Indonesian Journal of Electrical Engineering and Computer Science
Vol. 29, No. 2, February 2023, pp. 875~882
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v29.i2.pp875-882

An investigation of machine learning techniques in speech emotion recognition

Anu Saini[1], Amit Ramesh Khaparde[1], Sunita Kumari[1], Salim Shamsher[2], Jeevanandam Joteeswaran[3], Seifedine Kadry[4]
[1]Department of Computer Science and Engineering, G. B. Pant DSEU Okhla I Campus, New Delhi, India
[2]School of Commerce, Narsee Monjee Institute of Management Studies, Navi Mumbai, India
[3]Department of Computer Science and Engineering, GITAM University, Vishakhapatnam, India
[4]Data Science, Noroff University College, Oslo, Norway

Article Info

Article history:

Received Jun 26, 2022
Revised Oct 4, 2022
Accepted Oct 17, 2022

Keywords:

Deep learning
Emotions
Machine learning
Sentiment
Speech emotion recognition

ABSTRACT

The natural languages are medium of communication from the inception of civilization. As the technology improves, the text messages, voice messages and videos are the addons in medium of communication. In long distance communication, the analysis of expression is modern area of research. The parameters of assessment are subjective hence the emotion recognition is challenging task. This article furnishes the investigation of various machine learning techniques and novel methods for speech emotion recognition (SER) to determine the feeling/sentiments in a speech. Here, we investigate the three machine learning methods named multinominal Naive Bayes (MNB), logistic regression (LR), and linear support vector machine (LSVM). Further, these techniques are incorporated with the proposed method. The performance of these machine learning techniques is investigated on two different datasets. The datasets consist of voice and text data samples. The prosed method is trained and tested on these datasets. As per the experimentation, it has been observed that the LSVM has outperformed the other two machine learning techniques.

Corresponding Author:

Amit Ramesh Khaparde
Department of Computer Science and Engineering
G. B. Pant Delhi Skill and Entrepreneurship University Okhla I Campus
New Delhi, 110020, India
Email: khaparde.amit@gmail.com

1. INTRODUCTION

In today's data overload environment, organizations have typically accumulated mountains of customer feedback without much categorization. It is difficult for people to manually examine it without bias or inaccuracy. speech emotion recognition (SER), is the solution to this issue. Voice emotion recognition will use audio input from human speech to convert it to text, evaluate the text to determine whether the sentiment of the statement is positive, negative, or neutral, and then provide the analyzed emotion. Providing a substantial chunk of structured data instead of just human intuition, which isn't always accurate, improves in enhancing decision-making in a wide range of sectors. It examines whether the customer's comment or feedback is positive, negative, or neutral. SER can be implemented using many different approaches such as a hybrid technique, a deep learning strategy, or a Lexicon-Based method. Here, human emotions in speech are recognized using a machine learning approach.

In this paper, SER using various machines learning techniques (MLT) is tested on two different datasets. The datasets consist of voice and text data. The tweets from Twitter is the part of the considered data

sets. This paper is organized into five sections. first is introduction. The second section discusses the related work. The third section provides an explanation of the implementation. In the fourth section, the findings are provided, and in the final portion, the findings are summarised.

2. RELATED WORK

The emotions are recognized in speech SER using deep learning (DL) and an attention mechanism based on a deep neural network (DNN) [1]. It provides an overview of the most current changes to SER and looks at how various attentional mechanisms affect SER performance. The SER using mel frequency cepstral coefficients (MFCC), mel spectrogram (MS), and chroma is discussed precisely [2]. The authors used the liberos package (Python) for the extraction of mandatory information to achieve the SER. The study of various speech emotion detection techniques, including hidden markov model (HMM), support vector machine (SVM), and others are presented in [3]. The voice recognition system contains four key components, named speech input, extraction of features, SVM Algorithm based grouping, and emotion output, just like other conventional recognition systems. The three steps in the SER are as follows:

