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 - 1

Submitted by:

SAI ARYA R B	**PES1UG22AM142**
SAMARTH SK	**PES1UG22AM144**
SIDDHARTH GANDHI	**PES1UG22AM157**
SUDARSHAN SRINIVASAN	**PES1UG22AM166**

Under the guidance of

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

August - December 2024

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 Fifth-semester Capstone Project Phase - 1 (UE22AM320A) in the Program of Study -Bachelor of Technology in CSE(AI&ML) under rules and regulations of PES University, Bengaluru during the period Aug. 2024 – Dec. 2024. 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 5th-semester academic requirements in respect of project work.

Signature Dr. Jayashree R Chairperson	Signature Dr. Jayashree R Chairperson	Signature Dr. B K Keshavan Dean of Faculty

External Viva

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

DECLARATION

We hereby declare that the Capstone Project Phase - 1 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 Aug – Dec 2024. 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 – 1.

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 Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro-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 1 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 report 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.

This report 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, this 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

A. S. B. Wazir et al., "Spectrogram-Based Classification of Spoken Foul Language Using Deep CNN,"[1]

3.1.2 Summary

The authors of this paper realised the problem of the proliferation of inappropriate content in the internet and its easy access especially to those who should not be watching or listening to such content. They also acknowledged the difficulty in moderating and censoring content throughout the internet and aimed to design a method to automate censorship of profanity.

The authors planned to use transfer learning on CNN to train on the segments of spectral images generated from audio. This was the approach chosen by the authors. For feature extraction they chose to use **Mel-Frequency Cepstral Coefficients, Discrete Wavelet Transforms, spectral images**, etc. For classification they used **Deep Convolutional Networks, Recurrent Neural Networks, K-nearest neighbours**, etc.

The dataset was created by the authors themselves due to how uncommon foul language datasets are. They made sure to include **noise, accents and other transformations**. The dataset was then annotated in two ways. First a general **Foul** and **Normal** way. Then the **Foul** dataset was further classified into **9 different subtypes** of cuss words.

The CNN architectures that were chosen are: **ResNet50, VGG16, Googlenet and Alexnet**. The transfer was done by freezing some of the layer weights or fine-tuning the layers. For the spectrogram it was done using the vectors of consecutive spectral coefficients. This was achieved using the combination of a window and overlapping window for extraction.

The **ResNet50** model performed best in both the **2-way classification and 10-way classification**, showing an **F1 score** and **error rate** of **98.54%** and **1.24** for **2-way classification** and **94.20%** and **5.49** for **10-way classification**.

3.1.3 Key Insights

Reliable Automated Detection

This paper showed us a method of automated vulgar language detection that provides reliable and faster results.

Efficient Classification of profanity

This model uses a CNN trained on spectrogram depictions of audio segments which gives an effective method of detecting curse words.

Practical Applications

This paper also showed us a stable and scalable method that can be integrated in online and streaming platforms seamlessly to flag inappropriate content.

