Profanity Detector

Lidia Gomez

Motivation and Literature Research

The theoretical motivation for this project derives from the existing propensity to use profanity in daily language. While the use of profanity amongst adults is arguably not directly harmful, our global culture maintains societal standards for communication and decency, specifically in business settings. Moreover, the FCC maintains federal guidelines for obscenity, indecency, and profanity that inform our speech in the public sector[1].

The practical inception of the idea behind the creation of a Profanity Regulator came from a conversation with a friend who practices corporate litigation. In the conversation, said friend expressed her own proclivity to use profanity in regular conversation, and the fact that the regular usage of such language is becoming increasingly difficult to control in professional environments.

Although there is currently no precedent for a wearable device that measures the use of profane words in the wearer's daily speech, Amazon Halo employs machine learning to discern "tone" and detect the emotional affect of a user's speech.

Project Categories

Artificial Intelligence Application
General Event Detection Application
Audio Classification
Wearable Device Using Voice Over Wi-Fi

Description

The Profanity Detector is an application - deployed on a wearable device - that can correctly recognize when profane words are said in continuous language, and determine the word that is said. An end-product might be a wearable device that uses haptics to vibrate whenever the user says something profane. The MVP is a model that can correctly recognize when two specific profane words are said ('fuck' and 'bitch'), and distinguish between the words with more than 90% accuracy in real-time.

A neural network was trained on an extensive dataset of profane words to recognize when those words are said in regular speech. This model was tested and validated, and deployed to the embedded device (ST B-L475E-IOT01A) using the EdgeImpulse platform.

[1] https://www.fcc.gov/consumers/guides/obscene-indecent-and-profane-broadcasts

Objectives

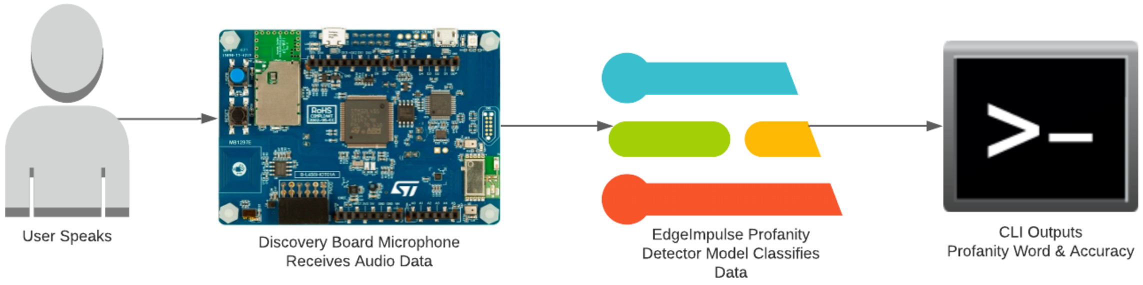
The main objective of the Profanity Regulator is to train a neural network on enough profanity audio data so that it may recognize those words in regular speech. The ultimate objective of this project is to deploy a system on a wearable device that captures continuous speech and is able to enumerate the amount of times certain profane words are spoken.
The objective of the MVP is to train a model to recognize two distinct profane words and to distinguish between the two with more than 90% accuracy. It must be noted that there are real-world limitations to this project as it stands, including hardware and software limitations.

Equipment

(Hardware)
ST B-L475E-IOT01A Discovery IOT Node
This particular board was chosen because it has an audio sensor, as well as wifi and bluetooth modules, and is compatible with Edge Impulse.

(Software)
For the MVP, EdgeImpulse[2] was used to acquire/upload data securely to build datasets, design ML algorithms for classification, test and validate the ML model with real-time data, build an optimized embedded inference and deploy it to the device.

System Architecture


User Speaks
Discovery Board Microphone Receives Audio Data
EdgeImpulse Profanity Detector Model Classifies Data
CLI Outputs Profanity Word & Accuracy

Implementation

Data

For proof of concept, the acquisition of data was restricted to two words: ‘fuck’ and ‘bitch’. Data was collected from two sources: the abuse project, which is an open dataset consisting of various audio snippets of an extensive list of profane words, and my own dataset. Because a more robust open-source dataset of spoken profanity does not exist, I employed the help of various volunteers to build a more varied dataset of naturally spoken profanity. The data

[2] https://docs.edgeimpulse.com/docs

collected amounted to a total of 4 minutes and 31 seconds - although this may not seem like much, it is important to note that most of the audio snippets in the dataset are about 1 second long, so this was sufficient to train the model. The 313 audio samples yielded very distinct features, which made for a promising result.

Model

The model was designed using MFCC to extract features from audio signals, as it is better than MFE for human voices, and Keras is used for classification, which works great for audio recognition. The neural network architecture included the input layer, a reshape layer (13 columns), a 1D Convolutional/Pooling layer (8 neurons, 3 kernel size, 1 layer), a dropout layer (rate 0.25), another 1D Convolutional/Pooling layer (16 neurons, 3 kernel size, 1 layer), a second dropout layer (rate 0.25), a flatten layer, and the output layer (2 classes).
The model was trained for 500 epochs, and yielded a 96.8% training accuracy.

Testing the Model

The model was tested using both the test data and the live classification functionality of EdgeImpulse. Testing yielded an accuracy of 97.37%, which is well above the proposed accuracy to meet the initial MVP goal. While there were a few misclassified samples for both words, the accuracy result is good enough to deploy the model on the discovery board and complete live testing. The WiFi capability of the hardware was used to project the results to the command line. During live testing, the words ‘fuck’ and ‘bitch’ were said five times in different intonations. Live testing yielded 100% accuracy on almost every iteration of testing.

Proof of Concept or MVP

The MVP is a model that can correctly recognize when two specific profane words are said (‘fuck’ and ‘bitch’), and distinguish between the words with more than 95% accuracy in real-time.

Final Project Deliverable

The final project deliverable includes: Final GitHub Public Repository that contains the final report, slides, live demo video, code, and the data.

https://github.com/LidiaGomez/profanity-detector

Conclusions

Ultimately, the Minimum Viable Product performed great - but it is not without limitations.
The current model correctly distinguishes between two profane words. This is only a microcosmic example of the capabilities of using Keras for audio classification.

Although the use of EdgeImpulse provided a complete platform for the training and deployment of this model, it is limited in its capabilities. Furthermore, the use of EdgeImpulse limits the ability of the hardware that may be used.

Future Plans

The aforementioned limitations can be mitigated by foregoing EdgeImpulse and programming the NN using Python, Librosa Sound Processing Library, Tensorflow, and Keras.

The sound data is processed using Librosa & FFMPEG, transformed from .wav audio files into MFCCs, which are then organized as numpy arrays with which to train the model.
This would provide more flexibility and control in the architecture of the model's layers, and deployment would no longer be dependent on specific devices, enhancing portability. In the future, cloud services would be leveraged to provide more robust continuous audio processing, recognition, and storage, and the resulting data would be displayed on a mobile application.