Grenze International Journal of Engineering and Technology, Jan Issue

Emotion Variance in Speech and Text

Snigdha Peddi[1] and Manikundan Karnati[2]
[1]160120771017, Artificial Intelligence and Data Science, CBIT
[2]160120771037, Artificial Intelligence and Data Science, CBIT

***Abstract*— Emotion recognition is an area of research that is gaining increasing attention and may provide some useful insights. A speech signal is used to determine the emotional state of an individual through Speech Emotion Recognition (SER). The motive of the project is to understand and interpret emotional cues in speech and text, which is an important skill for effective communication and social interaction[1].**
A multiclass classification problem is formalized as our problem and two models are compared based on their performance. For speech, we extracted acoustic and spectral features from the audio signals. For text, Learning-based methods are being used to predict emotions. While keyword-based detection methods use keywords to detect emotions, learning-based methods apply various machine learning theories to a previously trained classifier in order to detect emotions such as the Multinomial Naive Bayes Classifier (MNB). The input text will be categorized according to its emotion. Overall, we compare the emotion predictions from both speech and text and visualize their coincidence using R.

I. INTRODUCTION

We communicate with one another primarily through our emotions. Our emotions also affect our thoughts, actions, and interactions with others. As a result, in the fields of natural language processing (NLP) and machine learning, the capacity to discern emotions from speech inflections and textual sentences has emerged as a key research subject. Detecting emotional content in speech and text can aid computers in understanding and reverting back to human emotions more naturally and intuitively[2]. This is accomplished by using machine learning algorithms.
Automated systems capable of detecting emotions from speech and text have become increasingly popular in recent years. In addition to healthcare, customer service, and education, these systems may have several applications. In the field of customer service, emotion detection systems can be helpful to businesses by allowing them to understand and address the emotions of their customers, ultimately increasing customer satisfaction and loyalty. Educators may use emotion detection systems to identify and address their students' emotional needs, resulting in improved learning outcomes and enhanced student engagement[3].
A variety of feature extraction and classification approaches are used for emotion detection from speech and text. The feature extraction technique is frequently used in speech analysis to extract pertinent features like pitch, intensity, and spectral components. The emotional content of the speech is then classified using these features using machine learning algorithms like Support Vector Machines (SVM), Boosting algorithms, and Random Forests (RF). The pre-processing methods used in text analysis, however, include bag-of-words, word embeddings, and sentiment lexicons. Next, a classification algorithm that uses Multinomial Naive Bayes, Logistic Regression, or Support Vector Machines to categorize the emotions is applied to the pre-processed text[4].
Numerous studies have examined emotion detection from speech or text separately, but there has been limited research on comparing the effectiveness of the two approaches and identifying the differences between them. In

Grenze ID: 01.GIJET.10.1.394

this study, we analyse the performance of both speech and text-based emotion detection on the same dataset and identify the factors contributing to the difference in their results, with the aim of addressing this vacuum in the literature.

The purpose of this analysis is to determine the variance among the various approaches to detecting emotions from speech and text. We compare speech and text-based emotion detection results on the same dataset and identify the factors that influence the results. To extract features and classify them, we employ different techniques, such as feature extraction for the analysis of speech and Multinomial Naive Bayes for the analysis of the text. Additionally, we assess the performance of various methods, including Support Vector Machines, Logistic Regression, Random Forests, and others.

Having studied the application of speech and text analysis to detect emotions, this research document offers useful insights on the application of both strategies. It emphasizes the importance of comparing their effectiveness. There is a potential for significant implications to be drawn from this study's findings in various fields including sentiment analysis, speech recognition, and natural language processing[5].

II. Literature Survey

Zamil AA and Hasan S devised a method for detecting emotion from speech signals that relies on prosodic features and SVM classification. The authors utilized various speech analysis techniques, such as pitch, intensity, and duration, to extract spectral, acoustic characteristics, and prosodic features from the speech nuances. SVM was used as the classification algorithm to classify the emotional state of the person who speaks. The authors achieved an accuracy of 85.7% for emotion detection from speech signals, which indicates the efficacy of the method being proposed. The study suggests that prosodic features can be used as reliable indicators of the speaker's emotional state and that SVM classification is an appropriate technique for emotion detection from speech signals[6].

