Enhanced Speech Emotion Recognition based on Convolution Neural Networks using MLP Classifier and KNN

Dr. K. Sundravadivelu[1], Dr. M. Sornalakshmi[2], Mrs. M. Lakshmi Priya[3], Ms. G. Kahvya[4]

[1]*Assistant Professor, Department of Computer Science, Madurai Kamaraj University, Madurai, Tamil Nadu, India.*
svadiveluk2021@gmail.com
[2]*Assistant Professor, PG Department of Computer Science, Arulmigu Kalasalingam College of Arts and Science, Krishnankoil, Tamil Nadu India.*
sorna.jesus@gmail.com
[3]*Assistant Professor, Department of Computer Science, Bangalore City College, Bangalore, Karnataka, India.*
srilakshmipriya11.1990@gmail.com
[4] *PG Student, Department of Computer Science, Madurai Kamaraj University, Madurai, Tamil Nadu, India.*
kahvyaganesan2002@gmail.com

KEYWORDS
CNN, Neural Network, Emotion Recognition, RAVDESS, EmoDB, etc.

ABSTRACT
The paper aims to recognize emotions from speech using CNNs, leveraging well-known databases such as EmoDB and RAVDESS. It proposes a CNN-based architecture for emotion recognition. The proposed model is trained and evaluated in both speaker-dependent and speaker-independent settings. Speaker-dependent means the model is trained and tested on the same speaker's data, while speaker-independent means testing is done on speakers not seen during training. Techniques such as data augmentation and advanced pre-processing are included to enhance model performance. The results indicate that the proposed CNN architecture achieves comparable performance with state-of-the-art methods. It surpasses traditional systems in terms of accuracy or other relevant metrics. The paper contributes by proposing a CNN-based approach for speech emotion recognition using MLP classifier, which is evaluated on established databases and shows promising results compared to existing methods. It also explores various aspects of dataset properties, speech signal analysis, and classifier methods relevant to emotion recognition tasksdocuments in various Scopus indexed journals was examined bibliometrically, for the years 2014-2024.

1. Introduction

Discourse is just the most well-known technique for conveying as individuals. It is simply considered normal by then simply regular then to reach out this correspondence medium to PC applications. We portray discourse feeling acknowledgment (SER) as a collection of frameworks that technique and arrange discourse signs to identify the implanted feelings. In basic words, it is the demonstration of endeavouring to perceive human inclination and full of feeling states from discourse. This is the framework that will fundamentally go after the way that voice as often as possible reflects stowed away sentiments through tone and pitch. SER is extreme since feelings are abstract and it is trying to clarify sound. By utilizing this framework, we can recognize the human inclination like miserable, bright, quiet, furious, blissful, unfortunate, lament, and so on by their discourse or voice or we can say utilizing some sound [1-5].

In human machine interface application, feeling acknowledgment from the discourse signal has been an examination paper subject since numerous years. To recognize the feelings from the discourse signal, numerous frameworks have been created. In this work discourse feeling acknowledgment in view of the past advances which involves various classifiers for the

feeling acknowledgment is checked on. Discourse has a few trademark highlights, for example, effortlessness and productivity, which makes it an appealing point of interaction medium. Communicating feelings and mentalities through speech is conceivable.

