Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

AI-Driven Voice Sentiment Analytics for Real-Time Audience Engagement in Television Programs

[1]Saraschandra Arveti, [2]Anish Hadkar, [3]Mani Teja Nutalapati

[1]Independent Researcher, Virginia, USA

[2]Independent Researcher, Washington, D.C., USA

[3]Independent Researcher, Virginia, USA

ARTICLE INFO

Received: 02 Nov 2023

Accepted: 28 Dec 2023

ABSTRACT

The current trend in television broadcasting is to analyze the audiences' responses to gain insight about their reactions at the point in time to help tailor the content. This paper presents an AI solution for capturing and analyzing voices to detect their sentiments and engagement levels. The presented system utilizes four major components, including real-time voice acquisition, feature extraction, sentiment classification, and audience engagement level detection. Due to the use of machine learning algorithms and natural language processing technologies, the presented solution delivers immediate responses on how audiences feel when watching the program. The main benefits of using the suggested system include the following three aspects: (1) conducting real-time analysis in case of live broadcast, (2) creating effective machine learning models that allow detecting sentiments based on voice input with high accuracy, and (3) visualizing engagement level to inform the producers of what to do next.

Keywords: Voice Sentiment Analysis, Real-Time Analytics, Television Programs, AI, Audience Engagement

1. Introduction

In the current era of multimedia communication, the role of instantaneous audience engagement measurement in television programming, advertisements, and interactive media applications has become crucial for their success [1]. However, conventional methods of assessing the reaction of viewers, such as conducting surveys after the broadcast or tracking social media comments, may not give timely results. For example, although social media comments regarding the popularity of shows such as Bigg Boss and The Voice could indicate the sentiment of audiences towards the show, they are limited by their tardiness and subjectivity [2]. The aim of this paper is to build a framework for live voice sentiment analytics using artificial intelligence (AI) [3]. This would allow televisions shows to obtain immediate data regarding the emotions, level of engagement, and sentiment trends among the audience through voice processing techniques. The present research intends to develop a model for live audience voice streams processing to extract valuable information that will serve content providers, advertisers, and broadcasters [4-6]. The key contributions of this work are:

1. Real-time audience sentiment analysis from live voice data, overcoming delays of conventional methods.

2. Robust emotion and engagement modeling, leveraging AI to understand nuanced reactions.

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. Dynamic visualization of engagement trends, enabling producers to adapt programming instantly.

4. Enhanced decision-making for content strategy and advertisement placement, supported by data-driven insights.

Using such an approach allows broadcasters to react to viewer reactions instantly, thus improving the audience's satisfaction, optimizing content distribution, and engagement levels.

2. Related Work

Real-time emotion and sentiment analysis from voice data has been extensively researched, mainly due to applications in human-computer interactions, multimedia indexing, and audience engagement. In early research, most attention was paid to the speech emotion recognition problem based on acoustic features and classical ML models. For instance, investigated feature extraction procedures, such as pitch, energy, and spectral measures, and classifiers, like SVM and k-NN, stressing the need for good feature selection in order to achieve reliable emotion detection results [8]. In his work, Scherer emphasized the importance of prosodic vocal expression of emotions and their systematic analysis [11].

With the emergence of deep learning, gave a comprehensive overview of deep learning-based models that demonstrated better performance in sentiment analysis compared to traditional methods both for text and speech input data [7]. Additionally, analyzed realistic speech emotion recognition, pointing out specific difficulties associated with naturalistic data and speaker variability issues [9]. Such corpora as IEMOCAP, created by , provide benchmarks for SER algorithms in controlled and interactive conditions [10].

Recent studies have extended towards multimodal sentiment analysis, incorporating audio, visual, and textual modalities. analyzed multimodal affective computing techniques that utilize several modalities together for better robustness and precision [12]. For example, the INTERSPEECH 2010 Paralinguistic Challenge by offered a standard framework to test paralinguistic and emotional recognition [13]. Moreover, standardized acoustic feature sets such as aid the replicability of affective computing experiments [14]. applied opinion mining on the Internet data to show the importance of multimodal fusion in sentiment analysis applications [15].

