Audio Profanity Detection and Reduction Using MLP

[1]Selvaraj A, [2]Sharukesh B, [3]Trinity E, [4]Yogeshwaran P

[1]AP (Sr. Gr.)/IT, [2,3,4]Student,
[1]Department of Information Technology,
[1]K.L.N College of Engineering, Sivagangai, Tamil Nadu, India
[2,3,4]Department of Information Technology,
[2,3,4]K.L.N College of Engineering, Sivagangai, Tamil Nadu, India

Abstract: The aim of this paper is to review machine learning algorithms and techniques for hate speech detection in given audio file. Hate speech problem is normally modeled as an audio classification task. In input audio signal data, there have been improvements in ML algorithms that were employed for hate speech detection over time. In this process, the model is going to find or detect speech recognition such as hate words, offensive words, profanity words and standard words. Then they are going to be the class variable of our dataset classes. In this approach, machine learning algorithms will be implemented to analyze the process to find the accuracy, precision, and f1-score.

Index Terms: Machine learning, Speech detection, Audio classification, Profanity.

I. INTRODUCTION:

Filtering audio and video content has become a social concern with the high exposure of many young adults to portable and immediate screen time sources. There is a risk a person could be exposed to substantial amounts of offensive and foul language incorporated within entertainment videos and movies displayed on online platforms and broadcasting channels. For example, all movies contain foul words, and have kept increasing through the years, while foul language is known to bring a negative effect on society. Furthermore, broadcasting companies and media-sharing platforms have been held accountable to provide appropriate content through censorship tools. Censorship is a complex phenomenon in filtering and providing language content worthy of viewers due to the constraints in personnel, cost, time, and human fatigue that could lead to the misdetection of unwanted content. The proposed study aimed to construct an astute, competent, and mechanized censorship system for identifying undesirable spoken terms (profane language) within audio signals, such as stand-alone audio files and signals assimilated in prominent and accessible video-sharing websites and broadcasting media. In this regard, neural networks facilitated audio censorship in videos (movies and entertainment shows) and reported intriguing characteristics of the techniques. Speech recognition through deep learning has recently gained popularity. Specifically, speech identification systems operated by identifying various utterance types (spontaneous and continuous speeches and connected and isolated words). This study served to build an astute censorship system to precisely identify unfavorable speech content, given that literature on intelligent speech identification models using deep learning has solely emphasized inoffensive language identification. For example, recent studies have used conversational and read speech datasets that are clean of foul language utterances such as LibriSpeech.

II. RELATED WORKS:

Excessive content of profanity in audio and video files has proven to shape one's character and behavior. Currently, conventional methods of manual detection and censorship are being used. Manual censorship method is time consuming and prone to misdetection of foul language. This paper proposed an intelligent model for foul language censorship through automated and robust detection by deep convolutional neural networks (CNNs). A dataset of foul language was collected and processed for the computation of audio spectrogram images that serve as an input to evaluate the classification of foul language. The proposed model was first tested for 2-class (Foul vs Normal) classification problem, the foul class is then further decomposed into a 10-class classification problem for exact detection of profanity. Experimental results show the viability of the proposed system by demonstrating high performance of curse words classification with 1.24-2.71 Error Rate (ER) for 2-class and 5.49-8.30 F1- score. Proposed ResNet50 architecture outperforms other models in terms of accuracy, sensitivity, specificity, F1-score. [1] Attention-based encoder-decoder architectures such as Listen, Attend, and Spell (LAS), subsume the acoustic, pronunciation and language model components of a traditional automatic speech recognition (ASR) system into a single neural network. In our previous work, we have shown that such architectures are comparable to state-of-the-art ASR systems on dictation tasks, but it was not clear if such architectures would be practical for more challenging tasks such as voice search. In this work, we explore a variety of structural and optimization improvements to our LAS model which significantly improve performance. On the structural side, we show that word piece models can be used instead of graphemes. We introduce a novel multi-head attention architecture, which offers improvements over the commonly used single-head attention. On the optimization side, we explore techniques such as synchronous training, scheduled sampling, label smoothing, and applying minimum word error rate optimization, which are all shown to improve accuracy. We present results with a unidirectional LSTM encoder for streaming recognition. On a 12,500~hour voice search task, we find that the proposed changes improve the WER of the LAS system from 9.2% to 5.8%, which corresponds to a 13% relative improvement over the best conventional system which achieves 6.7% WER. [2] Describes an audio dataset of spoken words designed to help train and evaluate keyword spotting systems. Discusses why this task is an interesting challenge, and why it requires a specialized dataset that is different from conventional datasets used for automatic speech recognition of full sentences. Suggests a methodology for reproducible and comparable accuracy metrics for this task. Describes how the data was collected and verified, what it contains, previous versions and properties. Concludes by reporting baseline results of models trained on this dataset. [4] We show that an end-to-end deep learning approach can be used to recognize either English or Mandarin Chinese speech--two vastly

