Emotion Recognition of Multi-modal YouTube Data

MSc Research Project
Data Analytics

Ankur Ghogale
Student ID: x19193866

School of Computing
National College of Ireland

Supervisor: Prof. Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ankur Ghogale
Student ID:	x19193866
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Prof. Noel Cosgrave
Submission Due Date:	17/12/2020
Project Title:	Emotion Recognition of Multi-modal YouTube Data
Word Count:	5973
Page Count:	18

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:	
Date:	1st February 2021

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):	

Emotion Recognition of Multi-modal YouTube Data

Ankur Ghogale
x19193866

Abstract

Nowadays, watching videos has become one of the most common leisure activity and video streaming platforms are facilitating audience to access wide range of video categories. YouTube, one of the famous platforms for online videos does not consider emotions in their video categorization models. This research aims at classifying the videos into emotion classes (i.e., anger, happy, sad, disgust,fear, neutral, and surprise) based on video, audio and text modality using One-Minute gradual emotion dataset. This dataset comprises of transcription, links for YouTube videos, emotions, along with the arousal and valence values. It also discusses data collection, cleaning, and pre-processing techniques for all the three modalities. We have opted for ensemble method approach to recognize emotions in the video by fusing the model's output which were implemented on the different modalities individually. The results show that emotion recognition becomes more accurate by using multiple modalities.

1 Introduction

With the advent of 4G and 5G, internet speed and bandwidth has grown drastically, resulting in extensive development in video streaming web sites and making video information sharing an usual and fun activity. Owing to this reason, watching online videos is becoming the new trending leisure activity among the internet users. Several organizations are working hard to introduce new technologies which can improve user experience and satisfy rising demand.

Given the success of video streaming platforms, several platforms have emerged in the recent years but the most popular one which is ranked top among the video streaming websites is YouTube. YouTube has billions registered users worldwide and every day billions of hours are spent watching YouTube videos by the users collectively, generating millions and billions of views each day.As of 2019 statistics, 30 thousand hours of videos are uploaded each hour and around 9 hours of video for every second. Such a high volume of video data sparks a key question: how an user can obtain a filtered results which could match his mood and emotion?

It is vital to create an user-friendly browsing system which can filter the videos from the vast collection of videos as per the need of user. YouTube allows uploaders to tag the videos into 15 categories, helping user to search the videos using keywords. Moreover, it also uses video categorization methods to help user search the videos they are interested in. YouTube's video categorization model focuses on the content description, tags, comments from the viewers, and also number of likes and views. Higher the number of likes and views of a video, topper it will go in the results. However, the current video categorization technique does not consider emotion as a parameter for filtering the results. Therefore, we propose an approach to categorize YouTube videos based on the emotions which can help users to generate more appropriate responses for their search.

Emotion analysis has been an emerging topic in the last few decades due to its high application value for providing human-machine interactions and facilitating machines to provide value added services. This study has been motivated from the need for enhancing user's video browsing experience by categorizing YouTube videos into seven emotion categories which may produce following benefits:

1. Enhancing user experience by augmenting YouTube's search results filtering techniques. Emotion can play an important role in enabling user to get results which are more suitable to their expectations.

2. Enhancing effectiveness of YouTube recommendation system. With the addition of emotion feature, recommendation system will be suggesting videos which match closely to user's emotion choice.

3. Improving business and product advertisement. Business advertisements can be shown to the users as per their choice of emotion which can enhance the effectiveness of advertisement.

This research paper addresses the following questions: "How well can the combined classifier of the three modalities (Text, Audio, Video) perform better than the individual classifiers on recognizing emotions using YouTube data?"

The paper is structured as follows: Section 2 describes recent work done in the field of multimodal deep learning and discusses the approach which can be implemented in this project. Section 3 explains methodology implemented in this project. Section 4 shows design specification of the project framework. Section 5 shows the model implementation and system configuration for this project. Section 6 discusses the evaluation techniques on the three modalities and finally the discusssion and results.

2 Related Work

The author Kaya et al. (2017) describes multimodal approach chosen by the author for emotion recognition. The author used video and audio data out of laboratory data which represents real life and realistic problems. The various methods used for emotion recognition are image purification using PCA, visual descriptors, local phase quantization, local binary patterns, and CNN. Pre-trained CNN model and transfer learning was used to avoid problems like over-fitting. The author mainly focused on cleaning and aligning data for the better detection of face by the model. CNN model is a good model for emotion classification but it is a challenge to explain CNN results in real life application.

The researcher Herzig et al. (2017) has shown comparative study of embedding techniques on multiple datasets to get a better idea about the impact of these techniques on results. The author considers word embedding an effective approach for emotion recognition using deep learning and used pretrained representations like word2vec and GloVe algorithms due to limited data. After embedding, the author performed multiple emotion classification on binary classification models using one-vs-rest and ensemble methods. The results show that GloVe gave better results in terms of accuracy. However, this approach is gives low accuracy for domain specific data even after implementing domain adaptation and the results can be better if deep learning models would have been used.

The data was obtained from multimodal recording setup where facial videos, vocal expressions, eye gaze, and psychological signals were considered. The main intention of this research is to evaluate the impact of EEG signals, pupillary reflexes, nonverbal cues and other bodily responses. Here the ground truth was taken by calculating the median of arousal and valence for each instance. Initially one way ANOVA test was implemented for feature selection and significant feature extraction for pupillary responses was done by using PCA. For this multiclass classification, the author Soleymani et al. (2012) used binary classification SVM model and converted it multiclass by taking one-vs-all approach using a sigmoid fit. The author calculated F1 score for all the classifiers separately to evaluate the performance of each class. Further parameter tuning of the model was done using leave-one-participant-out cross-validation scheme where C value for the model was empirically set to 1. However, the data collected for this research was very limited but it sufficed the need of research.

This paper investigates combination of different auto-encoders and its effect on emotion recognition. The author Bairaju et al. (2019) used combination of traditional, convolutional auto-encoder and recently developed convolutional pooling layers, and VGG-Net 16 for feature extraction and emotion classification. Series and parallel combination these techniques were used on the JAFFE dataset for this research. The best performance was obtained for parallel combination of traditional auto encoders and VGGNet-16 with accuracy range of about 35-60 percent and the series combination of deep learning gave the lowest accuracy for this dataset. Since not much information about data cleaning is mentioned we don't know how well the data was prepared and what are the anomalies in the dataset.

