results from both text- and image-based sentiment analysis through multimodal fusion techniques is crucial for capturing the complete sentiment spectrum in YouTube review videos. This approach leverages the strengths of both modalities and ensures a comprehensive sentiment analysis, which aligns with objectives of the project.

By combining deep learning for text sentiment analysis, computer vision techniques for image sentiment analysis, and multimodal fusion for integration, the proposed technology approach is suited to understand and visualise the challenges in it.

3. Methodology

I have developed a thorough approach that combines a range of tools, design methodologies, data gathering strategies, and algorithms with the aim of developing an effective sentiment analysis. This section describes my methodology while emphasising the reasons for my decisions. The entire project is going to divide into three parts such sentiment analysis on text, video, and visualisation.

Sentimental Analysis on the Text Part:

1. Data Collection and Pre-processing:

Strong data collection and pre-processing are the basis of any sentiment analysis endeavour. For this research, new datasets and YouTube transcriptions are used to guarantee the quality and diversity of the datasets. The dataset with reviews will enable me to capture real-world reactions to movies. The selection of diverse data is crucial to creating a model that generalises well across different domains. To build a model that generalises successfully across several domains, diverse data must be carefully chosen. For pre-processing a dataset, we will use Python modules like Pandas and the Natural Language Toolkit (NLTK) and investigate APIs for social media platform data extraction.

2. Text Pre-processing and Tokenization:

To ensure that the textual data is ready for analysis, we will perform pre-processing steps such as lowercase, removal of special characters, and tokenization. The NLTK libraries will be employed for these tasks. Furthermore, we will experiment with comparing the values and filtering the dataset to improve the impact on model performance. Tokenization of the dataset is important for analysing word by word.

3. Embeddings and Model Architectures:

Incorporating pre-trained word embeddings plays a pivotal role in enhancing the performance of sentiment analysis models. The approach will encompass the utilisation of any of the popular word embeddings such as Word2Vec, GloVe, or FastText [5]. By capturing the semantic connections between words, these embeddings help the model understand context. For model architectures, we will implement a Simple Neural Network (SNN), Convolutional Neural Network (CNN), and Long Short-Term Memory Networks (LSTM) as the primary components of my analysis tool. After analysing the accuracy and test score, we will try using VADER and a pre-trained model such as roBERTa.

4. Model Training and Evaluation:

A split dataset will be used for the training process, with a sizable chunk left aside for validation. Overfitting is avoided by using dropout layers and batch training. The results of the literature analysis

will be taken into consideration when choosing optimizers like Adam and learning rates [3]. Based on criteria like accuracy, precision, recall, and F1-score, we will assess the model's performance.

5. Ensemble Approaches and Thresholding:

We will investigate ensemble strategies to further improve accuracy by combining predictions from various model architectures and embeddings. In addition, we will test several thresholding methods to find the most effective threshold for sentiment classification.

Sentimental analysis of the video part:

1. Dataset Collection:

For the evaluation of video sentiment, we need to compile a sizable collection of pictures of people with emotions that can be classified as neutral, positive, or negative. we must make sure that the dataset is varied and inclusive of the various emotional expressions that belong to the various classifications as required.

2. Data Pre-processing:

Normalizing and pre-processing the collected dataset to ensure consistency in lighting conditions, size, and orientation are required as pre-processing methods. Augmentation techniques like rotation, flipping, and brightness adjustment can help improve model performance.

3. Model Training:

Using the pre-processed image dataset, a convolutional neural network (CNN) model for emotion classification needs to be built. Create training and testing sets from the dataset. To avoid overfitting, train the CNN model on the training set while keeping an eye on accuracy and loss on the validation set. To improve model performance, experiment with hyperparameters, architectures, and regularization methods.

4. Model Evaluation:

Evaluate the trained model's performance on the testing set using accuracy, precision, recall, F1-score, and confusion matrix. These metrics will provide insights into the model's ability to accurately predict emotions.

5. YouTube Video Data Collection and Processing:

Download a YouTube video for emotion detection. Extract frames from the video at a regular interval. Perform face detection on each frame to locate human faces.

6. Emotion Detection on Video:

Apply the trained CNN model to classify emotions in the detected faces. The CNN should predict whether the emotion is positive, negative, or neutral in each frame.

7. Visualization and Analysis:

Visualize the findings from the sentimental analysis of the text and emotion parts of the video. Create visualizations (e.g., graphs, charts) to showcase the emotion detection and sentiment analysis results side by side on the video timeline.

In summary, proposed methodology for producing a sentiment analysis artifact combines data collection, pre-processing, model selection, training, and evaluation on both the textual and visual parts of the video. We need to build a piece of work by using an efficient approach that not only captures the subtleties of emotion across various textual data but also shows the precision and depth of our technological knowledge.

Project Management

For this project, several tools are used to ensure efficient and successful project execution. Here are some key tools and techniques, along with the reasons for their selection:

- Gantt chart: A Gantt chart shows the project schedule graphically. It shows the tasks that need to be completed, the order in which they need to be completed, and the estimated time required for each task. Initially, I used the Canva web platform to prepare it, and then I moved to Microsoft Excel for Gantt chart preparation.
- Weekly meetings: conducting regular weekly meetings with the project supervisor to communicate and update the project activities covered and yet to be done. Microsoft Teams is used for scheduling the online meeting. After the meeting, a weekly report is to be prepared to keep track of the progress of the project.
- Task breakdown and prioritization: breaking down the project tasks into smaller, manageable units and prioritizing them based on their dependencies and importance is essential. This approach facilitates a clear understanding of the project scope. For this purpose, I usually use a handbook to note down the steps I need to take to keep up with the Gantt chart.
- Integrated Development Environment (IDE): For the initial stages of the project, I used both Jupyter Notebook and Google Colab for implementation. As the Google Colab was taking a long time to train the models, I started using only Jupyter Notebook.
- Documentation Tools: For documentation purposes, Microsoft Word and Excel are mainly used. These are the most user-friendly and effective tools for documentation.

Project Report Delivery Schedule	Deadline Date
Literature - Technology Review	29th June 2023
Introduction	25th July 2023
Methodology	4th August 2023
Implementation and Results • Evaluation • Related Work	25th July 2023
Conclusion • Reflection • Future Work	4th August 2023
References	24th August 2023
Appendices	24th August 2023
Abstract	24th August 2023
Declaration	24th August 2023
Acknowledgements	24th August 2023

Table 1: Project Report Delivery Schedule

Artefact Delivery Schedule	Deadline Date
Artefact Planning and Resourcing	29th June 2023
Artefact Design	1st July 2023
Artefact Procurement Activities (e.g., data collection, source framework etc.)	5th July 2023
Artefact Development, Deployment, Implementation	14th August 2023
Artefact Evaluation and Testing	20th August 2023
Artefact Presentation and Demonstration	6th September 2023
Artefact Screencast	8th September 2023

Table 2: Artefact Delivery Schedule

4. Implementation

In this project, the implementation is divided into three phases, such as sentimental analysis on text, emotion detection from the video, and visualization. The first two phases consist of data collection, pre-processing, model selection, training, and evaluation. Here, we are considering YouTube videos for the testing as they are easier to get and use for this project.

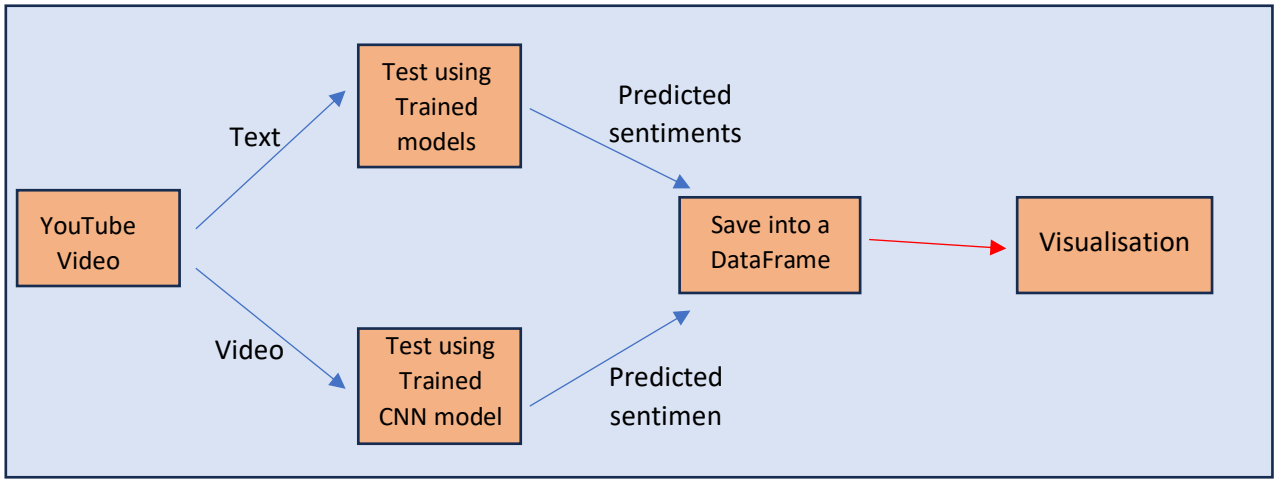


Figure 1: Simple workflow of the project

Sentimental Analysis on Text:

In the initial phase of the project, we focus on sentiment analysis of text data. This task aimed to understand the sentiment and opinions expressed in textual content, such as positive, negative, or neutral, which was crucial for later combining it with emotion detection on video data and visualization. Therefore, we need to train different models to compare the accuracy of the model and apply it to the transcription extracted from the YouTube video.