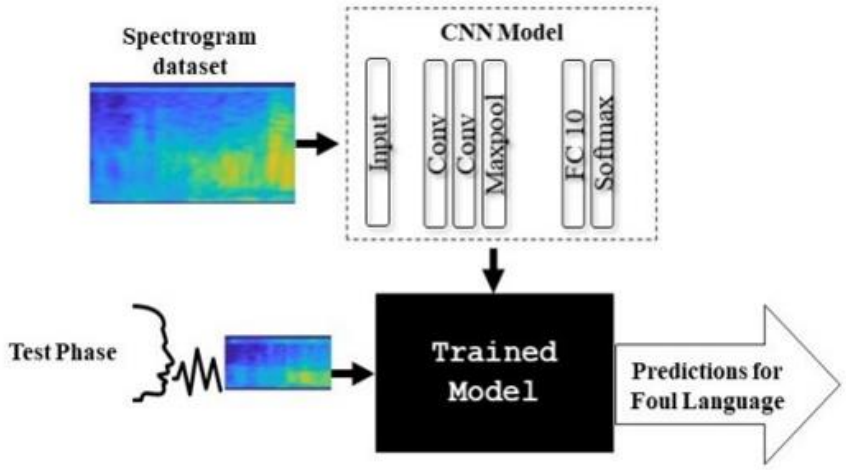

Fig. 3.1.1 Illustration of foul language model

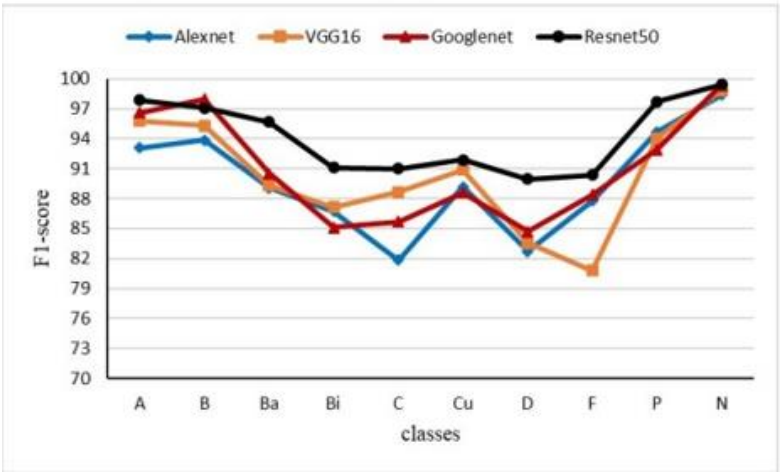

Fig. 3.1.2 F1-score comparison for 10-class models

3.2 Research Paper – 2

3.2.1 Paper Title

A.S. B. Wazir, H. A. Karim, H. S. Lyn, M. F. Ahmad Fauzi, S. Mansor and M. H. Lye, "Deep Learning-Based Detection of Inappropriate Speech Content for Film Censorship," [2]

3.2.2 Summary

This paper addresses the issue of established censorship methods, which require manual labour and may be computationally expensive and costly and also subject to human errors by leveraging deep learning to automate the Process. The Authors have used deep convolutional neural networks for this approach and have also used a manually annotated dataset for specific words and speech. The Paper tested the system and calculated metrics like accuracy, F1 score and compared this model to baseline models that were ASR-Based (Automatic Speech Recognition) and explained the improvements of the proposed CNN approach.

3.2.3 Key Insights

MMUTM and **TAPAD** datasets consisting of audio samples, around 5 offensive words per minute with the timestamps manually annotated for each class of profanity. the sizes of the datasets were increased by augmentation and the samples were classified as **'Foul'** and **'Normal'**. After training, testing was performed on continuous audio samples from videos.

Log-Mel spectrograms for Audio Pre-Processing and feature Extraction

The authors have used transformed audio samples into Log-Mel spectrograms which are a 2-D representation of audio data, and used these as input features for the CNN. Essentially, the model treats the Log-Mel spectrograms as images.

The authors have highlighted the benefits of using Log-Mel spectrograms as compared to traditional automatic speech recognition methods as:

- Log-Mel spectrograms capture both the spectral and temporal features of audio signals, making it effective to associate acoustic patterns with foul language.
- The complexity of ASR systems can lead to significant computational overhead.
- ASR systems require prior knowledge of the language being spoken while Log-Mel spectrograms only focus on acoustic properties of speech.

End-to-End Convolutional Neural Networks as a key method

The authors have emphasized the use of E2E CNN as compared to traditional computationally heavy ASR systems with the use of lower kernels, filters and a milder architecture which has a faster process for classification of inappropriate speech which used images from Log-Mel spectrograms and performed feature learning for classification.

Comparing performance with existing Baseline Models

The paper has stated that the model has high accuracy, with a **macro average AUC** of **93.85 %** and **weighted average AUC** of **94.58 %** and has outperformed all state-of-the-art baseline models.