- Identification of signals in pitch/energy. This task is performed by Language Development System.
- Feature reduction, i,e, quantities are encapsulated in fewer features.
- Mapping the characteristics with emotion, i.e. using the sample data, the characteristics are mapped to the emotions

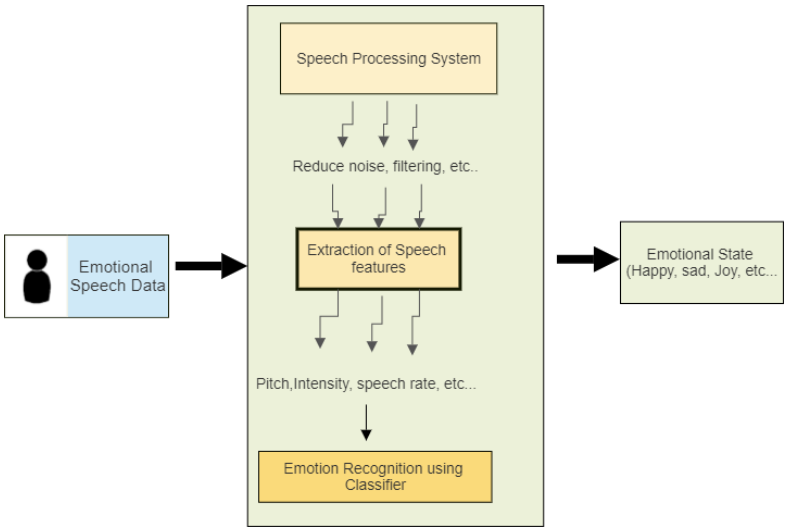

Figure 1. Speech emotion recognition system

Figure 1, depicts the fundamental structure of the human emotion recognition system. Speech signals supplies an input signal. The input signal is pre-processing to reduces the noise. Pre-emphasize, spatial and temporal filtering are used for pre-processing. Next, the common features are discovered by using feature extraction algorithms. Pitch, intensity, and speech rate are considered as the features in the speech signal. Finally, classifiers are employed to categories the input signal based on features into emotional states (happy, sad, joy, angry, surprise). artificial neural networks (ANN), gaussian mixture models, K-nearest neighbor, HMM, vector quantization, and SVMs are the most popular classification algorithms.

In [4] the author has discussed the emotions are and related theories. Here the author has discussed the six different theories from start to till date. Three categories of emotion theories physiological, neurological, and cognitive theories can be made [5]. Identifying the sentiments of person based on his communication is known as sentimental analysis [6]. It can be aspect-based or fine-grained. The fundamental of sentimental analysis and associated challenges are discussed in [7]. Here, the sentimental analyses are explained in terms of natural language processing.

Computational models such as Naive Bayes (NB), SVM and N-gram are successfully providing the desired results in sentimental analysis. These models are thoroughly explained in [8]. The analysis of various computational models on social networking platform called Twitter are performed [9] and categories the tweet

in positive, negative and neural categories. Here, the authors have considered the millions of tweets form the twitter. The database is created for the task. The NB method [10] and LR as a text classifier [11] are discussed; the methodologies, mathematical understanding and other variants are clearly discussed. linear support vector classifier (LSVC) for text classification and the confusion matrix function are discussed in [12]. Fit and transform functions can be used to train models are discussed in [13]. In [14] the authors are discussed K-Means clustering for grouping of data and multiple algorithms for emotion recognition like LR, random forest, NB, Linear svc, and decision tree. A systematic evaluation of hidden markov models for sentiment analysis is discussed in [15]. The lexicon-based methodology for sentiment analysis discussed in [16], it entails gauging sentiment from the text's use of words or phrases with different semantic connotations. The speech-to-text conversion using Google API is developed in [17]. It goes over the differences between synchronous and asynchronous audio files, it uses real-time speech input. Sphinx4, Google Speech API, and Bing speech API with weighted workload ratings (WWR) as the performance metric can be used for speech recognition [18]. For speaker recognition, MFCC is used as a feature and dynamic time warping (DTW) for feature matching, with several distance computation methods such as euclidean, correlation, and canberra, and recognition rate as the performance parameter. The HMM and neural networks are employed for speech-to-text conversion, [19]. The IEMOCAP corpus database is employed for feature extraction [20]. To create an emotion recognition model, Auther employed the inception net v3 model. Inception is based on the GoogLeNet Architecture.