different languages. Because it replaces entire pipelines of hand-engineered components with neural networks, end-to-end learning allows us to handle a diverse variety of speech including noisy environments, accents, and different languages. Key to our approach is our application of HPC techniques, resulting in a 7x speedup over our previous system. Because of this efficiency, experiments that previously took weeks now run in days. This enables us to iterate more quickly to identify superior architectures and algorithms. As a result, in several cases, our system is competitive with the transcription of human workers when benchmarked on standard datasets. Finally, using a technique called Batch Dispatch with GPUs in the data center, we show that our system can be inexpensively deployed in an online setting, delivering low latency when serving users at scale. [5] Recent studies have demonstrated the potential of unsupervised feature learning for sound classification. In this paper we further explore the application of the spherical k-means algorithm for feature learning from audio signals, here in the domain of urban sound classification. Spherical k-means is a relatively simple technique that has recently been shown to be competitive with other more complex and time-consuming approaches. We study how different parts of the processing pipeline influence performance, taking into account the specificities of the urban sonic environment. We evaluate our approach on the largest public dataset of urban sound sources available for research and compare it to a baseline system based on MFCCs. We show that feature learning can outperform the baseline approach by configuring it to capture the temporal dynamics of urban sources. The results are complemented with error analysis and some proposals for future research. [7] Humans can merge information from multiple perceptional modalities and formulate a coherent representation of the world. Our thesis is that robots need to do the same to operate robustly and autonomously in an unstructured environment. It has also been shown in several fields that multiple sources of information can complement each other, overcoming the limitations of a single perceptual modality. Hence, in this paper we introduce a data set of actions that includes both visual data (RGB-D video and 6DOF object pose estimation) and acoustic data. We also propose a method for recognizing and segmenting actions from continuous audiovisual data. The proposed method is employed for extensive evaluation of the descriptive power of the two modalities, and we discuss how they can be used jointly to infer a coherent interpretation of the recorded action. [8] Recognition of isolated spoken digits is the core procedure for a large and important number of applications in telephone-based services, such as dialing, airline reservation, bank transaction and price quotation, only using speech. Spoken digit recognition is a challenging task since the signals last for a brief period and often some digits are acoustically remarkably like each other. The objective of this paper is to investigate the use of machine learning algorithms for digit recognition. We focus on the recognition of digits spoken in Portuguese. [9] We propose the use of the line spectral frequency (LSF) features for emotion recognition from speech, which have not been previously employed for emotion recognition to the best of our knowledge. Spectral features such as Mel-scaled cepstral coefficients have already been successfully used for the parameterization of speech signals for emotion recognition. The LSF features also offer a spectral representation for speech, moreover they carry intrinsic information on the formant structure as well, which are related to the emotional state of the speaker. [10]

III. METHODOLOGY:

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. Data points are clustered based on feature similarity.

