*Dissertation on*

# "Real-Time Inappropriate Speech Detection"

*Submitted in partial fulfilment of the requirements for the award of the degree of*

**Bachelor of Technology**

**In**

**CSE(AI&ML)**

**UE22AM320A – Capstone Project Phase - II**

***Submitted by:***


| | |
|---|---|
| **SAI ARYA R B** | **PES1UG22AM142** |
| **SAMARTH S KULKARNI** | **PES1UG22AM144** |
| **SIDDHARTH GANDHI** | **PES1UG22AM157** |
| **SUDARSHAN SRINIVASAN** | **PES1UG22AM166** |


*Under the guidance of*


**Dr. Jayashree R**
**Professor & Chairperson**
Department of CSE (AI &ML)
PES University


**January - May 2025**


**DEPARTMENT OF CSE(AI&ML)**
FACULTY OF ENGINEERING
**PES UNIVERSITY**
(Established under Karnataka Act No. 16 of 2013)
100 feet Ring Road, BSK 3rd stage, Hosakerehalli, Bengaluru – 560085

# PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

**FACULTY OF ENGINEERING**

# CERTIFICATE

*This is to certify that the dissertation entitled*

## ‘Real-Time Inappropriate Speech Detection’

*is a Bonafide work carried out by*

| | |
|---|---|
| **SAI ARYA R B** | **PES1UG22AM142** |
| **SAMARTH S KULKARNI** | **PES1UG22AM144** |
| **SIDDHARTH GANDHI** | **PES1UG22AM157** |
| **SUDARSHAN SRINIVASAN** | **PES1UG22AM166** |

In partial fulfilment for the completion of Sixth-semester Capstone Project Phase - II (UE22AM320A) in the Program of Study -Bachelor of Technology in CSE(AI&ML) under rules and regulations of PES University, Bengaluru during the period January 2025 – May 2025. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6$^{th}$-semester academic requirements in respect of project work.

| | | |
|---|---|---|
| Signature<br>Dr. Jayashree R<br>Chairperson | Signature<br>Dr. Jayashree R<br>Guide | Signature<br>Dean, Engineering and Technology |

**External Viva**

| **Name of the Examiners** | **Signature with Date** |
|---|---|
| **1.** _____ | _____ |
| **2.** _____ | _____ |

# DECLARATION

We hereby declare that the Capstone Project Phase - II entitled **"Real-Time Inappropriate Speech Detection"** has been carried out by us under the guidance of **Dr Jayashree R, Professor & Chairperson Department of CSE(AI&ML)** and submitted in partial fulfilment of the course requirements for the award of the degree of **Bachelor of Technology** in **CSE(AI&ML)** of **PES University, Bengaluru** during the academic semester January – May 2025. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

| | |
|---|---|
| **PES1UG22AM142** | **SAI ARYA R B** |
| **PES1UG22AM144** | **SAMARTH S KULKARNI** |
| **PES1UG22AM157** | **SIDDHARTH GANDHI** |
| **PES1UG22AM166** | **SUDARSHAN SRINIVASAN** |

# ACKNOWLEDGEMENT

I would like to express my gratitude to **Dr. Jayashree R, Professor & Chairperson** Department of CSE(AI&ML), PES University, for her continuous guidance, assistance, and encouragement throughout the development of UE22AM320A - Capstone Project Phase – II.

I am grateful to Capstone Project Coordinator, Prof. Shwetha K N for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Jayashree R, Professor & Chairperson, Department of CSE(AI&ML), PES University, for all the knowledge and support I have received from the department. I would like to thank Dean, Engineering and Technology, PES University for his help.

I am deeply grateful to Prof. Jawahar Doreswamy, Chancellor, PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University, for providing me with various opportunities and enlightenment every step of the way. Finally, Phase - II of the project could not have been completed without the continual support and encouragement I have received from my family and friends.

# ABSTRACT

There have been serious challenges to keep online platforms safe and respectful due to unmoderated content because of increasing incidents of explicit speech. The importance of real-time inappropriate speech detection lies in its ability to censor profanity which makes online platforms safe for all demographics.

The objective of this work is to clearly define our problem statement and its scope, display our learnings from our literature survey, highlighting research and technological gaps, define our objectives and conclude our findings while planning for future work.

Our work focuses on the problem of detecting inappropriate speech in real-time, which has not been explored yet.

# TABLE OF CONTENTS

# LIST OF FIGURES

# Chapter 1

# INTRODUCTION

## 1.1 The increasing need for Content moderation in online Platforms

Due to recent increase in popularity of Live Streaming platforms, there exists a need for establishing a safe environment for all age demographics by moderating use of profanity that may or may not be intentional. Prediction of occurrence of inappropriate speech has not yet been explored in depth as compared to Traditional moderation techniques which rely on text or audio solely and are not used for Real-Time detection.

## 1.2 The Approach of Multimodal Models in Predicting Profanity

Multimodal models can offer a nuanced approach by simultaneous analysis of audio, video and text to overcome the limitations of unimodal systems. The model can extract temporal sequences and spatial features. By incorporating the use of deep learning architectures like Convolutional Neural Network and Recurrent Neural Networks, we can increase efficiency.

## 1.3 Scope of Proposed Framework

We are proposing the use of a multimodal deep learning framework for detecting inappropriate language in real-time systems. By integrating audio features with visual cues and utilizing cross - modal attention mechanisms, the model will improve its ability to predict inappropriate content by focusing on relevant features across different modalities.

The end goal is to develop a context - aware system that can enable platforms to establish a safe environment.

# Chapter 2

# PROBLEM STATEMENT

## 2.1 The Challenges of Moderating Live Streaming Platforms

It is difficult to moderate content on live streaming platforms because everything is unscripted. Live streams require continuous monitoring because of its unpredictable nature, which makes it harder to catch inappropriate content such as curse words. Also, real-time live streams are not just about one kind of data type but include audio, video, and text all at once. This variety makes moderating even more challenging to do unless you have a sophisticated model capable of analysing all inputs simultaneously.

## 2.2 The Problem of Predicting Inappropriate Language

Foul words need not always be spoken out but also can also depend on the body language and facial expression. Along with this sarcasm and irony can also be difficult to predict only on either spoken words or body language.

## 2.3 Why We Need Multimodal Analysis

A complete picture is difficult to get with traditional single-modal models. When audio, video, and text are all analysed together as a multimodal approach, we can improve the accuracy of identifying inappropriate language and behaviour.

It is easier to figure out the full context of what is happening when we combine speech with visual data. Also, by using many modalities of data, the system can get a better understanding and become more reliable. Similarly, only multimodal systems can notice and predict even the most subtle forms of curse words.

# CHAPTER 3
# LITERATURE SURVEY

## 3.1 Research Paper – 1

### 3.1.1 Paper Title

I. Smirnov and A. Laushkina, "Multimodal prediction of profanity based on speech analysis," [1]

### 3.1.2 Summary

The key idea behind this paper was to develop a real-time profanity prediction model, using audio features, to identify swearing before the full word is spoken. This helps with the task of early censorship. They found that using audio features would be better than Automatic Speech Recognition (ASR) as it can often lag and is prone to errors in certain conditions. This provides possibilities to combine with facial expression analysis to build a multimodal model which gives better results. The audio features were extracted using methods like MFCC and Wav2Vec. For the speech recognition task, they used Whisper ASR for transcribing to speech, and also label timestamps showcasing profanity. It created 3 class labels – Silence, Normal speech, Profanity. The CommonVoice and CMU MOSEI dataset were used with deep learning models like, LSTM and Bidirectional LSTM with Attention mechanism, this model improves prediction by understanding the context compared to LSTM which only predicts the upcoming words in speech. The real-time processing was executed by a pipeline which the authors implemented by using Voice Activity Detection (VAD), which initiates the prediction process upon detecting active speech, and Sliding window processing to analyse chunks of continuous audio. The best model – LSTM + Wav2Vec + Attention, achieved a F1-score of 92%. The idea of incorporating facial analysis was proposed by the authors to build even more robust models.

| Approach | F1 score (The main dataset) | F1 score (MELD dataset) | F1 score (Picked records) | Latency (sec) | Weights (mb) |
|---|---|---|---|---|---|
| LSTM (MFCC) | 86.6 | 65.2 | 73.7 | 0.347* | 51.53 |
| LSTM (MFCC) + Attention | 87.1 | 65.9 | 76.7 | 0.348 | 231.16 |
| LSTM Wav2Vec + Attention | **92.0** | 71.4 | 86.6 | 1.171 | 432.25** |
| **LSTM (MFCC) + ASR + Attention** | 90.3 | 67.6 | 74.6 | **2.148** | 231.19 |

* - the average duration of the element is 0.34 sec
** - it is also necessary to use Wav2Vec feature extractor

**Fig. 3.1.1** Comparison of different LSTM architectures

## 3.2 Research Paper – 2

### 3.2.1 Paper Title

A. Chaudhari, P. Davda, M. Dand and S. Dholay, "Profanity Detection and Removal in Videos using Machine Learning," [2]

### 3.2.2 Summary

The authors tackled the issue of automatically finding and getting rid of swear words in videos. As manually doing this task consumes time, they came up with a system using machine learning to handle it, aiming for real-time detection and censorship. They start by using an approach that gets the audio from video and then converts it into text. Then, using a pre-made list of profanities they check these words. On appearance of a swear word, two things were done: they mute that specific part of the audio track, and they also visually censor the speaker's mouth by pixelating their lips. To do the lip pixelation, they first detect the face in the video, find the lip landmarks (using stuff like HOG and SVM), and then apply a blur effect specifically to the lips during the time the swear word is spoken. To figure out exactly when the profanity occurs, they first break the video into 3-second chunks and then narrow it down to 1-second intervals. This was done for better accuracy. For testing, instead of using a standard dataset, they gathered their own videos containing profanity from places like YouTube and Facebook. Anyway, they tried it out on 50 videos, which apparently had around 187 swear words scattered through them. The system actually did pretty well, managing to catch and block the profanity correctly about 82% of the time (specifically, they reported 82.35% accuracy). They noted that using both the sound muting and the lip blurring together really did the trick for getting rid of the unwanted content.