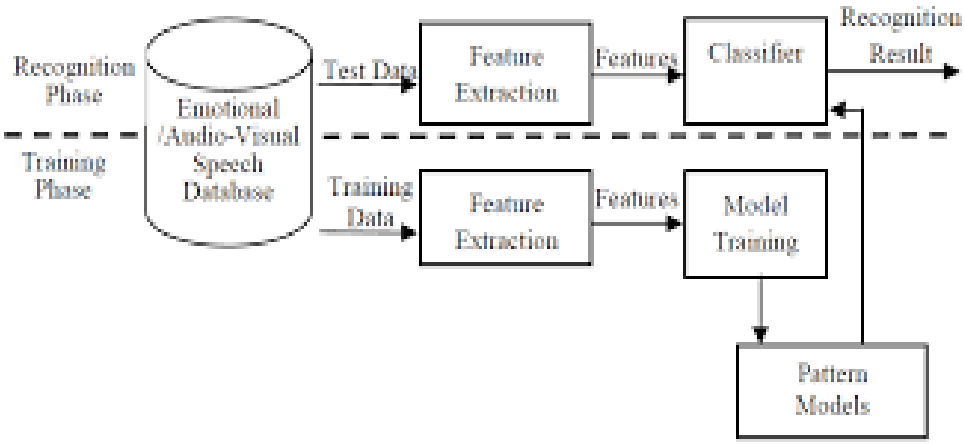


Fig1. Overall Proposed work

Here in this paper, concentrate on has been helped on a mission to perceive human feeling through discourse utilizing the MFCCs. To perceive feeling through discourse different discourse highlights were separated. In light of these discourse highlights Grouping of the feelings has been finished and the order execution of Convolution brain organization is examined. Here feeling acknowledgment is finished for various feelings like nonpartisan, cheerful, miserable, fatigue outrage, and dread. The characterization execution depends on extricated highlights. Deduction about the presentation and restriction of discourse feeling acknowledgment frameworks in view of the various classifiers are likewise examined [5-15]. Mel-recurrence cepstrum coefficients (MFCC) and tweak unearthly (MS) highlights are removed from the discourse flags and used to prepare various classifiers. Fig 1. Portrayed generally speaking proposed work.

2. Proposed Work

The thought behind making this examination paper was to fabricate an AI model that could recognize feelings from the discourse we have with one another constantly. These days personalization is something required in everything we experience consistently. Distinguishing feelings is one of the main advertising techniques in this day and age. You could customize various things for an individual explicitly to suit their advantage. Therefore, we concluded to do an exploration paper where we could recognize an individual's feelings just by their voice which will allow us to oversee numerous computer based intelligence related applications. A few models could be including call focuses to play music when one is furious on the call. Another could be a savvy vehicle dialing back when one is irate or unfortunate. Thus this sort of utilization has a lot of potential on the planet that would help organizations and furthermore even wellbeing to customers.

This work uses the libraries librosa, sound record, and sklearn to build a model .We would utilize MFCCs to be our feedback highlight. This will have the choice to see sentiments from sound archives. By then, we'll instate the CNN model with Keras and develop it with 7 layers — 6 Conv1D layers followed by a thick layer and train the model.

1. This examination paper requires some data on focuses like Python, sklearn, librosa, etc.

2. Python is not difficult to learn and chip away at with the language. It is a raised level, extensively valuable programming and significantly interfered with language.

3. Librosa is a Python library for looking at sound and music. It has a commendation group design, organized connection points, and names, backward similitude, estimated limits, and significant/decipherable code.

4. Highlight be noted, "Developers can investigate various roads viewing the programming language as shown by their solace level and information" and can change the as of late referred to language as per them.

5. CNN: Utilizing Convolutional Brain Organization to perceive feeling from the sound recording.

2.1 The proposed research work aimed to:

Create and Test new effective element extraction strategies for a programmed acknowledgment and order of pressure and feeling in discourse. Decide whether the new laryngological review showing a nonlinear person of discourse creation can be utilized to determine effective component extraction techniques for a programmed acknowledgment and grouping of pressure and feeling in discourse.

Package Required

glob2==0.6, ipython==7.4.0p, keras==2.3.1, librosa==0.6.3, matplotlib==3.0.2, numpy==1.16.2, pandas==0.25.1, plotly==4.6.0, scipy==1.2.1, seaborn==0.9.0, sklearn==0.20.1, tensorflow==1.13.1, tqdm==4.29.0

Data Description:

Data Used:

The first website contains speech data which is available in three different formats.
1. Audio Visual – Video with speech
2. Speech – Audio only
3. Visual – Video only

Data Set:

The Ryerson General media Data set of Profound Discourse and Tune (RAVDESS)
Entertainers and Entertainers recorded discourse and melody forms separately. Feeling Repugnance, Unbiased and Shocked are excluded from the melody form information. We went with the Sound just compress document since we are managing tracking down feelings from discourse. [7] The compress record comprised of around 1500 sound documents which were in wav design. The subsequent site contains around 500 sound addresses from four unique entertainers with various feelings. The following stage includes putting together the sound documents. Every sound document has an exceptional identifier at the sixth place of the record

name which can be utilized to decide the feeling the sound document comprises of. We have 5 distinct feelings in our dataset.

1. Calm 2. Happy 3. Sad 4. Angry 5. Fearful

We involved Librosa library in Python to process and concentrate highlights from the sound records. Librosa is a python bundle for music and sound investigation. It gives the structure blocks important to make music data recovery frameworks. Utilizing the librosa library we had the option to remove highlights i.e MFCC (Mel Recurrence Cepstral Coefficient). MFCCs are an element generally utilized in programmed discourse and speaker acknowledgment. We additionally isolated out the females and male's voices by utilizing the identifiers gave in the site. This was on the grounds that as an examination we figured out those isolating male and female voices expanded by 15%. It very well may be on the grounds that the pitch of the voice was influencing the outcomes.