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. Decision Tree Simple to understand and to interpret. Trees can be visualized. Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values. The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It is easy to implement and understand but has a major drawback of becoming significantly slower as the size of that data in use grows. MLP: hidden_layer_sizes: it is a tuple where each element represents one layer, and its value represents the number of neurons on each hidden layer. learning_rate_init: It is used to control the step-size in updating the weights. activation: Activation function for the hidden layer. Examples, identity, logistic, tanh, and relu. By default, relu is used as an activation function. random_state: It defines the random number for weights and bias initialization. verbose: It used to print progress messages to standard output.

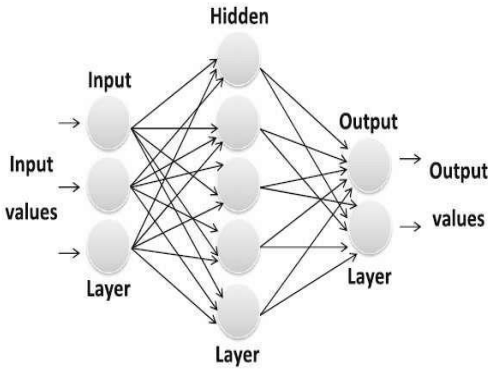

Figure 1: Working methodology of MLP

IV. IMPLEMENTATION AND ANALYSIS:

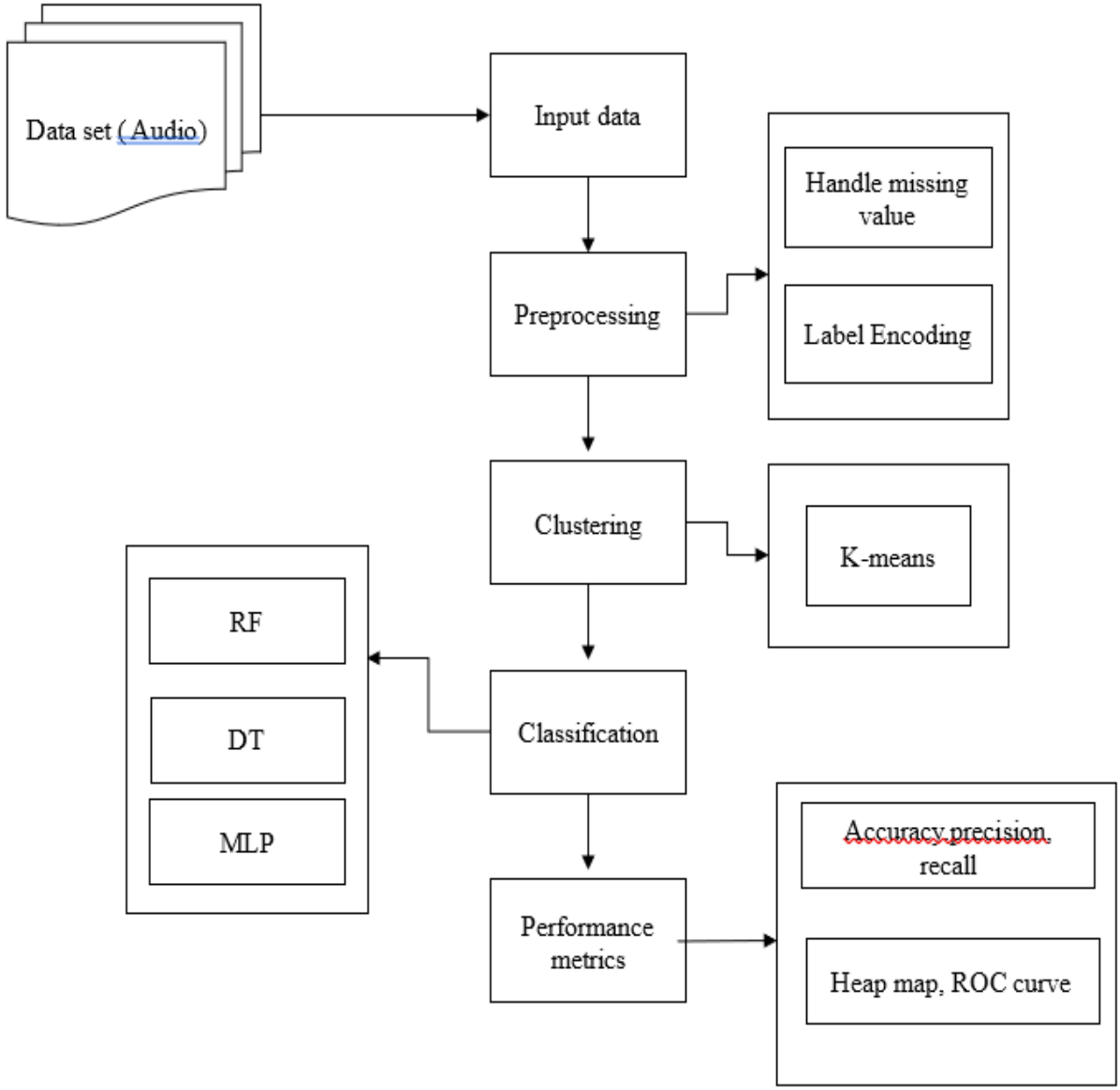

Figure 2: Software architecture.

MODULE:

- Data Preprocessing
- MLP Model Building