The research by the author Shen et al. (2020) focuses on multimodal emotion reasoning in videos using audio, text and visual signals and moreover it also introduces a new dataset called MEmoR. This dataset covers videos with 14 emotions including Ekman's 6 basic emotions. Initially, the annotated dataset is separated into three datasets depending upon the features like audio, video, text and personality. These datasets are further treated differently with different techniques for preprocessing. Then the author has used encoders to encode the multiple modalities into compact representations. He also compares different baseline models and their implementation with the proposed attention-based model on the given four modalities. The model was evaluated using micro-F1, macro-F1, and weighted-F1 metrics. The results showed that the multi-modal fusion method Bi-LSTM performs better for primary emotions but fails to classify fine-grained emotions. However, the proposed model outperformed all the models.

The main objective of this research was to create a framework for sequential multimodal data comprising of text, visual and audio data. The author Chandra and Hsu (2019) proposed a new method called Deep Attentive Residual Disconnected Recurrent Neural Network (DADRNN) which uses extracted deeper features by attending information from previous signals. The mechanism of DADRNN helps it to follow and remember long sequence of data. The experiment was carried out on LSTM baseline model and three variants of DRNN which are DRNN, Attentive DRNN, Deep Attentive Residual DRNN. The results show DRNN outperformed baseline LSTM model.

This paper focuses on the challenge researchers face while understanding and handling multimodal data for creating an effective multimodal deep learning model. The challenges addressed are representation, translation, alignment of different modalities, fusion, co-learning of data. The author Baltrusaitis et al. (2017) suggests different ways of handling these challenges using machine learning and deep learning algorithms. He explained two types of multimodal representation which are joint and coordinated. The important challenge which is faced during emotion recognition is fusion of multimodal data. This problem can be addressed using two approaches which are model agnostic and model-based approach. These are machine learning and deep learning based approaches which fuses different modalities at different stages.

This research is based on enhancing emotion recognition accuracy of the model by considering the human expression as well as the gesture and place in the visual. The author Lee et al. (2019) introduced a novel framework called Context-Aware Emotion Recognition Networks (CAER-Net) which exploits facial expression, scene contexts in a joint and boosting manner. The CAER-Net was trained on video data taken from TV shows and annotated manually. 16 random frames were extracted from every training video with 10 frames per second sampling speed. This model was implemented using pytorch library with learning rate initialized as 5 10-3 and dropped after every 4 epochs and cross-entropy taken as loss function. The results given by this framework were better than all the latest high performing frameworks however this idea cannot be valid for all the cases. Background scenes can sometimes be misleading so in that case tone of the person's audio can be used to validate the instance.

The author Xu et al. (2019) initially uses Google speech recognition API to extract text data from audio and uses fusion and attention layer to align them together for effective and efficient emotion recognition. The model architecture comprises of a speech encoder, a text encoder, a multimodal fusion network including attention and LSTM layers for multiclass classification. The speech encoder extracts feature by converting speech signals into frames and the attention mechanism was used to decode the necessary encoded text and speech signals and learns alignment weights between text and speech data. The model showed good performance but other state-of-the-art models were not implemented on the same data to check their performance. Hence, conclusion cannot be reached about this model depending upon the comparative study provided by the author.

The research proposes hybrid fusion method which deals with the problem of redundant information in audio and visual modalities and this is done by visualizing and analyzing correlations between visual and audio modalities using latent space fusion algorithms like cross-modal factor analysis (CFA), canonical correlation analysis (CCA), and marginal Fisher analysis (MFA). Feature extraction was performed on all three modalities. Features like shot length, color, motion and lighting key were extracted from visual data and Mel-Frequency Cepstrum Coefficient (MFCC), Zero Crossing Rate (ZCR), energy and pitch were extracted from audio data and n-gram features, TF-IDF, part of speech, and lexical features were used from textual data by the author Nemati et al. (2019). Naïve Baye's and SVM models were used to check the performance of hybrid fusion method and it was confirmed that the fusing audio and video modalities give better performance than simple concatenation of these modalities.

This paper addresses challenges due to unstructured data comprising of different modalities namely audio, video and text by proposing a new framework. Author Seng and Ang (2019) also introduced new feature extraction techniques for Big Data Analytics which are Divide and Conquer PCA (Div-Con PCA) and Divide and Conquer LDA (Div-Con LDA) proposed specially for unstructured data. The author suggests various machine learning and deep learning models for single modality and different combination of modalities. He investigated SVM, naïve Bayes, and maximum entropy for classifying sentiments from textual data and found that SVM performs the best whereas naïve Bayes gave worst performance. Similarly, for video modality he investigated different techniques for video classification like calculating pupillary response and gaze distance to identify one of the seven basic emotions using multiple kernel support vector machine.

This paper proposes an user-independent approach for emotion recognition based on multimodal physiological data collected during a Virtual Reality elicitation protocol. The author Pinto et al. (2019) has used valence-arousal dimensional model introduced by Lang for classifying emotions using Support Vector Machine (SVM) classifier. In this model emotions are considered as varying degree of valence and arousal ranging from negative to positive. The model performed well with recognition rates of 78.4 percent and 61.8 percent for three and four emotions respectively. However, we cannot tell the model's performance for high number of emotions. Moreover, using valence and arousal for classifying emotions is not an effective as these models can fail while identifying sarcastic comments.

This research detects emotions among Facebook diabetes communities using Multinomial Naïve Bayes algorithm. Around 1500 Facebook posts were crawled to get the textual data to train the model by the author Balakrishnan and Kaur (2019). The main purpose of the naïve bayes algorithm is to convert the text into vectors and map the near words to the identified word and identify the emotion based on bag of words concept. Multinomial type of naïve bayes is largely used for solving text classification problem and can be more effective for small corpuses. Owing to the fact that we had small sentences and corpus of textual data, we also implemented the same methodology of first converting the text into vectors and then fed the features to multinomial naïve bayes model.

This study comprises of emotion recognition on the twitter dataset which contains small sized tweets. Here, the researcher Sharupa et al. (2020) has used Naïve Bayes classifier with unigram models which gave a good performance for all the classes in the dataset. Owing to the fact that tweets are limited to 140 words and has lots of noisy data, we can say that words per row would be as low as our dataset and we opt to use Naïve Bayes classifier for this research study.