Dataset Selection and Pre-processing:

To train the sentiment analysis models, I carefully selected the IMDB movie review dataset, which offered a diverse range of opinions expressed by users. This dataset served as a valuable resource for capturing the nuances of human sentiment in language. Then we conducted thorough data pre-processing steps, including text cleaning, tokenization, and handling of imbalanced classes, to ensure the dataset's suitability for model training.

In the selected dataset, there were only two classes such as positive and negative. Therefore, a VADER-lexicon analysis has been done to get positive, negative, and neutral sentiments from that dataset. After the analysis, compared the pre-defined sentiments and calculated sentiments. Then removed the rows where the values are not equal and kept the rows with predicted value as neutral. Then saved it as a new data.

As many of the reviews are very lengthy, ‘textwrap’ library is used to split the reviews as each cell can only contain 40 characters. If it exceeds, that word will move to the next row. This will help improve the model's accuracy. After splitting, the total number of rows in the dataset changed from 24,311 to 523,911. Then again applied Vader Model on it and gave a threshold value to the Vader compound score to divide it into three classes, such as positive, negative, and neutral. Before deep learning model training, the classes are converted into integers such as 0,1 and 2 which represents ‘neutral’, ‘positive’ and ‘negative’ respectively. The resultant dataset is plotted as a WordCloud visualisation.

Figure 2: Wordcloud visualisation of ready to train IMDB dataset.

Model Training and Selection:

The own Models for Sentimental analysis are trained using three different models, such as a Simple Neural Network, Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM).

We started our analysis by implementing a Simple Neural Network. Before training the data has gone through several steps such as, splitting the dataset into training and testing data, vocabulary size is calculated for tokenizing each word, and ‘maxlen’ is set to ‘100’. The words in each row of the ‘Speech’ column will be padded as 100 values as shown below.

```
# Display the padded sequence of the third training example
print(X_train[2, :])
```

```
[975 112 834 129   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0]
```

Figure 3: Example for padded sequence.

Then the words in each row will be embedded using pre-trained Glove file. This file contains embeddings for each word in the dataset and the shape of the 'embedding_matrix' will become (67543,100).

- **Simple Neural Network (SNN):**

First, class labels are encoded using one-hot encoding, and class weights are computed to handle class imbalance. The model architecture begins with an Embedding layer, which uses pre-trained word embeddings to represent input text. These embeddings are frozen, making them non-trainable. A Flatten layer reshapes the data, followed by a Dense layer with a softmax activation function for multi-class classification. The model is compiled with the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric, resulting in a summary that displays the model's structure. This SNN is suitable for text classification tasks where the goal is to predict one of multiple classes based on input text data. Then model is trained using the training data with a batch size of 128 and for 100 epochs. After training, we evaluated the SNN model on the test data and recorded its performance metrics.

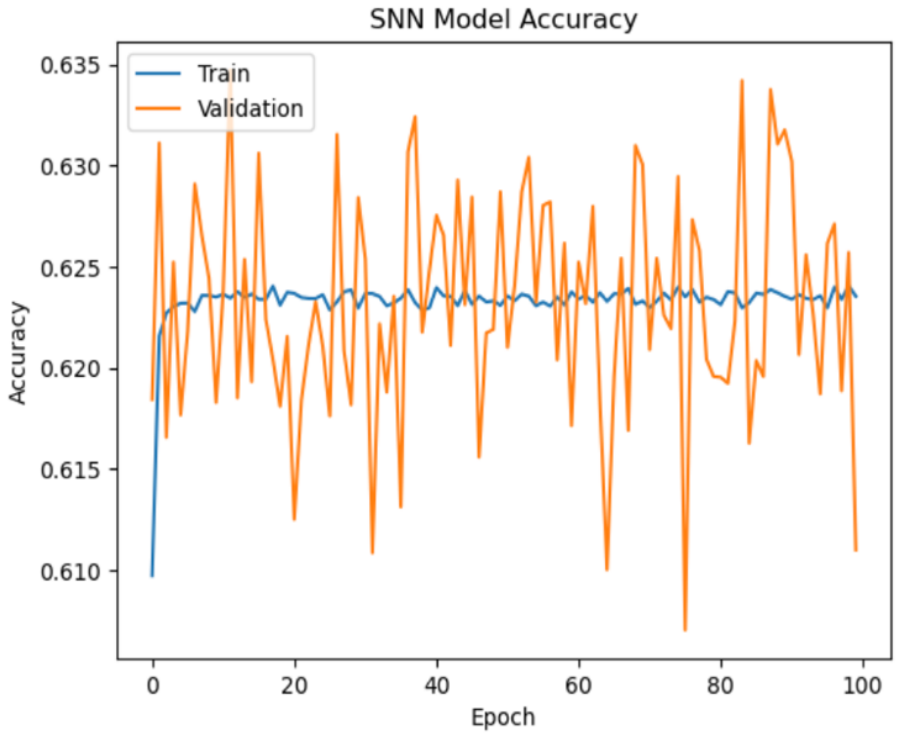


Figure 4: Epoch-Accuracy graph of SNN model.

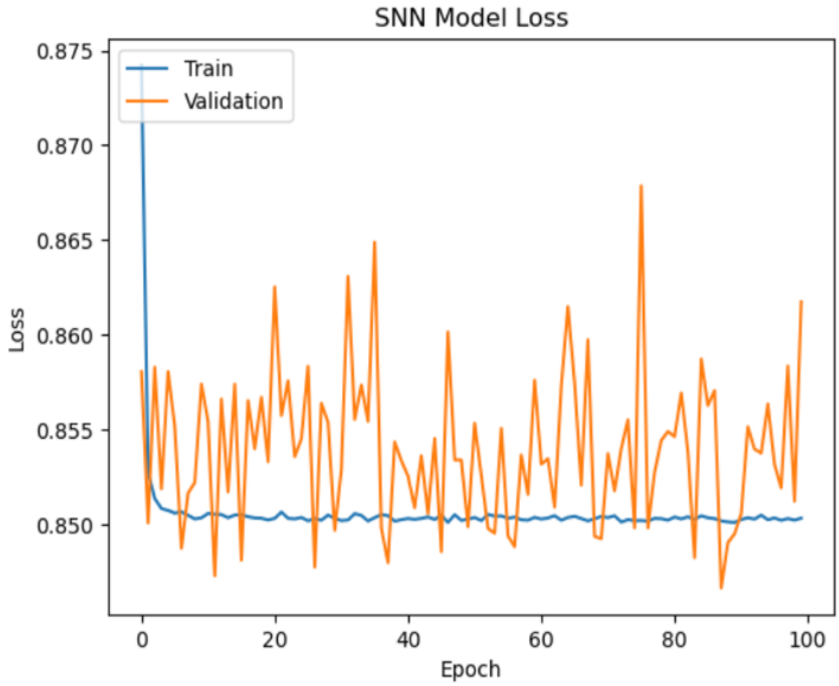


Figure 5: Epoch-Loss graph of SNN model

The model's performance can be assessed primarily based on the provided training and validation metrics:

Accuracy: The accuracy on the statistics starts at approximately 60.97% and regularly will increase. However, appears to stabilize around 62.36% after several epochs. On the validation statistics, the accuracy starts off evolved at about 61.84% and follows a comparable trend, stabilizing round 61.10%. It is clearly represented in *figure 4.*

Loss: The training loss starts at 0.8742 and decreases over time, attaining 0.8503 by means of the give up of training. The validation loss starts at 0.8581 and fluctuates barely, finishing at 0.8617. It is clearly represented in *figure 5.*

Overall, the model's accuracy is quite uniform and does not appear to improve considerably after a certain point in training. The loss values for both training and validation also level, indicating that additional training epochs may not assist the model significantly.

- **Convolutional Neural Network (CNN):**

Next, I implemented a Convolutional Neural Network (CNN) for sentiment analysis. First, class labels are encoded using one-hot encoding, and class weights are computed to address class imbalance. The CNN model architecture starts with an Embedding layer that utilizes pre-trained word embeddings, keeping them non-trainable. A 1D Convolutional layer with 128 filters and ReLU activation is added to capture text features. A Global Max Pooling layer extracts the most significant features, followed by a Dense layer with a softmax activation for multi-class classification. The model is compiled with the Adam optimizer, categorical cross-entropy loss, and multiple evaluation metrics.