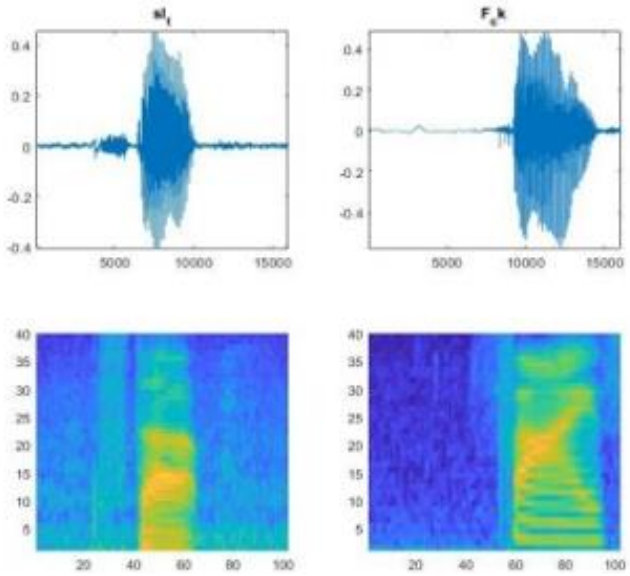

Fig. 3.2.1 Raw signals vs Spectrograms

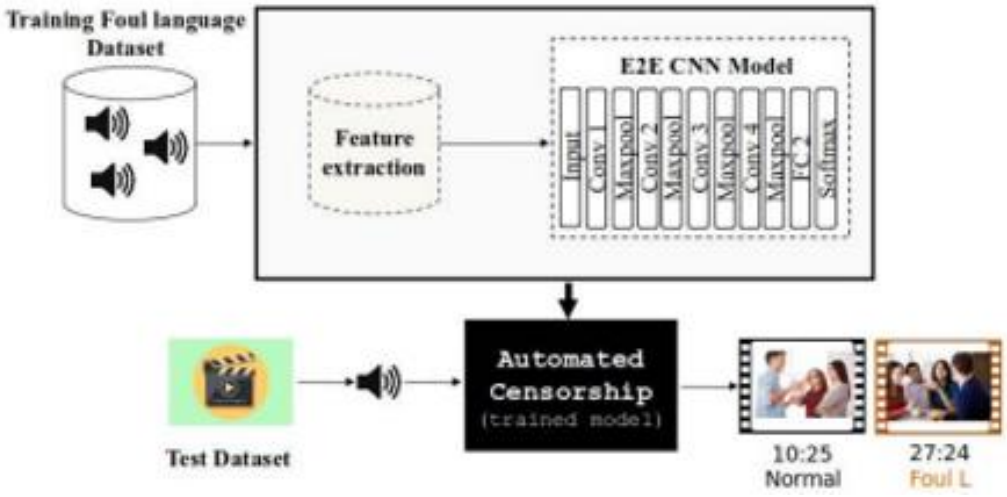

Fig. 3.2.2 System architecture for inappropriate language detection

3.3 Research Paper – 3

3.3.1 Paper Title

Q. Wei, X. Huang and Y. Zhang, "FV2ES: A Fully End2End Multimodal System for Fast Yet Effective Video Emotion Recognition Inference,"[3]

3.3.2 Summary

The paper introduces FV2ES model that leverages multimodal data to infer emotional states. The authors believe that people like to show their emotions in social networks through text, speech and facial expressions. They observed that Multimodal Integration by combining the multimodal data to enrich emotional context understanding. The authors decided to build the above fully end-to-end model as the lack of such a model hinders its application on real world and practical instances.

3.3.3 Key Insights

The authors devised the following methods in their FV2ES model:

- Usage of a Hierarchical Architecture for Nuanced insights. The authors used this method to particularly address the issue of Acoustic Modality. This issue is resolved by maximizing the acoustic modality's contribution.
- Usage of Multi-Scale Visual Feature Extraction. This method was used by them as it demonstrates how capturing fine-grained details can improve accuracy without increasing computational overhead.
- Usage of Integrated Preprocessing in Multimodal Learning. The authors focused on optimizing the performance by embedding preprocessing into model pipeline and hence reducing storage and computational costs.

3.3.4 Relevance to Project

- The method of Hierarchical Architecture could be adapted to enhance the accuracy of detecting subtle variations in tone, inflection, or other audio cues, indicative of inappropriate speech. These properties are derived from the sound spectra.
- Adapting a Multi-Scale Visual Feature Extraction can be used on text or speech features and refine our system's real-time detection capabilities while maintaining efficiency.
- Integrated Preprocessing and Model Pipelining can directly benefit our system by enabling faster real-time detection and minimizing resource requirements. This improves scalability to online platforms.

3.3.5 Technological Gaps Addressed

- The author's innovation on acoustic modality can be used to address the challenges of subtle or hidden explicit speech in audio formats, where features like tone and context matter significantly.
- The Scalable Design Principles adopted by the authors can efficiently process large-scale real-time data while maintaining low latency and storage requirements.

