Comparison on Sentiment Analysis on User Speech Specific Data Using Different Algorithms

Yadnesh Deshpande, *IT, WCE,Sangli,,India*
yadnesh.deshpande@walchandsangli.ac.in
Sumeet Chavan, *IT, WCE,Sangli, India*
sumeet.chavan@walchandsangli.ac.in ,
Mousmi Suryawanshi *IT, WCE,Sangli, India*
mousmi.suryawanshi@walchandsangli.ac.in ,
Pankaj Korke I*T, WCE,Sangli, India*
pankaj.korke@walchandsangli.ac.in ,
Amol Dongarwar *IT, WCE,Sangli, ,India*
amol.dongarwar@walchandsangli.ac.in ,
Rushikesh Pedge *IT, WCE,Sangli, India*
rushikesh.pedge@walchandsangli.ac.in

***Abstract*-Sentiment evaluation is the study of the human's feelings in any verbal exchange on subjects or preferences. It can be used in many areas, it can be a valuable tool, allowing companies as well as call centres to monitor the interaction that occurs between customers and company representatives. By identifying and correcting situations that are driving negative customer attitudes that improve overall customer satisfaction.There are two commonly used approaches to solve this problem- Acoustic approach and Linguistic approach. In acoustic approach the focus is on audio characteristics like pitch, tone, intensity, etc. In the linguistic approach we need to convert audio to text and build models to determine emotion. Our paper is only based on Acoustic approach. The technique attended within the paper investigates tests and techniques to hold out audio sentiment analysis on audio recordings with the usage of speaker & speech quality. Further, sentiment analysis is finished on the speaker precise speech facts which enables the models to recognize what the human beings were speaking about and the way they sense.**

***Index Terms*- Sentiment analysis, Natural Language Processing, Data Analytics, Speech data**

I. INTRODUCTION

Sentiment analysis can be termed as a way of using models to investigate not so organised data in a useful piece of work which can be used in improving various features for our product .Almost if we encountered today, we can expect around 80 percent of unorganised data, which is a big issue in today's world. However, it is very tough to look at, apprehend, and understand these records and the method is highly-priced and time-consuming. A lot of junk, in fact, is produced every day that includes e- mails messages, Whatsapp or Instagram chats , reviews, surveys and different online documents or materials. Sentiment evaluation, then again, enables in making sense of all these formless facts by robotically processing and classifying it into diverse emotional classes. In a world full of technology of networks, people have enormous data to train.Wherever we go we can find tons of sites, content with data, features, for example, in doing analysis of online products. A lot of people rely on online help for any situation. These social communications help people build a good and better perspective of their requirements. Social media is giving as well as consuming a lot of this data that people are in need from news to memes. Social platforms moreover do ads and take polls of their product or service suggestions to generate data. This allows them to make a great choice based upon the input given by the user. In today's world, any person surfs online for reviews since he/she has better accessibility to features of the product but, the quantity of information generated on the line on a day by day foundation is mannered too.

Prosodic capabilities and acoustic features inclusive of electricity and pitch make contributions to the sentiment variation.Support Vector Machine(SVM) was used for appearing emotion classification. In this strength, pitch and its second-order derivatives were used as features for appearing classification. But in some cases, the audio can also incorporate multiple resources in an unmarried channel. The audio sources will be tune segments, noise, clapping or more than one speaker. In cases when there are many speakers like a telephonic communique, the sentiment of the audio may get ruled with the sentiment of the speaker speakme for longer duration. Isolating those sources is crucial in these cases. Isolating the speakers from an audio conversation, the audio system is speaking on a flip, the aid-of-turn basis with little overlap is Speaker Diarization. It entails getting rid of the non-speech information like noise, pause-in between the communication to split the segments into distinct chunks with voiced and unvoiced facts. For this reason, audio may involve an extraordinary range of speakers, using agglomerative clustering.

Doing polling or filling forms generates a tracking system from various websites that people feel in every field. This type of technique of getting lots of information from people is itself an art in which sentiment analysis plays a vital role. There are tons of sentiment measurements through which tracking succeeds for social sites, analysing emotions and management. Automatic sentiment analysis, apart from programming, is a success in the field of apps, the web and cloud platforms[34]. The enormous large statistics for corporations can cause them to take selections thus. Our focus was not only limited to text data as might be misunderstood from above para but specific audio also generated from these techniques. Nowadays, with improved technology,The large data generated is very complex. Research people are doing Audial statistics for improvement. However, research concerned with other records, including image, speech, content has been restricted. Multi-modal emotion detection is the hottest topic people are interested in deep understanding evaluation[34][26]. Researchers are trying hard to find different techniques to go deeper in multimodal analysis. In multi algo sentiment analysis we implement it with the audio for different techniques to extract features thus can be well used in analysing youtube music and melancholy.We are going to try to find out in this research.