It is then trained on the provided training data for 50 epochs with a batch size of 64, using class weights for balancing.

The performance of the CNN model as per the accuracy and loss matrices.

Accuracy: It is the measure of closeness of the predicted value with the actual value. In this model, the training accuracy starts at around 71.4% and gradually increases to approximately 78.9% over 50 epochs. But the validation accuracy is decreasing over the epochs. It is clearly represented in *figure 6.*

Loss: The loss in the model training represents how well the model's prediction matches the actual values or labels. Here, the training loss is decreasing, and the validation loss is increasing over the epochs. It is clearly represented in *figure 7.*

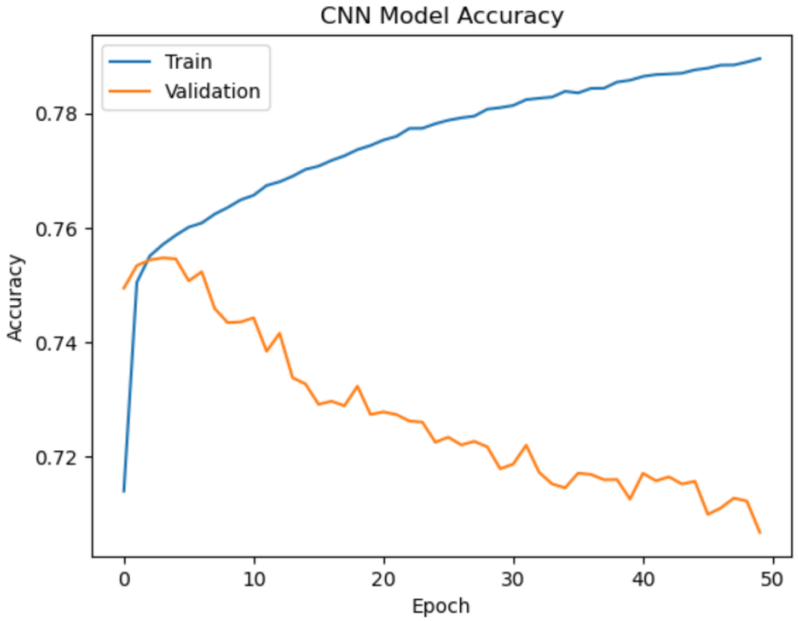


Figure 6: Epoch-Accuracy graph of CNN model.

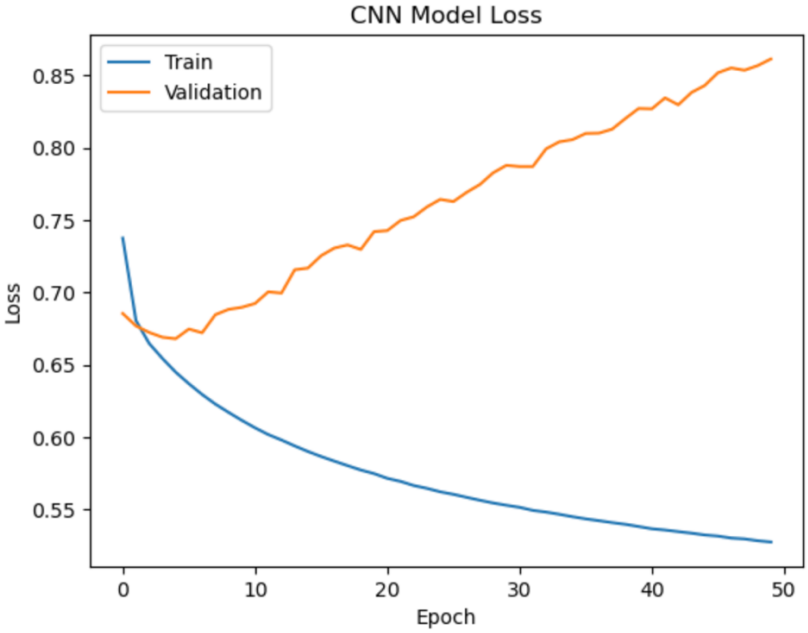


Figure 7: Epoch-Loss graph of CNN model.

- **Long Short-Term Memory (LSTM):**

Finally, we implemented a deep learning model for text classification using a Sequential LSTM neural network architecture. To prepare the text data, we encoded the class labels and applied one-hot encoding to facilitate multi-class classification. To address class imbalance, we computed class weights and incorporated them into the training process as we did in the previous model trainings. Our LSTM model comprises multiple layers, including pre-trained word embeddings, two LSTM layers with dropout for sequence processing, global max-pooling for feature extraction, and dense layers for classification. The model was compiled using the Adam optimizer and categorical cross-entropy loss, with metrics. It is then trained on the provided training data for 10 epochs with a batch size of 64, using class weights for balancing.

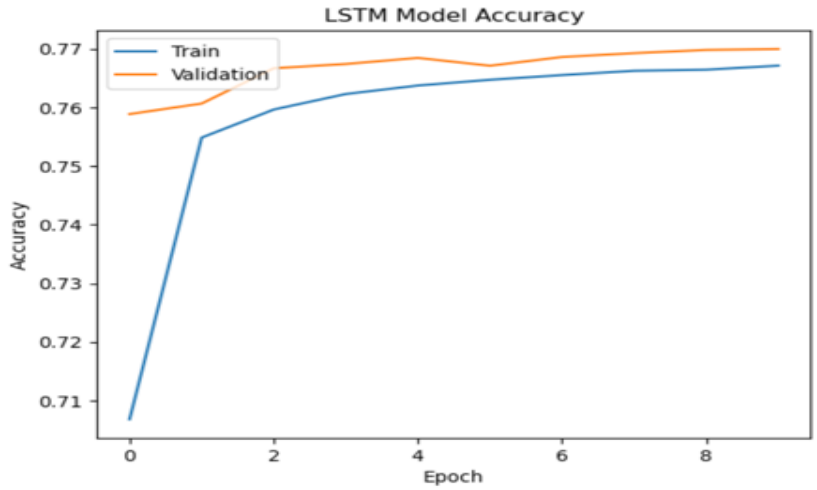


Figure 8: Epoch-Accuracy graph of LSTM model.

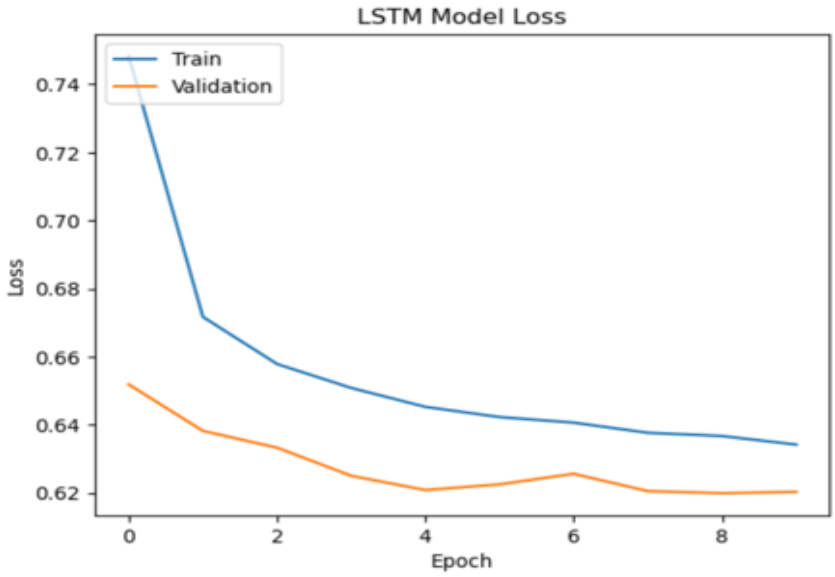


Figure 9: Epoch-Loss graph of LSTM model.

The accuracy and loss matrices of LSTM model can consider as quite promising. From the graph we can understand that,

Accuracy: In the beginning, the model's training accuracy over the epochs increased from 70.68% to an approximate value of 76.71% after 10 epochs. This indicates the model is training and improving its predictions on the training data. The validation accuracy also shows improvement over the epochs, reaching around 77.00% after 10 epochs. It is clearly represented in *figure 8.*
Loss: Here also we can see a positive sign that the training loss decreases from 0.7482 to 0.6341 over the 10 epochs. Lower training loss indicates the model is fitting the training data better. The decrease in validation loss from 0.6518 to 0.6203 indicates that the model is not only improving on the training data but also on the validation data as well. It is clearly represented in *figure 9.*
As a result, we can see that comparatively LSTM model is better among the three model which we have trained for textual sentimental analysis for the project. For further comparisons, this project is using more models such as pre-trained roBERTa model and VADER (Valence Aware Dictionary and sEntiment Reasoner) tool. VADER is a lexicon and rule-based sentiment analysis tool which we have used in pre-processing of the data.