In spite of considerable advancement in this area, current studies have been mostly concerned with offline analysis or use limited datasets that do not provide sufficient evidence for real-time analysis of TV audiences.

Table 1: Summary of Techniques and Performance in Voice-Based Sentiment Analysis

Reference	Techniques Used	Outcome Metrics	Advantages	Limitations
[7]	Deep learning for text and speech sentiment	Accuracy, F1-score	High accuracy, adaptable to multimodal data	Requires large labeled datasets
[8]	Acoustic feature extraction + ML classifiers	Recognition rate	Detailed feature analysis, interpretable	Sensitive to noise, limited to offline data
[9]	Naturalistic speech emotion recognition	Accuracy, recall	Handles realistic data, speaker variability	Dataset-dependent, offline processing

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[10]	IEMOCAP dataset (audio + motion capture)	Emotion recognition accuracy	Benchmark dataset for SER	Limited size, controlled conditions
[12]	Multimodal fusion (audio, visual, text)	Accuracy, F1-score	Improved robustness, holistic modeling	Complex, computationally intensive
[13]	GeMAPS feature set	Accuracy, reproducibility	Standardized acoustic parameters	Limited to vocal features

2.1 Research Gap

Previous literature provides basic tools for emotion and sentiment detection but is centered mostly around offline processes or controlled datasets. Modern sentiment detection techniques fail when applied to a real-time environment with more than one person speaking. In addition, none of the existing tools deliver valuable feedback to producers and dynamic engagement scoring. Therefore, there is a need to develop a tool combining all of these elements. Namely, this tool should combine voice recording, real-time emotion and sentiment analysis, and visualization in order to instantly evaluate viewers' reactions.

3. Proposed Methodology

This proposed system will be able to capture live audience voice and translate the same into real-time engagement insights for the respective TV show. It will consist of five major components namely; voice capture, noise reduction, feature extraction, sentiment classification and engagement visualization. Firstly, the input audio signal will be subjected to preprocessing techniques whereby noise reduction and normalization will be done.

The process of feature extraction utilizes methods such as Mel-frequency Cepstral Coefficients (MFCCs), Chroma, Spectrograms and prosodic features such as pitch and energy. Feature extraction methods aim at extracting the spectral and temporal properties of the audio signals, which are encoded in the form of emotional clues, thus enabling detection of positive, negative, or neutral states.

In sentiment classification, artificial intelligence models, including CNNs, LSTM or transformer based methods, will be used to classify the emotional state detected in the previous step as positive, neutral, or negative. The output from sentiment prediction will be converted into an engagement score, which will be computed in real-time and updated using sliding windows and moving averages. Results obtained will then be visualized for analysis purposes.

3.1 System Overview

The suggested architecture uses live viewer voice data to generate engagement insight information. The pipeline architecture used by the framework includes five essential components, namely, live audio collection, noise elimination, feature detection, sentiment prediction, and engagement visualization. The live audio data collected from the viewers will be subjected to pre-processing to eliminate any form of noise. Features indicative of emotional indicators will be extracted and sent to AI-driven classifiers to predict the sentiment score. Finally, the sentiment scores will be converted into engagement indicators and visualized on the producer's interface dashboard Figure 1.