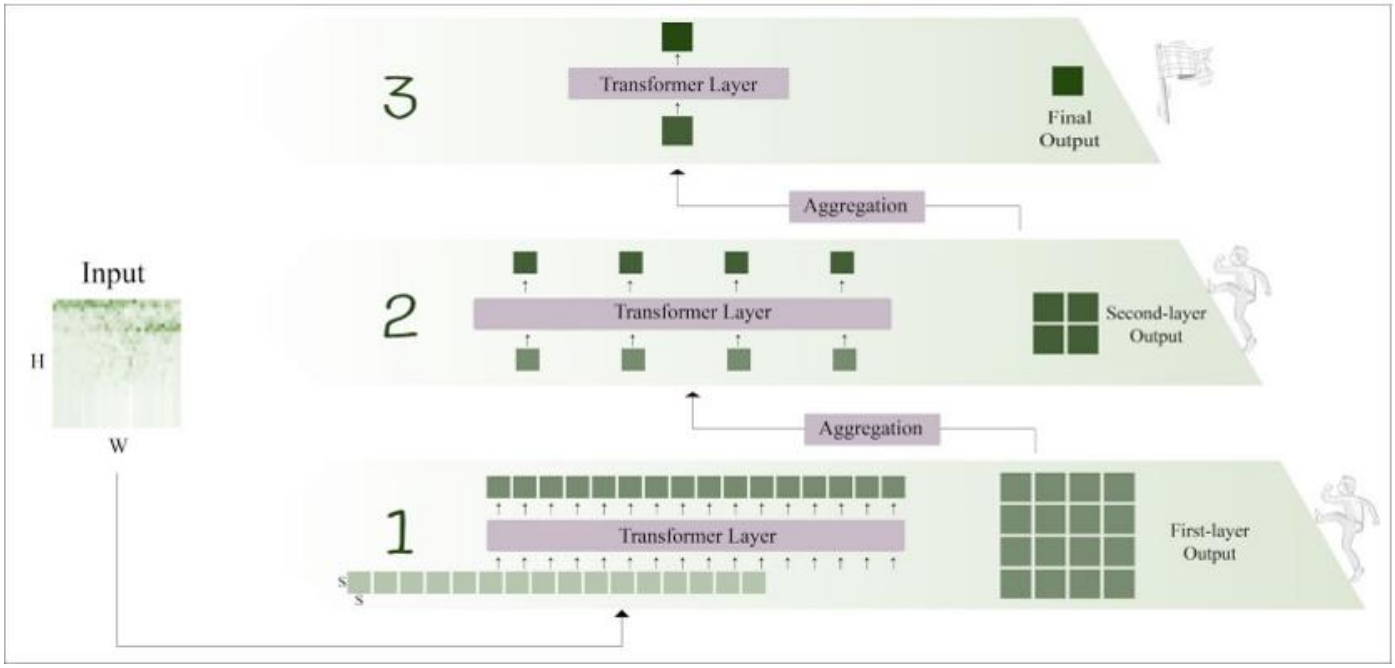

Fig. 3.3.1 Depiction of Hierarchical attention

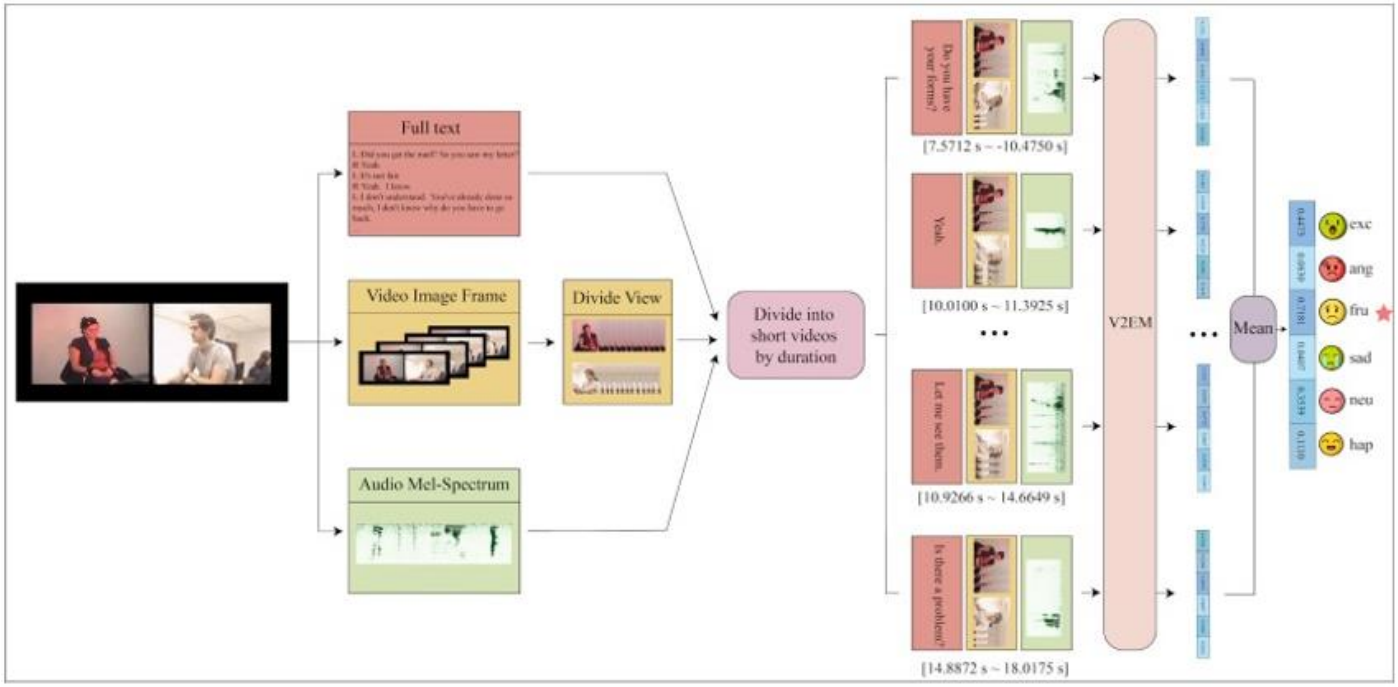

Fig. 3.3.2 The Architecture of FV2ES

3.4 Research Paper – 4

3.4.1 Paper Title

Bukhari, Nimra & Hussain, Shabir & Ayoub, Muhammad & Yu, Yang & Khan, Akmal. (2022). Deep Learning based Framework for Emotion Recognition using Facial Expression.[4]

3.4.2 Summary

The authors of this paper presented a deep learning system that identify feelings using facial expressions. The authors used a Convolutional Neural Network for feature extraction instead of traditional methods. **ResNet50**, **VGG16** and **InceptionV3** are the pre-trained models used. The results showed that **InceptionV3** was the best performer with an accuracy of **97.93%** on the **FER-2013** dataset, whereas **ResNet50** gave **92.91%** accuracy on the **CK+** dataset.

The preprocessing was done by including resizing grayscale images, to improve input quality. The model was evaluated using metrics like **precision, recall, F1-score, and confusion matrices**. The paper shows that deep learning works better than traditional methods for recognizing emotions from facial expressions.

3.4.3 Key Insights

Using deep learning models over traditional models

The authors show that using deep learning models is significantly better than traditional models.

Using pre-defined models

The authors demonstrated that using a predefined model like **VGG-16, ResNet-50, or Inception-V3** gives a **better accuracy** score than **CNN**. It is also faster as we don't need to train a model from scratch.

Model Performance Evaluation

We can adopt **precision, recall, and F1-score** metrics to evaluate inappropriate speech detection models.