Extracting the YouTube transcription and performing the sentimental analysis:

In this phase, transcription is extracted and analysed from a YouTube video. The code begins with inputting a YouTube link of a movie-reaction or product-review, and it is the data that used on the text part of the entire project.

Currently, YouTube has two types of videos: regular long-duration landscape videos and short-duration portrait videos called YouTube Shorts. Here, a 'youtube_transcript_api' library for extracting the transcription from the video is used. This library is not applicable to all the videos, as many of the YouTube videos do not have subtitles/caption. However, as a first step, this will assist in attaining a precise result and can perhaps get better in the future. As the video's ID is only needed for the transcription, an if' condition is used, which can remove the prefix from the video link and get only the video ID, even if it is a YouTube Shorts video link. Then, with the help of the 'youtube_transcript_api' library, the retrieved transcription and timestamp are saved into a DataFrame without the word '[Music]' because the YouTube transcription will have this word whenever there is only background music.

Then the transcription dataset will pass through a calculation, where the code helps to organize spoken segments into 3-second intervals, which can be helpful in analysing, visualizing, and combining with the emotion in the video part. With the cleaned transcript data in hand, we proceeded to perform sentiment analysis using the VADER sentiment analysis tool. To facilitate this, we imported the 'SentimentIntensityAnalyzer' from the NLTK library. VADER is a lexicon and rule-based sentiment analysis tool designed for analysing the sentiment or emotional tone of text data, particularly in the context of social media and short, informal text such as tweets, reviews, and comments. It was developed by researchers at the Georgia Institute of Technology.

For each spoken segment in the transcript, sentiment scores are computed using the VADER analyser. The sentiment scores are positive, negative, neutral, and compound scores, which collectively capture the sentiment of each spoken word. These scores provide valuable insights into the emotional tone or sentiment expressed in the YouTube video we are considering.

The sentiment analysis findings were added to a Pandas DataFrame, which contained the text, VADER sentiment scores, and a derived sentiment label. We assigned the sentiment label based on predetermined threshold values for the compound sentiment score. Segments with compound scores of 0.1 or higher were labelled as 1; segments with scores of 0.1 or lower were labelled as 2; and the remaining segments were labelled as 0, which represents the positive, negative, and neutral sentiments, respectively.

VADER sentiment scores are calculated according to the sentiments for each word, excluding stopwords, unwanted multiple characters and punctuation. There is a chance of compromising the accuracy when it comes to a long-script or content in a long-duration video. In that case, we can consider using a pre-trained RoBERTa (A Robustly Optimized BERT Pretraining Approach) model for better accuracy.

RoBERTa is a state-of-the-art NLP model that builds upon the success of BERT (Bidirectional Encoder Representations from Transformers). RoBERTa represents a significant advancement in pre-trained language models and has achieved outstanding performance on a wide range of NLP tasks. The code used in this project utilizes the Hugging Face Transformers library, and it loads a pre-trained RoBERTa model for sentiment analysis specifically trained on Twitter data: "cardiffnlp/twitter-roberta-base-sentiment" [9]. This is a base model of RoBERTa that is trained on about 58 million tweets and fine-tuned. Therefore, this model will have better accuracy in terms of human reviews and opinions. At the end of the sentimental analysis on text, a dataset with timestamps from the video, transcript speech text, and result of predictions using Simple neural network model, CNN model, LSTM model, VADER and RoBERTa model is achieved.

Sentimental Analysis on Video:

In this phase, the emotion expressed by the content creator or the person in the video is detected using CNN, and the recognised emotion from the video frames will be collected into a new dataset for visualisation. The dataset for training the model is collected from the Kaggle platform [10].

Data Pre-processing: This section began by organising the dataset into training and testing sets. The dataset contained more than 21000 images representing different emotions, such as positive, negative, and neutral. Data augmentation techniques, such as rescaling, shearing, zooming, and

horizontal flipping, were applied to the training data. This enhanced the model's ability to generalise and recognise emotions in various image conditions.

CNN architecture: In Keras, a sequential model is a linear stack of layers, which makes it easy to define and build deep learning architectures layer by layer. The first layer added to the classifier is a convolutional layer (Conv2D). It uses 16 filters with a size of 3x3 pixels.

The ‘input_shape’ parameter is set to (128, 128, 3), which means the expected input images have a size of 128x128 pixels and three colour channels (red, green, blue). To introduce non-linearity, the activation function called ‘ReLU’ is applied. After that, a max-pooling layer (MaxPooling2D) is added. The ‘pool_size’ parameter is set to (2, 2), which indicates that the operation will reduce the spatial dimensions of the feature maps by taking the maximum value within a 2x2 window.

Subsequently, another convolutional layer is incorporated to capture additional intricate image features, leveraging 32 filters at a 3x3 pixel scale. Following this, another round of max pooling with identical parameters is applied. To bridge the gap between convolutional and fully connected layers, a flattening layer (Flatten) restructures the 2D feature maps into a 1D vector, preparing the data for subsequent processing.

Then two fully connected, dense layers are added. The first layer has 128 units and uses the 'ReLU' activation function, and the final dense layer has 3 units, corresponding to the three classes (positive, negative, and neutral), and it uses the 'softmax' activation function. 'Softmax' is commonly used for multi-class classification problems as it outputs probabilities for each class.

The compile method is called to configure the model for training. The 'adam' optimizer, which is a popular choice for gradient-based optimisation, is used. The loss function is set to 'categorical_crossentropy,' which is suitable for multi-class classification tasks, and the 'accuracy' metric is chosen to monitor the model's performance during training.

Training and Testing: The model was trained using the training dataset, with hyperparameters tuned for optimal performance. Testing was performed using a separate testing dataset to evaluate the model's accuracy and performance. Then the trained model is saved for future use and evaluated on the testing dataset. After saving the model, the next step is to make the video ready for testing.

In the beginning of the training the accuracy was 50% and increased to an approximate training accuracy of 80% and validation accuracy of 70% at the end of the training. Similarly, the training loss of the model was high at the beginning and decreased at the end of the training. It means the model is minimising error during training. The graph represents accuracy and loss are plotted in *figure 10* and *figure 11* respectively.