- Audio Censoring
- Model Evaluation

MODULE DESCRIPTION:

DATA PREPROCESSING:

An audio file is collected from the user in form of .mp3 format. Then the .mp3 file is converted to .wav file format. Then the file and dataset are split into trainset and test set. In straightforward way two pre-processing are done, they are,

- Noise removal
- Audio to text Conversion on data

MLP MODEL BUILDING:

An MLP (Multilayer Perceptron) algorithm model is built. MLP networks are used for supervised learning format. It can be used to solve complex non-linear problems. It handles substantial amounts of input data well. The model conducts an accuracy test and gives a false positive rate as the output. The model checks the performance parameters.

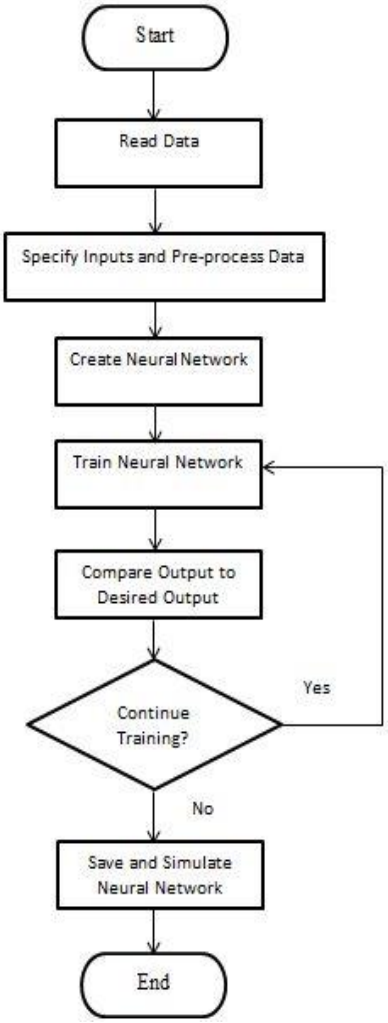

Figure 3: MLP model building flow chart.

AUDIO CENSORING:

The MLP algorithm changes the .wav file format into text format. The text format is evaluated with the dataset. When the model finds the profanity words in the text format, reduction of that word is done. Then the text format is again turned back to audio format. The output audio given is in the form of machine voice audio.