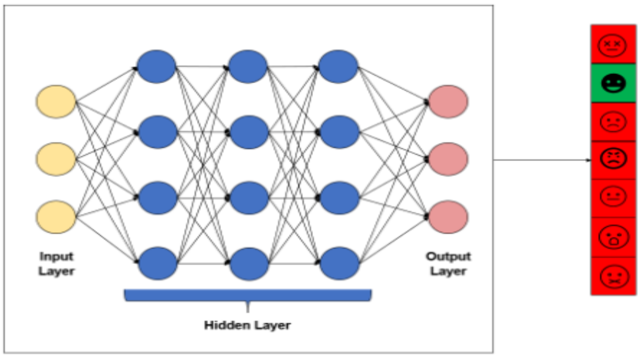
II. RELATED WORK IN THE SYSTEM

A. Overview

Our machine is specifically divided into elements, supply separation and type. Once we've got the audio document of the communique of patron and contact centre agent, we pass it through the first segment of our version. It calls for us to first phase the audio into components and as soon as segmentation is executed using Google's webrtc detector. Those segments are exceeded to the Adaptive Modularised Adaptive Processing(MAP) estimation for Speaker Verification.Clustering algorithm in each iteration with the sound surpassing the generated vector's symphony. The second one component includes the Audio Sentiment category. The Audio Sentiment classification is accomplished on datasets and the generated model is used to predict sentiment on the chunks generated in the first phase of the machine.

B. Classification of Audio Data

Counselled ways to use the acoustic capabilities extracted from audio signs so one can discover the emotional status of the speaker.The speech input given to the automatic detector which detects the behaviour of audio and the indication of speech which uses extraction of features . The audio changed into then fed to Automatice recognition(ASR) model and speaker discrimination version for identifying the information and speaker-identification. The ASR model then classified the voices with unique speaker-ids. The voices were then transformed to text with the assistance of a computerised Speech popularity device. Then the speaker Ids were further matched with the transformed text. ASR machine only generates unique audio as was investigated in reference [4], was a sizable feature to predict the emotion expressed with the aid of distinctive speakers. Studies, show ASR models had been used to mix acoustic features with textual content for the sentiment category. Right here, the RNN-T model was applied to carry out cease-to-give up speech recognition, then the result changed into fed to the sentiment decoder. Sentiment Decoder consisted of Bi-LSTM, interest version and softmax classifier. To reduce overfitting, spectrogram augmentation becomes implied. SVM and XGBOOST for speech recognition, acoustic functions use the HMM method as given in reference [2]. Final output is produced after combining the effects acquired from each sort of classifications It showed the processing of acoustic_feature and word root processing. Sentiment identification of audio facts.



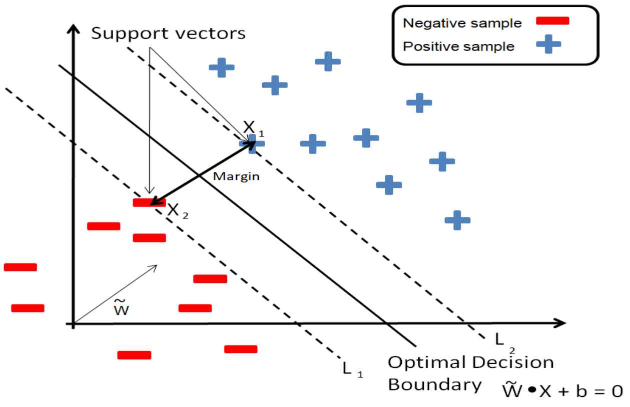


Fig. 1. A neural network used in the model to decide the emotion in output column as shown

Fig. 2. A linear SVM which used to separate positives and negatives in data

III. IMPLEMENTATION

A. Algorithms

1) XGBoost: This algorithm was used for speech data. Regularised, hardware optimised & out-of-core available for disk space. The technique uses cross validation which doesn't need external programming, estimating the performance of Algorithm having low variance in a single training dataset.