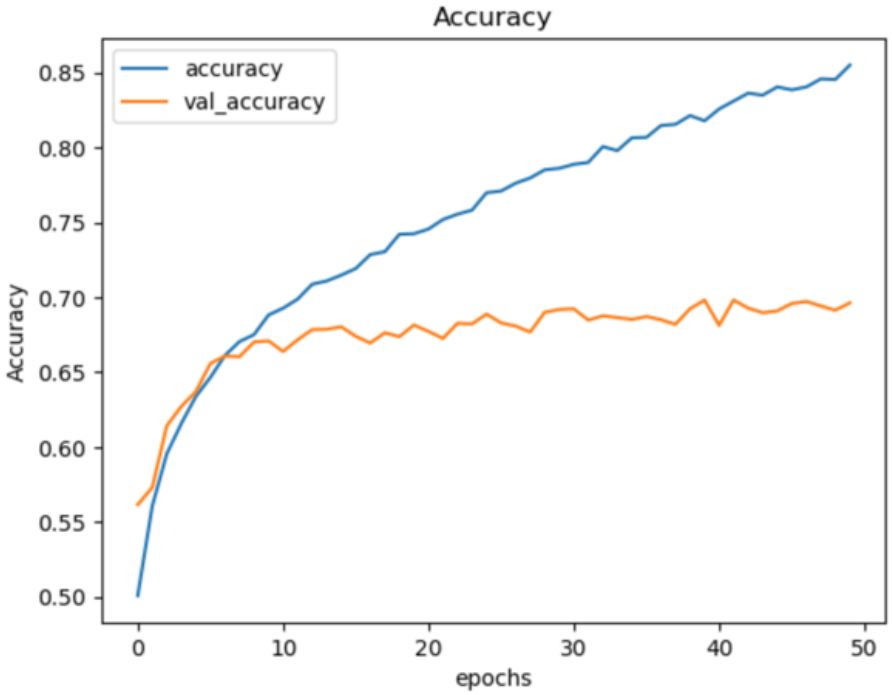


Figure 10: Epoch-Accuracy graph of video sentiment analysis model

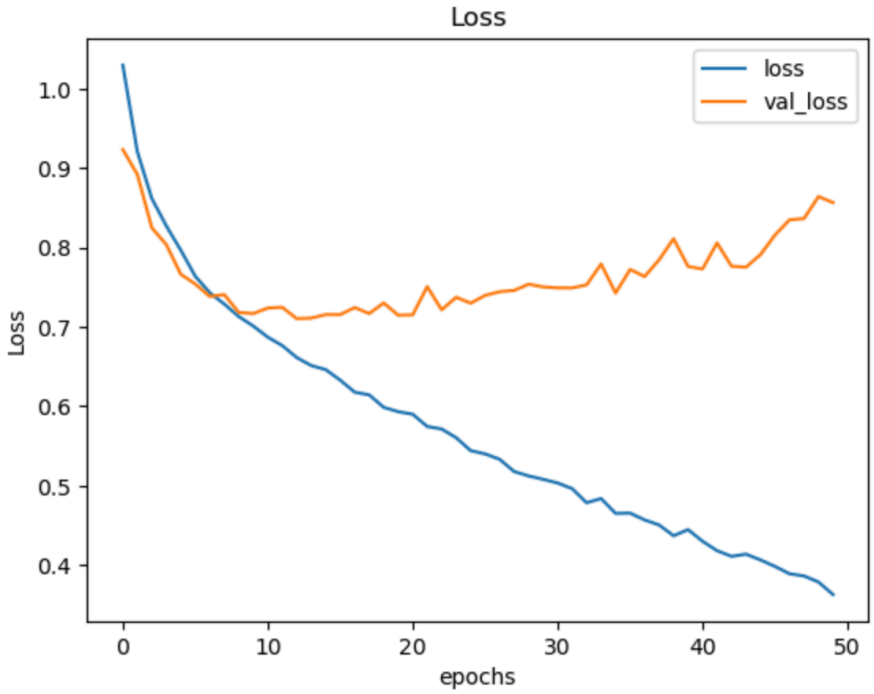


Figure 11: Epoch-Loss graph of video sentiment analysis model

In the testing phase, the process began with the extraction of a YouTube video using the 'PyTube' library, enabling the retrieval of video content from a provided URL. The video was then downloaded and stored for analysis. Next, the video's duration was calculated by examining its total frames and frames per second (fps) for visualisation and further analysis purposes.

Then, during the video preparation, the saved deep learning model was loaded to predict sentiment or emotion from video frames. Each frame of the video was examined individually, and predictions

were produced for each frame. The predictions were timed to measure the processing speed of the model. To enhance the interpretability of the results, a time-based analysis and calculations were performed, and a dataset with the mean value of the sentimental score was exported by considering a specific time interval of 3 seconds. This step was important for visualising the results from textual and video sentiment analysis. After the sentiment analysis on text and video, the predicted results were combined into a single DataFrame.

Visualisation:

The outcome dataset is subjected to several visualisations. These visualisations allow us to comprehend how each model interprets the sentiment from the textual and visual data. The first visualisation is a Wordcloud made up of the speaker's words from the video which is represented as *Figure 12*.

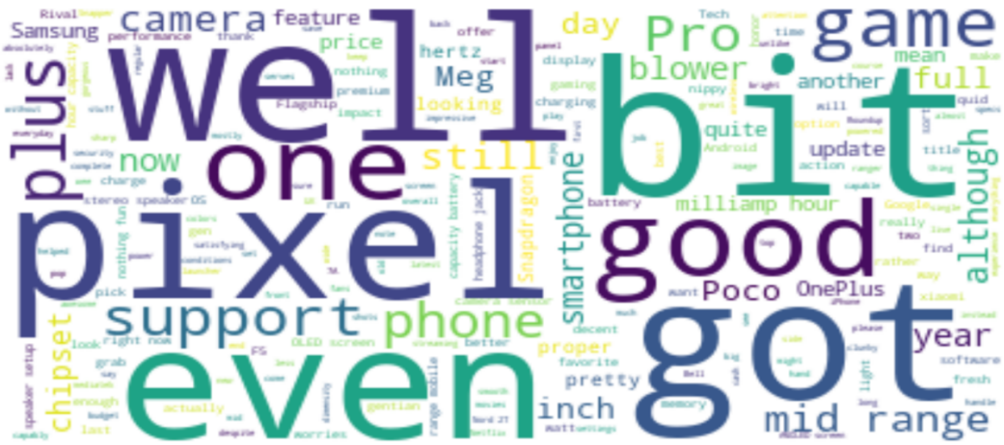


Figure 12: Wordcloud visualisation of the example YouTube Video

From this visualisation, the larger words will be the words which are most used by the content-maker in the selected video.

As there are six results in the result dataset, the entire count of number of classes is printed as summary_table as shown below in *figure 13.*

```
In [469]: # Example: Create a summary table
summary_table = pd.DataFrame({
    'roBERTa_Sentiment': count_rob,
    'Vader_Sentiment': count_vad,
    'LSTM_sentiment': count_lstm,
    'CNN_sentiment': count_cnn,
    'SimpleNN_sentiment': count_snn,
    'Video_Emotion': emotion_vid
})

print(summary_table)
```

```
          roBERTa_Sentiment  Vader_Sentiment  LSTM_sentiment  CNN_sentiment  \
negative                 41               36             124             90
neutral                 244              193              83            115
positive                135              191             213            215

          SimpleNN_sentiment  Video_Emotion
negative                 105            147
neutral                   98            171
positive                 217            102
```

Figure 13: summary_table

5. Result

Evaluation

In this project, the primary aim was to develop a model that accurately discerns the sentiment expressed in video reviews for various products or movies. To achieve this, YouTube has chosen as the platform for data collection, given its extensive repository of video reviews. The initial step involves building models for sentimental analysis on text data.

For sentiment analysis on text, we experimented with various deep learning techniques, including Natural Language Processing (NLP), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). These steps were very time-consuming and but obtained a good accuracy with some of the models. Among the own-trained models, LSTM model provided the better accuracy in predictions.

Through visualizations of the results, we can see that all the models are giving different sentiments from the entire video which is represented in *figure 14*. Therefore, a pre -trained roBERTa model is used, which was trained on 58 million tweets. Then extended the analysis to include emotion detection from the video content, classifying sentiments as negative, positive, or neutral. By training a large image dataset using CNN, I got a model with an accuracy of more than 69%.

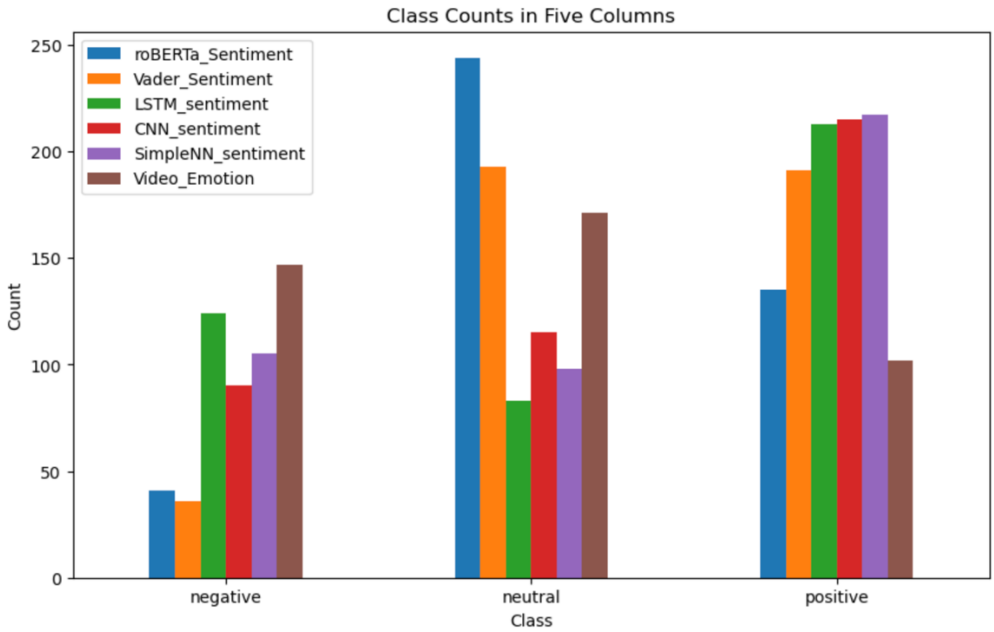


Figure 14: The count of predicted sentiments from 5 textually analysed models and video emotion model.

As the efficient models are, LSTM, roBERTa and the video emotion detected CNN model. The value count representation is done on these three models' results and represented in *figure 15*.