2.1. Approaches of sentimental analysis

The novel fake comment detection method is proposed to detect fake comments on E-commerce platforms [21]. The proposed method used the n-gram and term frequency-inverse document frequency (TF-IDF) approach to extract features in computation. The method is tested on a hotel review dataset. The proposed method performs better than another method. The sentiment analysis plays vital role in video conferencing. In the era of the pandemic video conferencing is the mode of communication. COVID-based analysis of video conferencing platforms [22] is presented. The platforms are analyzed in four stages. The analysis is supported by respective applications. The sentiment analysis plays a vital role in social media also. A novel algorithm is proposed in blockchain technology [23]. It creates a secure connection to transfer the information between sender and receiver. The application of the proposed methodology is live-streaming on social media. The proposed algorithm's performance is compared with other cutting-edge algorithms and found outstanding. The factors affecting the performance of wireless proxy internet protocol are addressed. The key [24] exchange protocol is negotiated. The experimentation is conducted by considering the safety of the TSL protocol. The evolutionary algorithms play a very vital sentiment analysis and speech recognition. An evolutionary method using the bee algorithm [25] is proposed to optimize the wireless sensor network. The results are compared with the genetic algorithm. The proposed method performed better than the genetic algorithm.

3. METHOD

The SER system takes human speech as an input and transforms it into text then it analyzes this text, at last, it determines if the sentiments of the input statement are positive, negative, or neutral. The proposed method collects real-time voice input from users, converts it to text, and analyses whether the user's statement is positive, negative, or neutral. " SER is a process of recognizing human emotions. natural language processing and machine learning techniques are the driving force of SER. The source of input in SER is real-time human speech, in the proposed method an additional input option-text format is also considered. A 2-way input acceptor is in implementation. The input method can be explained as follows:

```
Case 1.
If Input_signal = < text>
 Then
 Apply Machine Learning Algorithm
 Detect emotions
End if
```

```
Case 2.
If Input_signal = < Voice>
 Then
 Convert Voice into Text
 Apply Machine Learning Algorithm
 Detect emotions
End if
```

To identify the best machine learning techniques (MLT), there are six MLT are tested on given dataset, named, LR, Linear SVC, valence aware dictionary for sentiment reasoning (VADER), Text Blob, and nave bayes. As mentioned earlier, in real time the speech signal is converted into text. Then further processing take place. Whereas, the user has a facility to directly provide the text which can be directly process to categories into any one category of the emotion. For text classification, a deep learning model is used. As shown in Figure 2, model is trained to determine the sentiments of the inputted text as positive, negative, or neutral. The following is a detailed description of the proposed approach.