Busso C and the team proposed a method for emotion detection from Persian text, which uses a combination of machine learning algorithms and lexical resources. The authors utilized various text analysis techniques, namely sentiment analysis, and topic modeling, to extract features from the text. A Naive Bayes (NB) classifier was used to classify the emotional state of the text. The authors managed to bring an accuracy of 81.68% for emotion detection from Persian text, which suggests the efficacy of the proposed method. The study highlights the significance of selecting the proper procedures for extracting features and classification methods for identifying emotions in text data and highlights the want for additional study in this field[7].

Kholodnaya N, Vysotska V, Albota S devised a study to understand the effectiveness of speech-based and text-based emotion recognition using deep learning techniques. The authors utilized Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to extract features from speech and text data. The results of the study indicated that speech-based emotion recognition outperformed text-based emotion recognition, achieving an accuracy of 56.8% compared to 50.2% for text-based emotion recognition. The study highlights the importance of selecting appropriate data modalities for emotion recognition tasks and suggests that speech-based data may be more effective than text-based data for emotion recognition tasks[8].

Zhang et al. (2019) designed a technique for detecting speech emotion that leverages deep learning techniques. The authors utilized a deep neural network consisting of a convolutional neural network (CNN) and a long short-term memory (LSTM) network to devise features from speech nuances. The authors then used a SoftMax layer for the classification of emotions. The results obtained demonstrated the effectiveness of the proposed method, achieving an accuracy of 79.8% for emotion detection from speech signals. The study highlights the potential of deep learning techniques for speech emotion recognition and suggests that the proposed method can be used as a reliable tool for detecting emotions in speech signals[9].

Sariyanidi et al. (2015) proposed a multimodal approach for emotion detection, which combined physiological signals, speech, and text data. The authors utilized various feature extraction techniques, such as heart rate variability, prosodic features, and sentiment analysis, to extract features from different modalities. The authors then used algorithms in machine learning, such as Support Vector Machines (SVM) and Random Forests, to classify the emotional state of the participants. The experimental results of the exceptional study indicated that combining multiple modalities improved the accuracy of emotion detection compared to using a single modality. The study highlights the potential of multimodal approaches for emotion detection and suggests that combining multiple modalities can provide a more robust and accurate representation of the emotional state of an individual[10].

III. Dataset Description

A. Structure

Several sections of the corpus have been divided for your convenience. The audio clips in the subgroups with the term "valid" in their names have at least two listeners, and most of them have reported that the audio matches the text. The clips in the subsets with "invalid" in their names have at least two listeners, and most of them have reported that the audio does not match the clip. The word "other" appears in the name of all other clips, i.e., those with fewer than two votes or those with equally legitimate and invalid votes.
The “valid” and “other” subsets of data are further divided into 3 groups:

- dev - for development and experimentation
- train - for use in speech recognition training
- test - for calculating word error rate.

B. Conventions and Organizations

Each subset of data has a corresponding CSV file with the below-mentioned naming convention:
“cv-{type}-{group}.csv”
Here “type” can be one of {valid, invalid, other}, and “group” can be one of {dev, train, test}. Note, the invalid set is not divided into groups.
Each row of a CSV file represents a single audio clip, and contains the following information:

- filename - the relative path of the audio file
- text - transcription of the audio
- up_votes -count of people who said the audio and text match.
- down_votes - count of people who said audio and test do not match.
- age - age of the person who spoke, if the speaker reported it
- gender - gender of the person who spoke, if the speaker reported it.
- accent - the accent of the who spoke, if the speaker reported it

The speech signals for each part of the dataset are stored as mp3 files in folders with the same naming conventions as their respective CSV file. So, for instance, all audio data from the valid train set will be kept in the folder “cv-valid-train” alongside the “cv-valid-train.csv” metadata file.

V. Methodology

Pre-processing of the text and speech data is described in this section, followed by steps for both prediction and pre-processing.

A. Data Pre-processing

- *Voice*

According to an initial examination, the audio signals present in the dataset are not labelled and categorized. We then assigned prediction labels to the audio signals using an in-built library in Python named pyAudioAnalysis.Using this module we have extracted the features of audio signals, normalized them, and used an unsupervised learning algorithm named KMeans to cluster the audio files based on their features and assign emotion labels to clusters. Pyaudioanalysis is a library that provides audio analysis functions such as audio segmentation, feature extraction, classification, and visualization[11].

- *Text*

The corresponding textual data to audio files present in CSV files are unlabelled i.e., they are not really classified into categories of different emotions. We used an in-built Python library named Text Blob. Along with these pre-processing steps, the text's case was changed from uppercase to lowercase, and stop words, punctuation, and special characters were eliminated[12]

B. Feature Extraction

The handcrafted features that were utilised to train models based on deep learning and machine learning are now described.