This research is done on the classification of human made non-verbal sounds. In this study, author Chabot et al. (2021) has implemented Mel-frequency Cepstral Coefficient (MFCC) feature along with per-channel energy normalization and auditory inspired amplitude modulation features. He has also described the comparative study of SVM and Random Forest classifier along with clustering techniques like GMM and K-means. The results show both the classifier performed equally good for GMM clustering. This shows that Random forest classifier can perform well for audio data with small duration and limited audio patterns. Considering this study results and assuming the data to be of similar pattern we can also use Random Forest classifier along with MFCC feature for the audio-based emotion classification.

3 Methodology

This section describes the methodology adopted for the implementation of the project. In this research, since we deal with three different types of data, building an effective and efficient model is a significant task. This task can be achieved using different data analytics framework which consists of well-defined processes for successful product development. The framework we opted for is Knowledge Discovery in Database (KDD) because its set of processes closely align with the purpose of our research. KDD focuses mainly focuses on process execution and product development rather than project management. Figure

1 below describes the processes of KDD methodology which are implemented for all the three modalities separately.

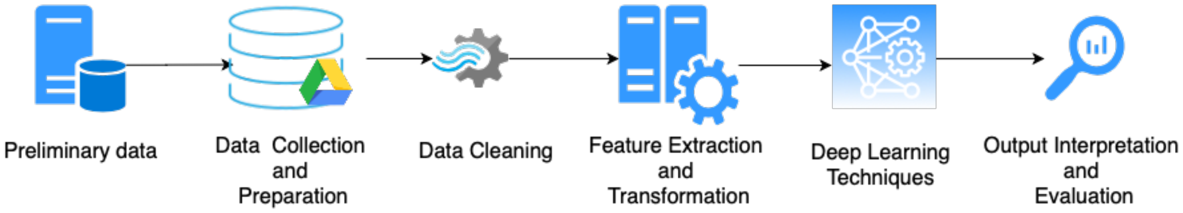

Preliminary data
Data Collection and Preparation
Data Cleaning
Feature Extraction and Transformation
Deep Learning Techniques
Output Interpretation and Evaluation

Figure 1: KDD Process Flow

3.1 Preprocessing and Feature Extraction of Text data

Initially, the traditional text mining methods are used to make the text data compatible and effective for modelling. The sentences in each row are divided into words using tokenization and cleaned by removing punctuations, short forms, numbers, and latter lemmatized to convert all the words into its basic form. Lemmatization is chosen over stemming as it uses word database and changes the inflected words more accurately ensuring the root word belongs to the language. Then stopwords like prepositions and determiners were removed from the text as they don't add enough meanings. We used Bag Of Words for analyzing and getting an idea of all the words present. POS tagging was also done to analyze the data but it was not used for further modelling as vectorizer approach was preferred. Vectorization approach is opted over lexicon dictionary approach because of less number of words per row and after analyzing all the words for each emotion, it was observed that words present in the data are not strong enough for using dictionary approach. Vectorization is a process of converting text into vectors as machine learning models can read only numeric input. Due to less number of words per row and important words, Count vectorizer is preferred over TF-IDF. We also used Synthetic Minority Oversampling Technique (SMOTE) for balancing all the classes in the dataset. As suggested by the author Balakrishnan and Kaur (2019) we have preferred to use Multinomial Naive bayes classifier for emotion classification as it performs efficienty for text classification on small text corpus and we also tried deep learning models like Bi-directional LSTM along with GloVe embedding and TFIDF but the results were biased for some emotion categories.

3.2 Preprocessing and Feature Extraction of Video data

Video data was not available beforehand and hence, it is gathered through the links mentioned in the csv file. 'Pafy' a video fetching library built using youtube-dl module is used to fetch YouTube videos with the best video quality. All the videos fetched are in mp4 format which facilitated higher and better quality of frames. Further the video files are clipped based on the start and end timestamps and sorted into emotion categories based on the emotions present in the csv file. As the deep learning model cannot read the complete videos, we converted the clipped videos into frames and to balance all the emotion categories, different number of frames per video were extracted. Further, for the better performance of the model, we excluded the background and cropped faces from the extracted frames and resized them using open cv library. Before reading the images and feeding them into model, image data augmentation is performed where the images are rescaled with appropriate zoom range and rotation range. We tried various pre-trained and complex neural network models but preferred to use 2D Convolutional Neural network with minimum complexity because the results achieved by this model were far better on our data.

3.3 Preprocessing and Feature Extraction of Audio data

The audio files are extracted from the clipped video data and sorted into different emotion directories. The next step after data sorting is feature extraction and transforming the audio files into numeric and vector form. The features used for modelling are Mel-Frequency Cepstral Coefficient (MFCC). Pitch is the important feature of voice signals and it is observed that the human's pitch of voice tends to vary with their emotions e.g., fearful speech has higher pitch compared to neutral speech. Pitch is evaluated as the frequency of speech signal and Mel scale is an effective way that evaluates perceived frequency of a signal to the actual frequency of signal. MFCC uses Mel scale to process audio signals and we have extracted the first 13 coefficients of MFCC as they represent envelop of spectra and the higher coefficients represents the spectral details but for emotion recognition envelopes are enough to recognize distinct pitch of signals. The audio features extracted were imbalanced for some categories which was further over-sampled using Synthetic Minority Oversampling Technique (SMOTE). We preferred to use Random Forest classifier over neural network models because of small number of observations present for each emotion category as mentioned by Chabot et al. (2021).

4 Design Specification

In this section, we discuss the framework we built and the approach we preferred to get effective results for emotion recognition for all the three modalities. Also, a detailed description of models built on individual modality and their working is mentioned here.

4.1 Ensemble Method Approach

Ensemble method is the technique which first makes multiple models and then combines these multiple models to get improved output. We have built three different classifiers for three different modalities which would take only a single modality as an input and classify the emotion category. The models classify emotions on the basis of probability. Once the input is passed to the classifier, it calculates probability for all the categories and the class with the highest probability wins and displayed in the output. Below Figure 2 shows the structure of ensemble approach and also an overview of framework.

We have used majority voting ensemble method where each model will predict a emotion category for test instances and the final output emotion category is the one which has more than half of votes. If any two models predict the same emotion it will be the final output but if all the three models predict three different emotions, then that instance will be discarded as it doesn't have a stable prediction.