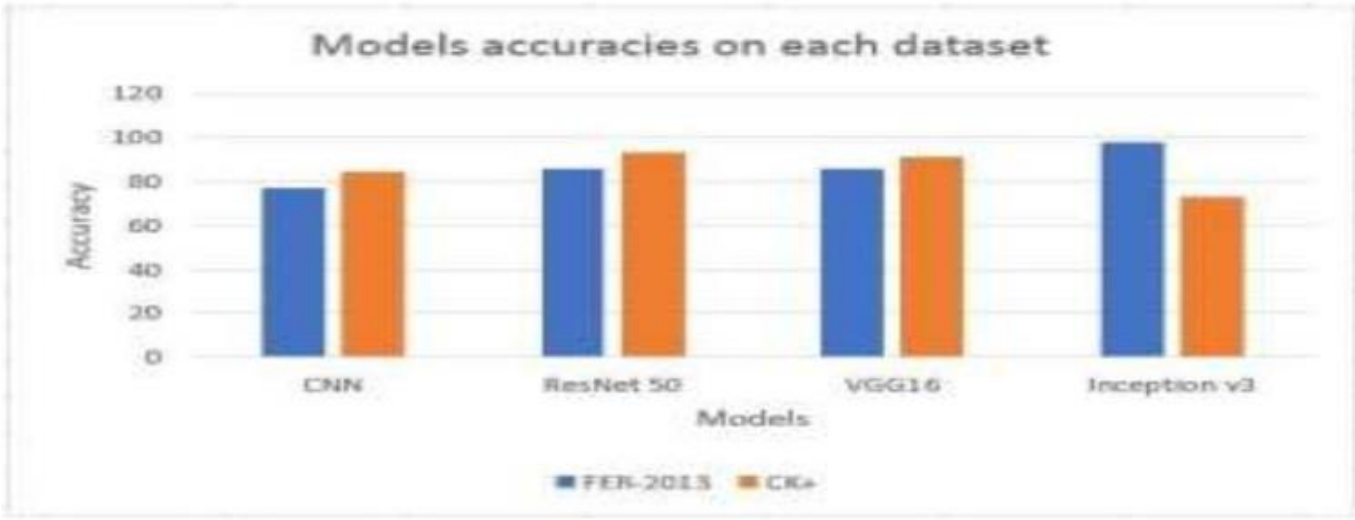

Fig. 3.4.1 Comparative accuracies of proposed models

CHAPTER 4

RESEARCH GAPS AND CHALLENGES

4.1 Research gaps

The field of inappropriate language detection in live-streaming videos is unexplored, particularly in addressing both text and speech features. We observe a limited amount of study that integrate both audio and visual modalities in real-time. Current work focuses more on individual modalities rather than a combined approach, leaving a significant gap in using their strengths. This allows space for exploration in cross-modal techniques.

Real-Time systems have not been explored, particularly in cases where input features could be corrupt or missing. Another gap lies in the interpretability and explainability of multimodal models for inappropriate language detection. This is observed mainly because current systems lack the ability to understand context effectively.

There is also a gap due to lack of publicly available datasets, this limits the ability to train and benchmark models.

4.2 Challenges

We encounter many challenges in Real-Time Inappropriate Speech Detection. The most challenging one is Real-Time Performance which leads to high latency when processing audio and visual features together. Another challenge occurs when we have to combine audio and visual features into a multimodal model, it could lead to addition of noise. This shows that Cross-modal techniques need optimization for Real-Time processing. For ensuring systems operate in real-time for live streaming scenarios, the system has to be computationally lightweight. This reduces delays while trying to maintain accuracy. Delays are important as if the delay is high then it renders our system useless for live moderation. This means that the system must work with high prediction accuracy while minimizing latency. We must ensure proper temporal alignment of audio and visual streams as it is crucial for the accurate fusion of features, and could lead to reduced performance.

CHAPTER 5

OBJECTIVES

5.1 Dataset Creation and Pre-processing Augmentation

We need to create a complete and balanced Dataset of foul and normal words to train our model. We can create our Dataset from multiple live streaming platforms such as YouTube and Twitch as sources. We need to make sure the data is varied with different augmentations such as noises, accents and other variations.

5.2 Optimisations and Comparisons

Throughout our architecture, we have many points that we need to do comparative studies and optimisations. For example, the CNN and RNN models need to be chosen from a wide variety of models, and the hyper parameters need to be tuned. The techniques for audio and video feature extraction needs to be experimented with as well. Different weighting for the cross-modal attention systems can also be tried.

5.3 Using Cross-Modal Attention

We also want our model to utilise cross-modal attention mechanisms that chooses the features that are the best and most relevant from both the modalities to increase performance as much as possible. These attended features which make use of temporal dependencies using RNNs or Transformers are then combined and sent through a Sequence Modelling Layer, Finally, the fused and refined representations are fed into a prediction layer to classify content as appropriate or inappropriate.