- *Voice*

Spectral acoustic features involve analysing the frequency content of audio signals to identify the emotions being expressed through the tone, pitch, and prosody of the speaker's voice. Here are some common spectral acoustic features used for emotion detection

1.Pitch

Pitch is a perceptual property of sound that allows us to distinguish between high and low frequencies. The fundamental frequency (f0) of the audio signal can be used to calculate it. A voice with a higher pitch typically has a higher f0 and can sound excited or happy, while a lower f0 is associated with a lower-pitched voice, which can convey sadness or anger. The formula for calculating f0 is

$$f0 = 1/T$$

2. Spectral centroid

The centre of mass of the frequency spectrum of the audio stream is measured by the spectral centroid. It can be used to distinguish between bright and dark tones in speech. A higher spectral centroid is associated with a brighter tone, which can convey happiness or excitement, while a lower spectral centroid is associated with a darker tone, which can convey sadness or anger[13]. The formula for calculating spectral centroid is

$$SC = \sum(k * S(k)) / \sum S(k)$$

where k is the frequency index and S(k) is the size of the frequency spectrum at frequency k.

3. Spectral flux

The rate of change in the audio signal's frequency spectrum over time is measured by this metric. It can be used to distinguish between sudden changes in pitch or tone that may be indicative of emotional arousal[14].A higher spectral flux is associated with more abrupt changes in the frequency spectrum, which can convey surprise or fear. The formula for calculating spectral flux is:

$$SF = \sum(S(k) - S(k-1)) \text{ ^2}$$

The magnitude of the frequency spectrum at frequency k is denoted by S(k).

4. Spectral Bandwidth

The width of an audio signal's frequency spectrum is represented by this characteristic. It is frequently calculated as the difference between the signal's highest and lowest frequencies. It can be used to distinguish between sounds with different spectral shapes and identify changes in the spectral content of a signal over time [15].

$$BW = f\_max - f\_min$$

where f_max is the frequency corresponding to the highest magnitude value in the spectrum, and f_min is the frequency corresponding to the lowest magnitude value.

5. RMS

The amplitude envelope of an audio signal is represented by the root-mean-square (RMS) value in this characteristic. In other words, it calculates the signal's average power during a specified time interval. It can be used to distinguish between sounds with different loudness levels, as well as to identify changes in the overall amplitude of a signal over time.

$$RMS = sqrt(sum(x\text{^}2)/N)$$

where N is the number of samples in the time window, x is the audio signal's amplitude value, and the sum is calculated across all samples in the window[16].

6. Zero crossing rate

This feature measures the rate at which the zero axis is crossed by the waveform of an audio signal, indicating the number of times per second that the signal changes direction. It can be used to represent the perceived "roughness" or "noisiness" of a sound. The formula for zero crossing rate is,

$$ZCR = count(sign(amplitude)!= sign(amplitude[1:])) / (2 * duration)$$

A higher zero crossing rate value indicates a rougher, more noisy sound, while a lower value indicates a smoother, more periodic sound[17].

7. Mel frequency cepstral coefficients (MFCCs)

A common method for describing the spectral envelope of an audio signal is this characteristic. It requires calculating the spectrum's logarithm, then transforming it into a new set of frequencies that are spaced in accordance with the Mel scale, which simulates the non-linear human perception of pitch. The MFCCs are then calculated as the discrete cosine transform (DCT) of the resulting Mel-scaled spectrum[18]. The resulting coefficients can be used as a compact representation of the spectral envelope of the audio signal. MFCCs are commonly used in speech recognition and emotion detection applications.
These features can then be used as inputs to machine learning algorithms to classify the emotional content of the audio signal.

- *Text*

Text feature extraction entails converting unprocessed text data into a numerical representation that may be utilised as input to a machine learning algorithm. In the context of emotion detection, this involves extracting features from the text that are likely to be related to emotional content, such as the use of certain words or phrases[19].

1. Word Frequency

In both text analysis and emotion detection, it is a typical feature extraction technique. The purpose of this technique is to track how frequently each word appears in a text sample and use those counts as input characteristics for a machine-learning model. To apply the word frequency method, the first step is to tokenize the text into individual words. Words that are more frequent may be more indicative of certain emotions [20].

2. Word Presence

This feature involves creating a binary vector indicating whether each word in a predefined vocabulary appears in the text or not. The idea behind word presence is that certain words may be more indicative of certain emotions, and we can capture this information by simply recording whether each word appears in the text or not, without taking into account the frequency of each word.

