tinyML for Crime Prevention: Detecting Violent Conversations

Amna Anwar & Prof. Eiman Kanjo
amna.anwar@ntu.ac.uk
Smart Sensing Lab, Department of Computer Science, Nottingham Trent University, United Kingdom

Introduction

Our work focuses on using tinyML for detecting violent language on edge devices, such as mobile phones and wearables, in the context of preventing domestic violence. Our multimodal algorithm uses natural language processing (NLP) and audio processing to detect violent conversation from short audio-text segments.

The algorithm is converted to a tinyML model using TensorFlow Lite, which enables the detection to run on edge devices such as mobile phones and smart home sensors. Our mobile application enables near-real-time detection of violent conversations to enable victims or potential victims to detect and report crimes to close or trust contacts.

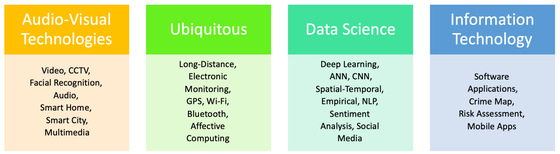

Figure 1 – A classification of technologies for crime prevention.

Objectives

1. To develop and determine the most effective method of inferring violent and aggressive content from audio recordings using a binary classification model.
2. To investigate the results of natural language processing models, and determine the most accurate combination to be used in the research project.
3. To develop a data fusion text model that uses binary classification to accurately identify violent or aggressive language from transcribed conversations.
4. To effectively use data fusion to combine the previously identified audio and natural language processing models, which can classify results at a high accuracy.
5. To extract a lite version of the data fusion model to implement on smart home and mobile edge devices to detect violent or aggressive language without high computational power.
6. To explore potential further applications and research relating to the data fusion audio text model, and highlight any future opportunities.

Methodology

Figure 2 shows our overall framework for the proposed fusion model. Multimodal information integration is achieved by the concatenation of features and embeddings from:

1. BERT [1]
2. Bi-LSTM applied to LIWC data (e.g., [2])
3. Audio Time Domain
4. MFCC after applying CNN

This is then fed into three-layer Fully Connected (FC) networks, followed by a Softmax layer which assesses the type of label. The audio and text concatenation captures short-term, as well as long-term acoustic and linguistic characteristics to detect violence level in conversations.

The framework of the proposed method is shown in Figure 2 below:

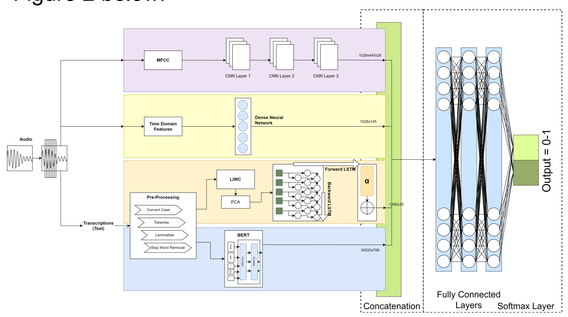

Figure 2 - Diagram of the framework for the proposed fusion model, presenting the audio segmentation and data processing techniques proposed. The extracted features are then merged and used to determine the probability of violent language.

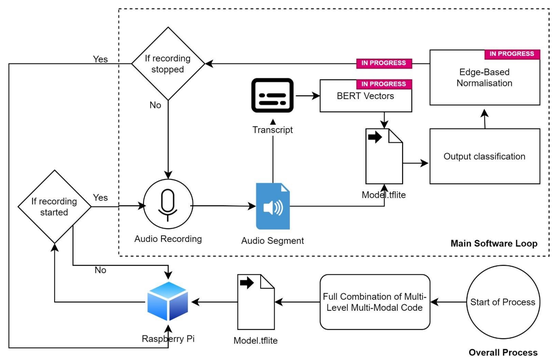

Figure 3 – The process of exporting the model and overall edge system

Results

- NLP features: Utilise Bidirectional Encoder Representations from Transformers (BERT) and Linguistic Inquiry and Word Count (LIWC) features.
- Fusion algorithm: Concatenate NLP features through a Convolutional Neural Network (CNN).
- Audio elements: Incorporate Mel-Frequency Cepstral Coefficient (MFCC) features and time domain features for audio-violent conversation processing.
- Model performance: Achieved effective results of 0.87, indicating the potential usefulness of the system.
- Applications: Detect potential violent conversations in domestic violence contexts and monitor public spaces for suspicious language.
- Model optimization: Compressed the model to TensorFlow Lite and deployed it on edge devices using tinyML.
- Implementation: Utilised TensorFlow Lite for deployment on mobile phones, simulated smart home devices, and wearable watches.

Model	F1 Score
Baseline Model (Random Forest - all features)	0.7469
MFCC + Time Domain + BERT	0.7790
MFCC + Time Domain + LIWC	0.6934
MFCC + Time Domain + LIWC + BERT	**0.8454**

Figure 4 - Overall model performance of the multi-modal features.

Figure 5 – Raspberry pi showing the transcription.

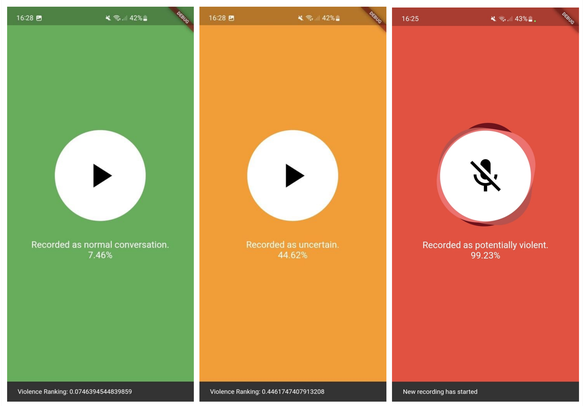

Figure 6 – Screenshots of application displaying various outputs of the violent language detection.

Discussion

- Future work focuses on co-designing and developing innovative edge devices (wearable or fixed) using multi-modal sensing.
- Data collection for violence detection models will be conducted in real-world, free-living scenarios instead of controlled settings.
- The edge devices will incorporate embedded electronics for intervention and be informed by close-to-market research.
- Information fusion will be enabled through Edge Computing (EC), embedded sensors, wireless communications, data fusion, and deep learning.
- Utilising text modalities in addition to audio features has shown improved results for violence detection.
- Future experiments will explore labelling specific types of violence and text categories to provide more contextual assessment of violence.

Conclusion

In this poster, a multilevel multimodal fusion approach was proposed for the detection of violent language in conversations. The proposed system uses data fusion to combine MFCC acoustic audio features and text-based features that have been processed using BERT and LIWC. The resulting algorithm enables a multimodal classification of violent language in both audio and text form to be performed. The early results suggest that the combination of these features is possible and that the proposed fusion model can effectively extract violent language information from audio and text features with an F1 score = 0.85. The research will now investigate opportunities related to applying this system to edge devices such as wearable or mobile technologies and investigate the experimental implementation and evaluation in edge scenarios.

References

1. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova,"Bert: Pre-training of deep bidirectional. Transformers for language understanding," 2018.

2. A. Bisht, A. Singh, H. S. Bhadauria, J. Virmani, and Kriti, Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. Singapore: Springer Singapore, 2020, pp. 243–264.

3. A. Alrehili, "Automatic hate speech detection on social media: A brief survey," in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), 2019, pp. 1–6.