Multimodal Fusion for Abusive Speech Detection Using Liquid Neural Networks and Convolution Neural Network

K S Paval, Vishnu Radhakrishnan, KM Krishnan, G. Jyothish Lal, B Premjith
Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
g_jyothishlal@cb.amrita.edu

***Abstract*—The recent surge in the use of social media has created vast spaces for viral content that attracts attention of large crowds. This has paved the way to the misuse of these platforms making them breading grounds for toxicity and harassment. Hence there is a need for effective abuse detection methods. In our research, we leverage the ADIMA dataset to investigate abuse detection methodologies, aiming to enhance the effectiveness of existing systems. We propose a multimodal, multilingual abuse detection system that includes three main aspects: the utilization of multimodal fusion techniques for abuse detection, the application of Liquid Neural Networks (LNN) in identifying abusive text content, the use of Convolutional Neural Networks (CNN) in identifying abusive audio utterances and the extension of multimodal fusion abuse detection to cross-lingual settings. This enabled our system to detect abuses in 10 Indian languages. Our approach takes the existing works a step further as it is able to accommodate multiple modalities and multiple languages. We use Convolutional Neural Networks to analyze sound patterns from melspectrograms and Liquid Neural Networks to process text information. To consolidate the strengths of both the representation, late fusion is applied to combine the results, resulting in an ensemble model which achieves an accuracy of 77.47%, and an AUC of 77.89% on the multilingual test set.**

***Index Terms*—Abuse-Detection, Late Fusion, Liquid Neural Networks, Abusive Speech, Convolutional Neural Networks.**

I. Introduction

Traditional Social Media has long undergone a metamorphosis. Long-form posts and meticulously curated feeds are giving way to a relentless stream of bite-sized content. With the rise of six-second loop videos from tik-tok in 2013 that captured a plethora of audience, the competition was forced to move into short form content with Instagram and Youtube launching their respective short-form content. [1]. A recent data survey projects 5.42 billion social media users by the end of 2025. [2]. While this widespread accessibility has undoubtedly democratized communication, it has also raised concerns about the ease with which hate speech and communal unrest can be propagated. The seamless dissemination of information across these platforms presents a double-edged sword, offering unprecedented reach while also amplifying the potential for negative societal consequences.

The Internet's massive user base has led to an explosion of all sorts of content. But alongside this surge of online expression, there's a darker side that's emerged - one marked by a rise in abusive language and harmful speech. Across different languages, including those spoken in India, online platforms have become hotbeds for all kinds of harmful content, from hate speech to cyberbullying and discrimination. In a diverse country like India, where people speak multiple languages, the impact of online content, especially hate speech, hits harder when it's in the language both the creator and the audience understand best. According to recent studies, content that is in line with the linguistic and cultural preferences of its target audience has a deeper and more immediate impact. Language is an effective means of communicating thoughts and emotions that are integral to both our personal and societal identities. Hate speech hits even more when it is said in the native language since it not only capitalizes on linguistic nuances but also makes connections with common cultural references. This linguistic connection creates a stronger emotional bond and strengthens existing biases or prejudices. So, the increase in hate speech in regional languages poses a big challenge for efforts to promote social harmony and fight online toxicity. It highlights the need for smart strategies that can handle the diversity of languages in online conversations [3]. In response to this pressing issue, the task of detecting and mitigating abusive speech in Indic languages has garnered significant attention, seeking to uphold the integrity of online discourse and cultivate safer digital environments.

The identification and successful mitigation of dangerous content is facilitated by abuse detection, which is a critical component in preserving the safety and integrity of online environments. Algorithms for detecting abuse have traditionally concentrated on text-based analysis, focusing on keywords and explicit language patterns that indicate abusive behavior. Although text-based methods have shown to be beneficial, they frequently fail to fully capture the intricacy and subtlety of abusive content [4]. In our research we work on a multimodal approach that uses speech and text in abuse detection. Speech as a extra modality can help pick up on vocal cues, body language, and context that text lacks, revealing the true intent behind words. This allows for better detection of sarcasm, veiled threats, and non-verbal aggression. Hence, in the proposed work, we employ a multimodal approach that consists of a Convolutional Neural Network (CNN) that handles the audio modality and a Liquid Neural Network (LNN) that handles the text modality. The Mel-spectrograms extracted from the audio files are given as input to the CNN whereas the embeddings

extracted from the text transcriptions are given to the LNN. Further, Late fusion is applied to combine the classification results of both modalities.