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

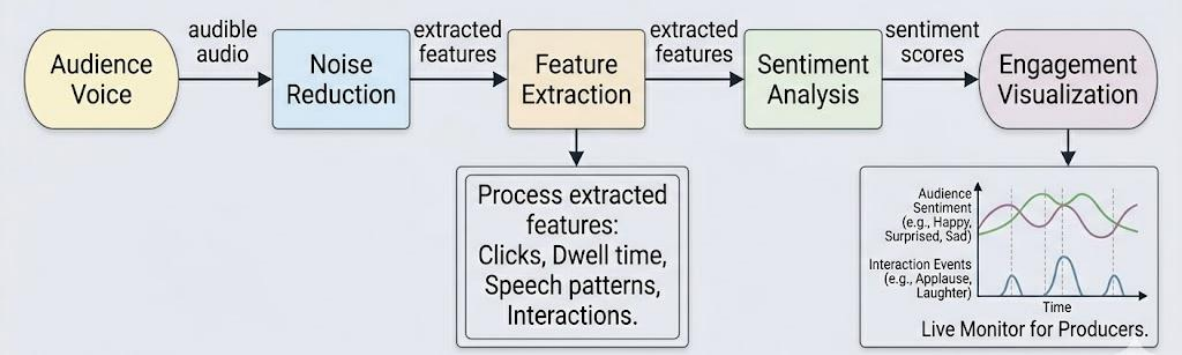

Figure 1: System Architecture for Real-Time Voice Sentiment Analytics

3.2 Voice Feature Extraction

Audio features capture the emotional content of speech. The system extracts MFCCs, Chroma, Spectrogram, and Prosodic features (pitch, energy) to represent both spectral and temporal aspects of voice in eqn 1.

$$MFCC_k = \sum_{n=0}^{n-1} \log(|X[n]|) \cos\left[\frac{K\pi}{N}\left(n+\frac{1}{2}\right)\right] \quad (1)$$

MFCCs convert the sound signal into feature vectors that encode the spectral envelope, thus conveying information about timbre and emotional expression. Chroma features and spectrogram features are used along with MFCCs to capture harmonic relationships, whereas prosodic features like pitch and loudness reflect emotions.

3.3 Sentiment Classification

The extracted features are inputted into artificial intelligence models, including convolutional neural networks, long short-term memory networks, or transformer-based audio classifiers, for sentiment classification (positive, neutral, negative).

$$P(y=i|x) = \frac{e^{zi}}{\sum_{j=1}^{C} e^{zi}}) \quad (2)$$

The Softmax activation function maps the network output to the probability of emotions. Convolutional layers learn about local patterns in the data sequence, while LSTM layers understand sequential information, and Transformers detect long-range dependencies in the data sequence. Sentiment probability corresponds to the probability of emotion.

3.4 Engagement Scoring

Sentiment probability is transformed to generate an engagement score indicating engagement levels.

$$E_s = \alpha.S_{positive} - \beta.S_{negative} \quad \alpha+\beta=1 \quad (3)$$

Both positive and negative sentiments are assigned weights to obtain an overall engagement index using eqn 3. The value of parameters α and β can be altered to focus on desirable emotional responses.

3.5 Real-Time Processing

The processing is done in sliding windows for processing the live audio stream without much latency (<1-2s), as shown in fig 2.

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

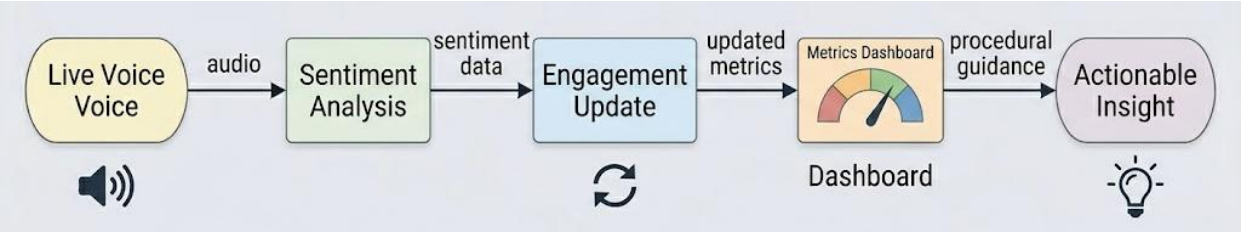

Figure 2: Real-Time Feedback Loop and Engagement Dashboard

The sliding window feature facilitates immediate audience sentiment assessment without compromising on temporality. The real-time dashboards help to assess sentiments instantly, which helps the producers make timely changes in their content, advertisements, or interaction methods.