Every sound record gave us many highlights which were essentially a variety of many qualities. These highlights were then affixed by the names which we made in the past step.

The subsequent stage included managing the missing highlights for some sound documents which were more limited long. We expanded the inspecting rate by two times to get the one of a kind element of each profound discourse. We didn't expand the inspecting recurrence considerably moresinceit could gather clamor in this way influencing the outcomes.

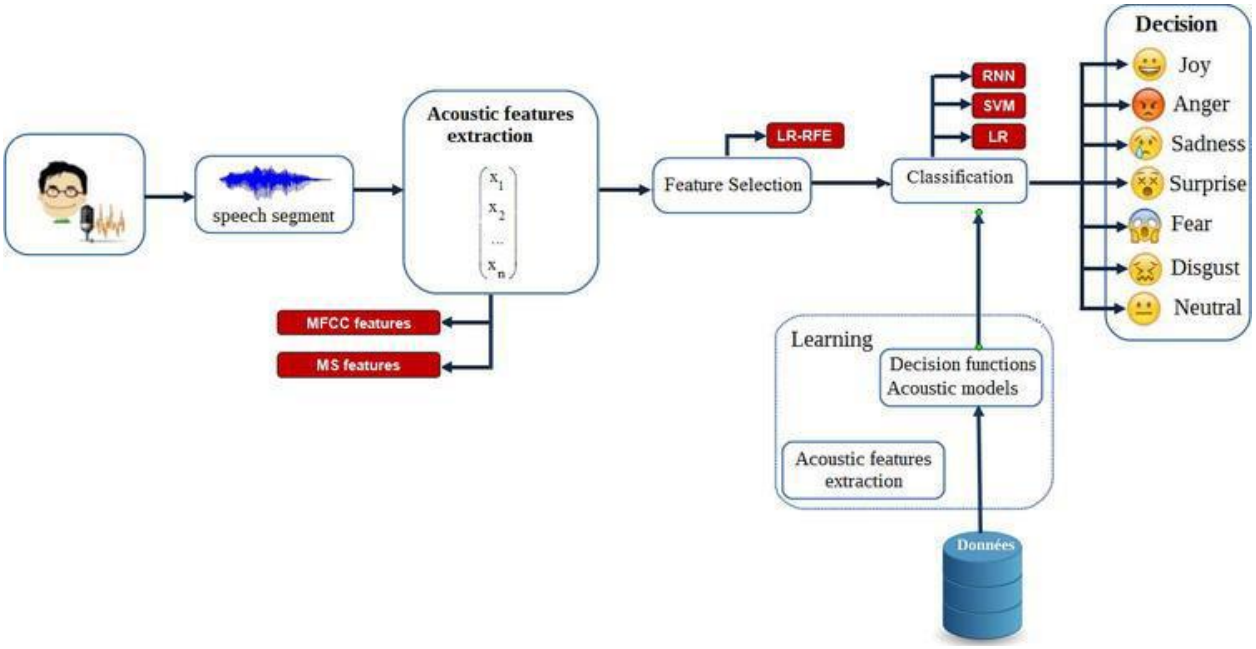


Fig.2 Testing out live Voices

3. Results and Discussion

In this Python research paper; we figured out how to perceive feelings from discourse. We involved a MLPClassifier for this and utilized the sound record library to peruse the sound document, and the librosa library to remove highlights from it. As you'll see, the model conveyed an exactness of