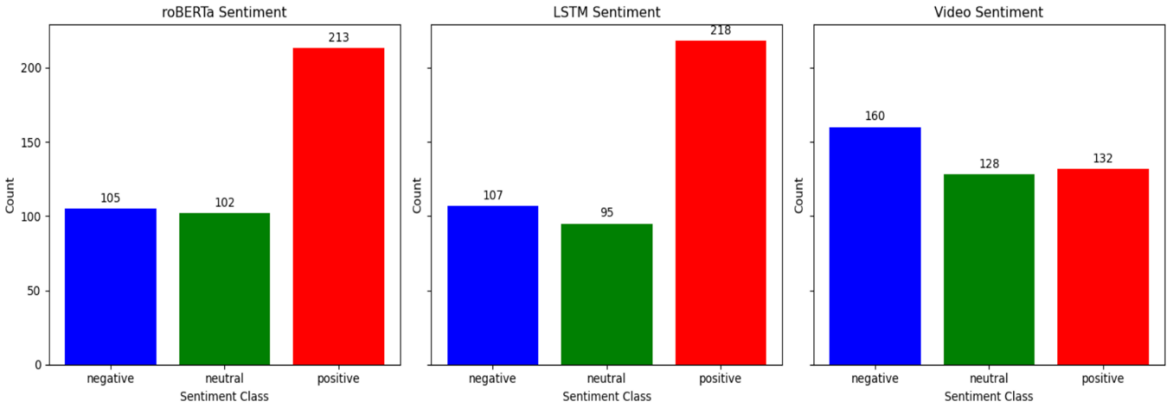


Figure 15: Value cpunt representation of LSTM, roBERTa and Video emotion models.

To address this limitation, the insights obtained from emotion detection and sentiment analysis in text is integrated. By fusing and analysing both text and video-based visualizations, I aimed to obtain a comprehensive understanding of the reviewers' overall impressions of the products. As both results from the sentimental analysis on text and video have an attribute time, it was easy to combine the visualizations, but with some inaccuracies to be studied. Visualisations on LSTM model result vs Video Emotion and roBERTa model result vs Video Emotion over first 90 seconds of the video are plotted in *figure 16* and *figure 17*.

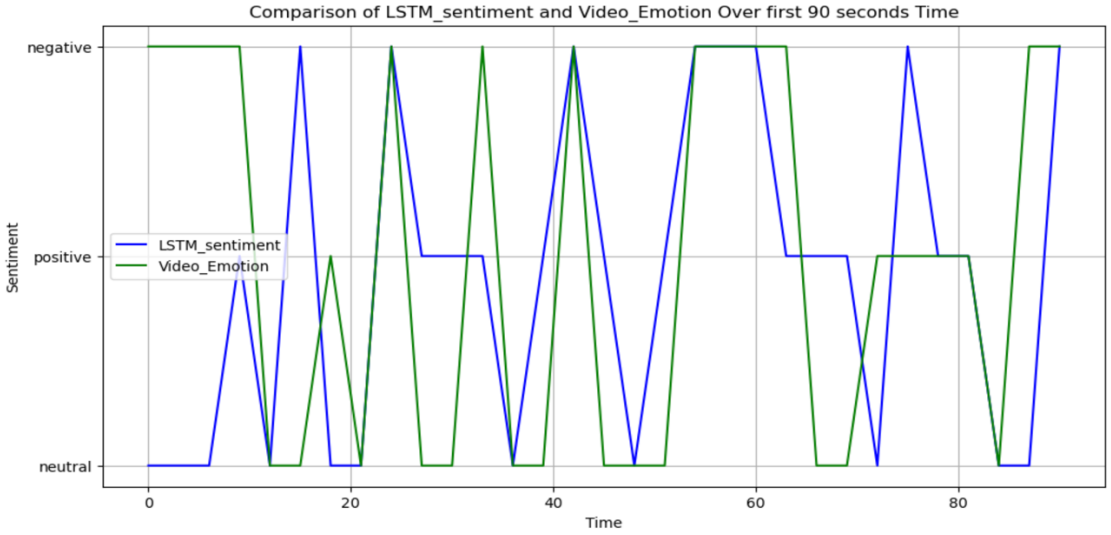


Figure 16: LSTM sentiment vs Video Emotion

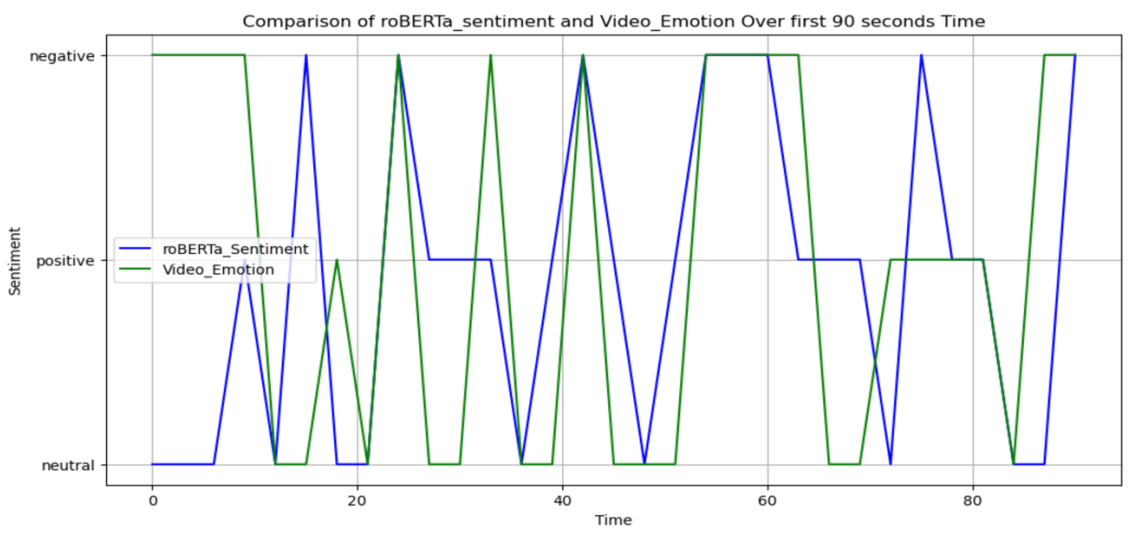


Figure 17: roBERTa result vs video emotion

When considering the emotions and textual sentiment, 'neutral' emotion from the visual sentiment results and any of the three sentiments from text can occur at the same time and vise-versa. For example, a person can speak negative, neutral and positive words in neutral expression. Similarly, a person can speak neutral words with neutral, positive, and negative emotions on face.

There are many loopholes and tricky considerations that need to be done in this topic as the same words can have different meaning and the way of using it by a person will be according to their place, accent, and facial orientation.

Related Work

Several illustrious academic publications have examined comparable topics in sentiment analysis in multimedia content. In a paper by Martin Wöllmer, Felix Weninger, Tobias Knaup, and Björn Schuller on the topic 'YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context', it investigates sentiment analysis within the context of YouTube movie reviews, which aligns closely with the essence of this project. It explores sentiment within audio-visual content, offering valuable insights into the challenges and opportunities associated with analysing emotions in video data. Comparatively, our project, which conducts sentiment analysis on YouTube videos, extends beyond movie reviews to a broader range of video content. However, both works share the common objective of understanding sentiment in multimedia, providing a foundation for our project's relevance in this research domain.

The data collection, pre-processing techniques, and training are different from this project. But when it comes to the whole idea, it is similar. Here we are more concentrated on the visualization part, where we can easily understand the sentiment of the video content.

Another paper on the topic ‘Multimodal Aspect-Level Sentiment Analysis based on Deep Neural Networks’ focuses on multimodal aspect-level sentiment analysis, incorporating deep neural networks to analyse sentiment across different aspects within multimedia content. Although our project primarily focuses on overall sentiment analysis within YouTube videos, this work aligns with the broader concept of sentiment analysis in multimedia. It highlights the significance of leveraging neural networks for enhanced sentiment understanding, which resonates with our project's use of deep learning techniques.

In comparison to these related works, our project extends its scope to perform sentiment analysis on YouTube videos, and it is mainly focused on the visualization of the results.

6. Conclusion

In conclusion, this project's primary goal of performing sentiment analysis on YouTube videos and offering insightful information about the emotional content expressed in the videos was successfully achieved. We successfully identified attitudes as Positive, Negative, or Neutral over the course of video lengths by combining deep learning algorithms, video processing, and natural language processing. This strategy not only improves our comprehension of the emotional dynamics in videos, but it also shows how sentiment analysis could be applied to multimedia content.

Reflection

In the initial stage, my plan was to directly train and test the dataset and compare the accuracies to proceed with the model for sentimental analysis on text. In the development stage, I learned that data pre-processing and feature engineering are pivotal steps that can significantly impact the success of a project. Challenges, such as varying video lengths and language nuances, underscored the need for robust data handling strategies lead to utilize a pre-trained model to accomplish my objective.

In the video emotion detection model, the collected dataset was huge and accurate. In hindsight, there are aspects we would approach differently. Enhanced project planning, particularly in terms of resource allocation and time management, could have streamlined certain phases of the project. Additionally, deeper exploration of advanced sentiment analysis techniques, including fine-grained emotion detection, could have enriched the analysis.