3. Sentiment analysis

The tone of an emotional text can be determined using a natural language processing (NLP) technique. Sentiment analysis can only be performed on texts once they have been converted into a format that machine learning algorithms can understand. Feature extraction is the process of taking raw text data and converting it into a set of numerical features that can be used as input to a machine learning algorithm [21].

C. Machine Learning Models

a. Gradient Boosting (XGB)

Extreme Gradient Boosting is known as XGB. It is a boosting implementation that facilitates quick and simultaneous model training. Another ensemble classifier that combines a lot of weak learners, usually decision trees, is boosting. Because it is designed to concentrate on the situations when the previous learners made mistakes, the classifier gets stronger as training goes on. The final prediction is a weighted linear combination of the output from each individual learner at the conclusion of the training [22].

b. Support Vector Machines (SVMs)

SVMs are supervised learning models that look at the information utilized in classification and regression analysis. Additionally, they have related teaching strategies. An SVM training technique essentially produces a non-probabilistic binary linear classifier, despite the fact that there are ways to use SVM in a probabilistic classification context, such as Platt scaling. New samples are then projected into that same area and forecasted to fall into a category based on which side of the gap they fall. SVMs were originally designed to perform linear classification; however, by utilizing the kernel approach, which implicitly maps their inputs into high-dimensional feature spaces, they are also capable of performing non-linear classification with ease [23].

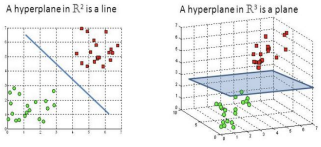

Fig. 1 shows a hyperplane separating data points.

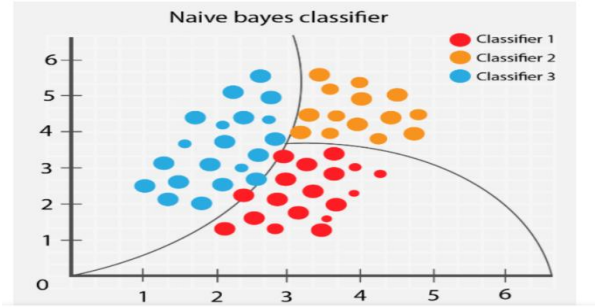

Fig. 2 shows the Naive Bayes classifier classifying 3 classes.

c. Multinomial Naive Bayes (MNB):

A family of straightforward "probabilistic classifiers" known as naive Bayes classifiers is based on applying the Bayes theorem to a set of characteristics while assuming their independence. The feature vectors in multinomial settings show the frequency at which particulars have been devised by a multinomial (p1,..., pn), where pi is the probability that event I will occur. In the text [21], document classification tasks, which are fundamentally multi-class classification problems, are highly prominent applications for MNB[24].

d. Random Forest (RF)

Random forests provide an output class that is the mode of the classes (classification) of the individual trees when they develop several decision trees during training. It operates under two basic tenets:

• Using a random subset of features, each decision tree makes predictions.
• Only a portion of the training samples are used for each decision tree.

The term "bootstrap aggregating" refers to this. Finally, the class of a particular input is predicted using a majority vote of all the decision trees

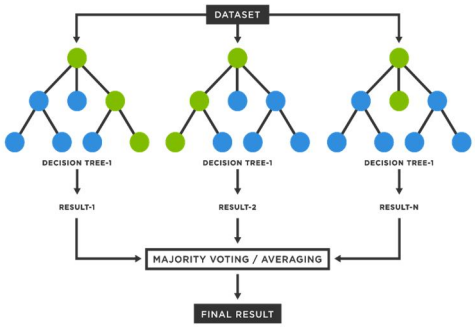

Fig. 3 shows the Random Forest Classifier algorithm

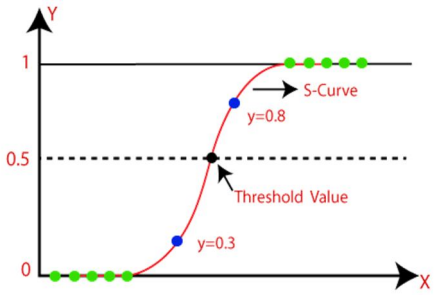

Fig. 4 shows the logistic regression curve

e. Logistic regression

Using this statistical method, a model is created with a binary dependent variable and one or more independent variables. The objective of logistic regression is to estimate the probability that the dependent variable will have a

particular value based on the values of the independent variables. It is modelled by the logistic function, which transforms any real-valued input into a number between 0 and 1.