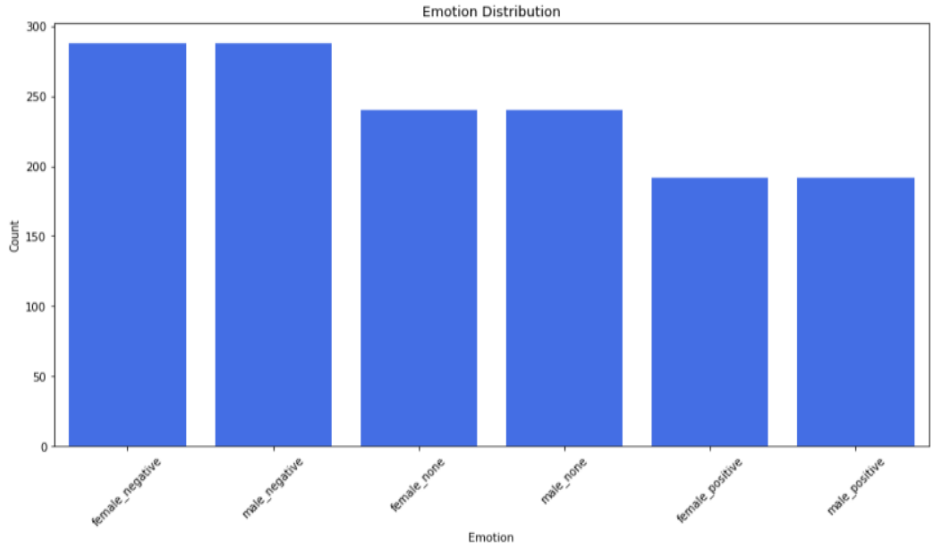
75.21%.That's adequate as far as we're concerned at this point. Precision Score: 75.21%.Model has an exactness of 75.21%, which is viewed as a decent one portrayed in fig 3.

jupyter Speech_Emotion_Recognition Last Checkpoint: 12/21/2020 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [25]:
```
a = data_df.label.value_counts()
plot_emotion_dist(a, "#2962FF", "Emotion Distribution")
```



jupyter Speech_Emotion_Recognition Last Checkpoint: 12/21/2020 (autosaved)

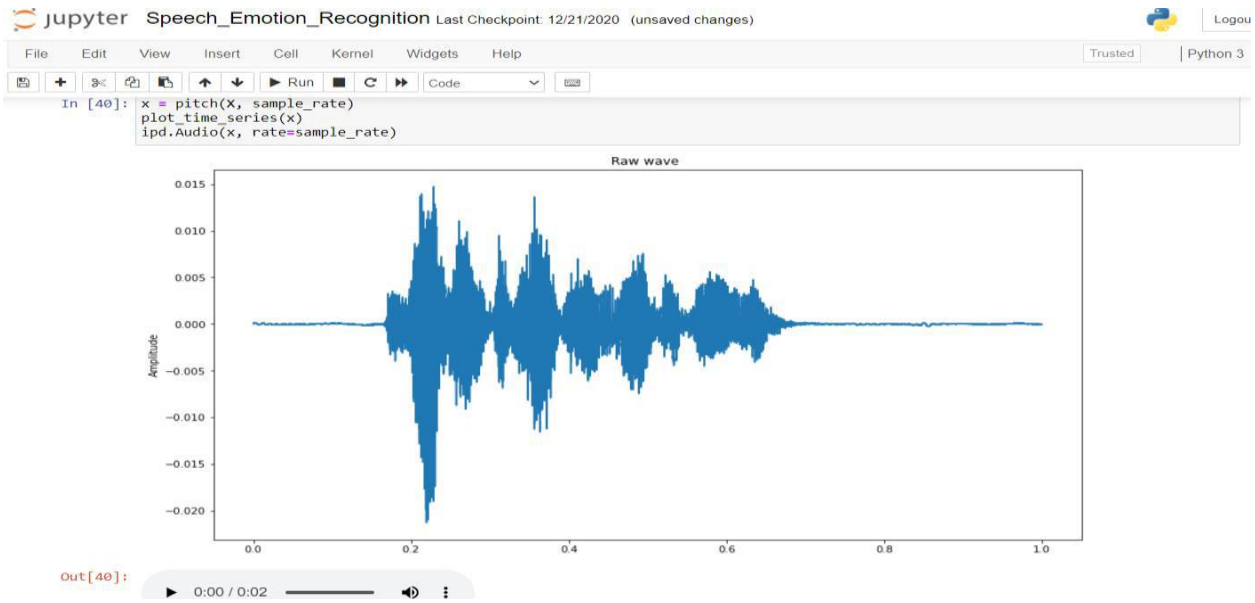
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

VIII. Changing dimension for CNN model

In [57]:
```
x_traincnn = np.expand_dims(X_train, axis=2)
x_testcnn = np.expand_dims(X_test, axis=2)
```