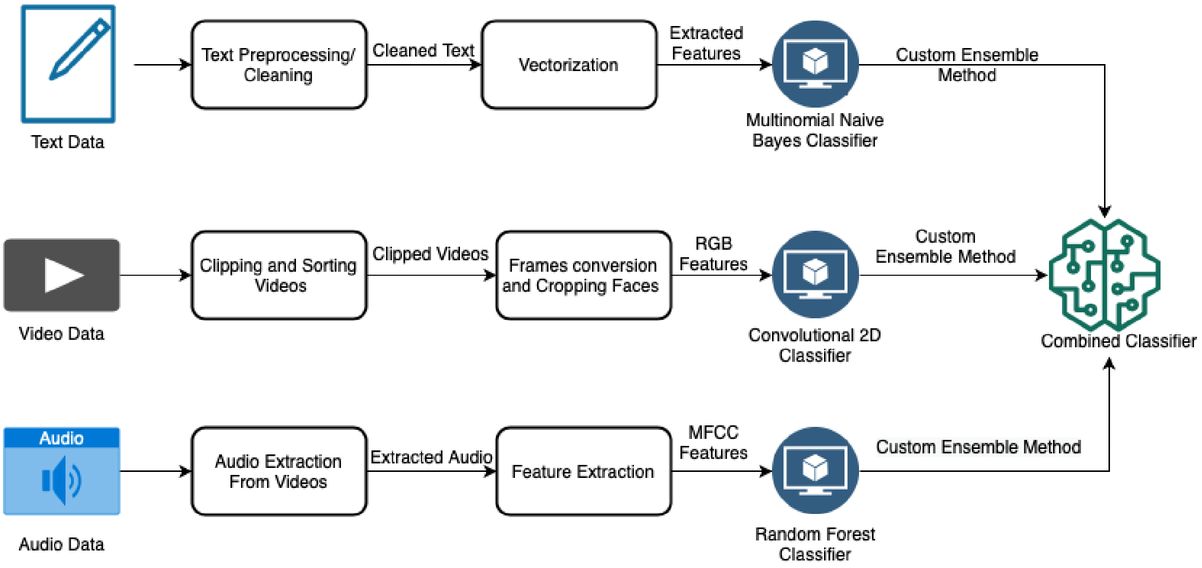


Figure 2: Project Framework

4.2 Multinomial Naive Baye's classifier for Text Data

Multinomial Naive Baye's (MNB) is a simple yet fast, reliable and efficient model for NLP tasks and relies on word representation technique called Bag-Of-Words representation. The multinomial terms describes that the features follow a multinomial distribution and therefore it counts the relative frequency of each feature which helps it to classify emotions based on the relative frequency with which a word present in the corpus. Moreover, being a probabilistic classifier, it calculates probability using Bayes probability theorem and the class with the highest probability is given in the output.This feature of MNB classifier makes it one of the good choice for implementing it in an ensemble method approach.

4.3 Random Forest classifier for Audio Data

The Random Forest classifier combines numerous decision trees, trains each one on a random subset of observations and split nodes in each decision tree considering best features from the random set of features. The final output of the classifier is based on soft voting i.e. the probability vector of each predicted class (for all the classifiers) is averaged and class with the highest value is the winning class. The parameter used in our model is n_estimator which is used to assign number of trees in the model and we have used 1000 trees in our model for better because higher the number of trees the better model learns the data.

4.4 Convolutional Neural Network (CNN) for Image Data

Convolutional Neural network is a special architecture of neural network which uses features of the visual cortex for image classification. The algorithm sees an image as array of pixels. If the size of an image is 48x48, then the size of the array for a grayscale image would be 48x48x2 where 2 is grayscale channel. A value ranging from 0 to 255 is assigned to every point which defines the intensity of that pixel at each point. The convolutional layer is the first layer of our CNN model and the image (matrix with pixel values) is fed to this layer with the correct size and dimension. Now the algorithm starts reading the input matrix and converts it into a smaller matrix which is called as filter or neuron. Its task is to multiply its values with the original pixel matrix, and then sum up all these multiplications untill one number is obtained. It further performs similarly operation moving along the image and after passing through all the points a matrix is obtained which is smaller than the input matrix. This operation is analogous to finding the curvature and boundaries of the image but in order to recognize higher order features it is important to use a full network of such layers. After convolutional network we used pooling layer and the task of this layer is to down-sample the image, resulting into reduction in the image volume. This means if some of the features in the image are already identified in the previous convolutional process, then the detailed image is no longer needed for further processing. Below Figure 3 shows the architecture of our 2D CNN model implemented on image data.

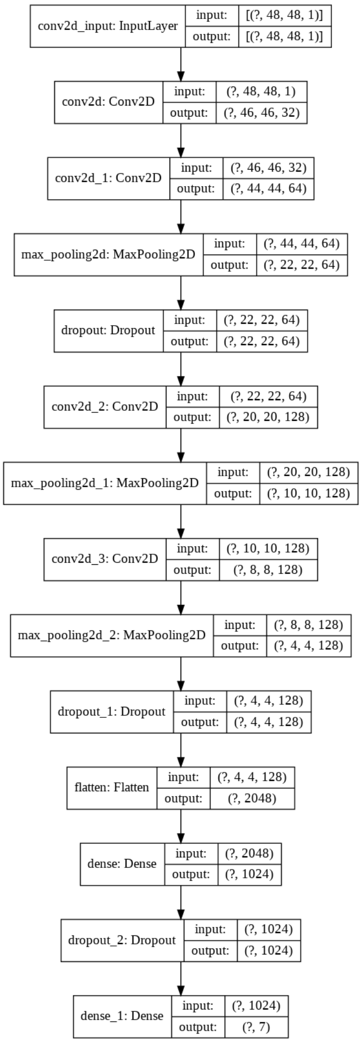


Figure 3: CNN Architecture

5 Implementation

In this section we discuss the environmental setup, various software, tools, modules and libraries implemented in the project. Also, a detailed description of dataset and working of the project is mentioned here.

5.1 Environmental Setup and System Configuration

The complete project is built using Python programming language (version 3.6.9). We opted python over other programming language because it is easy to use, has wide range of APIs for videos, audios and NLP which can be easily imported and has a good support from online communities and forums. Initially, all the experiments were tried on IDEs (Jupyter notebook, Jupyter Lab, Pycharm, Spider) installed in the local system with configuration i5 6th generation, 64-bit Mac OS Catalina (version 10.15.2), 8GB RAM and 250GB HD. Due to the poor and time-consuming performance, we opted to use paid version of Google Colaboratory (Colab Pro) which gives access to higher RAM, GPU and TPU. The GPU offered by Google Colab Pro are Nvidia K80, Tesla T4 and P100 and includes RAM with maximum limit up to 24 GB. We preferred to use Google Drive One which is subscribed version of Google drive due to high volume of video, images and audio data.