The main contributions of the study are listed below.

- Developed a Multilingual, Multimodal Abuse detection system, bridging the gap in hate-speech detection on low-resource Indic Languages.
- Evaluate the classification framework to cross-lingual settings to assess its capabilities across diverse linguistic context.

II. Literature Survey

Numerous studies are being conducted to address the intricate problem of identifying abuse in internet communication. These studies investigate several methods and strategies for recognising abusive content across different languages and media formats. A particular area of study is multilingual abuse detection. For instance, the study by Sharon et al. [5] suggests MADA, a technique that uses text, emotion and audio information to efficiently detect abusive activity on social media. The study also finds a link between abusive conduct and the emotions conveyed in audio. Their findings are complemented by the work of Thakran et al. [6] who introduces ACMAD, a method for detecting abusive speech directly from audio in multiple languages. Unlike prior approaches using generic pre-trained models, ACMAD focuses on extracting language and emotion features from the audio using specialized models. These features are then combined and fed into a lightweight CNN for classification. The study demonstrates that ACMAD achieves state-of-the-art performance on a multilingual abuse detection benchmark, highlighting the importance of using specific acoustic cues for this task.

While these are multilingual abuse detection, another area of interest is improving hate speech detection, particularly in resource-scarce languages. Roychowdhury et al. [7] propose EasyMixup, an input-level data augmentation technique that leverages the observation that the label of a hateful instance is preserved on concatenation with another hateful or non-hateful instance. The authors also reformulate hate speech detection as an entailment-style problem, finding this approach yields better performance than English-based entailment. Finally, they explore the relationship between explicit and implicit hate speech, and find that leveraging the correlation between the two can improve hate speech detection.

Recognizing the limitations of text-based approaches, researchers are exploring multimodal detection methods. The study by Chen et al. [8] tackles offensive language detection using a video dataset with offensive and normal content then employing a model, CLS-CNN, that transcribes videos, analyzes the text using BERT, and extracts features using CNNs. CLS-CNN achieves 88% accuracy on their data and surpasses other methods on public text datasets, demonstrating its effectiveness in this task. Similarly, Cheruvu et al. [9] utilizes YOLOv5 for object detection and Optical Character Recognition (OCR) to identify abusive content within images and videos. However, as highlighted by Kaur et al. [10], a major challenge lies in effectively handling non-textual content for abuse detection.

Some studies delve into the acoustic properties of abusive speech. Spiesberger et al. [11] Machine learning models which analyzed audio properties like pitch and loudness to identify abusive speech in recordings across ten languages. Using a mix of classifiers(Logistic Regression, XGBoost, Support Vector Machine, Random Forest) and feature sets, the models achieved moderate accuracy up to 84% in detecting abuse, even when tested on different languages than they were trained on. This suggests some universality in how abusive speech sounds across languages. While promising for combating online abuse, the method requires further development to handle subtleties like sarcasm and variations in speaker characteristics.

Deep learning techniques are showing promise in abuse detection. Kaur et al. [10] explores methods for abusive content detection in online user-generated data. The authors analyze various approaches, including machine learning with features like n-grams and sentiment analysis, lexicon-based methods using predefined offensive word lists, and rule-based techniques with manually crafted rules. While machine learning shows promise, challenges include the lack of standard datasets and the difficulty of defining and detecting nuanced abuse like sarcasm. The study highlights deep learning as an emerging powerful technique and calls for future research on areas like fine-grained abuse classification, handling non-textual content, and establishing common benchmarks for better evaluation.