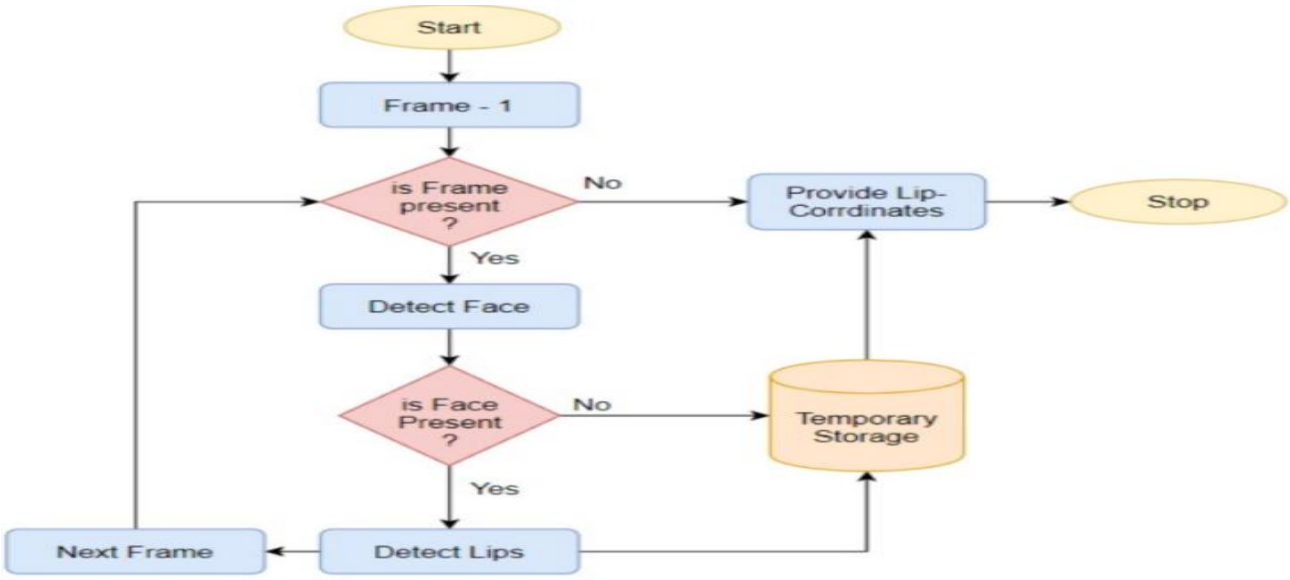


**Fig. 3.2.1** Lip-Landmarks Detection methodology

## 3.3 Research Paper – 3

### 3.3.1 Paper Title

Rana, Aneri and Sonali Jha, “Emotion Based Hate Speech Detection using Multimodal Learning,” [3]

### 3.3.2 Summary

The authors of this paper created the first multimodal deep learning framework that combines audio (emotion detection) and text features to detect hate speech in videos. This is really hard to do because it needs processing of different types of data simultaneously. Their model does 2 things - the first one is text processing using BERT & ALBERT (Transformer models) trained on hate speech datasets from Twitter and Reddit, and the second thing is emotion detection in speech using a multi-task deep learning model trained on IEMOCAP dataset to detect valence (positive/negative emotion), arousal (intensity), and dominance (speaker control). They combined these features using a Multilayer Perceptron model for final classification. They also made their own Hate Speech Detection Video Dataset (HSDVD) with videos from Twitter & YouTube. This approach proposed by the authors is useful because it can help identify moments when a speaker might swear based on tone and intensity, which could enable real-time censoring. The results of this paper showed that the multimodal learning (which was a combination of text and speech features) performed better than text-only models. The BERT + Speech Emotion features combo gave the best results. An interesting find was that their model made fewer mistakes by correctly identifying sarcasm and neutral speech that text only models wrongly identified as hate speech.

## 3.4 Research Paper – 4

### 3.4.1 Paper Title

Ba Wazir, A.S.; Karim, H.A.; Abdullah, M.H.L.; AlDahoul, N.; Mansor, S.; Fauzi, M.F.A.; See, J.; Naim, A.S., “Design and Implementation of Fast Spoken Foul Language Recognition with Different End-to-End Deep Neural Network Architectures,” [4]

### 3.4.2 Summary

The authors of this paper developed lightweight deep learning models (which were CNNs & RNNs) for detecting bad language in speech quickly, with the goal of making real-time censorship actually possible. Most existing models are way too bulky for live use, which is a big problem. They tested two types of models: first was CNNs trained on Mel-Spectrograms to get visual patterns from speech, and second was RNNs with LSTM Cells that can handle speech as it happens over time. The dataset the authors worked with was the MMUTM foul language dataset which has 9 types of offensive words plus regular speech from 117 different people in a variety of situations. To make their system more robust, they altered the data by changing pitch, shifting frequencies, and adding noise - basically making the dataset 9x bigger which helped it to deal with background noise better. The best part is probably their super compact CNN model with just 57K parameters, which means it can actually run in real-time situations as it is pretty lightweight. They tested in different noise conditions from clean to really noisy (0db) to see how well the models were. Their CNN model got a 96.11% F1-score which beat the RNN model, which got a 94.85% F1-score. Their little CNN performed better than huge pre-trained models like AlexNet and ResNet50 while using lesser parameters also (100x lesser). Under noisy situations the efficiency dropped by 10%, but the data augmentation done by the authors helped reduce the impact. This work is very relevant because the LSTM approach can catch swear words as they're being spoken, not after, which is exactly what you'd need for live video censoring.

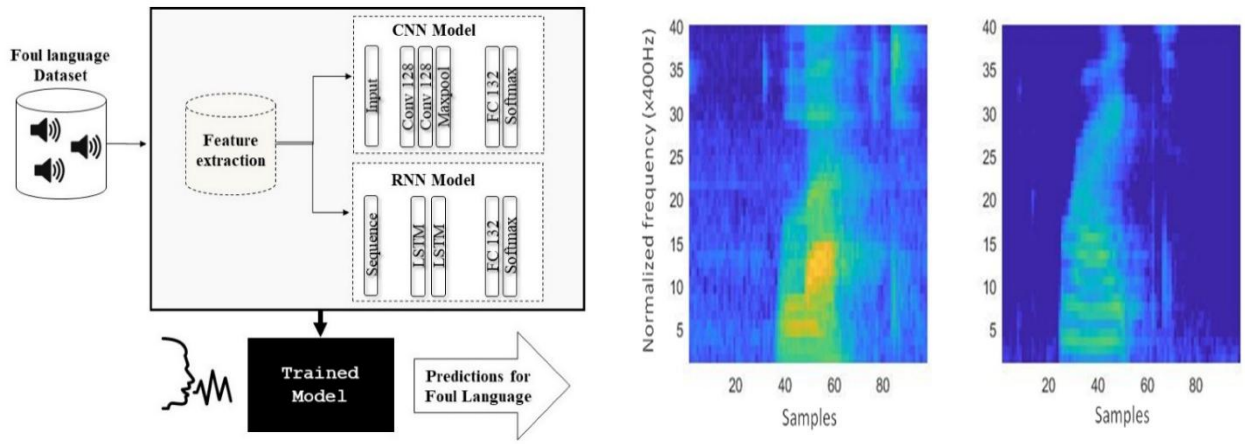


**Fig. 3.4.1** Flow and Samples used in Foul Language Detection model

## 3.5 Research Paper – 5

### 3.5.1 Paper Title

J. Pan, W. Fang, Z. Zhang, B. Chen, Z. Zhang and S. Wang, "Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG," [5]

### 3.5.2 Summary

The authors came up with this system called Deep-Emotion that basically combines face expressions, speech, and brain signals (EEG) to perceive people's emotions better. The other available systems only use one or two data types. That's not as good. For analysing faces, the authors used an improved GhostNet CNN model that catches emotion features better and overfits less. For speech emotion part, the authors made this lightweight network called LFCNN that's small enough to be able to work in real-time (which is important). They also threw in EEG analysis with this Tree-like LSTM model to check brain activity. The highlights about this paper are probably how they combined everything - they used this Decision-Level Fusion Method with optimized weights to bring all three predictions together. Their results were pretty good - face model got 98.27% accuracy (which was better than other available methods), their speech model hit 94.36%, and the EEG model worked well too. This showed that combining different signals definitely works better than just using one. This research is relevant because the face and speech emotion analysis could help identify when someone's getting frustrated or angry right before they swear. Their fusion approach could definitely be adapted to mix speech and facial cues to predict swear words in videos, and since their speech model is small, it could actually work for real-time processing.

## 3.6 Research Paper – 6

### 3.6.1 Paper Title

H. M. Fayek, M. Lech and L. Cavedon, "Towards real-time Speech Emotion Recognition using deep neural networks," [6]

### 3.6.2 Summary

The authors developed a model for real time speech emotion recognition using a deep learning model to skip the steps of data preprocessing such a complex feature engineering due to the fact that deep learning models do their own feature analysis. They suggested using one second frames of raw speech spectrograms and also implemented techniques such regularization and data augmentation to combat overfitting as well as improve generalizability. The model achieved good accuracy on both the eNTERFACE and SAVEE datasets.