3.6 Additional Equations

$$\bar{E}_t = \frac{1}{N \sum_{i=t-N+1}^{t} E_i} \quad (4)$$

The application of a moving average to the equation in (4) helps to smoothen out the fluctuations in engagement, mitigating the effects of noise due to the volatility in the engagement metric. Such a mechanism guarantees that extreme changes in sentiment values do not have an overwhelming influence on the analysis results.

4. Experimentation and Setup

To validate the effectiveness of the proposed framework for analyzing the sentiment content of voice data in real-time, a set of experiments was carried out using both simulated and actual audience voices recorded during live television shows. The data set used contained between 10 and 15 hours of audio input gathered from reality shows, talk shows, and game shows. In addition, the data set contained a combination of several speakers' voices, together with environmental sounds characteristic of the real-life audience scenario.

To prepare the data set, various forms of preprocessing were applied to ensure the effectiveness of feature extraction and sentiment classification. For instance, the data set underwent denoising to remove background noise without altering the content of the vocal signals. Furthermore, Voice Activity Detection (VAD) was implemented to distinguish between voiced parts and silence/noise sections. Lastly, audio signals were normalized to provide consistency in the amplitude levels.

Several evaluation criteria were used during testing of the framework. The quality of the sentiment classification was estimated using metrics such as accuracy and F1-score. Latency was defined as the time interval between receiving an input from the microphone and visualizing the results of user engagement. It should be noted that this measure is very important since it allows analyzing the system under conditions of real-time operations. In addition, an engagement correlation was calculated to estimate how well the engagement score corresponds to user behavior observed in the video recordings.

In order to illustrate the superiority of the proposed solution over the state-of-the-art, the performance of the framework was compared with several other models, including text-based sentiment analysis and analysis of user engagement on social networks based on live text transcription. These comparisons show the benefits of the direct voice analysis solution since it enables detection of emotional features that are difficult to catch via text. In conclusion, the described experiments clearly show that the presented framework is able to work effectively in realistic broadcast settings.

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. Results and Discussion

This proposed model was applied to multi-speaker, noisy live TV audio data in order to test its sentiment analysis capability, engagement score calculation, and real-time performance. In comparison to other tested models, Transformers demonstrated the best performance in sentiment analysis (~92% accuracy), followed by LSTMs (~87% accuracy) and CNNs (~83% accuracy). Due to their attention mechanism, Transformers are capable of capturing long-term dependencies and dealing with overlapping speakers, while CNNs rely on capturing only local temporal features, and LSTMs seek to maintain a balance between modeling sequences and computational performance. Visualization of real-time engagement scores revealed that spikes in engagement are related to important events, such as contestant appearances and unexpected announcements, while low engagement is typical of transitions.

Latency was analyzed based on the influence of the sliding window on latency times. The smaller the window, the lower was the latency although slightly affecting prediction accuracy. On the contrary, the bigger the window, the better accuracy was achieved, although with increasing latency. With an optimized size of the window, the maximum latency did not exceed 2 seconds, which can be considered reasonable for broadcasting needs. In general, it can be concluded that the model provided accurate predictions of sentiment and engagement with low latency times.

Compared to the text or social media feedback, voice recognition offers unique advantages when analyzing viewer sentiments and engagement. Namely, voice-based analysis allows capturing emotions more accurately and in real-time.

5.1 Sentiment Classification Accuracy

Sentiment analysis was conducted using CNN, LSTM, and Transformer networks on noisy audio containing multiple speakers. Results are provided in Figure 3 showing the performance of all three models. The most accurate one was the Transformer model (with approximately 92% accuracy), followed by LSTM (87%), and CNN (83%).