III. Dataset Description

Table I: Data Description

Data Description	Value
Number of Languages	10
Total Sample	11775
Total Abusive Samples	5108
Total Non-abusive Samples	6667
Number of Unique Users	6446
Total Duration	65 hours
Average Duration	20 $\pm$ 3 seconds
Min-Max Duration	5/58 seconds

The dataset used for our reasearch is the ADIMA dataset [12]. The Abuse Detection In Multilingual Audio (ADIMA) dataset comprises 11,775 audio recordings sourced from ShareChat chatrooms. It has a total of 65 hours of recordings. The recordings are avalable for 10 Indian languages which are Hindi (Hi), Bengali (Be), Punjabi (Pu), Haryanvi (Ha), Kannada (Ka), Odia (Od), Bhojpuri (Bh), Gujarati (Gu), Tamil (Ta), and Malayalam (Ma). The dataset is balanced across these languages. Around 6446 users contribute to the dataset increasing the user diversity as well.The dataset is well balanced (43.38%) with 6,667 non-abusive and 5,108 abusive recordings. The recordings are sampled at 16kHz, mono-channel and range from 5-60 seconds with an average duration of 20 seconds. Dataset has aldready been split in 70:30 ratio as

Train:Test Split. Following table I gives a detailed description of the dataset.

IV. Methodology

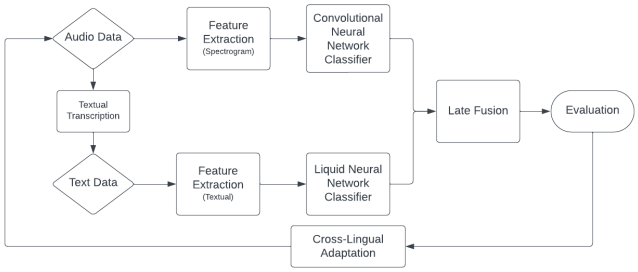

Figure 1: Methodology

The block diagram of the proposed methodology is given in Fig. 1. The description of the different stages are decribed in the following subsections.

A. Transcription from Audio Files

The Google Web Speech API, particularly the `recognize_google` function, was employed to transcribe audio data of various Indian languages from the dataset. This API facilitates the conversion of speech to text and is integrated into the Python package for speech recognition. Once the transcriptions are generated, they are stored in a structured format, in a CSV file.

B. Tokenization of text data

The tokenization process entails breaking down each transcription into individual tokens or words, thereby converting the continuous stream of text into a structured sequence of discrete elements. We employ the `preprocessing.text.Tokenizer` module, a component of the TensorFlow library, to perform tokenization efficiently. Setting constraints such as `num_words` and `max_len` allows us to control the vocabulary size and sequence length, ensuring that the tokenized representations remain manageable and computationally feasible.

Upon fitting the tokenizer on the training data (`X_train_text`), it learns the vocabulary of words present in the transcriptions and assigns a unique integer index to each word. Subsequently, the `texts_to_sequences` method converts each transcription into a sequence of integers, where each integer corresponds to the index of the respective word in the vocabulary.

C. Audio Feature Representation

This part focuses on the process of extracting melspectrograms from the audio recordings. It is a meaningful representation of audio by mapping frequencies to the Mel scale, which closely aligns with how humans perceive pitch. Here, we use the librosa package to get the melspectrograms from the audio files and further gave it as input to the Convolutional Neural Network (CNN) for classification.

D. Training an LNN model on text data

In the proposed methodology, we use Liquid Neural Network (LNNs) [13] to train the text data. Unlike tradition NN with fixed connections, LNN have dynamic connections that allows it to adapt to change in variations in time-series data. Therefore, textual representation and categorization, particularly in abuse detection and contextual understanding can benefit from this due to its dynamic architecture. By integrating convolutional and recurrent layers, it adeptly captures both local and temporal dependencies within textual data. The embedding layer initializes word vectors for text comprehension, while convolutional layers extract hierarchical features, and recurrent layers capture sequential patterns. The 'Liquidity' of this architecture allows it to learn even after training, making it adept at working with multiple languages and dialects. This architecture's synergy of convolutional and recurrent layers equips it to discern nuanced language aspects like irony and ambiguity, essential for precise abuse identification.