5.2 Dataset Description

The preliminary dataset comprised of a csv files taken from the One-Minute Gradual Emotion Challenge which is split into training, testing and validation data. It's a public dataset used by researchers to study on multimodal data. It comprises of links of the YouTube video along with the timestamps to split them as per the seven emotions (Anger, Sad, Neutral, Happy, Disgust, Fear, surprise) mentioned in it. All the attributes and their descriptions are summarized in the below Table 1.

Table 1: Dataset Description

Attribute	**Description**
Link	Links of all the YouTube videos
Start	Start timestamp of the video
End	End timestamp of the video
video	Video id of the YouTube video
utterance	Division of individual videos as per Emotions
Transcript	Transcribed textual data of the audio signal present in the video
EmotionMaxVote	Emotions assigned to all the rows

5.3 Model Implementation

First of all three csv files were obtained from the official site conducting the research for emotion recognition. Using the data from this file videos were fetched from the YouTube site and further all the three modalities (video, audio, text) were cleaned and pre-processed individually. Below sub-sections explain the implementation of all three modalities.

5.3.1 Implementation of model for text data

Initially, the transcription of videos present in the csv file is cleaned and pre-processed. First, the sentences are divided into words to clean it by removing punctuations, numbers, abbreviations which are of no use for the model. The output of this technique is then fed for lemmatization to convert the words into its basic form. Finally a customised list of stopwords is passed to clean prepositions and determiners and converted to vectors to make it compatible for the multinomial naive bayes model. Using these vectors, model finds the probability of each emotion categories and presents the output.

5.3.2 Implementation of model for video data

Video data was fetched from YouTube using the links mentioned in the csv file. Once all the videos were extracted, we used video editor library from python to clip the videos as per the timestamps mentioned in the csv file and later sorted as per their corresponding emotions. Once the fetched videos are converted to frames, we crop the faces from the frames and further augment the images by rescaling so that the model could extract important features and classify emotions from the facial expressions.

5.3.3 Implementation of model for audio data

The audio is extracted from the sorted video file and saved in the .mp3 format in their corresponding emotion directory. Once the audio files are loaded, mel-frequency cepstral coefficient (mfcc) feature is extracted and passed to the model for emotion classification.

The output of all the three models are combined and a single emotion category is decided using majority voting ensemble method and presented in the output.

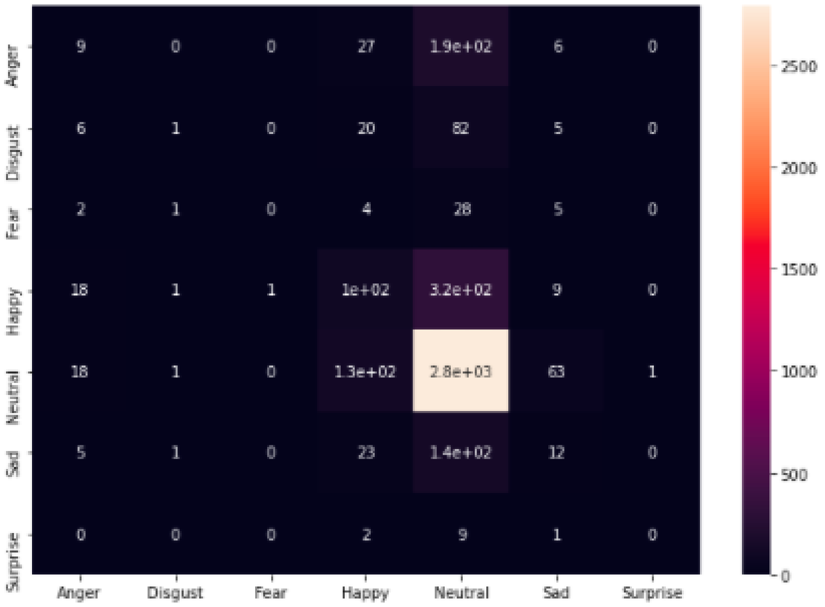
6 Evaluation

The purpose of this section is to provide a comprehensive analysis of the results and main findings of the study. The section is divided according to the different experiment performed on all the three modalities. Also, the graphs and charts of the results obtained are mentioned to support our findings.

6.1 Experiment on Text data

The dataset is obtained from a one-minute gradual (OMG) emotion challenge and it was already split into training, testing and evaluation set. Due to imbalanced training set we have to merge it with the testing set and once the cleaning was done, it was shuffled and split using stratified split method therefore all the categories were evenly split into training and test set. After running the model on testing data, we got an accuracy of 72.37 percent on the testing set and further the model performance is evaluated on the basis of confusion matrix to analyze the model performance for all the seven emotion categories. Also, precision, recall, and f1-score are analyzed for the detailed evaluation of model. Below figure 4 show the confusion matrix and we can observe that model performed better for 'Happy' and 'Neutral' categories whereas for the other categories it could classify only a few correct emotion categories.

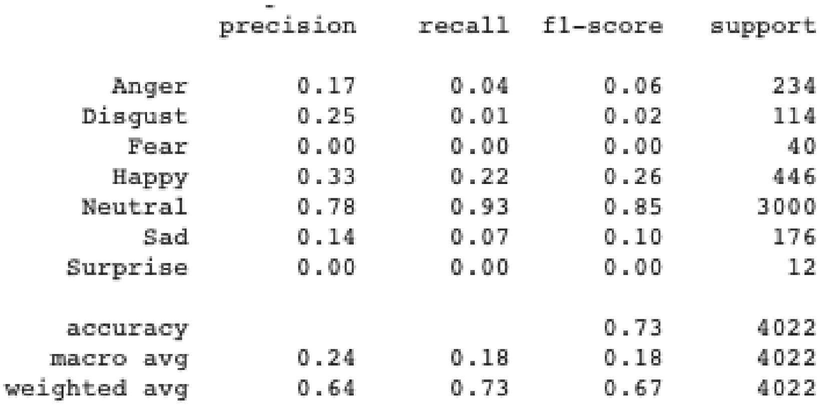
Below figure 5 show precision, recall, and f1-score for all the emotion categories on the test data. We can see that the classifier performed better for the classes higher number