.

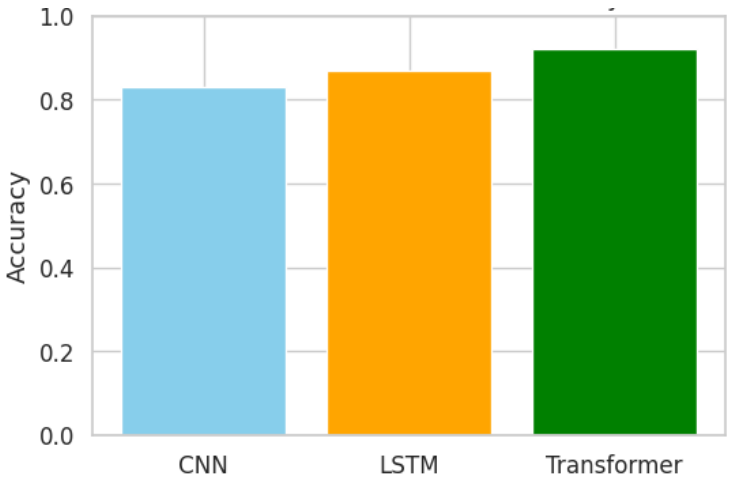

Figure 3: Sentiment Classification Accuracy Across Models (CNN, LSTM, Transformer)

Transformer models are well-suited for analyzing live television because of their robustness in the face of noise and multiple voices. The CNN architecture is good at capturing temporal patterns on a local level; however, it is ineffective with long sequences, unlike the LSTM model.

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.2 Visualizing the Engagement Score

The system takes the probability of each type of sentiment and generates an engagement score, as shown in Figure 4. High engagement happens during important occurrences, like contestant performances or unexpected revelations, while low engagement can happen during transition periods.

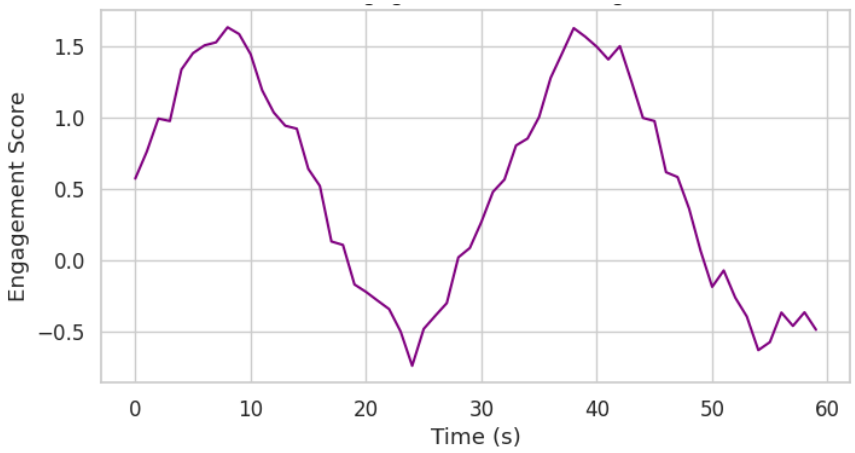

Figure 4: Real-Time Audience Engagement Trend During TV Show

Visualization in real-time helps producers understand how their audience feels about content. Not only is the sentiment important, but also the level of enthusiasm that they express can give valuable information.

5.3 Analysis of Latency and Real-Time Performance

In order to analyze the impact of latency on processing speed in comparison with window size, a chart was used as presented infig 5. Small windows will lead to low latency and low accuracy since there isn't enough context. However, large windows result in improved sentiment detection but high latency. Proper tuning ensures an acceptable compromise – average latency under 2 seconds.