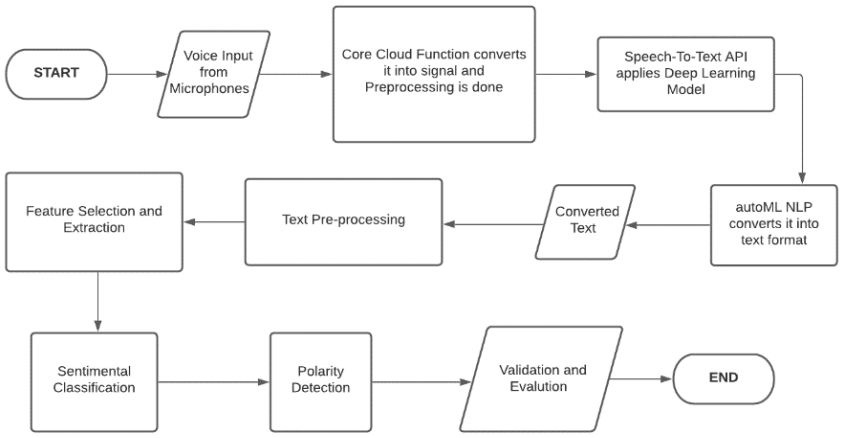

Figure 2. Proposed method

3.1. Speech to text

In this module, a deep learning model accepts the audio signal and convert the audio into the text. The Google API is used recognized the audio language. Using Google API, the speech signals are converted into electronic signals. The next step is pre-processing, in pre-processing, the electronic signals are normalized. next, the output of the pre-processing supplies to speech-to-text. The speech-to-text module employs the deep learning computational model. The text's context is provided by the deep learning module. Finally, the output of the speech-to-text provided to auto ML-NLP and auto ML-NLP generates the final text.

3.2. Text pre-processing

It includes data cleaning, suppressing words, tokenization, and stemming/lemmatization. In data cleaning, the symbols and punctuations are removed from the text. All the natural languages have some common words. These common words take valuable computation time in data pre-processing. These common words generally do not provide any meaning to the text. Hence, they can be avoided from the text. These common words are known as stop world. Suppressing words included removing the stop words from the provided text. The smallest unit of the text is the words. The words can further be divided into chunks, which are referred to as tokens in natural language processing (NLP). The tokens are a useful unit for the semantic analysis. The process of splitting the text into a token is known as tokenization. Finally, the words are normalized using stemming/lemmatization. The stemming/lemmatization provides the root words. These root words are further used in processing.

3.3. Feature extraction

In this module, the features are obtained from the text. The features are extracted using word bag method. In word bag method, the words are searched into the text and tagged to the corresponding class. The features of the text are extracted via text vectorization.

3.4. Sentiment classification using Naive Bayes classifier

Classification and predication are important aspects of ML. NB is simple and powerful algorithm. It is a classification technique that relies on the independent predictor assumption and the bayes theorem. It has two phases: a naïve period and a bayes phase. Even if these features are reliant on one another or on the existence of the other features, the NB classifier assumes that the presence of one feature in a class has no influence on the presence of any other feature.

The NB classification model is straightforward to construct and is particularly beneficial for very big datasets in probability theory and statistics-based theorems. It expresses the likelihood of an event based on the prior knowledge of possible confounding variables. The Bayes theorem uses conditional probability. The conditional probability is the likelihood of an event occurring in the presence of one or more other events. The proposed method is tested on dataset of 17K Twitter tweets.

3.5. Training and testing

The training and testing are done by using all 17K Twitter tweets. The NB classifier is trained on 55 percent of the dataset. Whereas, it has been tested on the remaining 45 percent of data. The advantage of the

NB classifier is that it is extremely fast and only requires one pass through the data. The NB classifier has the advantage of being extremely fast and just requiring one pass over the data.

4. RESULTS AND DISCUSSION

As discussed earlier, 17K Twitter tweets are considered for the training Naive-Bayes classifier. These 17K tweets are the text datasets. Apart from these, other texts, audio files, and real-time voice input from Kaggle.com are also considered for the SER.

4.1. Data collection

We must first collect input samples, which can be in the form of text, audio files, or real-time voice input, in order to implement the SER system. It will be used to hone the model's skills. Wav or mp3 files are commonly used as audio input. We used two datasets based on Twitter tweets that included positive, negative, and neutral reviews.

4.1.1 Analyzing dataset

The data of the dataset is labeled in three categories. These categories are positive, negative, and neutral. The labeling of the dataset is done on the basis of the words and the context of the words. NLP with NLTK is used to find the token and label the data. The pictorial representation of the labelled data is displayed in Figure 3.