Reflecting on the project's timeline, it became apparent that the allocated time felt rather limited considering the complexities and the novel techniques involved in this field. Therefore, this project should be seen as a first step, a base from which other projects can be constructed. There is plenty of space for development and extension, including a wider range of methodologies and improving the pre-processing techniques used on the data.

Future Work

In the realm of sentiment analysis within multimedia content, this project has paved the way for several promising avenues of future work. First and foremost, a deeper investigation of comprehensive sentiment analysis methods, such as fine-grained emotion identification, could offer deeper understandings of the emotional dynamics of video content. Using advanced pre-trained models like BERT and RoBERTa, customized for analyzing multiple data types, could greatly improve the accuracy and depth of sentiment classification.

Additionally, real-time sentiment analysis remains an intriguing area for further investigation. Developing a system capable of providing live sentiment feedback during video playback could have applications in content recommendation, user engagement analysis, and even mental health support. In conclusion, this project marks the beginning of an exciting journey rather than the end. The field of multimedia sentiment analysis has a lot of room for improvement, growth, and innovation in the future.

7. References

[1] E. Cambria, B. Schuller, Y. Xia, C. Havasi, and A. Cichocki, "New Avenues in Opinion Mining and Sentiment Analysis," in *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15-21, 2013.

[2] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746-1751, 2014.

[3] M. Wöllmer, F. Weninger, T. Knaup, and B. Schuller, "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context," *Technische Universität München*.

[4] M. Ban, L. Zong, J. Zhou, and Z. Xiao, "Multimodal Aspect-Level Sentiment Analysis based on Deep Neural Networks," in *School of Computer and Software Engineering, Xihua University, Chengdu, China*, 2021.

[5] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines," *Nanyang Technological University, Instituto Politecnico Nacional, National University of Singapore, Edinburgh Napier University*, 2017.

[6] Z. Liu, "Sentiment-Analysis of Review Text for Micro-video," *Information, Engineering Institute, Communication University of China, Beijing*, 2019.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *Proceedings of the Annual Conference on Neural Information Processing Systems,* Volume 32, 10492-10501.

[8] Liao, Wenxiong; Zeng, Bi; Yin, Xiuwen and Wei, Pengfei. (2021).” An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa” 10.1007/s10489-020-01964-1

[9] "Barbieri, Francesco and Camacho-Collados, Jose and Espinosa Anke, Luis and Neves, Leonardo", "*{T}weet{E}val: Unified Benchmark and Comparative Evaluation for Tweet Classification" (2020)* ,[Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment.

[10] ARES,“www.kaggle.com”(2020),[Online].Available:https://www.kaggle.com/datasets/ananthu017/emotion-detection-fer

8. Appendices

GitHub link: https://github.com/Akshaymohan7/Sentimental-analysis.git
Google Drive link:
https://drive.google.com/drive/folders/1Vp_wkdHooOQXnvRzahk28Hf7PcVWRXeY?usp=drive_link

Gantt Chart:

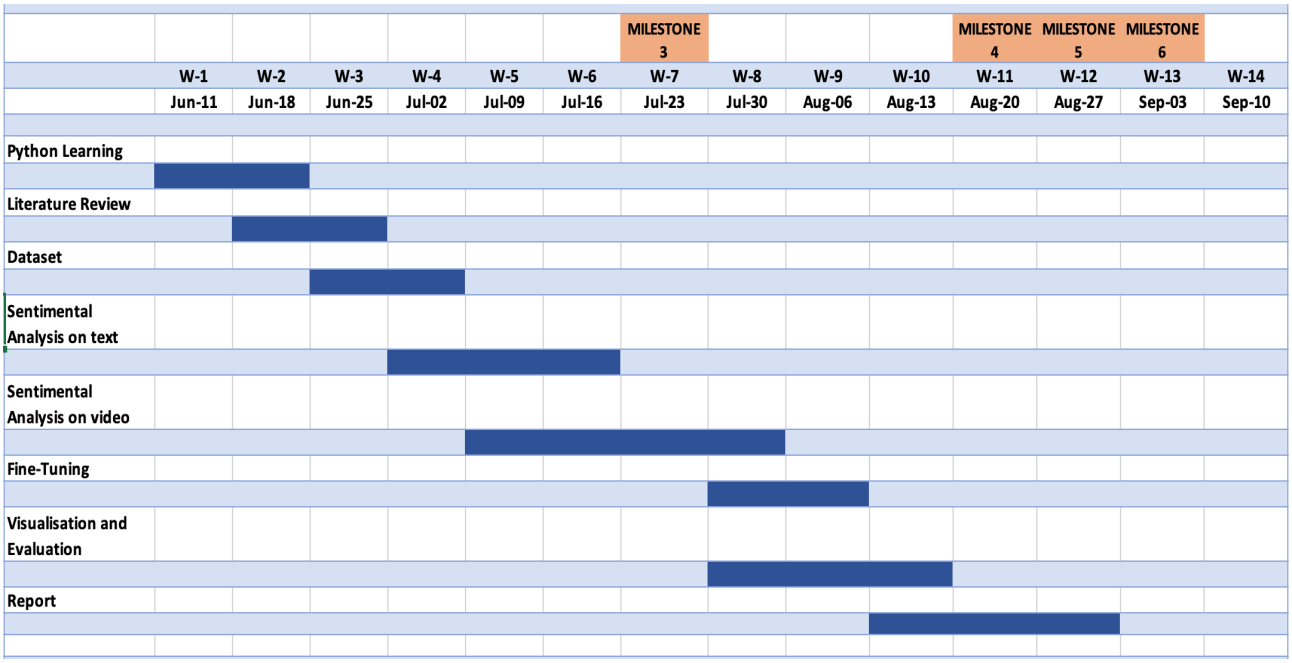


Figure 18: Gantt Chart

Project Proposal:

MILESTONE 02: Project Proposal

BSc & MSc Projects

Your first and last name

Akshay Mohan

Your Student ID

MOH21542963

What degree programme are you on?

MSc Data Science

What is the working title of your project (this can be changed at a later date)?

Sentimental Analysis on video

What is the principal problem that your project aims to resolve?

The aim of this project is to do a sentimental analysis on the video product reviews by evaluating the video part, transcription and visualize it.

Describe your approach to solving the principal problem, and the technologies that will be used?

Steps for solving this problem are,

Data collection in the form of video from various sources.
Video transcription and sentimental analysis on the transcripted text using NLP.
Emotion detection from the video.
Combining and visualizing the result from the emotion detetction and sentimental analysis.

Technologies:
Python programming
Natural Language Processing

Classify your project as a technology theme (i.e. How you want your project to be scrutinsed)

Artificial Intelligence & Machine Learning

If you selected "Other" above, please specify your theme below.

How will you test and evaluate your project?

from the visualization of the sentimental analysis, we will get the impression that is expressed from the video as positive , negative or neutral.

Figure 19: project proposal

Supervisor (First Marker)

Changjiang He

Second Marker (Second Supervisor)

Fakhreldin Saeed

List up to 3 aims of your project.
NOTE: An aim is an expected outcome of your project (e.g., issues it will address, how it might improve or enhance a situation for stakeholders, etc.)

1) Enhance viewer understanding: The project aims to improve viewers' understanding of sentiments and emotions expressed in videos by providing comprehensive sentiment analysis and emotion detection results.

2) Enable actionable insights: The project aims to generate actionable insights for various stakeholders, such as content creators, marketers, and social media analysts, by identifying and visualizing sentiments and em

3) Improve user experience: By developing a combined graph visualization, the project aims to enhance the user experience of exploring sentiment and emotion data in videos. The visualization will provide an intuitive a

List up to 4 key objectives of your project.
NOTE: Objectives are tangible tasks that you will complete. They are typically steps/activities that you must complete in order to deliver your project aims successfully.

1) Collect the dataset and preprocess the video content by extracting transcript to obtain the necessary textual data for sentimental analysis and visualization.

2) Implement natural language processing techniques and machine learning models to perform sentiment analysis on the video transcripts, aiming to accurately determine the sentimental impression on viewers.

3) Design and develop a visually appealing and informative combined graph visualization that effectively represents the results of sentiment analysis and emotion detection

Figure 20: Project proposal

List background/literature/technology review sources, that have been used to inform your project.

Python programming skill
Analysing and visualising skills
Machine learning
Knowledge on Natural Language Processing.

Describe any risks, ethical issues or other factors that are relevant to this project.

Privacy and Data Protection
Bias and Fairness
Informed consent
Data quality and representation
Time limit

Student and First Supervisor Project Sign Off

STUDENT: I agree to completing this project: ✔ **Date:** 06 / 06 / 2023

Student Name: Akshay Mohan

Student Signature: Akshay Mohan

SUPERVISOR: I approve this project proposal: **Date:** / /

Supervisor Name: Changjiang He

Supervisor Signature:

NOTE: It is the supervisor's responsibility to approve this project as meeting the requirements for the module. This includes professional body requirements, programme requirements, and module requirements. By signing the form, you are agreeing that you have validated the suitability of the project.