2) SVM: SVMs maximise the margin, across the setting apart hyperplane.The choice of characteristic is complete specified by a (typically very small subset of standard samples), the support vectors[29].This generates a Quadratic equation that is easily solvable, This is a major factor in deciding the overall accuracy for each algorithm.

3) MLP: Multilayer perceptron(MLP) uses neurons which are linked to each other depending on the input given and processing of nodes value processing through the feed forward neural networks abbreviated as ANN. The time period MLP is used ambiguously, occasionally loosely to intend any feedforward ANN, from time to time strictly to refer to networks composed of more than one layer of perceptrons (with threshold activation). It also has a different name called "vanilla" neural networks, in particular once they have a single hidden layer.[1]

Generally MLP has 3 layers of node network as shown in Figure 2: Input , middle one hidden, and last one output layer. The second layer uses the activation, in this case tanh. Also, required biases can be found out by backpropagation in MLP. Since layers using non-linear activation clarifies b/w what is a single perceptron and a multilayer perceptron. Moreover, it can make non- linearly separable to separable that distinguishes b/w positives and negatives, in this case different acoustic speech. MLP is a subset of DNN, with finite acyclic graphs. The activation functions used in the code and learning figured below:

$$\mathcal{E}(n) = \frac{1}{2}\sum_j e_j^2(n).$$

$$\mathcal{E}(n) = \frac{1}{2}\sum_j e_j^2(n).$$

Using gradient descent, the change in each weight is

$$\Delta w_{ji}(n) = -\eta\frac{\partial\mathcal{E}(n)}{\partial v_j(n)}y_i(n)$$

B. High Level Architecture

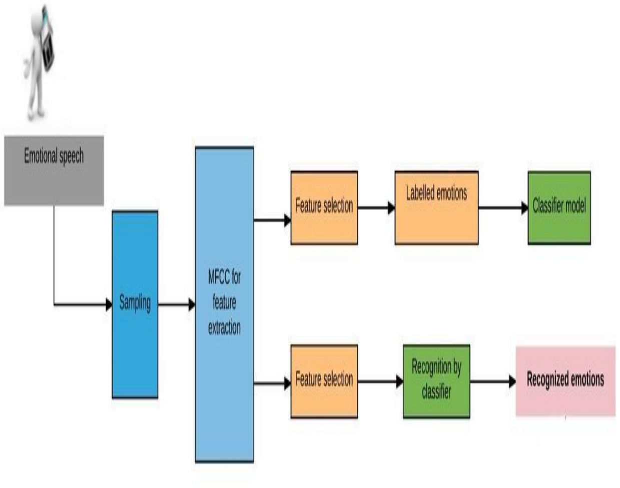


Fig. 3. WorkFlow Model used by the algo to output

C. A Simplified Mechanism

The algorithm applied in Section 3.1, can correspond to various sources for improving the model of dynamic data parallely with improving the accuracy of feature engineering & algorithm tuning using ensemble methods. feature engineering: This has the ability to define and extract more data features using the already present, thus, giving us more meaning and increasing the variance, hence the accuracy. It uses feature transformation as well as feature creation. The feature transformation uses data normalisation; removing skewness of variable, changing the original scale to vary. The feature creation uses the transitivity property of equality, since we can create a greater correlation variable which in terms makes no correlation in features.

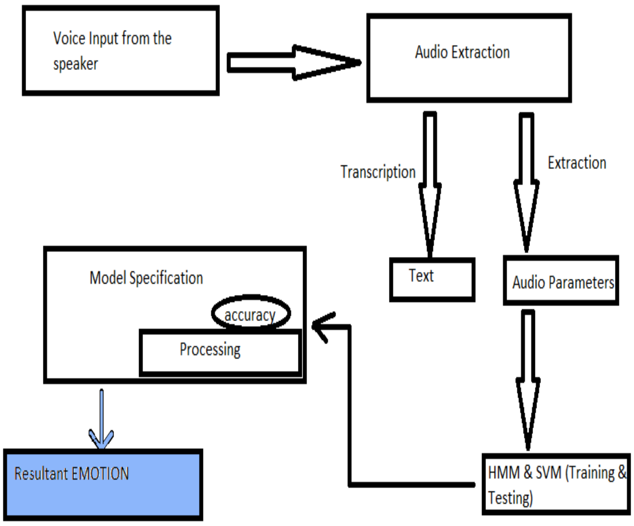


Fig. 4. A detailed resultant output for the workflow shown in (Fig3).