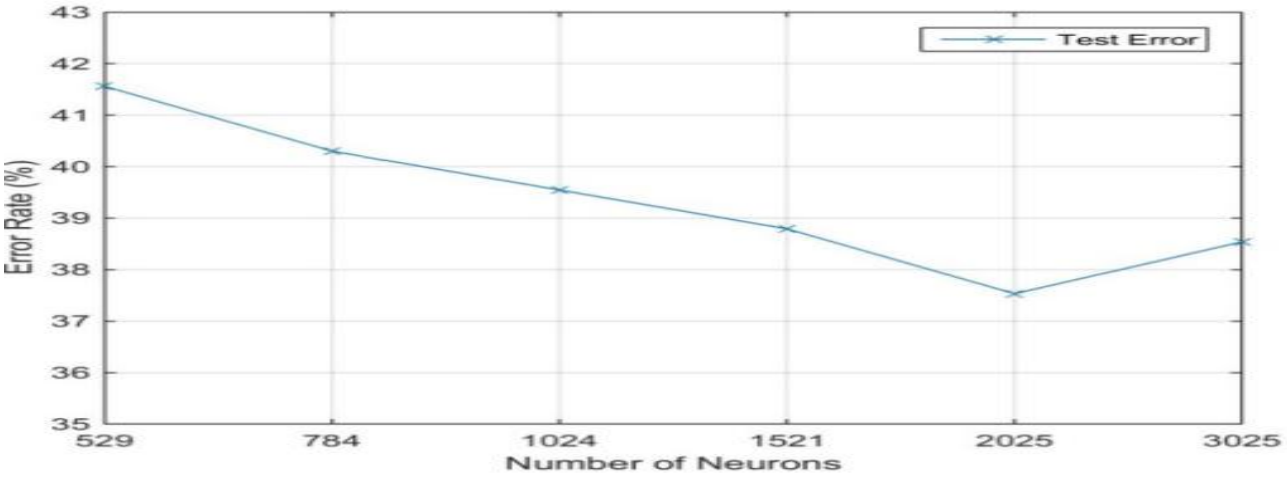


**Fig. 3.6.1** Test Error Rates with varying neurons and fixed layers

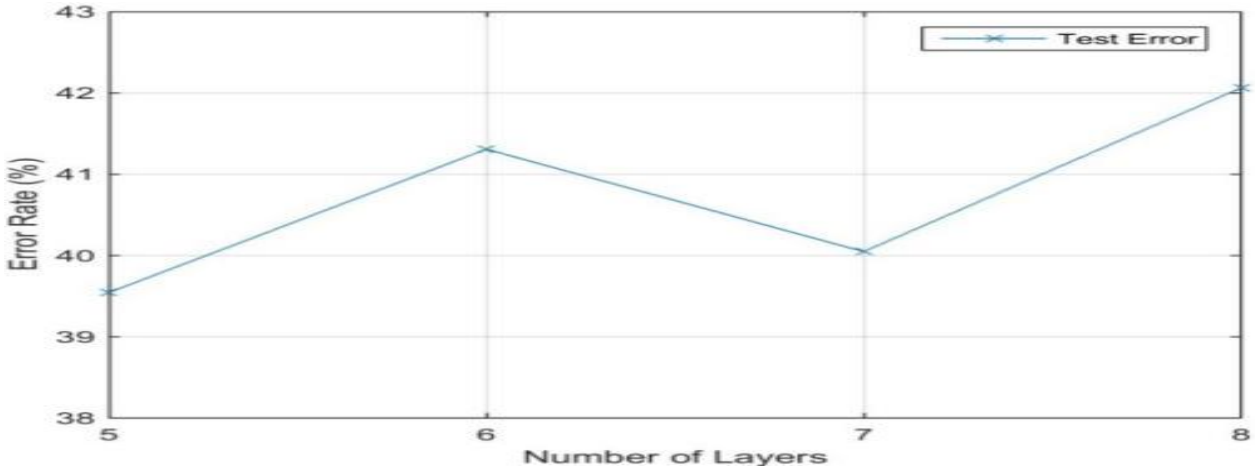


**Fig. 3.6.2** Test Error Rates with varying layers and fixed neurons

## 3.7 Research Paper – 7

### 3.7.1 Paper Title

H. Meng, T. Yan, F. Yuan and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," [7]

### 3.7.2 Summary

The authors use a deep learning architecture for speech emotion recognition that consists of CNN for spatial feature extraction and a BiLSTM model with attention for temporal feature extraction. Coming to the loss function they used a hybrid loss function of softmax and cross entropy. The model showed an improvement of 4.58% over other models.

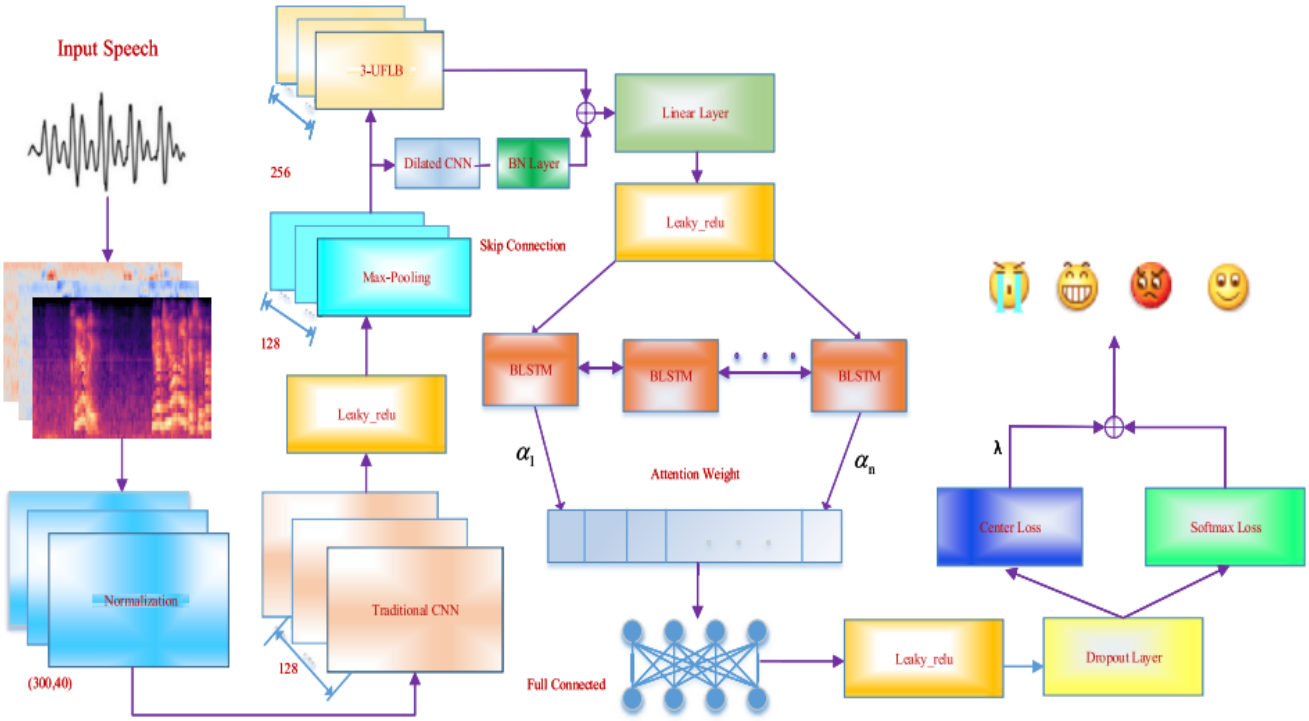


**Fig. 3.7.1** Structure of dilated CNN and BiLSTM

## 3.8 Research Paper – 8

### 3.8.1 Paper Title

D. Duncan, G. Shine, and C. English, "Facial Emotion Recognition in Real Time," [8]

### 3.8.2 Summary

The authors used CNNs to perform real time facial emotion recognition in video feeds. They implemented a pipeline to capture live video, detect facial landmarks, feed it to the CNN and recognise emotions such as happiness, anger, sadness, etc. The authors designed this pipeline to improve as much speed and performance as possible. The final findings were that CNNs are quite capable at real time emotion recognition.

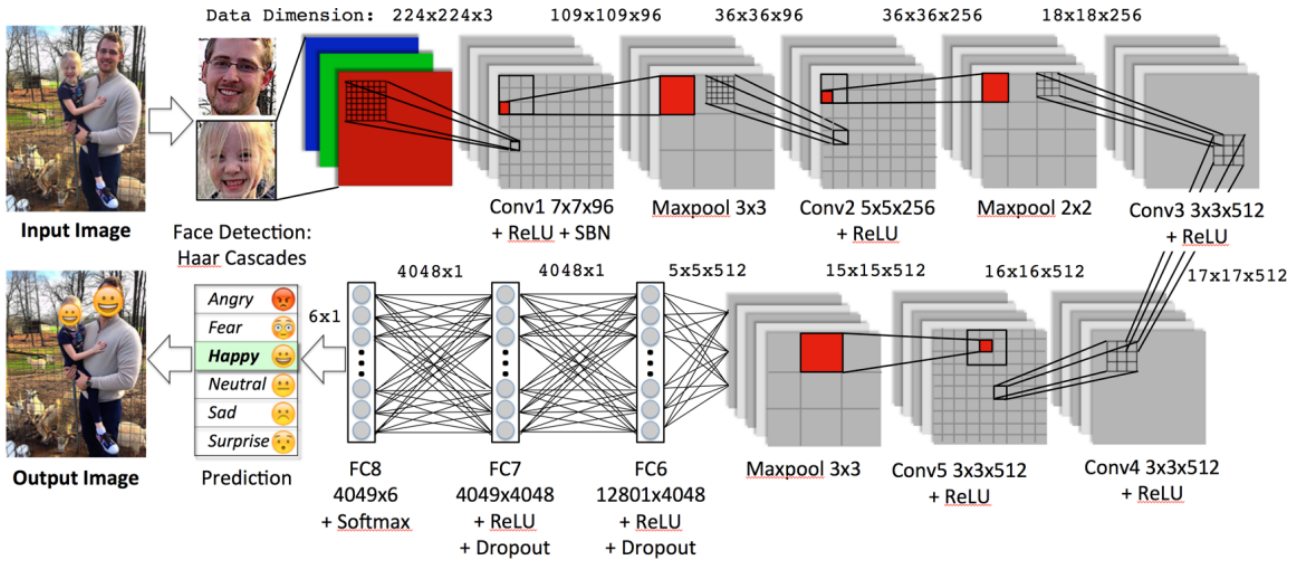


**Fig. 3.8.1** CNN architecture with a Haar-Cascade detector


**Fig. 3.8.2** Emotions detected by the CNN

## 3.9 Research Paper – 9

### 3.9.1 Paper Title

S. Dwijayanti, M. Iqbal and B. Y. Suprapto, "Real-Time Implementation of Face Recognition and Emotion Recognition in a Humanoid Robot Using a Convolutional Neural Network," [9]

### 3.9.2 Summary

The authors researched the possibilities of integrating facial emotion recognition in robots. They compared AlexNet, VGG-16 and a custom model for this task on data collected from human participants. The authors found that VGG-16 performed best on both the facial recognition and emotion recognition tasks with 100% and 73% accuracy respectively. The research concluded that VGG-16 is highly effective for facial emotion recognition tasks.