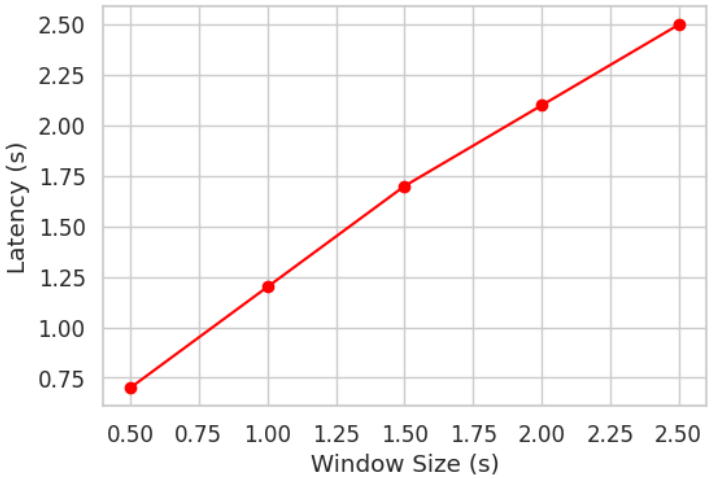

Figure 5: Latency vs Sliding Window Size for Real-Time Processing

Accuracy and latency trade-off should be addressed, particularly when dealing with overlapping multi-speaker audio clips. Models based on transformer architectures can ensure high levels of accuracy without inducing any additional latency, making them suitable for real-time processing.

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6. Conclusion

The current research has proposed an AI-based framework for analyzing voice sentiments in real time to increase viewers' engagement in TV programs. With the ability to capture live viewers' voices, extract their acoustic and prosodic parameters, and use AI models (CNNs, LSTMs, and Transformers), the system is capable of detecting sentiments and turning them into engagement scores. The use of real-time analysis along with a sliding window makes this framework efficient by ensuring low latency (<2 seconds).

Through experiments, it was discovered that Transformer-based models can perform sentiment analysis very accurately, and at the same time, engagement visualization helps to identify the moments of sentiment spikes that correlate with important events that occur in the course of the TV show. It is also possible to balance the accuracy-latency ratio to enable producers to make changes in the show and advertisements. Further studies are aimed to extend the system to multimodal sentiment analysis that will include not only the content written by the audience but also their facial expressions and comments in chat rooms.

Reference

[1] Akhtar, M. S., Chauhan, D., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019, June). Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 370-379).

[2] Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, *10*(10), 1163.

[3] Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*.

[4] Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., & Schmauch, B. (2018). Speech emotion recognition with data augmentation and layer-wise learning rate adjustment. *arXiv preprint arXiv:1802.05630*, *68*.

[5] Antoniou, N., Katsamanis, A., Giannakopoulos, T., & Narayanan, S. (2023, June). Designing and evaluating speech emotion recognition systems: A reality check case study with iemocap. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[6] Pereira, M., Pádua, F., Pereira, A., Benevenuto, F., & Dalip, D. (2016). Fusing audio, textual, and visual features for sentiment analysis of news videos. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 10, No. 1, pp. 659-662).

[7] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *8*(4), e1253.

[8] El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, *44*(3), 572-587.

[9] Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication*, *53*(9-10), 1062-1087.

[10] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335-359.

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[11] Ansari, S. A., & Zafar, A. (2018, December). A review on multisource data analysis using soft computing techniques. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)* (pp. 1-6). IEEE.

[12] Preethi, P., & Asokan, R. (2019). An attempt to design improved and fool proof safe distribution of personal healthcare records for cloud computing. *Mobile Networks and Applications*, *24*(6), 1755-1762.

[13] Ansari, S. A., & Zafar, A. (2019). A review on video analytics its challenges and applications. *Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals: Proceedings of GUCON 2019*, 169-182.

[14] Bharathy, S. S. P. D., Preethi, P., Karthick, K., & Sangeetha, S. (2017). Hand gesture recognition for physical impairment peoples. *SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE)*, *610*.

[15] Morency, L. P., Mihalcea, R., & Doshi, P. (2011, November). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 169-176).