IV. COLLECTION SETUP & EXECUTION

A. Speaker Diarization

An audio document with telephonic communique from Exotel changed into used. The dataset consists of 300 audio files & sentiment labelled for the entire conversation. The files are labelled with 4 emotions including unhappy, irritated, satisfied and impartial. After acting segmentation on the audio report via VAD detection, chunks are received[3]. Those chunks comprise voiced statistics which surpasses the adaptive MAP estimation to get an excellent vector. This awesome vector when handed to the clustering offers the chunk numbers that belong to Speaker zero and the chunks that belong to Speaker 1. Label zero or 1 is decided on the fact that which character starts off evolved to speak first, the first individual to talk is assigned label 0 and a different person.

B. Clustering

K-means and Spectral clustering had been used for clustering at the high-quality vector received after MAP estimation. The labels received after clustering were verified manually. Despite the fact that most of the separation changed into performed well, there have been few chunks in which overlap among the two human beings having a conversation can be located[18]. The fundamental purpose behind this overlap is because there has been no pause within verbal exchange; that's why exclusive chunks have been not created throughout the speaker diarization process[19][20]. The Overall performance of each of the clustering algorithms have been affected to a certain extent due to overlap. In previously noted clustering algorithms, spectral clustering gave excellent outcomes compared to Kmeans. For each of the algorithms, we take two clusters because we best have a communique of two people. For spectral clustering, affinity used becomes cosine considering that we are looking for the cosine similarity among the chunks to cluster them. After clustering is performed, we then pass the chunks of client audio to our type model.

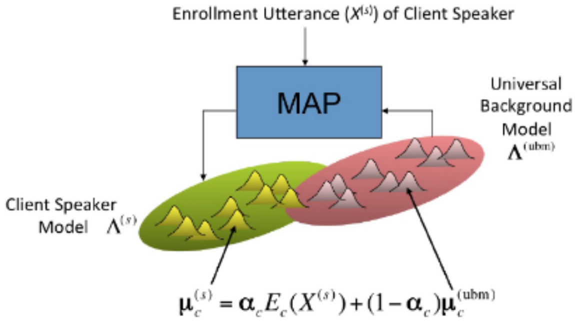

Enrollment Utterance ($X^{(s)}$) of Client Speaker
MAP
Universal Background Model $\Lambda^{(ubm)}$
Client Speaker Model $\Lambda^{(s)}$
$\mu_c^{(s)} = \alpha_c E_c(X^{(s)}) + (1-\alpha_c)\mu_c^{(ubm)}$

Fig. 5. MAP used in the model

C. Classification

Training on the speech datasets. These datasets contain audio files from different actors. These actors utter a sentence in 8 feelings: satisfied , angry, calm, unhappy, surprised, neutral, disgust, worried. We educate our version by extracting acoustic features of the audio clip. We use a complete 193 features of all the audio clips from our datasets for class reason. We use distinctive algorithms for classifying the sentiment of our input chunk. We first educate our version on datasets for which we at once pass our 193 features and labels of the sentiment in the audio clip. Then we make predictions.

V. RESULTS & ANALYSIS

TABLE I

ACCURACIES ON COMPARING DIFFERENT ALGO'S ON VARIOUS DATASET

Models	**Dataset Accuracy**		
	RAVDESS	**TESS**	**BOTH**
SVM	80%	78%	86%
XGBoost	75%	72%	80%
MLP	70%	65%	72%