E. Training a CNN model on spectrogram data

Convolutional Neural Networks (CNNs) have proven to be highly effective in the classification of Mel-spectrograms due to their inherent ability to capture spatial hierarchies of features within data. Mel-spectrograms, being two-dimensional representations of audio signals, exhibit spatial correlations where adjacent spectrogram elements (frequency bins and time frames) contain valuable contextual information. CNNs are well-suited to exploit these spatial relationships. Moreover, CNNs inherently possess translational invariance, meaning that they can identify patterns irrespective of their position within the input spectrogram. This property is particularly beneficial for Mel-spectrogram classification tasks, where the position of relevant audio features may vary across different audio samples. We use a simple CNN with 3 convolutional and 3 fullycoonnected dense layers for classification.

F. Late Fusion

In our research we use late fusion to combine the results of audio and text modalities. The late fusion approach combines these representations to make a final decision regarding the presence of abusive content. [14]. In our work, we combine data from many modalities using the late weighted average fusion which computes a weighted average of the output representations from each modality after giving each representation a weight.

Weights based on relative relevance or performance are assigned to each modality prior to fusing the output representations. These weights can be ascertained via experimentation, domain expertise, or optimization methods like cross-validation. Once the weights are assigned, the output representations of each modality are multiplied by their respective weights and then averaged together. Mathematically, the fused representation F can be computed as follows:

$$F = \sum_{i=1}^{N} w_i \times O_i$$

where O_i represents the output representation of modality i, w_i represents the weight assigned to modality i, and N

G. Cross Lingual Adaptation

Once our model is trained on one language, we evaluate its cross lingual capabilities. Since Indic languages come from a wide variety of regions, abusive content vary based on linguistics, cultures and dialects. Therefore, a cross lingual adaptation is crucial in making sure this methodology can effectively identify abusive content regardless of its language. The language the model is trained on is the 'source' language and the evaluation is done on all other 'target' languages.

V. Results

This section focuses on the results of abusive speech detection. Firstly, we will discuss the results obtained from transcribing audio utterances for all languages. This will be followed by the monolingual-multimodal abuse detection performed for all ten Indian languages. Lastly, we will discuss the results of multilingual-multimodal abuse detection.

A. Transcription

Table II shows the confidence scores of the transcriptions for all the languages. The validity of these transcriptions can be measured by their confidence scores, a value between 0 and 1. These values are calculated by aggregating the likelihood values assigned to each word in the audio. A higher number indicates a higher accuracy. These confidence scored are the averaged over all audios for a particular language. Languages like Haryanvi and Bhojpuri are transcribed in the Hindi language [15] since they do not have a script and are written in Devanagiri Script [16]. The confidence scores of the transcription of Bhojpuri, Haryanvi languages are calculated as an average under the Hindi language transcription.

Table II: Transcription Confidence Scores

Language	Confidence Score
Bengali	0.8845
Gujarati	0.8678
Kannada	0.8389
Punjabi	0.9649
Tamil	0.8502
Malayalam	0.6919
Hindi	0.7898
Odia	0.8235

B. Results for Monolingual Multimodal Abuse Detection

We have experimented our proposed methodology on the individual languages, separately for abuse detection. Here, we show the results obtained for one particular language, Bengali, alone because of page restrictions. Nevertheless, we observed similar results on other languages too. Table III shows the performance of the proposed model on Bengali language using standard metrics such as accuracy, F1-score and Area under the curve (AUC). Figure 2 gives the ROC curve which is the plot between the true positive and false positive rates of abuse labels for Bengali language. Table III also gives the area under this ROC curve. From the confusion matrix shown in Figure 3 we observe low false positive and false negative predictions, which indicates the robustness of our model.