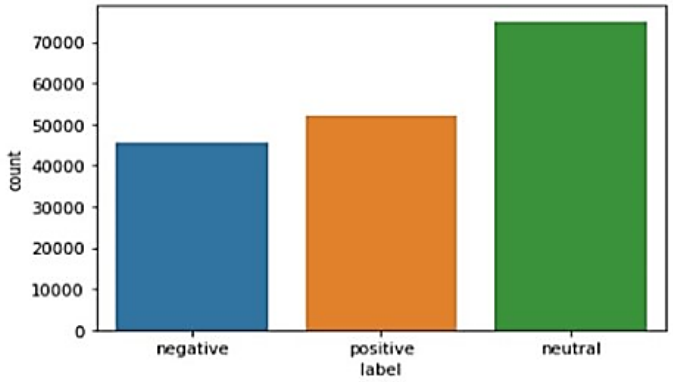

Figure 3. Dataset analysis

4.2. Training the model

Clear data, labeled data, and normalized data are the characteristics of a good database for speech emotion recognition. In the proposed method, sklearn model and split model of Python is used for the training. The training involves the following steps.

- Split the data into two sets: training and testing. The ratio of training and testing is 80 and 20 respectively.
- Extraction of features from the data i,e, converting textual contents into numerical logic.
- Removing the stop words, I,e data cleaning.
- Model fitting, Model fitting refers to generalizing of testing data with respect to training data by ML model.
- In the proposed method, there are three classifiers, named, multinomial Naive Bayes (MNB) model, LR model, and LSVM classifier model are used. The results of these classifiers are discussed in the next sections.

4.3. Multinomial Naive Bayes model

The proposed model employs the MNB Model. It has been tested in three classes. As per Figure 4, all three classes, the support factor is 34534. It shows, the data is well balanced. Therefore, it gives the satisfactory results in accuracy blends specificity and sensitivity, recall and precision. This model achieves an accuracy of 86% for dataset 1 as shown in Figure 4 and 87% for dataset 2 as shown Figure 5.

4.4. Logistic regression mode

The LR model is tested on two datasets. As per the results obtained, the data is balanced. The model's accuracy is 95% for both, dataset1 and dataset 2 as illustrated in Figures 6 and 7.

4.5. Linear support vector classifier model

As per other models, the linear support vector (LSV) classifier model is tested two datasets. On dataset-1, the model's accuracy is 96 percent, as shown in Figure 8. On dataset-2, the model's accuracy is 96 percent, as shown in Figure 9. Comparison of aforementioned methods are explained in further section.

4.6. Comparisons

The results of all the aforementioned classifiers are trained and tested on two different datasets. The summary of results is shown in Table 1. All the three classifiers are evaluated on four parameters, called, precision, recall, f1-score and accuracy. As per the results, it has been observed that linear support vector machine (LSVM) has outperformed all the other two classifiers. The precision value, recall value, F1-score and accuracy of Linear SVC are 96%. on both datasets. Whereas, the MNB has performed worst in experimentation. The graphical representation of accuracy is shown in Figure 8 and Figure 9.