	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	9	0	0	27	1.9e+02	6	0
Disgust	6	1	0	20	82	5	0
Fear	2	1	0	4	28	5	0
Happy	18	1	1	1e+02	3.2e+02	9	0
Neutral	18	1	0	1.3e+02	2.8e+03	63	1
Sad	5	1	0	23	1.4e+02	12	0
Surprise	0	0	0	2	9	1	0

Figure 4: Confusion Matrix for Text Modality

of records. The support column shows that the Happy and Neutral has significantly high number of records as compared to Fear and Surprise category. In the precision column we can see that Neutral and Happy are the top two with high number of correct classifications which is justified because of high number of observations for these class. Similarly, recall and f1-score shows that the same classes have highest fraction of positives that are correctly identified and weighted harmonic mean of recall and precision.



```
              precision    recall  f1-score   support

       Anger       0.17      0.04      0.06       234
     Disgust       0.25      0.01      0.02       114
        Fear       0.00      0.00      0.00        40
       Happy       0.33      0.22      0.26       446
     Neutral       0.78      0.93      0.85      3000
         Sad       0.14      0.07      0.10       176
    Surprise       0.00      0.00      0.00        12

    accuracy                           0.73      4022
   macro avg       0.24      0.18      0.18      4022
weighted avg       0.64      0.73      0.67      4022
```

Figure 5: Classification report for Text Modality

6.2 Experiment on video data

The video data once converted into frames are fed to the 2D CNN model with input size of 48x48, activation function relu for all the convolutional layers and softmax for the output layer. Following are the hyperparameters used: loss = categorical crossentropy, optimizer = Adam with learning rate 0.0001 and decay rate 1e-6, number of epochs = 50 and batch size = 64. The model is implemented on the test set and gave an accuracy of 68 percent and loss of 1.68. The performance of the classifier is evaluated on the basis of confusion matrix and classification report. Below confusion matrix heatmap shows that the model classified high number of correct positive instance correctly only for Disgust category and confused other classes with this category. Lets see if the classification report can give the reason behind this unevenness.

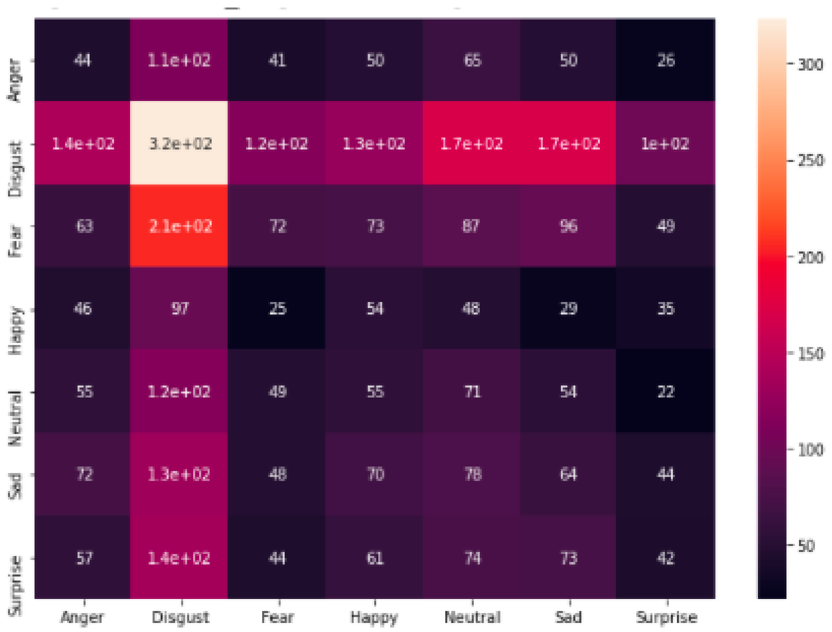
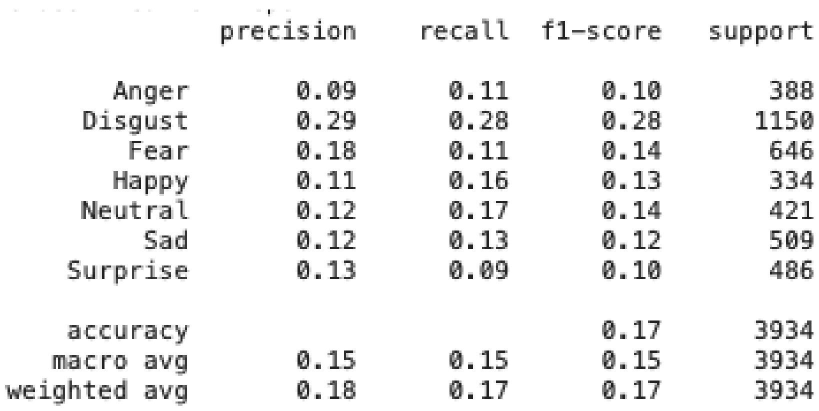


Figure 6: Confusion Matrix for Video Modality

The precision, recall, and f1-score show the best figures are for Disgust category and if we observe support column we can notice that the Disgust category has significantly high number of records but unlike text modality classifier for video modality gave results for all the emotion categories.



```
              precision    recall  f1-score   support

       Anger       0.09      0.11      0.10       388
     Disgust       0.29      0.28      0.28      1150
        Fear       0.18      0.11      0.14       646
       Happy       0.11      0.16      0.13       334
     Neutral       0.12      0.17      0.14       421
         Sad       0.12      0.13      0.12       509
    Surprise       0.13      0.09      0.10       486

    accuracy                           0.17      3934
   macro avg       0.15      0.15      0.15      3934
weighted avg       0.18      0.17      0.17      3934
```

Figure 7: Classification Report for Video Modality

6.3 Experiment on Audio modality

The audio data is extracted from videos and fed to random forest classifier in the form of mel-frequency cepstral coefficient with the parameter n_estimator set to 1000. The random forest gave an accuracy of 61 percent on the audio modality. We evaluated the results using confusion matrix and classification report and below are the figures representing both. The confusion matrix in the figure 8 shows that Neutral category showed highest number of correct predictions. The below classification report in the figure 9 shows that classes like Fear and Surprise did not classify any instance due to very less number of observations and the only class performing comparatively better is Neutral as it has high number of instances.