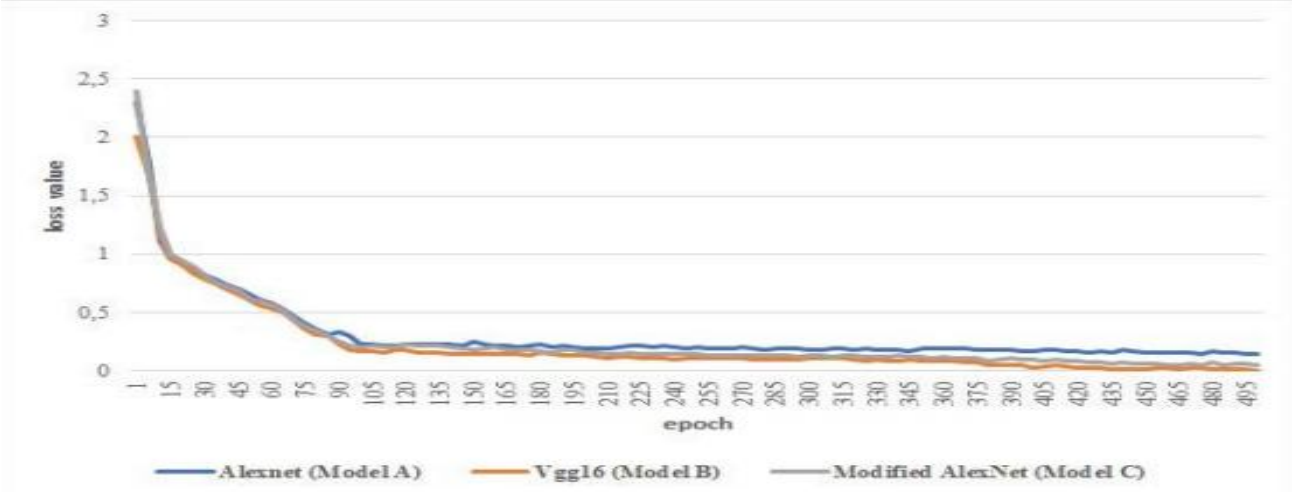


**Fig. 3.9.1** Training Loss b/w models in Face Recognition

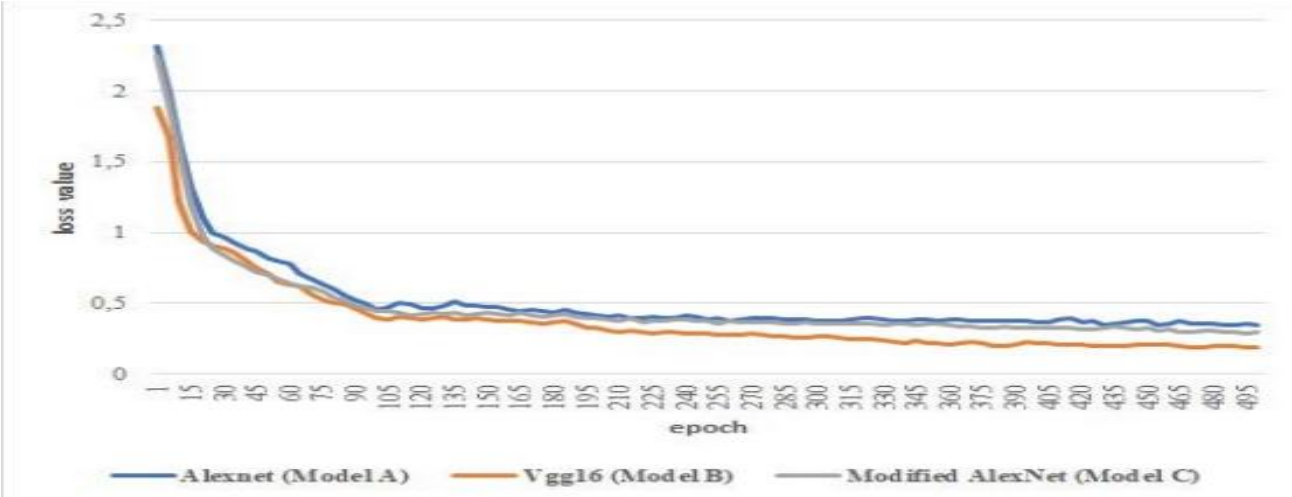


**Fig. 3.9.2** Training Loss b/w models in Emotion Recognition

## 3.10 Research Paper – 10

### 3.10.1 Paper Title

Joan Serrà, Ilias Leontiadis, Dimitris Spathis, Gianluca Stringhini, Jeremy Blackburn, and Athena Vakali. 2017, “Class-based Prediction Errors to Detect Hate Speech with Out-of-vocabulary Words,” [10]

### 3.10.2 Summary

The main challenge addressed by this paper involves detecting hate speech that includes Out-of-Vocabulary (OOV) words, which consists of offensive words that are intentionally misspelled and new unique slurs. Most of the existing methods were basic involving n-gram count and word embeddings, and these models struggle with such ever-evolving language. The authors use prediction errors generated from class-specific models in order to predict these OOV words. The methodology put forward includes Class-specific language models, Error signal extraction, and Neural Network Classifier. In class-specific LMs, the training involved individual character-level models for each label, like hate speech or non-hate speech. These class-specific LMs train on data of their respective class labels and learn to predict the next character. For Error signal extraction, a prediction error, such as Cross-Entropy loss, from each class-specific LM is computed for a given input text which results in a set of error signals showing how well each model makes predictions. These signals are passed to the Neural Network Classifier that provides the likelihood of the input text belonging to each class. According to the authors, this approach helped improve the model’s generalisation. The test was conducted on a dataset of abusive tweets with higher percentage of OOV words. This test showed that the model outperformed the old methods by 4% - 11% during classification tasks. We find that this paper shows the importance in trying to stay updated with the evolving linguistic terms used by and influencing a great part of the online society in everyday life.

# 3.11 Research Paper – 11

## 3.11.1 Paper Title

H. Bhin, Y. Lim and J. Choi, "Multimodal Personality Prediction: A Real-Time Recognition System for Social Robots with Data Acquisition," [11]

## 3.11.2 Summary

This paper sees the author aim to predict human personalities in real-time with a goal to enhance human-robot interactions. After analysing personalities, they believe robots can alter their responses based on each individual. This analysis helps in creating more natural response and behaviour to different situations. The authors find this crucial for social robots to interact in the same manner as humans, such as showing empathy. This is much more effective as it is not as time consuming as traditional methods and is feasible for real-time applications, with multimodality enabling robustness for such operations. The multimodal system consists of Data Acquisition followed by Feature Extraction before Model Training and finally implementing Real-Time Predictions. The initial step starts with collection of data like audio and video cues from human interactions. This is followed by extracting the relevant information which could indicate personality traits. The model is then trained on these features to make predictions of 5 class labels, namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This framework is integrated into a pipeline which makes the real-time system possible. Although there was no real method or benchmark to validate the results of the experiment, it is clear that this multimodal approach is feasible to a reasonable degree.

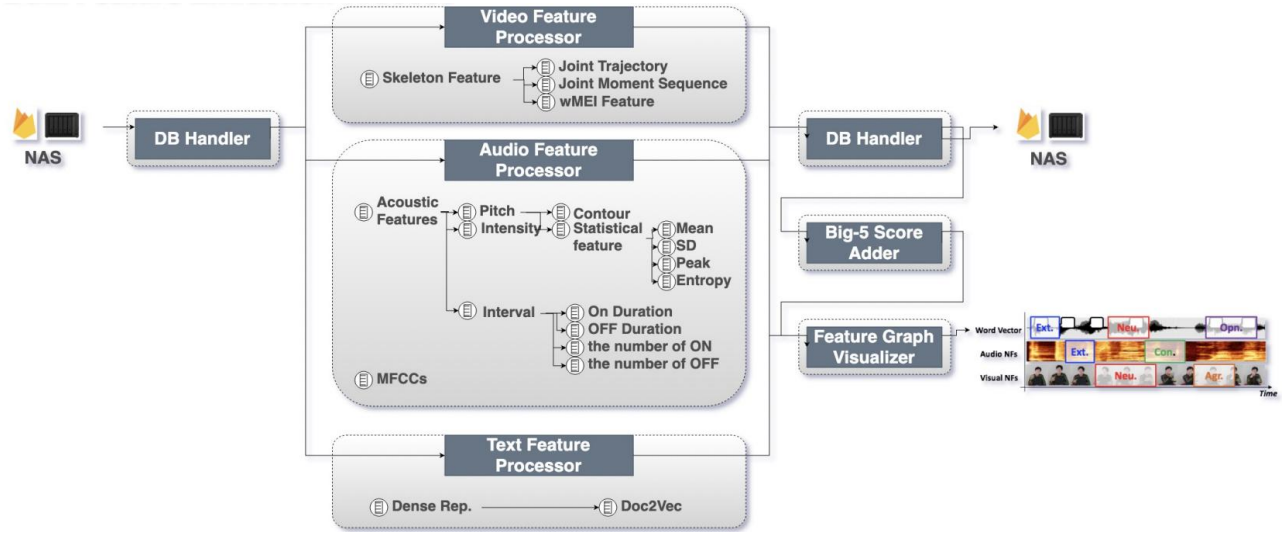


**Fig. 3.11.1** Feature Extraction Process

## 3.12 Research Paper – 12

### 3.12.1 Paper Title

Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023, “Towards Detecting Contextual Real-Time Toxicity for In-Game Chat,” [12]

### 3.12.2 Summary

The model ToxBuster was introduced by the authors. ToxBuster is a real-time toxicity detection model focused mainly on the challenges seen in gaming platforms. An ever-growing platform with rapid messages being exchanged, and slangs related to different games, which needs moderation to try and control the harmful interactions promptly. It is visible that context plays an important role and decisions cannot be made by analysing individual messages. In order to resolve these complexities, the ToxBuster system uses Chat Data to analyse the sequence of messages, which helps in understanding the context. The system then checks metadata to capture timestamps, events that took part in the game, and player roles to provide more context to the model. The system is then made efficient enough to analyse real-time messages in-game. ToxBuster was trained and evaluated on many games such as Rainbow Six Siege, For Honor, and DOTA 2. This model was able to flag 82.1% of reported players with a precision of 90%. It was also able to identify an additional 6% of unreported yet toxic players.