5.4 Creating a Safe Environment

We hope to make digital spaces safer and more inclusive by focusing on important problems related to the harmful content on live streaming platforms. One such incident is the permanent ban of popular streamer Adin Ross from Twitch for "hateful conduct" after a number of racist and antisemitic messages were sent in his unmoderated live chat, requiring proactive moderation tools now.

Another such incident is the Australia's Online Safety Amendment Bill 2024 which regulates the access of social media sites by children under the age of 16 years, and it introduces fines up to $32 million for corporations that do not implement these procedures. This shows that governments and agencies are also looking to implement measures to enhance cyber security online. Therefore, we see the emphasis on developing systems of real-time monitoring for security in young people's access online.

This system will hence not expose offensive material to all ages more so minors if platforms are enabled with effective mechanisms of detection and remedial mechanisms of inappropriate speech; thereby enhancing trust, satisfaction among users, and compliance with increasingly sophisticated safety standards.

CHAPTER 6

CONCLUSION OF PHASE – 1

In this phase of our project, Real-Time Inappropriate Speech Detection, we have been able to make tremendous advances by totally outlining the problem statement with approaches at a high-level, combining the audio and video modality, using spectrograms for audio but facial expression analysis for the video based and relying on each modality alone to improve the results. We have done broad literature review that includes foul language detection techniques, multi-modal integration methods, emotion detection research, and approaches toward audio data analysis. This has given us a very robust understanding of methodologies and challenges while trying to build a model that is reliable, accurate, and scalable. Going forward, we outlined the essential steps into the next phase of data acquisition and preprocessing, literature review expansion to incorporate actual application, determination of content types that can be targeted for consideration, search for more advanced models, cross-modality techniques, optimization, and scalability techniques. These will set us up to be ready to develop a robust and efficient system for the following phases.

CHAPTER 7

PLAN OF WORK FOR PHASE - 2

7.1 Gantt Chart and Insights

Fig. 7.1.1 Gantt Chart

The above figure is the Gantt Chart for our phase-2. For our phase-2 we would like to conduct a more comprehensive literature review, and in parallel start with data exploration. After the literature review is over, we would like to start a high-level implementation of the model.

REFERENCES

[1] A. S. B. Wazir et al., "Spectrogram-Based Classification of Spoken Foul Language Using Deep CNN," 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 2020, pp. 1-6, doi: 10.1109/MMSP48831.2020.9287133.

[2] A. S. B. Wazir, H. A. Karim, H. S. Lyn, M. F. Ahmad Fauzi, S. Mansor and M. H. Lye, "Deep Learning-Based Detection of Inappropriate Speech Content for Film Censorship," in *IEEE Access*, vol. 10, pp. 101697-101715, 2022, doi: 10.1109/ACCESS.2022.3208921.

[3] Q. Wei, X. Huang and Y. Zhang, "FV2ES: A Fully End2End Multimodal System for Fast Yet Effective Video Emotion Recognition Inference," in *IEEE Transactions on Broadcasting*, vol. 69, no. 1, pp. 10-20, March 2023, doi: 10.1109/TBC.2022.3215245.

[4] Bukhari, Nimra & Hussain, Shabir & Ayoub, Muhammad & Yu, Yang & Khan, Akmal. (2022). Deep Learning based Framework for Emotion Recognition using Facial Expression. Pakistan Journal of Engineering and Technology. 5. 51-57. 10.51846/vol5iss3pp51-57.

Samarth S K Phase1_check.pdf

ORIGINALITY REPORT

4%	2%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	digital.library.unt.edu Internet Source	1%
2	"Soft Computing and Its Engineering Applications", Springer Science and Business Media LLC, 2024 Publication	1%
3	Submitted to PES University Student Paper	1%
4	rrs.mmu.edu.my Internet Source	<1%
5	Submitted to University of Hull Student Paper	<1%
6	paperswithcode.com Internet Source	<1%
7	Shabir Hussain, Muhammad Ayoub, Ghulam Jilani, Yang Yu et al. "Aspect2Labels: A novelistic decision support system for higher educational institutions by using multi-layer topic modeling approach", Expert Systems with Applications, 2022	<1%

	Publication	
8	link.springer.com Internet Source	<1%
9	mafiadoc.com Internet Source	<1%
10	thesai.org Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		