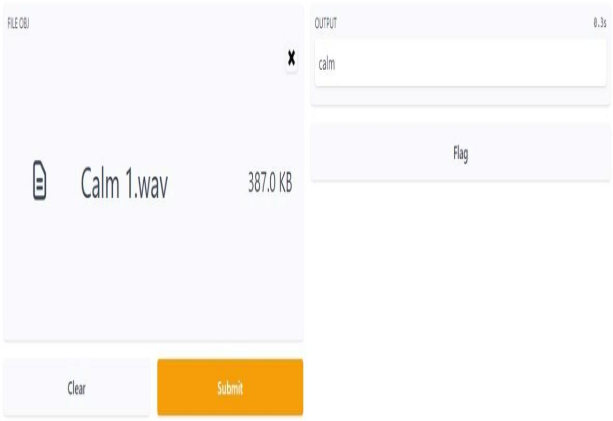
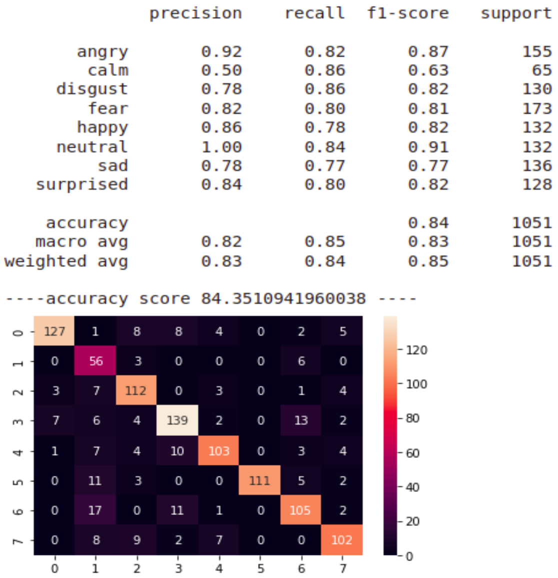


Fig. 5. Web UI screenshot for our emotion model



```
              precision    recall  f1-score   support

       angry       0.92      0.82      0.87       155
        calm       0.50      0.86      0.63        65
     disgust       0.78      0.86      0.82       130
        fear       0.82      0.80      0.81       173
       happy       0.86      0.78      0.82       132
     neutral       1.00      0.84      0.91       132
         sad       0.78      0.77      0.77       136
   surprised       0.84      0.80      0.82       128

    accuracy                           0.84      1051
   macro avg       0.82      0.85      0.83      1051
weighted avg       0.83      0.84      0.85      1051

----accuracy score 84.3510941960038 ----
```

Fig. 6. SVM increased accuracy output

VI. THE CONCLUSION

After going through the findings we have come to recognize that every one the experiments that have been finished yield both a fantastic or terrible sentiment, polarity or feelings which include happiness, anger, calm, fear and surprise as an output. Overall observation of the result leads to very important conclusions. SVM performs better in comparison to XGBoost and MLP when using both RAVDESS and TESS datasets, but the results are competitive when using individual datasets. If we observe carefully the result for individual datasets, RAVDESS as well as TESS are not that impressive for SVM as well as XGBoost and MLP, but if we combine both the datasets and apply the algorithms on them both at a time, the difference between results is quite large. In all, the use of acoustic only models rather than text based models are also producing great results with better datasets. With acoustic models the processing becomes easy to train and working with higher performing ML models is used. We therefore do executions for many iterations training the model. Studies show use of many function word phrases as well as learning models. Those models have additionally proven correct consequences as they integrate the best of strategies.

As evident from the desk, preventing term elimination has made a moderate boom in accuracy of truly all of the fashions. Support Vector device (SVM) completed the fine compared to Logistic Regression and Neural Networks. MLP has capability which prevents phase-elimination.

REFERENCES

[1] Maghilnan, S., and M. Rajesh Kumar. "Sentiment analysis on speaker specific speech data." In *2017 international conference on intelligent computing and control (I2C2)*, pp. 1-5. IEEE, 2017.

[2] Abburi, Harika, Eswar Sai Akhil Akkireddy, Suryakanth Gangashetti, and Radhika Mamidi. "Multimodal sentiment analysis of telugu songs." In *SAAIP@ IJCAI*. 2016.

[3] Gupta, Shilpi, and Anu Mehra. "Speech emotion recognition using svm with thresholding fusion." In *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 570-574. IEEE, 2015.

[4] Nandwani, Pansy, and Rupali Verma. "A review on sentiment analysis and emotion detection from text." *Social Network Analysis and Mining* 11, no. 1 (2021): 1-19.

[5] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in Interspeech, 2018, pp. 247–251.

[6] Ibáñez, Manuel López, Nahum Álvarez, and Federico Peinado. "Litsens: An improved architecture for adaptive music using text input and sentiment analysis." *In Proceedings of the C3GI Conference*, vol. 2017. 2017.

[7] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition," in Interspeech, 2020, pp. 3755–3759.

[8] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6675–6679.

[9] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1675–1685, 2019.

[10] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227–2231.