## 3.13 Research Paper – 13

### 3.13.1 Paper Title

Przemyslaw Lenkiewicz, Binyam Gebrekidan Gebre, Oliver Schreer, Stefano Masneri, Daniel Schneider, and Sebastian Tschöpel. 2012, “AVATecH — automated annotation through audio and video analysis,” [13]

### 3.13.2 Summary

The authors try the creation of annotations, which is really hard and takes 50 to 100 times the length of the annotated media, using automated annotations by learning through audio and video analysis. The model has 2 main roles, i.e., decrease the time taken to perform the annotations and using automation to increase the uniformity of the annotations, this can help maintain consistency. For audio and video analysis, the authors use Audio Segmentation, Speech Detection, Speaker clustering (label audio according to the speaker), Vowel and pitch contour detection (graphically use pitch contours to find similar patterns), Global motion detection (distinguish b/w different types of video content using motion), Hands and head tracking (use of a region-growing algo on seed points to create regions approximated by an ellipse). This method has helped reduce the average time taken to annotate by a factor of 0.23 to 0.38 as compared to manual annotating. The next focus of this method is to completely detect and track the movement of hands in a video and differentiate b/w the left and right hand.

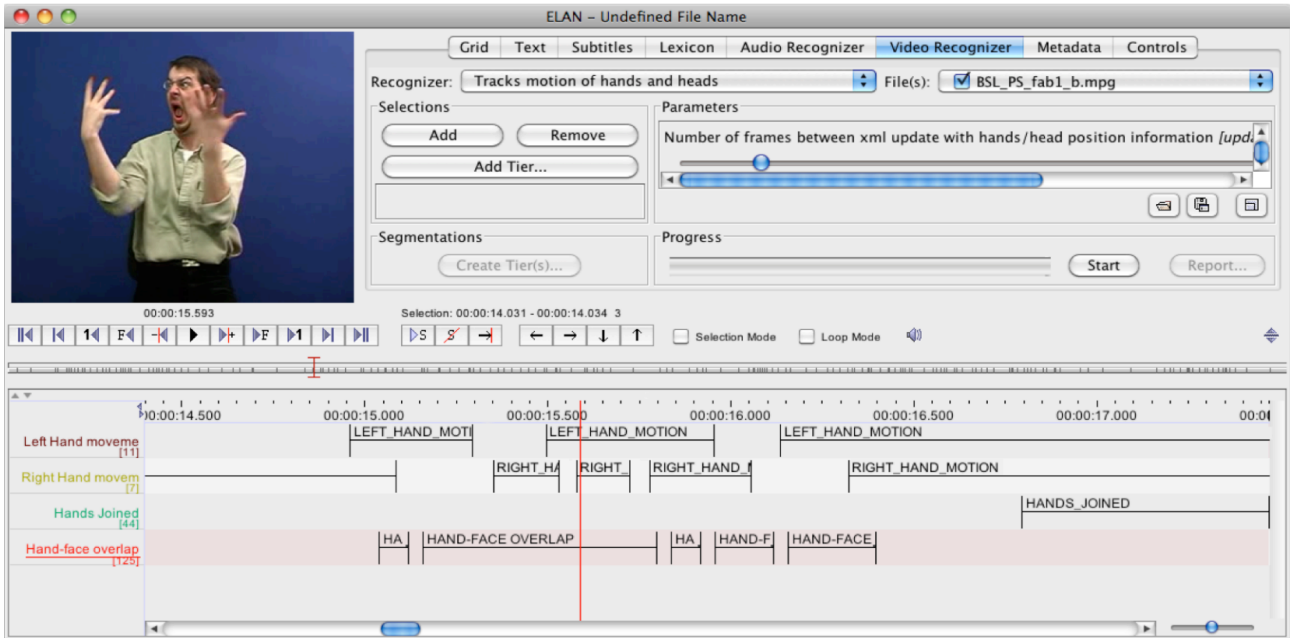


**Fig. 3.13.1** Sample detection of motion for Annotator

## 3.14 Research Paper – 14

### 3.14.1 Paper Title

Jihyung Moon, Dong-Ho Lee, Hyundong Cho, Woojeong Jin, Chan Park, Minwoo Kim, Jonathan May, Jay Pujara, and Sungjoon Park. 2023, "Analyzing Norm Violations in Live-Stream Chat," [14]

### 3.14.2 Summary

The goal of this paper is to find toxic language, such as hate speech, and norm violations in live streaming platforms like Twitch and YouTube Live. The study focuses on NLP to detect the aforementioned problem. The authors define norm violation categories in live stream chats and annotate 4,583 moderated comments (NormVio-RT is the dataset created by the authors) from Twitch. The model is trained based on informational context humans use in live stream moderation. The final results showed that using the right contextual information can increase the performance of moderation by 35%. The dataset contains moderated comments from the top 200 Twitch steamers, along with their chat rules. The main focus is on comments that triggered a moderation like user ban, or user timeout.

| Coarse | Fine-grained | # Rules | # Violates | | |
|---|---|---|---|---|---|
| | | | stage 1 | stage 2 | stage 3 |
| Discrimination | Discrimination | 13.98% (46) | 2.34% (104) | 2.25% (101) | 2.34% (105) |
| HIB | HIB | 22.49% (74) | 21.33% (947) | 26.55% (1,190) | 27.80% (1,246) |
| Privacy | Doxing | 0.60% (2) | 0.34% (15) | 0.36% (16) | 0.36% (16) |
| Inappropriate Contents | Spoiler | 0.60% (2) | 0.02% (1) | 0.02% (1) | 0.02% (1) |
| | NSFW | 1.82% (6) | 0.86% (38) | 0.85% (38) | 0.85% (38) |
| | Self-destructive | 1.21% (4) | 0.32% (14) | 0.29% (13) | 0.29% (13) |
| | Illegal | 0.30% (1) | 0.16% (7) | 0.07% (3) | 0.07% (3) |
| Off Topic | Controversial Topic | 5.47% (18) | 0.59% (26) | 0.85% (38) | 0.83% (37) |
| | Begging | 1.51% (5) | 1.44% (64) | 1.36% (61) | 1.36% (61) |
| Spam | Excessive & Repetitive | 11.24% (37) | 17.59% (781) | 21.64% (970) | 21.42% (960) |
| | Advertisements | 11.24% (37) | 4.64% (206) | 4.40% (197) | 4.42% (198) |
| Meta-Rules (Live streaming specific) | Mentioning other streamers | 14.28% (47) | 0.72% (32) | 10.62% (476) | 10.58% (474) |
| | Backseating & Tall order | 5.16% (17) | 3.45% (153) | 3.70% (166) | 3.77% (169) |
| | Specific language only | 10.03% (33) | 0.97% (43) | 6.94% (311) | 6.94% (311) |
| Incivility (Miscellaneous) | Incivility | - | 12.30% (546) | 11.57% (519) | 11.51% (516) |
| | Non-Identifiable | - | 32.93% (1,462) | 8.52% (382) | 7.45% (334) |
| Total | | 329 | 4,439 | 4,482 | 4,482 |

**Fig. 3.14.1** Statistics related to # of messages violating norms in different stages

# 3.15 Research Paper – 15

## 3.15.1 Paper Title

Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019, "Detecting harassment in real-time as conversations develop," [15]

## 3.15.2 Summary

The authors find a method to detect video game players that harass teammates or opponents in chat. This would simplify the process for gaming companies to intervene during games by issuing warnings, muting, or banning the player. Toxic players are detected as the conversation develops, as early as possible, making it possible for gaming companies to intervene. The machine learning task is that instances (e.g.: players) change over time, as more information about the instances (more utterances) becomes available. This leads to time as an extra dimension of interest for metrics like precision, recall and F-score: instead of presenting them as a single number, it should be represented how they change during the conversation. The dataset consists of 5000 conversations from League of Legends, with utterances of 48,512 players. The proposed framework is called HaRe (Harassment Recognizer), it keeps track of toxicity estimates for all participants separately, updating the estimate for each utterance. This is done by concatenating all utterances, separated by [NEW UTTERANCE] tags, and classifying the resulting text. The rate in which the sliding threshold should be increased depends on the size of the training set: the larger the training set, the slower the threshold can be increased.

## 3.16 Research Paper – 16

### 3.16.1 Paper Title

Xiaoda Yang, Xize Cheng, Jiaqi Duan, Hongshun Qiu, Minjie Hong, Minghui Fang, Shengpeng Ji, Jialong Zuo, Zhiqing Hong, Zhimeng Zhang, and Tao Jin. 2024, “AudioVSR: Enhancing Video Speech Recognition with Audio Data,” [16]

### 3.16.2 Summary

The authors develop a VSR to predict spoken words by analysing lip movement in videos. To enhance VSR, they use audio data from a generative model for data inflation, including synthetic data with authentic visual data, and try leveraging the alignment of the audio and video. In zero-shot and full-shot situations, an audio-to-lip model was used for data inflation, using the knowledge learned from the generative model to enhance the lip-reading model. For cross-linguistic scenarios, the impact of different language families was considered on lip-reading tasks and later used to train an audio-lip-alignment model using self-supervised learning. Using TFG for data inflation is implemented by using a sequence model, AV-HuBERT which highlights the importance of temporal correlations in video data for VSR tasks. Audio inherently carries temporal information, so by transferring audio knowledge to the corresponding video, the audio-to-lip model generates a video stream that contains the same temporal information. The model significantly improves the semantic understanding of video content by using the temporal information in audio to strengthen the flow amongst video frames. AudioVSR is a unified model that maps the audio and video of all languages into the same space. With the AudioVSR model, they input audio from entirely new datasets and fine-tune on downstream tasks, developing a model suitable for VSR tasks.

## 3.17 Research Paper – 17

### 3.17.1 Paper Title

Athanasios Lykartsis and Margarita Kotti. 2019, “Prediction of User Emotion and Dialogue Success Using Audio Spectrograms and Convolutional Neural Networks,” [17]

### 3.17.2 Summary

The paper focuses on predicting dialogue success and user satisfaction, as well as user emotion, using only audio spectrogram representations and CNNs. The approach allows for a fast-running basic-time system that can enhance spoken dialogue systems (SDS). The speech itself contains verbal cues and emotions that can indicate user satisfaction. The dataset used is Let’s Go V2, which has 5065 audio interactions, with 3 labels – user emotion, subjective success, objective success. The audio was segmented into 1-second and 2-second for testing and mel-spectograms were used. The best performing model was bottleneck CNN with 90% accuracy and the 1s spectrograms performed better than 2s.