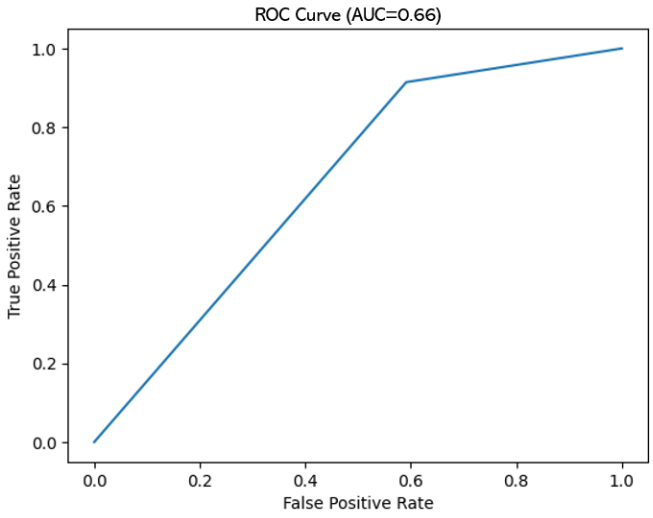

Figure 2: ROC Curve plot for Monolingual-Multimodal Model(Bengali)

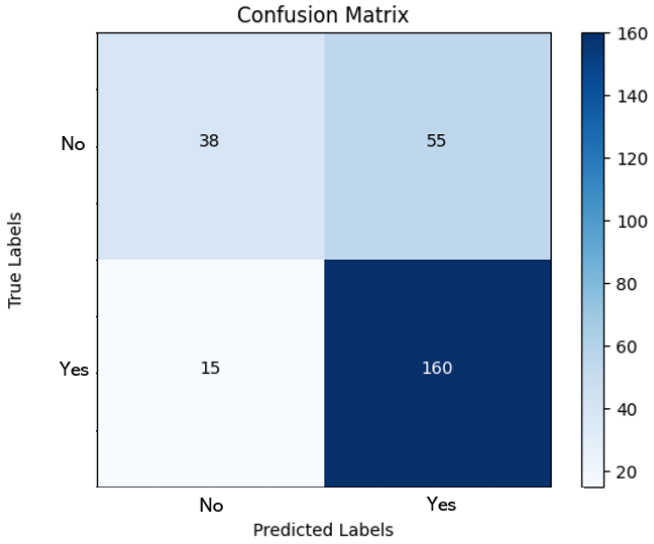

Figure 3: Confusion Matrix for Monolingual-Multimodal Model (Bengali). [Yes = Abusive, No = Not Abusive]

Table III: Evaluation Metrics for Monolingual-Multimodal Model (Bengali)

Metric	Value
Testing Accuracy	0.7389
Testing F1 Score	0.8205
Testing AUC	0.6615

C. Results for Multilingual Multimodal Abuse Detection

This section gives the result for the model that has been evaluated on multilingual settings. The model has been trained

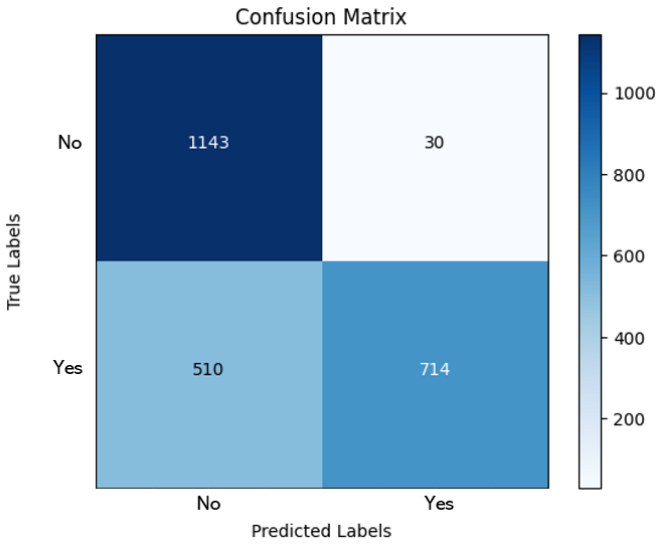

Figure 4: Confusion Matrix for Multilingual-Multimodal Model. [Yes = Abusive, No = Not Abusive]