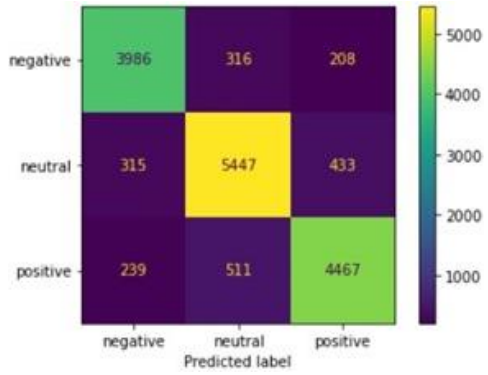

Figure 4. Multinomial Naïve Bayes on dataset-1

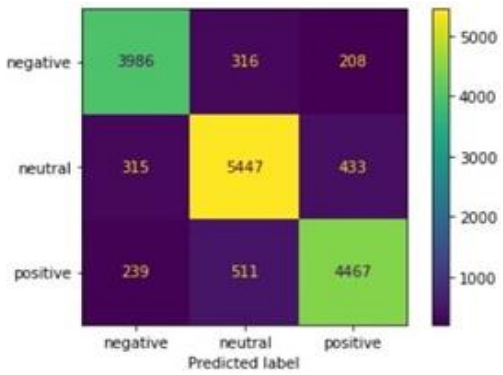

Figure 5**.** Multinomial Naïve Bayes on dataset-2

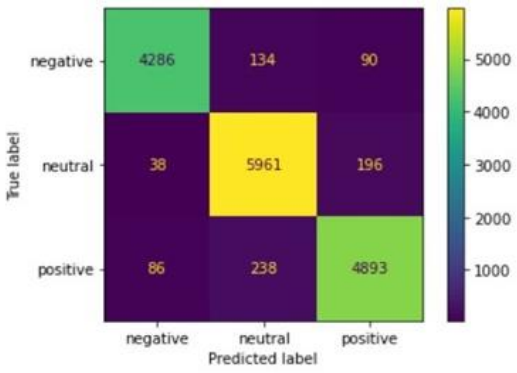

Figure 6. Logistic regression on dataset-1

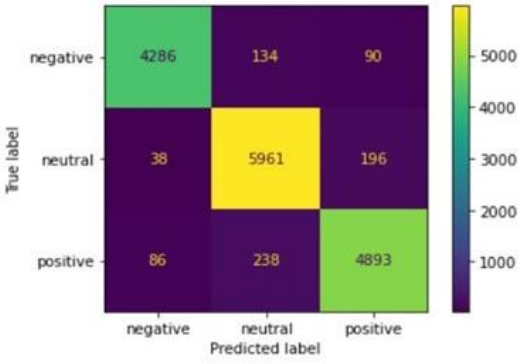

Figure 7. Logistic regression on dataset-2

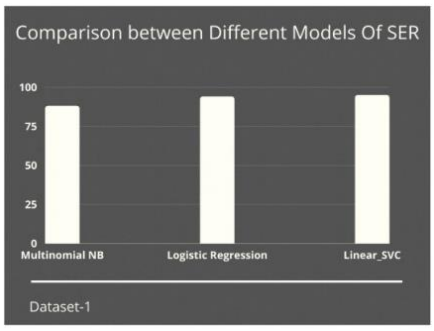

Figure 8. Graph depicting accuracy of all the three models on dataset-1

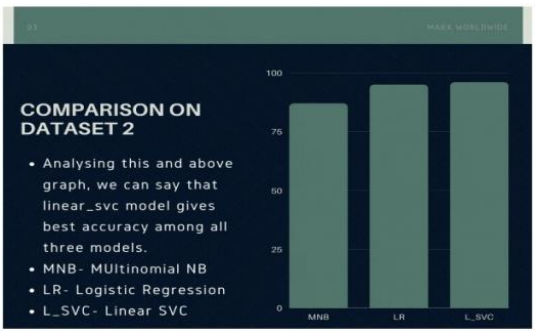

Figure 9. Graph depicting accuracy of all the three models on dataset-2

Table 1. Comparison table

Model	Precision	Recall	f1-Score	Accuracy
Multinomial NB on Dataset-1	86%	87%	87%	87%
Logistic Regression Model On dataset-1	95%	95%	95%	95%
Linear SVC on Dataset-1	96%	96%	96%	96%
Multinomial NB on Dataset-2	88%	88%	88%	87%
Logistic Regression Model On dataset-2	95%	95%	95%	95%
Linear SVC on Dataset-2	96%	96%	96%	96%

5. CONCLUSION

The analysis of various computational models that could be used as a classifier in SER is performed. Three different computational models named MNB, LR, and LSVM are trained and tested on two different datasets. The datasets consisted of video, audio, and text data. Eighty percent (80%) of test samples are used for training and twenty percent (20%) of samples are used for testing. In experimentation, it has been observed that MNB has an average accuracy of eighty seven percent (87%), LR has a ninty five percent (95%) average accuracy, and LSV has a ninety six percent (96%) average accuracy. Hence, LSVC is the most efficient of the three models and provides the highest level of emotion analysis accuracy.

REFERENCES

[1] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, May 2021, doi: 10.3390/electronics10101163.