Figure 21: Project proposal

Weekly Summary:

Weekly summary

Name: Akshay Mohan
Date: 22/06/2023
Week: 1 & 2
Supervisor: Changjiang He

1. What have you done for the past week?
   - I was learning and improving my skills in Python Programming.
   - Understanding currently available sentimental analysis projects.

2. What are the issues you are facing right now and how will you solve it?
   - No issues now.

3. What are you going to do for the next week?
   - I need to find more relevant research papers.

Figure 22: Weekly Report (week 1&2)

Date: 29/06/2023
Week: 3
Supervisor: Changjiang He

1. What have you done for the past week?
   - Literature Review.
   - Collecting dataset.

2. What are the issues you are facing right now and how will you solve it?
   - I need to find more relevant research papers.

3. What are you going to do for the next week?
   - Start with basic text sentimental analysis.

Figure 23: Weekly Report (week 3)

Date: 06/07/2023
Week: 4
Supervisor: Changjiang He

1. What have you done for the past week?
   - Extracted the transcription from a 'YouTube video' using the API.
   - Prepared a dataset using attributes such as time and the text content of that video.
   - Done the sentimental analysis on that text data using 'VADERS' and 'roBERTa'.

2. What are the issues you are facing right now and how will you solve it?
   - When I used the 'SpeechRecognition' python Library for videos without transcription. I was not able to get an accurate result from it.
   - Therefore, initially I'm trying to do this entire project on YouTube review videos and then with other videos.

3. What are you going to do for the next week?
   - Next week I'll continue working on the text sentiment analysis by using my own model and its visualisation.

Figure 24: Weekly Report (week 4)

Date: 13/07/2023
Week: 5
Supervisor: Changjiang He

1. What have you done for the past week?
   - Tried different approaches of deep learning on sentimental analysis of text such as CNN and RNN (LSTM).

2. What are the issues you are facing right now and how will you solve it?
   - Facing difficulties to come up with an innovative approach.

3. What are you going to do for the next week?
   - I need to compare the accuracy and performance of all the models I have tried on my dataset.

Figure 25: Weekly Report (week 5)

Date: 20/07/2023
Week: 6
Supervisor: Changjiang He

1. What have you done for the past week?
   - Compared the accuracy and performance of sentimental analysis on text using RNN and pre-trained models.
   - Started working on video emotion detection such as positive, negative and neutral.
   - Started doing the report.

2. What are the issues you are facing right now and how will you solve it?
   - Facing issues in improving the accuracy of the model.

3. What are you going to do for the next week?
   - Fine-tune the selected model.
   - Combine the visualizations of text and visual sentimental analysis.

Figure 26: Weekly Report (week 6)

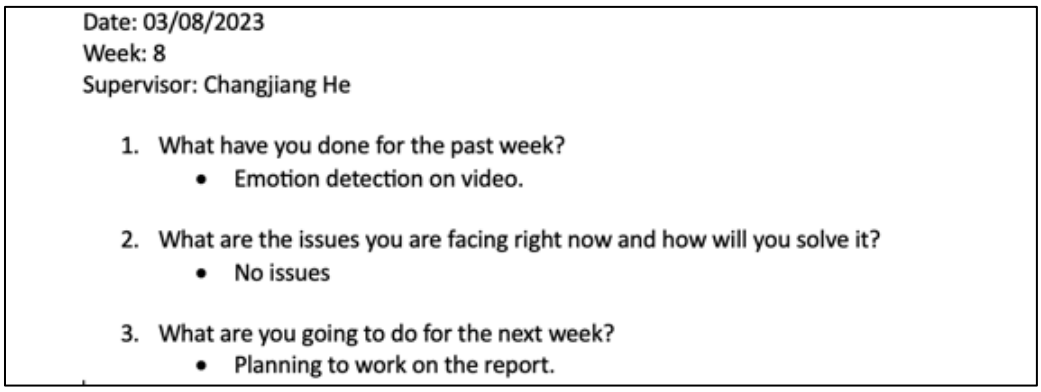
Date: 27/07/2023
Week: 7
Supervisor: Changjiang He

1. What have you done for the past week?
   - I was working on the mid-point review documentation.

2. What are the issues you are facing right now and how will you solve it?
   - Currently no issues

3. What are you going to do for the next week?
   - Fine-tune the selected model as planned for the current week.
   - I have to work on more visualizing ideas.

Figure 27: Weekly Report (week 7)



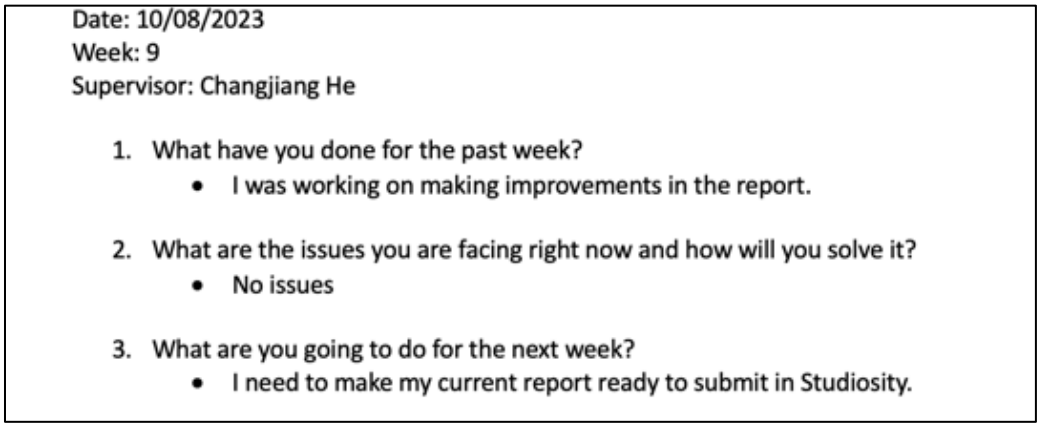
Date: 03/08/2023
Week: 8
Supervisor: Changjiang He

1. What have you done for the past week?
   - Emotion detection on video.

2. What are the issues you are facing right now and how will you solve it?
   - No issues

3. What are you going to do for the next week?
   - Planning to work on the report.

Figure 28: Weekly Report (week 8)



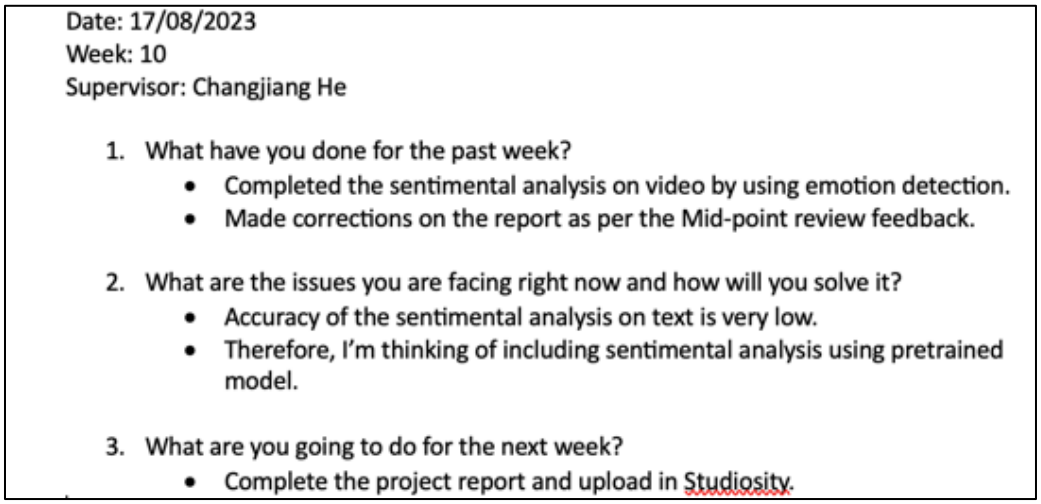
Date: 10/08/2023
Week: 9
Supervisor: Changjiang He

1. What have you done for the past week?
   - I was working on making improvements in the report.

2. What are the issues you are facing right now and how will you solve it?
   - No issues

3. What are you going to do for the next week?
   - I need to make my current report ready to submit in Studiosity.

Figure 29: Weekly Report (week 9)



Date: 17/08/2023
Week: 10
Supervisor: Changjiang He

1. What have you done for the past week?
   - Completed the sentimental analysis on video by using emotion detection.
   - Made corrections on the report as per the Mid-point review feedback.

2. What are the issues you are facing right now and how will you solve it?
   - Accuracy of the sentimental analysis on text is very low.
   - Therefore, I'm thinking of including sentimental analysis using pretrained model.

3. What are you going to do for the next week?
   - Complete the project report and upload in Studiosity.

Figure 30: Weekly Report (week 10)