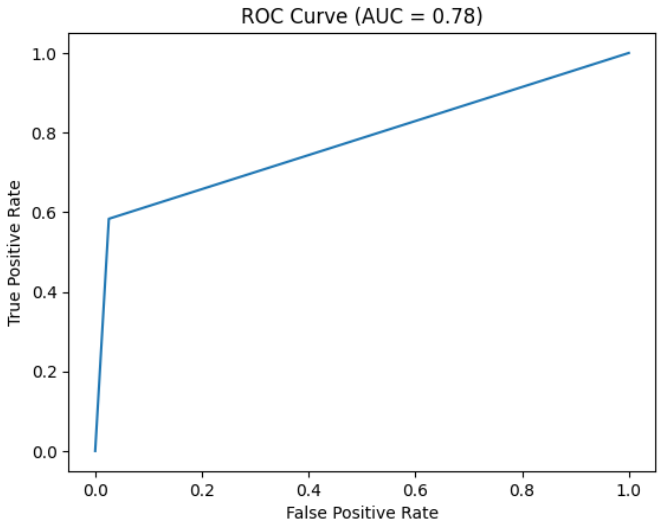

Figure 5: ROC curve for the Multilingual-Multimodal Model

on 10 indian languages. Table IV gives the performance of the proposed method in terms of standard measures such as aforesaid.

Table IV: Evaluation Metrics for Multilingual-Multimodal Model

Metric	Value
Testing Accuracy	0.7747
Testing F1 Score	0.7256
Testing AUC	0.7789

From the ROC curve, we can see that the curve is skewed to the left corner. ROC is a plot of True Positive vs False Positive and hence more left skewed means more proportion of observations were correctly classified. A similar observation can be made from the confusion matrix.

D. Results for Cross Lingual Multimodal Abuse Detection

The following results were obtained by training our model on one language and testing on all languages. This tells us about the versatility of the model in handling cross lingual scenarios. Accuracies have been provided in tabular format.

The Source Language is the Language the model was trained on and the Target Languages are the Languages that the model adapted to in a Cross Lingual manner. These results are shown in Table V and VI .

The findings of our combined cross-lingual adaptation show extraordinary consistency and robust performance over a wide range of language pairs. Our model achieves high accuracy ratings in both portions of the analysis, confirming its effectiveness in handling multilingual abuse detection tasks.

In the Part 1 results table, we find consistently high accuracy scores across language pairs, ranging from 0.6178 to 0.7333. Notably, languages such as Haryanvi, Tamil, and Odia consistently produce accuracy values more than 0.71,

Table V: Accuracy scores for multimodal Cross Lingual Adaptation (Part 1)

source/target	Bhojpuri	Gujarati	Haryanvi	Hindi	Kannada
Bhojpuri	0.6815	0.6813	0.6816	0.6814	0.6815
Gujarati	0.6906	0.6908	0.6905	0.6907	0.6906
Haryanvi	0.7104	0.7103	0.7105	0.7104	0.7105
Hindi	0.6179	0.6180	0.6178	0.6179	0.6178
Kannada	0.6585	0.6584	0.6586	0.6585	0.6586
Malayalam	0.7231	0.7230	0.7232	0.7231	0.7230
Odia	0.7288	0.7287	0.7289	0.7288	0.7289
Punjabi	0.6485	0.6486	0.6484	0.6485	0.6484
Tamil	0.7332	0.7333	0.7331	0.7332	0.7331
Bengali	0.6514	0.6513	0.6515	0.6514	0.6515

demonstrating the model's ability to detect abusive content in these languages. This constant performance demonstrates the model's adaptability and generalisation skills, which allow it to retain high accuracy levels across linguistic changes.

Table VI: Accuracy scores for multimodal Cross Lingual Adaptation (Part 2)

source/target	Malayalam	Odia	Punjabi	Tamil	Bengali
Bhojpuri	0.6817	0.6814	0.6816	0.6815	0.6816
Gujarati	0.6905	0.6907	0.6906	0.6907	0.6905
Haryanvi	0.7104	0.7105	0.7104	0.7103	0.7105
Hindi	0.6179	0.6178	0.6179	0.6178	0.6179
Kannada	0.6585	0.6586	0.6585	0.6586	0.6585
Malayalam	0.7232	0.7231	0.7232	0.7231	0.7232
Odia	0.7288	0.7289	0.7288	0.7289	0.7288
Punjabi	0.6486	0.6485	0.6486	0.6485	0.6484
Tamil	0.7333	0.7332	0.7331	0.7332	0.7333
Bengali	0.6513	0.6514	0.6515	0.6514	0.6513

Similarly, in Part 2 result table, our model maintains its high consistency and performance, with accuracy scores ranging from 0.6178 to 0.7333. Languages such as Malayalam, Tamil, and Odia once again show consistently good accuracy scores, demonstrating the model's dependability and efficiency in cross-lingual abuse detection tasks.

