# Audio-Based Hate Speech Detection for the Metaverse using CNN

Robin Matthew Medina, Judith Nkechinyere Njoku, and Dong-Seong Kim
IT Convergence Engineering, Kumoh National Institute of Technology, Korea

***Abstract*—Since the metaverse is like a virtual representation of the real world, bullying is bound to occur in many different ways. One such way is in the use of hate speech and offensive words. As a result, there is a need to detect, recognize, and censor these words. Motivated by this, we created a system that uses convolutional neural networks (CNN) to detect hate speech in the metaverse. The results show that the CNN model achieved a test accuracy of $96.875$ and a latency of $32$ms.**

***Index Terms*—metaverse, hate speech, speech recognition, CNN, audio speech**

## I. Introduction

The word ”meta,” which implies transcending, and the suffix ”verse,” are combined to form the term ”metaverse.” It describes a digitally recreated setting that is integrated with the real world. [1]. The Metaverse has tremendous potential for the future, but with its new worlds, ideas, and experiences—which elevate what Second Life, Roblox, and Minecraft have to offer—come problems like cyberbullying, privacy concerns, harassment, and other issues. The distinction between reality and virtual reality is said to be susceptible to blurring by the metaverse by many people. The sensory experience is enhanced by a constant, all-encompassing digital reality, which in turn intensifies the encounters with bullying, abuse, cyberbullying, and hate speech [2]. Since the metaverse is like a virtual representation of the real world, bullying may often happen in many different ways. Such as hate speech and offensive words, where it can be used as a medium to bully someone, there is a need to censor or filter these words. And since there are also a lot of papers regarding keyword spotting that revolve around text recognition as their datasets, we proposed using audio datasets in this paper, which can greatly help in training the model to be used. Since CNN has been successfully applied in so many speech related tasks [3], [4], and music information retrieval [5], we propose using the CNN algorithm to train a model to recognize audio datasets with high accuracy, which can then be used to detect hate speech within the metaverse.

## II. Methodology

The system model of the proposed hate speech detection is presented in Fig 1. After collecting the data from the metaverse audio or the gathered datasets, it is then converted into a spectrogram. The model will then train the converted data to find out the accuracy of the model. This paper focused on the detection of hate speech. It can only detect cursed words and words used in bullying.

### A. Dataset description and preprocessing

For this paper, we gathered hate speech audio data from audio recordings for now but will proceed with collecting the data within the metaverse. The dataset consists of a total of $533$ audio recordings. The dataset was split into a ratio of $80\% : 20\%$ for training and testing data respectively. Table. I shows the statistics of the entire dataset used in this paper.

TABLE I: Statistics of Hate Speech Data

| Label | Amount of Speech Data |
|---|---|
| Bastard | 127 |
| Damn | 134 |
| Rubbish | 131 |
| Shit | 141 |

### B. Feature extraction

The Short Time Fourier Transform (STFT) is a special flavor of the Fourier transform where we can see how the frequencies of the signal change through time. In this work, we convert the raw waveform of the hate speech data to STFTs, which slices up the signal into many small segments and takes the Fourier transform of each of these segments. A sample rate of $44,100$ was used for this task. The extracted STFT features are then forwarded to the CNN model for onward classification.

### C. CNN Model

After feature extraction, the CNN model [6] was trained to recognize the different hate speech from the datasets it was fed with. As shown in Fig. 2,the CNN model is composed of two 2D convolutional layers, followed by pooling layers, dropout layer with a rate of $0.5$, a flatten layer, a dense layer, another dropout layer of same rate and finally a dense layer.

## III. Performance Evaluation and Results

Fig.4 shows the training performance of the model in terms of accuracy. It was observed that as the model is being trained both the target and the predicted output converges closely until the end of $40$ epochs.

For the sake of simplicity and clarity, Fig.5 only shows the plot of the loss performance while training. The losses do not converge until the entire $40$ epochs and there is a considerable disparity between the training and validation loss. It takes $5.42$ seconds to train the model. The only machine learning model that been used is CNN, and based on the confusion matrix on Fig.6 its performance on the trained datasets have produced a

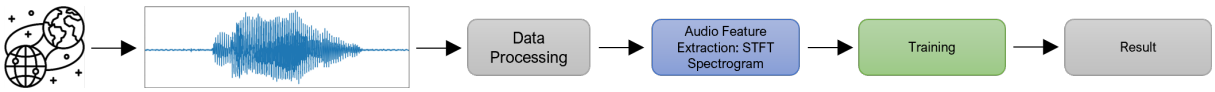


Fig. 1: Proposed System Model on Detecting Hate Speech


Fig. 2: Seven-layer network architecture used for audio hate speech recognition.