[2] H. Jaakkola, B. Thalheim, and J. Henno, "About the essence of intelligence-will artificial intelligence (Ever) cover human intelligence?," 2022.

[3] A. Joshi and R. Kaur, "A study of speech emotion recognition methods," *International Jo urnal of Computer Science and Mobile Computing*, vol. 2, no. 4, pp. 28–31, 2013.

[4] M. S. Mahmud, M. U. Talukder, and S. M. Rahman, "Does 'Fear of COVID-19' trigger future career anxiety? An empirical investigation considering depression from COVID-19 as a mediator.," *The International journal of social psychiatry*, vol. 67, no. 1, pp. 35–45, Feb. 2021, doi: 10.1177/0020764020935488.

[5] G. Dreisbach, "Using the theory of constructed emotion to inform the study of cognition-emotion interactions," *Psychonomic Bulletin & Review*, Sep. 2022, doi: 10.3758/s13423-022-02176-z.

[6] H. M. Zakir and S. V. Jinny, "Convolutional neural networks method for analysis of e-commerce customer reviews," in *AIP Conference Proceedings*, 2022, vol. 2444, p. 030004, doi: 10.1063/5.0078372.

[7] L. J. Sheela, "A review of sentiment analysis in twitter data using hadoop," *International Journal of Database Theory and Application*, vol. 9, no. 1, pp. 77–86, Jan. 2016, doi: 10.14257/ijdta.2016.9.1.07.

[8] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment analysis: a comparative study on different approaches," *Procedia Computer Science*, vol. 87, pp. 44–49, 2016, doi: 10.1016/j.procs.2016.05.124.

[9] Kaggle, "Emotion negative positive neutral text data," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/talhaaljunaid170221/emotion-negative-positive-neutral-text-data. Aug. 2021. (Accessed: 4 Oct 22) .

[10] G. Mostafa, I. Ahmed, and M. S. Junayed, "Investigation of different machine learning algorithms to determine human sentiment using twitter data," *International Journal of Information Technology and Computer Science*, vol. 13, no. 2, pp. 38–48, Apr. 2021, doi: 10.5815/ijitcs.2021.02.04.

[11] K. Ganesan, "Build your first text classifier in python with logistic regression, / AI implementation, hands-on NLP, machine learning, text classification," *K. Ganesan*, 2020. [Online]. Available: https://kavita-ganesan.com/news-classifier-with-logistic-regression-in-python/#.Yct5z2hBy3A. Aug. 2021 (Accessed: 4 Oct 22).

[12] M. E. Valle, "Reduced dilation-erosion perceptron for binary classification," *Mathematics*, vol. 8, no. 4, p. 512, Apr. 2020, doi: 10.3390/math8040512.

[13] E. Bisong, "Introduction to scikit-learn," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkeley, CA: Apress, 2019, pp. 215–229.

[14] N. Sundarprasad, "Speech emotion detection using machine learning techniques," M.S. thesis, Dept. Comput. Sci., San Jose State Univ., San Jose, CA, USA, 2018.

[15] V. Odumuyiwa and U. Osisiogu, "A systematic review on hidden markov models for sentiment analysis," in *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, 2019, pp. 1–7, doi: 10.1109/ICECCO48375.2019.9043297.

[16] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Security Informatics*, vol. 4, no. 1, p. 9, Dec. 2015, doi: 10.1186/s13388-015-0024-x.

[17] N. Shakhovska, O. Basystiuk, and K. Shakhovska, "Development of the speech-to-text chatbot interface based on google API," *CEUR Workshop Proceedings*, vol. 2386, pp. 212–221, 2019.

[18] S. Maghilnan and M. R. Kumar, "Sentiment analysis on speaker specific speech data," in *2017 International Conference on Intelligent Computing and Control (I2C2)*, 2017, vol. 2018-Janua, pp. 1–5, doi: 10.1109/I2C2.2017.8321795.