In [58]:
```
# Set up Keras util functions

from keras import backend as K

def precision(y_true, y_pred):
    true_positives = K.sum(K.round(K.clip(y_true * y_pred, 0, 1)))
    predicted_positives = K.sum(K.round(K.clip(y_pred, 0, 1)))
    precision = true_positives / (predicted_positives + K.epsilon())
    return precision


def recall(y_true, y_pred):
    true_positives = K.sum(K.round(K.clip(y_true * y_pred, 0, 1)))
    possible_positives = K.sum(K.round(K.clip(y_true, 0, 1)))
    recall = true_positives / (possible_positives + K.epsilon())
    return recall


def fscore(y_true, y_pred):
    if K.sum(K.round(K.clip(y_true, 0, 1))) == 0:
        return 0

    p = precision(y_true, y_pred)
    r = recall(y_true, y_pred)
    f_score = 2 * (p * r) / (p + r + K.epsilon())
    return f_score
```



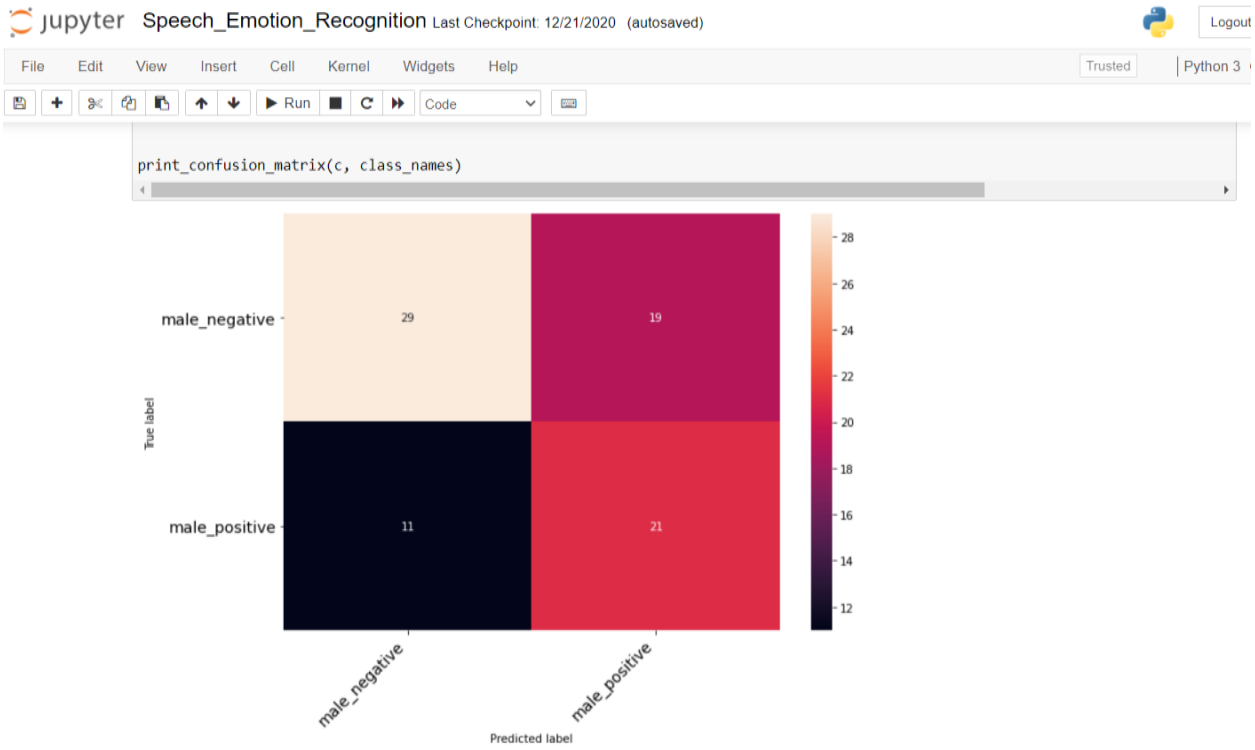


Fig.3 Model has an accuracy of 75.21%, which is considered a good one.

4. Conclusion

Eventually, have opportunity and energy to explore different avenues regarding the male informational collection. I re-split the information with defined mix split to ensure there is no information lopsidedness or information spillage issue. I tuned the model by exploring different avenues regarding the male dataset since I need to work on the model toward the start. I likewise tried with various objective name arrangements and expansion techniques. I figured out Commotion Adding and Moving for the imbalanced information could help in accomplishing an improved outcome. Building the model was a difficult undertaking as it included a great deal of experimentation techniques, tuning and so on. The model is very thoroughly prepared to recognize male and female voices and it recognizes with 100 percent precision. The model was tuned to distinguish feelings with over 75.21% exactness. Precision can be expanded by including more sound records for preparing. Future extension: Just chose the initial 3 seconds to be the info information since it would diminish the aspect, the first scratch pad utilized 2.5 sec as it were. I might want to utilize the full length of the sound to do the analysis. Pre-process the information like trimming quiet voice standardize the length by zero cushioning, and so forth. Analyze the Intermittent Brain Organization approach on this point. There are various ways that this exploration paper could be expanded. Maybe quite possibly of the most widely recognized apparatus in feeling identification frameworks is the brain organizations and secret markov model for programmed acknowledgment. Additionally the joining of different modalities, for example, video based or manual communication will be researched further.