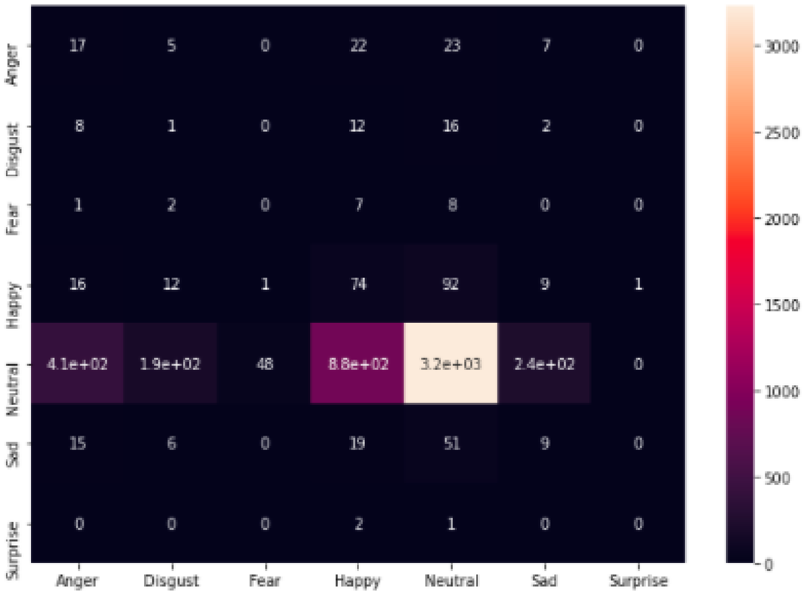
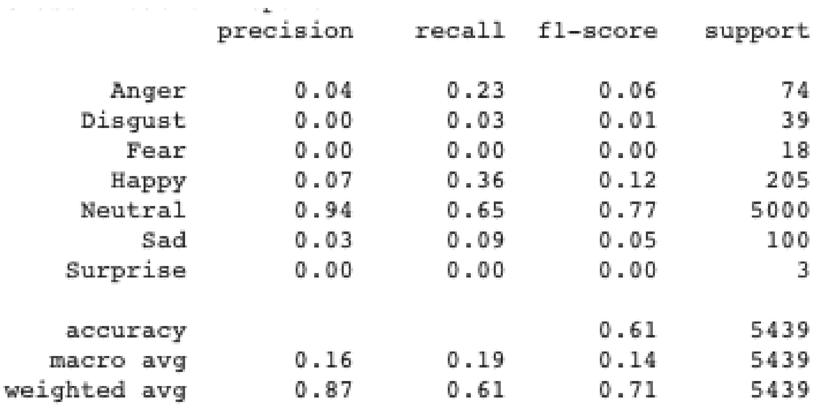


Figure 8: Confusion Matrix for Audio Modality

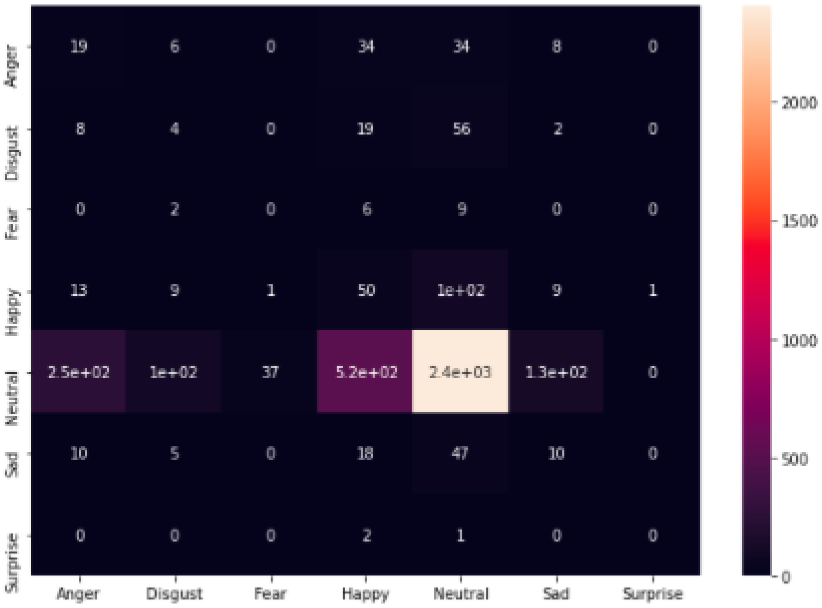


```
              precision    recall  f1-score   support

       Anger       0.04      0.23      0.06        74
     Disgust       0.00      0.03      0.01        39
        Fear       0.00      0.00      0.00        18
       Happy       0.07      0.36      0.12       205
     Neutral       0.94      0.65      0.77      5000
         Sad       0.03      0.09      0.05       100
    Surprise       0.00      0.00      0.00         3

    accuracy                           0.61      5439
   macro avg       0.16      0.19      0.14      5439
weighted avg       0.87      0.61      0.71      5439
```

Figure 9: Classification report for Audio Modality

6.4 Experiment on combining all the three modalities

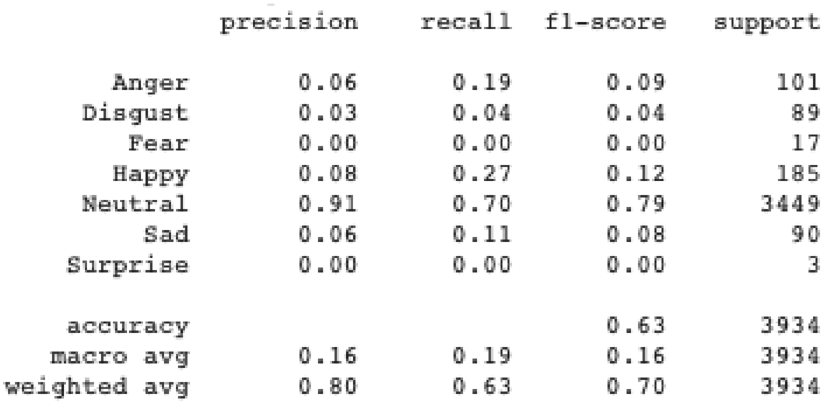
This experiment combines the output results from all the three classifiers and give final output which has majority votes which is done using ensemble method approach. Below figure 10 and figure 11 describes the results given by ensemble method. The confusion matrix in the figure 10 shows almost similar results to the individual models but ensemble method approach is also classifying other emotions apart from Surprise and Fear.