[11] Driscoll, Beth. "Sentiment analysis and the literary festival audience." *Continuum* 29, no. 6 (2015): 861-873.

[12] Tolstoukhov, D. E., D. P. Egorov, Y. V. Verina, and O. V. Kravchenko. "Hybrid Model for Sentiment Analysis Based on Both Text and Audio Data." In *Sentimental Analysis and Deep Learning*, pp. 993-1001. Springer, Singapore, 2022.

[13] Kharde, Vishal, and Prof Sonawane. "Sentiment analysis of twitter data: a survey of techniques." *arXiv preprint arXiv:1601.06971* (2016).

[14] Watanabe, Kanako, Yoko Greenberg, and Yoshinori Sagisaka. "Sentiment analysis of color attributes derived from vowel sound impression for multimodal expression." In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pp. 1-5. IEEE, 2014.

[15] Salas, Jessie. "Generating music from literature using topic extraction and sentiment analysis." *IEEE Potentials* 37, no. 1 (2018): 15-18.

[16] Feld, Steven. "Sound and sentiment." In *Sound and Sentiment*. Duke University Press, 2012.

[17] Jia, Yanan, and Sony SungChu. "A deep learning system for sentiment analysis of service calls." *arXiv preprint arXiv:2004.10320* (2020).

[18] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, no. 4, pp. 335–359, 2008

[19] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.

[20] Lakshmi, Magudilu Srishyla Kumar, Ayasakanta Rout, Ariana Morris, and Joseph Smaldino. "Consumer Opinion of Personal Sound Amplification Products: A Preliminary Sentiment Analysis." *American Journal of Audiology* 28, no. 2S (2019): 450-459.

[21] Samareh, Aven, Yan Jin, Zhangyang Wang, Xiangyu Chang, and Shuai Huang. "Detect depression from communication: how computer vision, signal processing, and sentiment analysis join forces." *IISE Transactions on Healthcare Systems Engineering* 8, no. 3 (2018): 196-208.

[22] Alsayat, Ahmed, and Nouh Elmitwally. "A comprehensive study for Arabic sentiment analysis (challenges and applications)." *Egyptian Informatics Journal* 21, no. 1 (2020): 7-12.

[23] Veenendaal, Anne, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, and Rahul S. Patwardhan. "Sentiment Analysis in Code Review Comments." *Computer Science and Emerging Research Journal* 3 (2015).

[24] Pereira, Moisés Henrique Ramos, Flávio Luis Cardeal Pádua, Adriano César Machado Pereira, Fabrício Benevenuto, and Daniel

Hasan Dalip. "Fusing audio, textual, and visual features for sentiment analysis of news videos." In *Tenth International AAAI Conference on Web and Social Media*. 2016.

[25] Napier, Kathleen, and Lior Shamir. "Quantitative sentiment analysis of lyrics in popular music." *Journal of Popular Music Studies* 30, no. 4 (2018): 161-176.

[26] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6720–6724

[27] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 4835–4839.

[28] Chakravarthi, Bharathi Raja, K. P. Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, and John P. McCrae. "DravidianMultiModality: A Dataset for Multi-modal Sentiment Analysis in Tamil and Malayalam." *arXiv preprint arXiv:2106.04853* (2021).

[29] Nuanmeesri, Sumitra. "Sentiment Analysis of Thai Sounds in Social Media Videos by using Support Vector Machine." *Indian Journal of Science and Technology* 12 (2019): 13.

[30] Mehmood, Khawar, Daryl Essam, Kamran Shafi, and Muhammad Kamran Malik. "An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis." *Information Processing & Management* 57, no. 6 (2020): 102368.

[31] Rodriguez, Axel, Yi-Ling Chen, and Carlos Argueta. "FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis." *IEEE Access* 10 (2022): 22400-22419.

[32] Gitari, Njagi Dennis, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. "A lexicon-based approach for hate speech detection." *International Journal of Multimedia and Ubiquitous Engineering* 10, no. 4 (2015): 215-230.

[33] Narrain, Siddharth. "Hate Speech, Hurt Sentiment, and the (Im) Possibility of Free Speech." *Economic and Political Weekly* (2016): 119-126.

[34] Balahur, Alexandra, and Ralf Steinberger. "Rethinking Sentiment Analysis in the News: from Theory to Practice and back." *Proceeding of WOMSA* 9 (2009): 1-12.