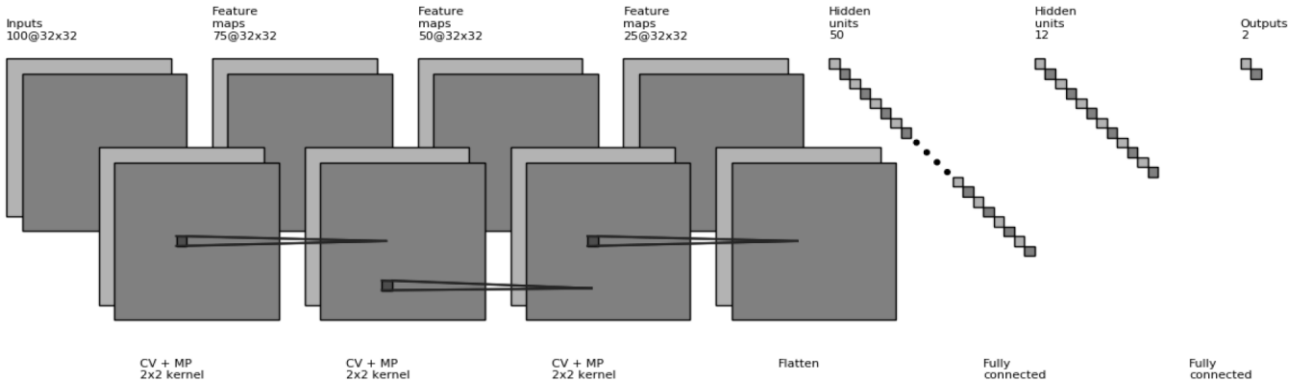


**Fig. 3.17.1** Bottleneck architecture flowchart

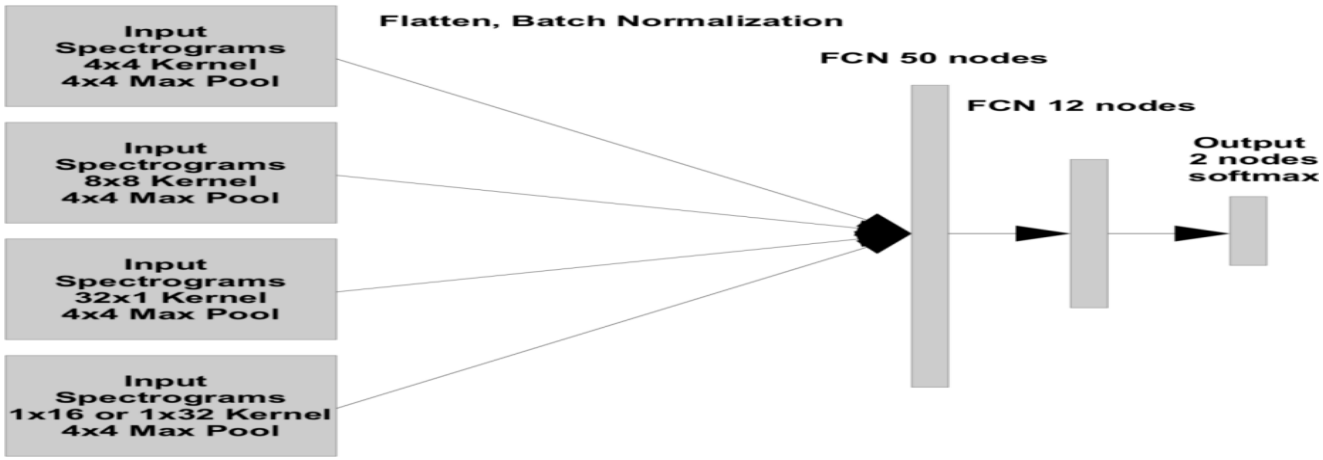


**Fig. 3.17.2** Parallel architecture flowchart

# CHAPTER 4
# HIGH LEVEL DESIGN

## 4.1 About the Design

The inputs provided to the model are real-time videos, and the information extracted from this input is 2 part – audio cues and visual cues. The audio cues are in waveforms and represented as Log-Mel spectrograms, capturing spectral and temporal features, i.e., it shows how the energy of a signal is distributed across frequencies over time. In this process of creating spectrograms, the Mel scale helps perceive the human pitch and the logarithmic scale is applied to reduce the impact of very high-energy frequencies and amplify the lower-energy frequencies. This is done because the human ears are more sensitive to lower frequencies. For visual cues, similar to the Log-Mel spectrograms for audio cues, the spatial and temporal features are represented using Facial Expression Analysis. With room for improvement, we can find a better framework that could help us improve the type of information needed and used by the model to work on visual cues. These inputs are then passed onto their own neural networks stack, which consists of CNNs, RNNs, or both. The output tensors from these stacks are then used to align the temporal features between audio and visual modalities. This is implemented using Cross-Modal Attention, where one modality guides the model's focus on the other modality's features. Example, cross-modal attention helps to highlight a part of the video, containing a type of facial expression, that corresponds to a type of tone or pitch in the audio. These attended features are then integrated to create a representation which can be used for classification or regression. This representation is passed into a Sequence Modelling Layer which is a type of neural network targeted for tasks using data such as time-series data, video frames, etc. This SML is implemented using architectures like RNNs, transformers, or both. The final prediction is carried out in different forms based on the type of classification, such as Binary Classification, Multi-Class Classification, or Sequence Predictions, with the end goal to detect and censor inappropriate speech at real-time.

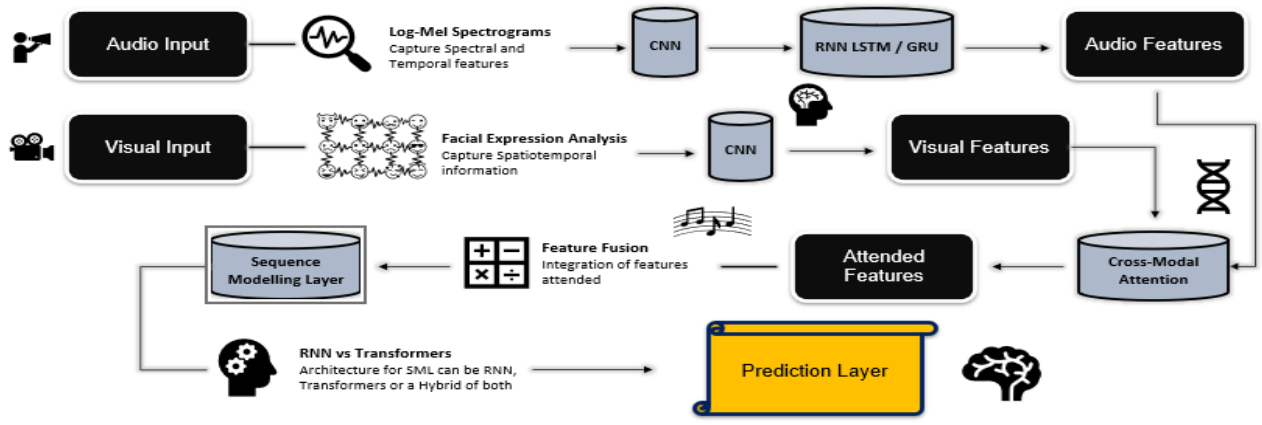


**Fig. 4.1.1** Design of Model's Architecture

# CHAPTER 5

# DATA PRE-PROCESSING

## 5.1 Audio Dataset

For Audio, the dataset at "https://huggingface.co/datasets/Lameus/en_spontaneous_profanity" [18] was used. This dataset consists of 2,332 rows of audio clips ranging from 2 to 8 seconds, and are obtained from the Common Voice dataset which has samples with explicit profanity. There are 2 labels used – 0 for Train data and 1 for Test data, and covers over 388 profane words.

## 5.2 Video Dataset

For Video, we are planning to use the FER and CMU MOSEI dataset. The FER dataset consists of 35,887 grayscale images of faces (48x48 pixels). It has seven emotion classes – Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. It is used in research for facial expression recognition tasks. CMU MOSEI is a very large dataset with multimodality for sentiment analysis. It contains over 23,500 video clips from YouTube, covering a variety of topics and speakers. It works with 3 types of data, namely textual, visual, and acoustic modality. This dataset has labels for sentiment ranging from -3 to +3, and six basic emotions – Happy, Sad, Anger, Surprise, Disgust, Fear. It is currently widely used in multimodal machine learning research. The potential use of these datasets would be Real-Time Sentiment Analysis in videos.

## 5.3 Pre-Processing

With an aim to predict the use of profanity by analysing sequence of speech leading up to the use of the actual profane word. We require the audio to be represented as raw waveforms. Hence, we have pre-processed the samples in 3 formats – Mel-Frequency Cepstral Coefficients, Log-Mel Spectrograms, Wav2Vec.

- Mel-Frequency Cepstral Coefficients captures observable features of audio signals.
- Log-Mel Spectrograms are spectrograms where frequency bins are mapped onto the Mel scale and converted to logarithmic values.
- Wav2Vec is a model that extracts information related to speech from raw audio waveforms.