	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	19	6	0	34	34	8	0
Disgust	8	4	0	19	56	2	0
Fear	0	2	0	6	9	0	0
Happy	13	9	1	50	1e+02	9	1
Neutral	2.5e+02	1e+02	37	5.2e+02	2.4e+03	1.3e+02	0
Sad	10	5	0	18	47	10	0
Surprise	0	0	0	2	1	0	0

Figure 10: Confusion Matrix for Ensemble method

The below figure 11 shows classification report for ensemble method. It shows that recall is comparatively higher for Happy and Neutral which means these classes have higher fraction of correctly identified positives. Precision is also higher for Neutral category but this can all be justified from the support column which shows that number of observations are significantly high for Neutral category.



```
              precision    recall  f1-score   support

       Anger       0.06      0.19      0.09       101
     Disgust       0.03      0.04      0.04        89
        Fear       0.00      0.00      0.00        17
       Happy       0.08      0.27      0.12       185
     Neutral       0.91      0.70      0.79      3449
         Sad       0.06      0.11      0.08        90
    Surprise       0.00      0.00      0.00         3

    accuracy                           0.63      3934
   macro avg       0.16      0.19      0.16      3934
weighted avg       0.80      0.63      0.70      3934
```

Figure 11: Classification report for Ensemble method

6.5 Discussion

A detailed discussion of the findings from the above experiments are discussed in this section. The results show that the model is biased for some classes and this is due to the high imbalance in the dataset. The csv file received had comparatively lower imbalance and therefore we can see a better results for this modality. However, after learning the detailed description of the dataset it is found that the data was completely annotated manually based on the video and audio modality. Therefore, the text data had very minimum strong words which could depict the emotions in the sentence. Moreover, due to division of individual videos based on the timestamps, there are not many words per row which makes it difficult for the model to learn the differences between the classes. Similarly, the video links mentioned in the csv file is not updated which restricts fetching of all the videos due to unavailable status of the videos. As a result, created more imbalance due to high number of videos being unavailable from the minority class which significantly increased the difference between majority and minority class. As the audios were fetched from these videos the imbalanced continued to the audio modality. SMOTE technique was implemented to make the data balance but it only duplicates the values and due to very less number of original observation present in the minority class the classifier gave no better results.

Due to unavailable status of high number of videos and less number observations, multimodal deep learning approach couldn't be implemented which gives better results as mentioned by Shen et al. (2020). We have taken an appropriate approach in our project owing to the constraints we had due to less amount of data.

The results can be improved by adding more instances of minority classes to make it balance which could help to implement deep learning for all the modalities. Moreover, the textual data should be containing longer sentences which can help to implement lexicon emotion dictionary or GloVe embedding approach which could highly improve the results for text modality as given by (Chabot et al.; 2021).

7 Conclusion and Future Work

In this research, we implemented an ensemble method to combine the three models and get a final output.The classifier for text, classifier for video and classifier for audio gave accuracy of 72.37, 68, and 61 percent respectively. But the ensemble method gave us the accuracy of 63 percent. Comparing the performance of individual classifier with the ensemble method we can state that the results in the section experiment shows both performed nearly similar and as per the study of related work we can say that the results can be improved with the increase in the data volume. As a part of future work, this emotion categorization of videos can be used as a parameter for recommendation system which can be proposed in the future work. This recommendation system can be useful for not only online video platforms but also for online music, online movies, and social media platforms. Also multimodal deep learning model approach can be implemented for improving the accuracy and performance of the classifier.

References

Bairaju, S. P. R., Ari, S. and Murthy Garimella, R. (2019). Emotion detection using visual information with deep auto-encoders, *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp. 1–5.

Balakrishnan, V. and Kaur, W. (2019). String-based multinomial naïve bayes for emotion detection among facebook diabetes community, *Procedia Comput. Sci.* **159**(C): 30–37.
URL: *https://doi.org/10.1016/j.procs.2019.09.157*

Baltrusaitis, T., Ahuja, C. and Morency, L. (2017). Multimodal machine learning: A survey and taxonomy, *CoRR* **abs/1705.09406**.
URL: *http://arxiv.org/abs/1705.09406*

Chabot, P., Bouserhal, R. E., Cardinal, P. and Voix, J. (2021). Detection and classification of human-produced nonverbal audio events, *Applied Acoustics* **171**: 107643.

Chandra, E. and Hsu, J. Y. (2019). Deep learning for multimodal emotion recognition-attentive residual disconnected rnn, *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 1–8.

Herzig, J., Shmueli-Scheuer, M. and Konopnicki, D. (2017). Emotion detection from text via ensemble classification using word embeddings, p. 269–272.
URL: *https://doi.org/10.1145/3121050.3121093*

Kaya, H., Grpnar, F. and Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion, *Image Vision Comput.* **65**(C): 66–75.
URL: *https://doi.org/10.1016/j.imavis.2017.01.012*

Lee, J., Kim, S., Kim, S., Park, J. and Sohn, K. (2019). Context-aware emotion recognition networks, pp. 10142–10151.

Nemati, S., Rohani, R., Basiri, M. E., Abdar, M., Yen, N. Y. and Makarenkov, V. (2019). A hybrid latent space data fusion method for multimodal emotion recognition, *IEEE Access* **7**: 172948–172964.

Pinto, J., Fred, A. and Silva, H. (2019). Biosignal-based multimodal emotion recognition in a valence-arousal affective framework applied to immersive video visualization, *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* pp. 3577–3583.

Seng, J. K. P. and Ang, K. L. (2019). Multimodal emotion and sentiment modeling from unstructured big data: Challenges, architecture, techniques, *IEEE Access* **7**: 90982–90998.

Sharupa, N. A., Rahman, M., Alvi, N., Raihan, M., Islam, A. and Raihan, T. (2020). Emotion detection of twitter post using multinomial naive bayes, *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6.

Shen, G., Wang, X., Duan, X., Li, H. and Zhu, W. (2020). Memor: A dataset for multimodal emotion reasoning in videos, *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, Association for Computing Machinery, New York, NY, USA, p. 493–502.
URL: *https://doi.org/10.1145/3394171.3413909*

Soleymani, M., Pantic, M. and Pun, T. (2012). Multimodal emotion recognition in response to videos, *IEEE Transactions on Affective Computing* **3**(2): 211–223.

Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y. and Li, X. (2019). Learning alignment for multimodal emotion recognition from speech, *CoRR* **abs/1909.05645**.
URL: *http://arxiv.org/abs/1909.05645*