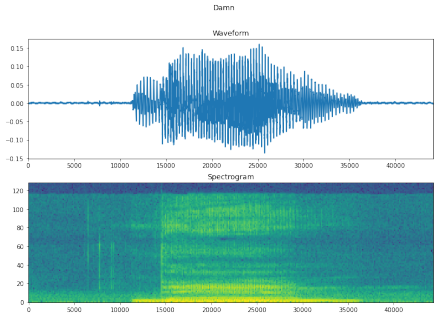


Fig. 3: Audio Data being converted to Spectrogram

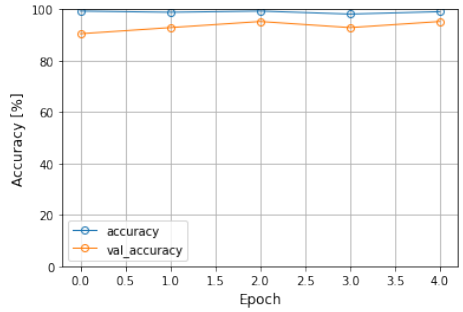


Fig. 4: Accuracy Plot of the CNN Model

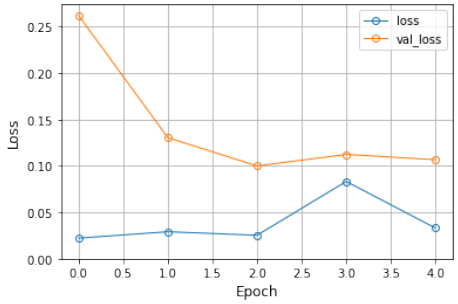


Fig. 5: Loss Plot of the CNN Model

high accuracy on a total of 54 test recordings. The Test Time it takes to train the model is 32 ms while having a Test Accuracy of 96.875.

## IV. Conclusions

This paper explored the detection of specific hate speech words said by users since the metaverse is prone to involving a lot of spoken speech and has a tendency to enable bullying

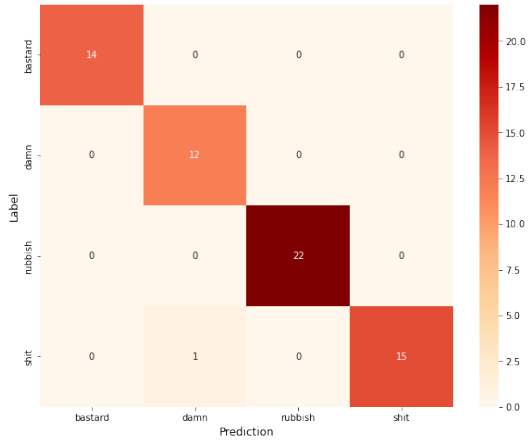


Fig. 6: Confusion Matrix using CNN Model

with the use of hate speech. To that end, a CNN model was trained on a small dataset of spoken hate speech. In the future, we will consider the censoring of such speeches and their implementation in a metaverse environment.

## Acknowledgments

This research was supported by MSIT, Korea, under the Grand Information Technology Research Center support program(IITP-2022-2020-0-01612) supervised by the IITP and Priority Research Centers Program through the NRF funded by the MEST (2018R1A6A1A03024003).

## References

[1] J. N. Njoku, C. I. Nwakanma, G. C. Amaizu, and D.-S. Kim, "Prospects and challenges of metaverse application in data-driven intelligent transportation systems," *IET Intelligent Transport Systems*, vol. n/a, no. n/a. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/itr2.12252

[2] Arti, "Cyberbullying is slowly moving from the internet to metaverse." [Online]. Available: https://www.analyticsinsight.net/cyberbullying-is-slowly-moving-from-the-internet-to-metaverse/

[3] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8614–8618.

[4] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[5] E. J. Humphrey and J. P. Bello, "Rethinking automatic chord recognition with convolutional neural networks," in *2012 11th International Conference on Machine Learning and Applications*, vol. 2, 2012, pp. 357–362.

[6] Y. LeCun, "Generalization and network design strategies," in *Connectionism in Perspective*, R. Pfeifer, Z. Schreter, F. Fogelman, and L. Steels, Eds. Zurich, Switzerland: Elsevier, 1989, an extended version was published as a technical report of the University of Toronto.