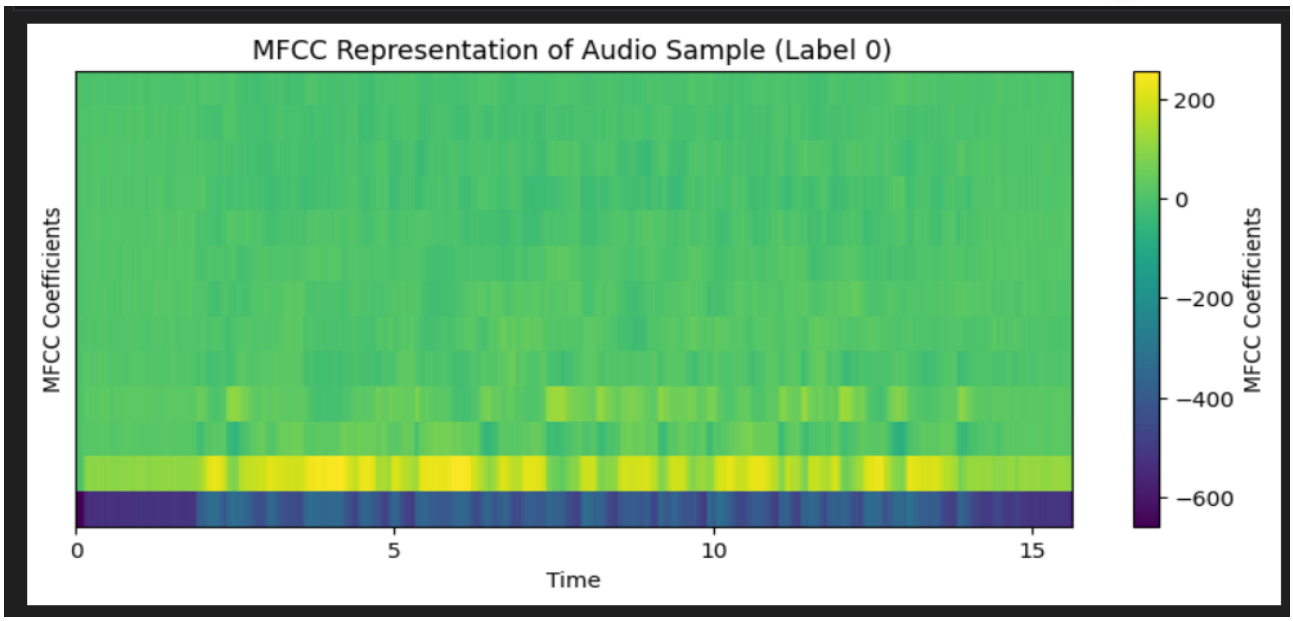


**Fig. 5.3.1** MFCC representation of audio samples

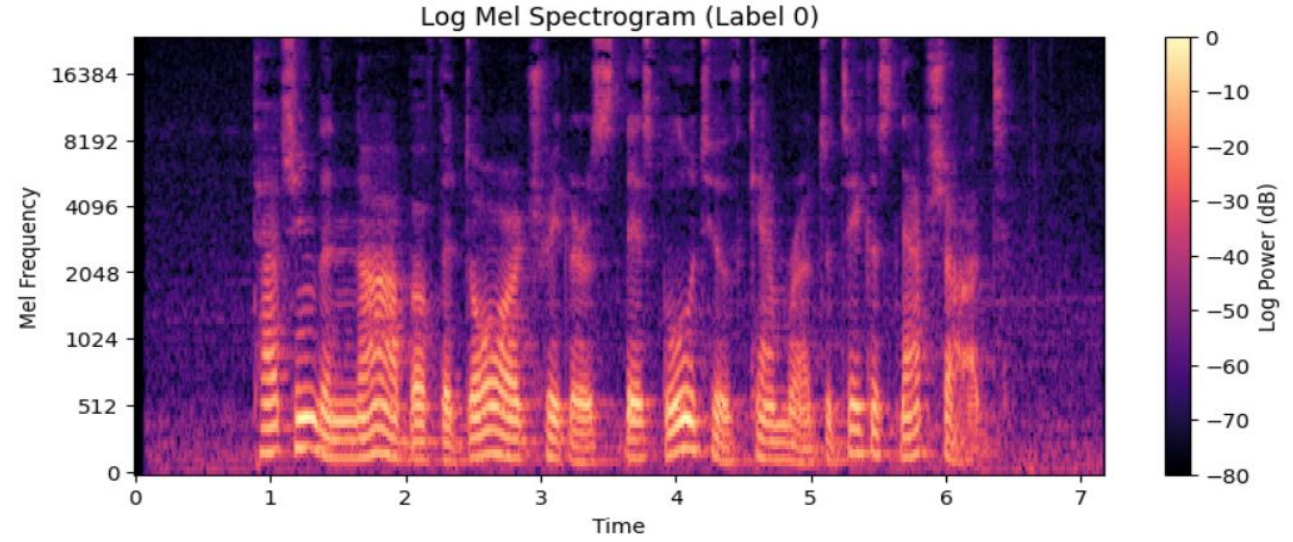


**Fig. 5.3.2** Log-Mel Spectrogram representation of audio samples

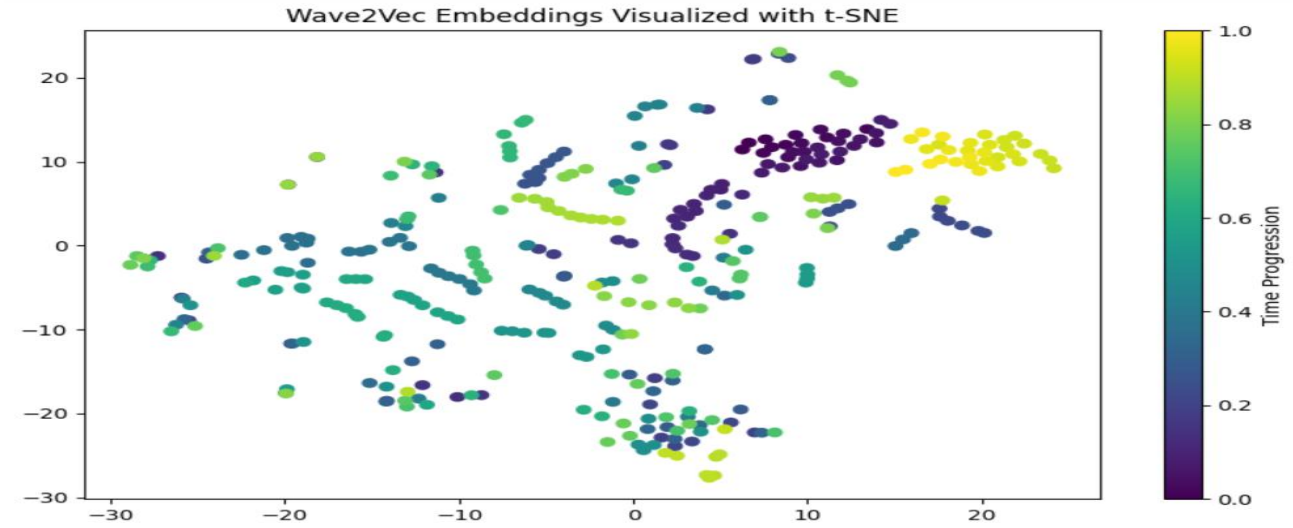


**Fig. 5.3.3** Wav2Vec representation of audio samples

# CHAPTER 6

# IMPLEMENTATION

## 6.1 Data Collection

The data for implementation was scraped from the IMSDB site [19]. This site covers the captions and scripts of various movies and shows, making it suitable for having text content that captures a variety of emotions and context that are likely to place in our everyday life. The figure 6.1.1 shows the frequency of swear words extracted from the site.

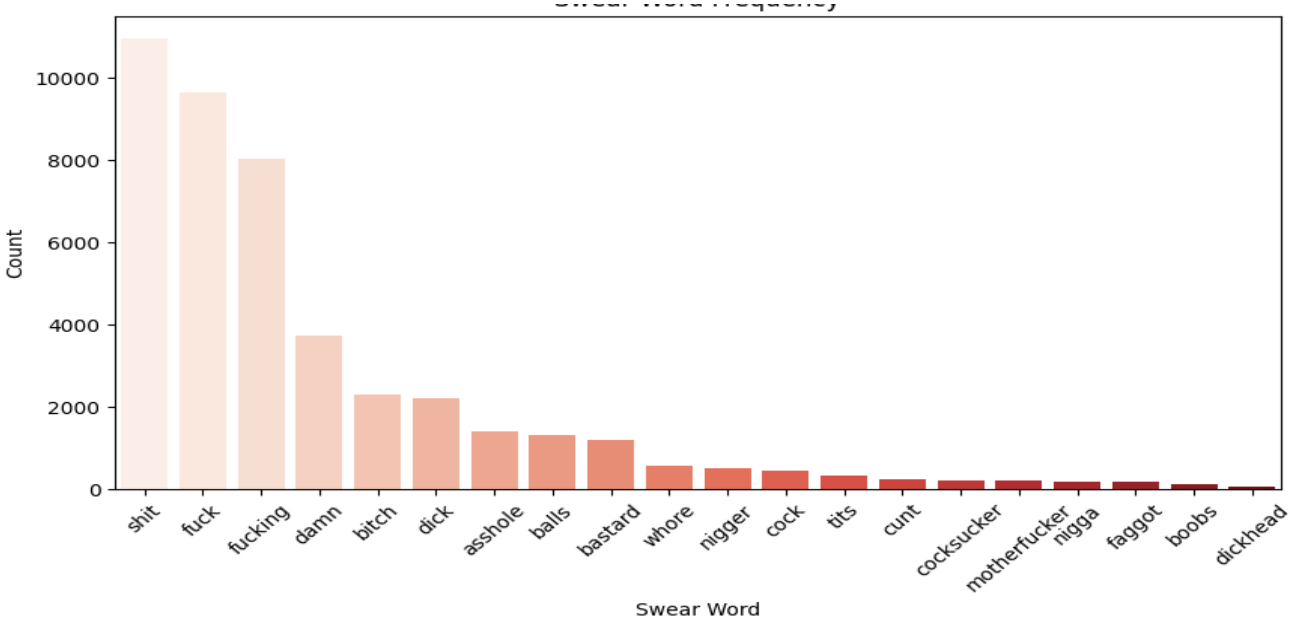


**Fig. 6.1.1** Frequency of swear words in collected data

## 6.2 Pre-processing

After analysis, we found that the data collected shows class imbalance with very few cases of swear word dialogues as compared to non-swear word dialogues. This was resolved through sampling a subset of the data and in this sampling, we try to get a fixed, and reasonable percentage of swear word dialogues.

```
Positive examples (swear words): 11429
Negative examples (non-swear words): 276975
Class balance: 3.96% positive
```

```
After oversampling: 311262 total examples
New class balance: 11.02% positive
```

**Fig. 6.2.1** Class imbalance before and after preprocessing

## 6.3 Results

The model uses the Out-of-Vocabulary tokenizer, with binary cross-entropy loss, and Adam optimizer. It was trained on batch size of 64, for 10 epochs, and executes early stopping if the validation loss doesn't improve for 3 continuous epochs. This code also ensures that the model restores back to its best weights if needed. Upon custom testing on likely swear phrases, neutral phrases, and ambiguous phrases, the model performs with an accuracy of **87.5%**.

```
==================================================
TEST SET EVALUATION
==================================================
              precision    recall  f1-score   support

   Non-Swear       0.96      0.98      0.97     55396
       Swear       0.78      0.68      0.73      6857

    accuracy                           0.94     62253
   macro avg       0.87      0.83      0.85     62253
weighted avg       0.94      0.94      0.94     62253

Confusion Matrix:
[[54113  1283]
 [ 2211  4646]]

Input: sykes you jus ' lay still , we goin figure out what the
True label: Swear
Predicted probability: 1.0000
Predicted label: Swear
```