[19] S. Nagdewani and A. Jain, "a review on methods for speech-to-text and text-to-speech conversion," *International Research Journal of Engineering and Technology*, no. May, pp. 4459–4464, 2020.

[20] D. Kumar and M. Rajaselvi, "Journal of speech to text conversion," *International Journal of Advance Research*, vol. 4, no. 2, pp. 867–872, 2018.

[21] S. N. Alsubari *et al.*, "Data analytics for the identification of fake reviews using supervised learning," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 3189–3204, 2022, doi: 10.32604/cmc.2022.019625.

[22] N. O. García, M. F. Díaz Velásquez, C. A. Tavera Romero, J. H. Ortiz Monedero, and O. I. Khalaf, "Remote Academic Platforms in Times of a Pandemic," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 21, p. 121, Nov. 2021, doi: 10.3991/ijet.v16i21.25377.

[23] O. I. Khalaf, G. M. Abdulsahib, H. D. Kasmaei, and K. A. Ogudo, "A new algorithm on application of blockchain technology in live stream video transmissions and telecommunications," *International Journal of e-Collaboration*, vol. 16, no. 1, pp. 16–32, Jan. 2020, doi: 10.4018/IJeC.2020010102.

[24] O. I. Khalaf and G. M. Abdulsahib, "Design and Performance Analysis of Wireless IPv6 for Data Exchange," *Journal of Information Science and Engineering*, vol. 37, no. 6, pp. 1335–1340, 2021, doi: 10.6688/JISE.202111_37(6).0008.
[25] O. I. Khalaf, G. M. Abdulsahib, and B. M. Sabbar, "Optimization of wireless sensor network coverage using the bee algorithm," *Journal of Information Science and Engineering*, vol. 36, no. 2, pp. 377–386, 2020, doi: 10.6688/JISE.202003_36(2).0015.

BIOGRAPHIES OF AUTHORS

Dr. Anu Saini received Btech, Mtech and PhD from reputed universities. She is an Assistant Professor in the Department of Computer Science and Engineering at G. B. Pant Govt. DSEU okhla I campus, New Delhi. Her research interests are in service oriented architecture, software engineering, big data, and recommendation systems. She has published more than 15 papers in the Journals and conferences of repute. She can be contacted at email: anuanu16@gmail.com.

Dr. Amit Ramesh Khaparde received Btech, Mtech and PhD from reputed universities. He is an Assistant Professor in Computer Science and Engg Dept. in G. B. Pant DSEU okhla I campsu. He is working in area of soft computing and machine translation. He has more than ten research publication various international and national journals and conferences. He can be contacted at email: khaparde.amit@gmail.com.

Ms. Sunita Kumari is currently working as an Assistant Professor in the Department of CSE, G. B. Pant DSEU Okhla-I Campus, Delhi, India. Her areas of interest are machine learning, databases, big data, operating systems, and natural language processing. She has several publications in SCOPUS/SCI-indexed International/National Journals and Conferences. She can be contacted at email: sunita.kumari@dseu.ac.in.

Dr. Salim Samsher is Associate Dean in NMIMS, Navi Mumbai. He has 22 years of experience. He has received his UG, PG and PhD from reputed universits. Dr. Shamsher has more then 20+ publications in reputed journals and confrences. He can be contacted at email: salim.shamsher@nmims.edu.

Jeevanandam Jotheeswaran is Senior Director-Centre for Faculty Development in GITAM University. He has 20+ industry and teaching experience. His research areas are cognitive science, data science and engineering, big data, social web mining and other disruptive technologies towards realising the vision of super intelligence. He can be contacted at email: jjothees@gitam.edu.

Dr. Seifedine Kadry has a Bachelor degree in 1999 from Lebanese University, MS degree in 2002 from Reims University (France) and EPFL (Lausanne), PhD in 2007 from Blaise Pascal University (France), HDR degree in 2017 from Rouen University. At present his research focuses on data science, education using technology, system prognostics, stochastic systems, and applied mathematics. He can be contacted at email: skadry@gmail.com.