REFERENCE

1. Ahmed, T.U., Hossain, M.S., Alam, M.J., Andersson, K.: An integrated CNN-RNN framework to assess road crack. In: 2019 22nd International Conference on Computer and Information Technology (ICCIT), pp. 1–6. IEEE (2019)Google Scholar
2. Alharbi, S.T., Hossain, M.S., Monrat, A.A.: A belief rule based expert system to assess autism under uncertainty. In: Proceedings of the World Congress on Engineering and Computer Science, vol. 1 (2015)Google Scholar
3. Aloufi, R., Haddadi, H., Boyle, D.: Emotionless: privacy-preserving speech analysis for voice assistants. arXiv preprint arXiv:1908.03632 (2019).
4. K. Sundravadivelu, Suraj Rajesh Karpe, Harish V Mekali, Shital Nalgirkar, K. Abdul Rasak, Dr. V S Narayana Tinnalur, Information Theory and Coding: Techniques for Error Control and Data Compression, Journal of Electrical Systems 20-10s (2024): 5665-5674.
5. Deepak Sharma, Kamatchi Sundravadivelu, Juhi Khengar, Dhaval J Thaker, Sulay N Patel, Priya Shah, Advancements in Natural Language Processing: Enhancing Machine Understanding of Human Language in Conversational AI Systems, Journal of Computational Analysis and Applications VOL. 33, NO. 6, 2024.
6. Sundravadivelu, K., Thangaraj, M., & Gnanambal, S. (2022). An extensive work on comparing sentiment patterns in twitter archives between two persons. International Journal of Health Sciences, 6(S7), 5170-5180. https://doi.org/10.53730/ijhs.v6nS7.13104.
7. Dr. Kamatchi Sundravadivelu,Mr. S. Muthukumar,Ms. A. Sirin Vifakga, Aditi Chaudhary,Dr. Shabnam Gulati, A Systematic Literature Review On Speech Emotion

Recognition Approaches Using Different Methodologies, Library Progress International Vol.44 No. 3, Jul-Dec 2024: P. 11487-11496.

8. K. Sundravadivelu, M. Thangaraj, "A Novel Approach for Discovering the Patterns by using PDBD Model in Big Data", Journal of Computer Science, (Science Publications), Volume 18 Issues 5, DOI: 10.3844 / jcssp.2022.382.395, ISSN: 1549-3636, pp.382-395, May 2022.
9. K. Sundravadivelu, M. Thangaraj, "Analyzing Educational Tweets using LDA Model", International Journal of Intelligent Systems and Applications in Engineering (IJISAE), Volume 10, Issue 4, ISSN:2147-6799, PP. 100- 104, Dec. 2022.
10. Thangaraj. M., & Sundaravadivelu, K.," Mining effective patterns from text data-a survey" International Journal of Scientific & Technology Research. ISSN-10: 2277-86161930 IJSTR, 2020.
11. S. Firoz, S. Raji, and A. Babu, "Discrete wavelet transforms and artificial neural networks for speech emotion recogni tion," International Journal of Computer :eory and Engi neering, vol. 2, no. 3, pp. 319–322, 2010.
12. S. Livingstone and F. Russo, "'e Ryerson audio-visual da tabase of emotional speech and song (RAVDESS): a dynamic," Multimodal Set of Facial and Vocal Expressions in North American English, vol. 13, 2018.
13. U. Raghu Vamsi, B. Yuvraj Chowdhary, M. Harshitha, S. Ravi 'eja, and J. Divya Udayan, "Speech emotion recognition(ser) using multilayer perceptron and deep learning techniques," IEEE Access, 2021.
14. M. Xu, F. Zhang, and W. Zhang, "Head fusion: improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," IEEE Access, vol. 9, pp. 74539–74549, 2021.
15. H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, and G. Dai, "Pre trained deep convolution neural network model with atten tion for speech emotion recognition," Frontiers in Physiology, vol. 12, Article ID 643202, 2021.
16. Z. Zhao, "Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition," Neural Networks, vol. 141, pp. 52–60, 2021.