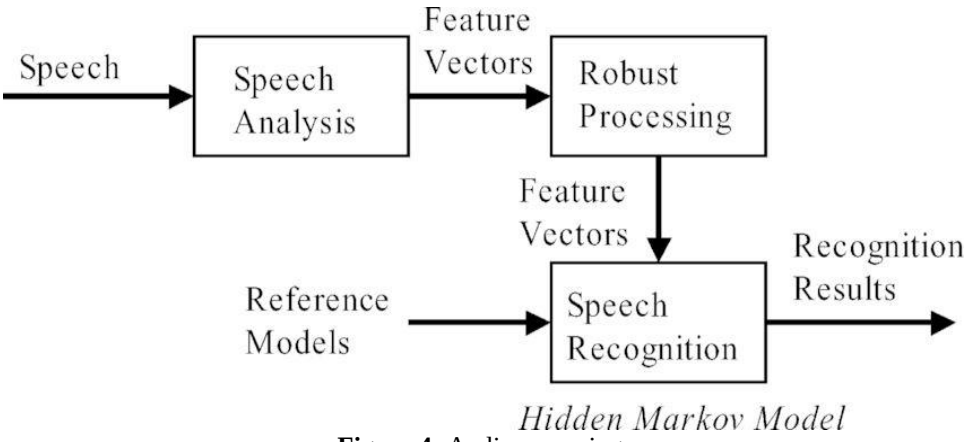

Figure 4: Audio censoring.

MODEL EVALUATION:

We use F1-Score, ROC, Confusion Matrix methods to evaluate our model. F1 score is a machine learning evaluation metric that measures a model's accuracy. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.

EQUATIONS:

Accuracy - AC= (TP+TN)/ (TP+TN+FP+FN)
Precision - Precision=TP/ (TP+FP)
Recall - Recall=TP/ (TP+FN)

V. Result and Conclusion:

The result of this concerned project is obtained via the web app by the process done by the MLP. Then a naïve user enters our web app, he uploads the audio file and within seconds he can download the censored audio. When the audio is received from the user, it is converted into a text file first, and then the classification is done via the MLP algorithm to remove the profanity, then the classified text is obtained, it is converted into an audio file using speech recognition. The interaction between the web app and MLP model and other business logic is done via flask. The F1 score of the MLP model is 1.91% and False Negative Rate is 1.57%.

VI. References:

1. Wazir, A.S.B.; Karim, H.A.; Abdullah, M.H.L.; Mansor, S.; AlDahoul, N.; Fauzi, M.F.A.; See, J. Spectrogram-based classification of spoken foul language using deep CNN. In Proceedings of the 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 21–24 September 2020; pp. 1–6.
2. Chiu, C.-C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-art speech recognition with sequence-to-sequence models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778.
3. Day, S. Cursing negatively affects society. In The Baker Orange; Baker University Media: Baldwin City, KS, USA, 2018.
4. Warden, P. Speech commands: A dataset for limited-vocabulary speech recognition. arXiv 2018, arXiv:1804.03209.
5. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; PMLR: New York, NY, USA, 2016; Volume 48, pp. 173–182. 5. Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakl, P.; Bengio, Y.
6. Han, K.; He, Y.; Bagchi, D.; Fosler-lussier, E.; Wang, D. Deep Neural Network Based Spectral Feature Mapping for Robust Speech Recognition. In Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech), Dresden, Germany, 6–10 September 2015; ISCA: Dresden, Germany, 2015; pp. 2484–2488.
7. Salamon, J.; Bello, J.P. Unsupervised feature learning for urban sound classification. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, Australia, 19–24 April 2015; pp. 171–175.
8. Pieropan, A.; Salvi, G.; Pauwels, K.; Kjellström, H.; Salvi, G. Audio-visual classification and detection of human manipulation actions. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 3045–3052.
9. Silva, D.F.; Souza, V.M.; Batista, G.E.A.P.A.; Giusti, R. Spoken digit recognition in Portuguese using line spectral frequencies. In Computer Vision; Springer Nature: Berlin/Heidelberg, Germany, 2012; Volume 7637, pp. 241–250.
10. Bozkurt, E.; Erzin, E.; Erdem, C.E.; Erdem, T. Use of line spectral frequencies for emotion recognition from speech. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 3708–3711.