These results demonstrate our model's versatility and robustness in dealing with linguistic diversity and unpredictability. By consistently maintaining excellent accuracy scores over a wide range of language pairs, our model demonstrates its capacity to recognise abusive content independent of linguistic nuances or differences.

VI. Conclusion

In conclusion, our study addresses the the need for an abuse detection model in social media platforms. Using our Novel methodology we utilise the processing capacities of both Convolutional and Liquid Neural networks to learn the intricacies of different Indic languages present in the ADIMA dataset which contains 10 Indian Languages, so that it is equipped to identify abusive content. Combining CNN's image processing powers to unravel hidden patterns in melspectrogram and LNN's dynamic capacities to handle variations in text data, we have shown how using multiple modalities is beneficial in this present work. This proposed multimodal approach was tested in multilingual, monolingual and cross-lingual settings. These experimental showed better performance in monolingual settings. Further the proposed methodology showed an average accuracy of 77.47%, and an AUC of 77.89% on the multilingual configurations. Moreover cross-lingual configurations has also shown comparatively good performance in all languages considered in the study.

References

[1] Y. Qin, B. Omar, and A. Musetti, "The addiction behavior of short-form video app tiktok: The information quality and system quality perspective," *Frontiers in Psychology*, vol. 13, p. 932805, 2022.

[2] Demand Sage, "Social media users statistics 2021 - the ultimate list of statistics." https://www.demandsage.com/social-media-users/, 2021. Accessed: April 28, 2024.

[3] F. Murtadho, "The effects of using mother tongue in delivering health protocol messages on health attitudes and behaviors: Do gender, age, and education level make any difference?," *Indonesian Journal of Applied Linguistics*, vol. 12, no. 2, pp. 348–360, 2022.

[4] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in *Proceedings of the 10th ACM conference on web science*, pp. 105–114, 2019.

[5] R. Sharon, H. Shah, D. Mukherjee, and V. Gupta, "Multilingual and multimodal abuse detection," *arXiv preprint arXiv:2204.02263*, 2022.

[6] Y. Thakran and V. Abrol, "Investigating acoustic cues for multilingual abuse detection,"

[7] S. Roychowdhury and V. Gupta, "Data-efficient methods for improving hate speech detection," in *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 125–132, 2023.

[8] X. Chen, X. Ye, M. Mohanty, and S. Manoharan, "Detecting offensive posts on social media," in *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pp. 1–6, IEEE, 2023.

[9] S. M. Cheruvu, D. Sesank, K. S. Sandilya, and M. Gupta, "Vituperative content detection: A multidomain architecture using opencv," in *International Conference on Advances and Applications of Artificial Intelligence and Machine Learning*, pp. 409–421, Springer, 2022.

[10] S. Kaur, S. Singh, and S. Kaushal, "Abusive content detection in online user-generated data: a survey," *Procedia Computer Science*, vol. 189, pp. 274–281, 2021.

[11] A. A. Spiesberger, A. Triantafyllopoulos, I. Tsangko, and B. W. Schuller, "Abusive speech detection in indic languages using acoustic features,"

[12] V. Gupta, R. Sharon, R. Sawhney, and D. Mukherjee, "Adima: Abuse detection in multilingual audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6172–6176, IEEE, 2022.

[13] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu, "Liquid time-constant networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 7657–7666, 2021.

[14] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective techniques for multimodal data fusion: A comparative analysis," *Sensors*, vol. 23, no. 5, p. 2381, 2023.

[15] D. of Linguistics, "About hindi," 2024. Accessed: 2024-05-15.

[16] R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Offline recognition of devanagari script: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 782–796, 2011.