**Fig. 6.3.1** Classification report, Confusion matrix and Sample of Test results

## 6.4 ASR Testing

During implementation, we used Google's Web Speech API and the Speech_Recognition python library to create a pipeline using our model's tokenizer and weights which converts spoken words, on the microphone, to text and feeds it to the model which then determines if the next word is a swear word or non-swear word.

```
Model loaded successfully.
Tokenizer loaded successfully.
--------------------------------------------------
Speech Recognition Configuration:
 - Sample Rate: 44100
 - Sample Width: 2
 - Using Microphone: Default
 - Recognizer Energy Threshold: 300
 - Prediction Threshold: 0.4000
--------------------------------------------------
Adjusting for ambient noise...
Ambient noise adjustment complete. Threshold: 2099.68
Ready! Start speaking (press Ctrl+C to stop)...

Listening...
Got audio, processing...
Recognized: "what the"
Current context: 'what the'
Prediction Probability: 0.9875
*** Prediction (Threshold 0.40): SWEAR WORD LIKELY NEXT! ***
```

**Fig. 6.4.1** Sample output of ASR Testing

# CHAPTER 7

# CONCLUSION OF CAPSTONE PROJECT PHASE – II

In this phase of our project, Real-Time Inappropriate Speech Detection, we have expanded our understanding of this problem statement to build a solid foundation to work with. Upon this, we have extended our literature survey by going through 17 additional research papers which have helped us create a basic idea on completing tasks from identifying the types of data which are suitable for our system, to different methods of extracting this data and preprocessing it. It also gave an outline on how each method impacts the type of information retrieved and the model's performance. It also gave an outline into model architecture choices.

Adding on this, we tested with Automatic Speech Recognition (ASR), which helped convert words spoken into a microphone to text. The model was able to detect the right class of the next word, after conversion, with an accuracy of 87.5%.

Future plans include shifting to a multimodal implementation by including video-based analysis. It includes analysing facial expressions to help improve the understanding of context. Next, we move to spectrographic form of audio, which is likely to capture more vocal characteristics. Finally, we intend to get a more diverse dataset which is more generalizable and robust.

# CHAPTER 8

# PLAN OF WORK FOR CAPTION PROJECT PHASE – III

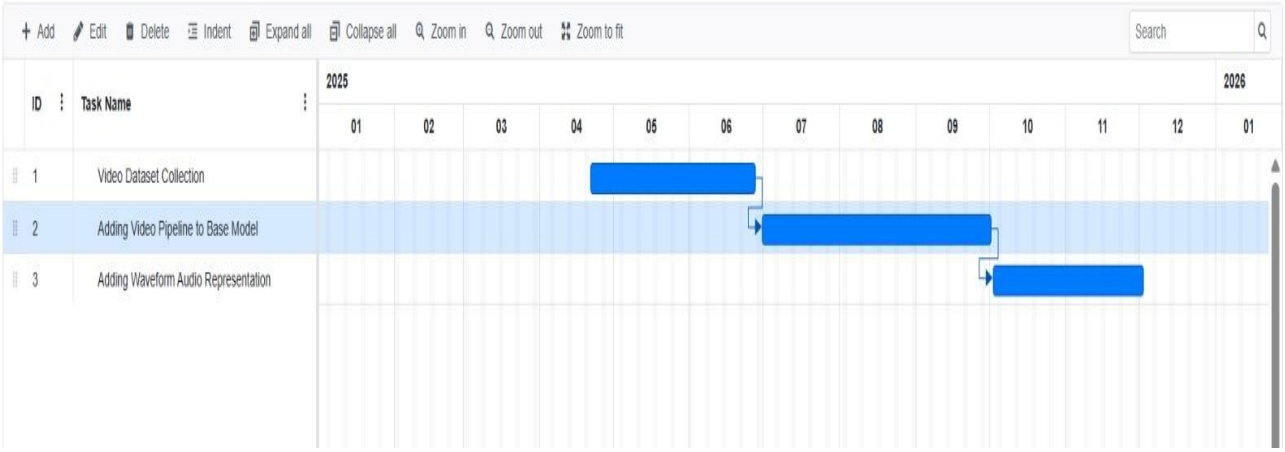


**Fig. 8.1** Gantt Chart

The above figure is the Gantt Chart for our Phase – III. For this phase, we would like to conduct a more systemised process for collection of video data, and then create a video pipeline for our base model. We are also looking to use waveforms for audio representations and build a good, composed low level design.

# REFERENCES

[1] I. Smirnov and A. Laushkina, "Multimodal prediction of profanity based on speech analysis," Procedia Computer Science, vol. 229, pp. 62–69, 2023.

[2] A. Chaudhari, P. Davda, M. Dand and S. Dholay, "Profanity Detection and Removal in Videos using Machine Learning," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp.572-576, doi:10.1109/ICICT50816.2021.9358624.

[3] Rana, Aneri and Sonali Jha. “Emotion Based Hate Speech Detection using Multimodal Learning,” ArXiv abs/2202.06218 (2022): n. pag.

[4] Ba Wazir, A.S.; Karim, H.A.; Abdullah, M.H.L.; AlDahoul, N.; Mansor, S.; Fauzi, M.F.A.; See, J.; Naim, A.S., “Design and Implementation of Fast Spoken Foul Language Recognition with Different End-to-End Deep Neural Network Architectures,” Sensors 2021, 21, 710. https://doi.org/10.3390/s21030710

[5] J. Pan, W. Fang, Z. Zhang, B. Chen, Z. Zhang and S. Wang, "Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG," in IEEE Open Journal of Engineering in Medicine and Biology, vol. 5, pp. 396-403, 2024, doi: 10.1109/OJEMB.2023.3240280.

[6] H. M. Fayek, M. Lech and L. Cavedon, "Towards real-time Speech Emotion Recognition using deep neural networks," 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), Cairns, QLD, Australia, 2015, pp. 1-5, doi: 10.1109/ICSPCS.2015.7391796.

[7] H. Meng, T. Yan, F. Yuan and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," in IEEE Access, vol. 7, pp. 125868-125881, 2019, doi: 10.1109/ACCESS.2019.2938007.

[8] D. Duncan, G. Shine, and C. English, "Facial Emotion Recognition in Real Time," Stanford University, 2016.

[9] S. Dwijayanti, M. Iqbal and B. Y. Suprapto, "Real-Time Implementation of Face Recognition and Emotion Recognition in a Humanoid Robot Using a Convolutional Neural Network," in IEEE Access, vol. 10, pp. 89876-89886, 2022, doi: 10.1109/ACCESS.2022.3200762.

[10] Joan Serrà, Ilias Leontiadis, Dimitris Spathis, Gianluca Stringhini, Jeremy Blackburn, and Athena Vakali. 2017, “Class-based Prediction Errors to Detect Hate Speech with Out-of-vocabulary Words,” In Proceedings of the First Workshop on Abusive Language Online, pages 36–40, Vancouver, BC, Canada. Association for Computational Linguistics.

[11] H. Bhin, Y. Lim and J. Choi, "Multimodal Personality Prediction: A Real-Time Recognition System for Social Robots with Data Acquisition," 2024 21st International Conference on Ubiquitous Robots (UR), New York, NY, USA, 2024, pp. 673-676, doi: 10.1109/UR61395.2024.10597440.

[12] Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023, “Towards Detecting Contextual Real-Time Toxicity for In-Game Chat,” In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9894–9906, Singapore. Association for Computational Linguistics.

[13] Przemyslaw Lenkiewicz, Binyam Gebrekidan Gebre, Oliver Schreer, Stefano Masneri, Daniel Schneider, and Sebastian Tschöpel. 2012, “AVATecH — automated annotation through audio and video analysis,” In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 209–214, Istanbul, Turkey. European Language Resources Association (ELRA).

[14] Jihyung Moon, Dong-Ho Lee, Hyundong Cho, Woojeong Jin, Chan Park, Minwoo Kim, Jonathan May, Jay Pujara, and Sungjoon Park. 2023, “Analyzing Norm Violations in Live-Stream Chat,” In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 852–868, Singapore. Association for Computational Linguistics.

[15] Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019, “Detecting harassment in real-time as conversations develop,” In Proceedings of the Third Workshop on Abusive Language Online, pages 19–24, Florence, Italy. Association for Computational Linguistics.

[16] Xiaoda Yang, Xize Cheng, Jiaqi Duan, Hongshun Qiu, Minjie Hong, Minghui Fang, Shengpeng Ji, Jialong Zuo, Zhiqing Hong, Zhimeng Zhang, and Tao Jin. 2024, "AudioVSR: Enhancing Video Speech Recognition with Audio Data," In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15352–15361, Miami, Florida, USA. Association for Computational Linguistics.

[17] Athanasios Lykartsis and Margarita Kotti. 2019, "Prediction of User Emotion and Dialogue Success Using Audio Spectrograms and Convolutional Neural Networks," In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, pages 336–344, Stockholm, Sweden. Association for Computational Linguistics.

[18] "https://huggingface.co/datasets/Lameus/en_spontaneous_profanity,"

[19] "https://imsdb.com/,"

# 6% Matches

| | | |
|---|---|---|
| 1 | Internet | |
| web.archive.org | | 2% |
| 2 | Internet | |
| aclanthology.org | | 1% |
| 3 | Internet | |
| preview.aclanthology.org | | <1% |
| 4 | Internet | |
| upcommons.upc.edu | | <1% |
| 5 | Internet | |
| www.researchgate.net | | <1% |
| 6 | Internet | |
| prr.hec.gov.pk | | <1% |
| 7 | Internet | |
| summit.sfu.ca | | <1% |
| 8 | Internet | |
| www.aclweb.org | | <1% |
| 9 | Internet | |
| www.isteonline.in | | <1% |
| 10 | Publication | |
| Iacopo Carnacina, Mawada Abdellatif, Manolia Andredaki, James Cooper, Darren ... | | <1% |

11 Internet

arxiv.org <1%

12 Internet

link.springer.com <1%

13 Internet

open.metu.edu.tr <1%

14 Internet

www.pccoer.com <1%