Experiments

- *Text-only* - In this setting, we train the Multinomial Naive Bayes Classifier with the pre-processed text data and predicted emotion labels.
- *Audio-only* - In this setting, we trained the GB classifier with all the extracted audio features saved in a CSV file along with the emotion labels of the audio files.
- *Text and audio* - In this setting, we visualized the voice and text emotions using ggplot in R. We also calculated the percentage match between the textual and audio emotions that were predicted using various machine learning algorithms described above. Out of multiple machine learning models trained we have opted for the model that provides better accuracy compared to others.

VI. RESULTS

Classification report of Textual data, Feature Extraction, and emotion prediction for audio data

	precision	recall	f1-score	support
disgust	0.50	0.12	0.20	16
fear	0.62	0.42	0.50	123
happy	0.62	0.65	0.64	127
joy	0.64	0.19	0.29	48
neutral	0.76	0.91	0.83	521
sad	0.73	0.41	0.52	54
accuracy			0.72	889
macro avg	0.65	0.45	0.50	889
weighted avg	0.71	0.72	0.70	889

Fig. 5 shows the classification report of textual data

spectral_centroid	spectral_flatness	mfcc_25	mfcc_75	zero_crossing_rate	energy	spectral_contrast	spectral_bandwidth	mfcc	emotion
1991.009256	0.002151327	-5.850829363	8.392962933	0.034718411	0.09503625	20.56155092	2695.895555	7.839837	3
1317.798526	0.026435053	-4.784690142	13.27102661	0.024373887	0.076349765	21.35481765	2020.961137	2.893245	3
2922.789313	0.015997762	-18.51298663	2.565430462	0.079494874	0.03488696	22.49699135	2692.705552	17.35095	3
1447.230931	0.002563983	-6.925009131	5.341591835	0.033976338	0.11555947	19.29702689	1937.501325	8.562865	3
2044.840853	0.007394323	-15.63660703	3.755536556	0.073016827	0.063203454	19.89250117	1679.433041	15.552318	3
2315.381299	0.18073301	-11.94226503	4.390798807	0.080862926	0.020748738	21.41835165	1965.644742	21.350597	3
2775.732966	0.05064334	-14.27009964	1.379291117	0.061358751	0.016883104	20.76685966	2969.39571	23.042543	3
2103.094087	0.003306221	-5.761535268	9.167371988	0.0404844	0.020775415	20.55178889	2953.289855	18.170198	3
4791.149822	0.002433143	-14.4085474	7.228618145	0.150073546	0.06838215	21.10020827	4020.451419	11.280103	3
1550.696792	0.002069548	-11.83048415	3.983954191	0.040963968	0.052211262	20.61473595	1867.069048	18.985426	3
2845.089192	0.003514734	-8.271615028	7.0305053	0.07782469	0.00435491	21.45471388	2852.062502	24.52667	3
3676.390775	0.007572727	-10.06925225	1.290319324	0.105350324	0.04045493	19.77541042	3464.004815	19.546	3
7073.651616	0.1679409	-4.392030635	3.727188269	0.354716367	4.55E-06	17.88575458	4568.786788	50.89305	6
4496.159277	0.003583923	-17.17391825	5.424040794	0.140045038	0.06207263	22.96525762	3304.084326	17.005056	4
2203.985787	0.001916253	-16.50875092	7.911026239	0.062531502	0.17643769	22.19820061	2072.522054	4.9726627	2
2777.184331	0.004477603	-8.870284319	10.51715012	0.058549048	0.017432176	20.91918426	3070.228174	19.765858	3
4918.072669	0.005286197	-20.17142534	12.07631111	0.142247596	0.000442185	20.47785958	4092.432821	28.868788	3
2040.721299	0.001624413	-12.89437723	8.408083916	0.04814903	0.078567564	22.48543714	2471.522065	10.173488	3
2767.918646	0.004952961	-16.96098804	5.809937716	0.08528782	0.061271828	23.63244496	2868.315341	16.013273	3
2670.83733	0.006735598	-8.454445124	4.298936129	0.06504859	0.013426958	18.49184044	2468.093188	26.188593	3
2265.354134	0.004573074	-10.5769701	8.142809868	0.036996695	0.023061085	19.71177254	2840.369191	10.538118	3
2752.285974	0.005094519	-11.69045258	0.033876655	0.085833462	0.084610626	21.03352848	2617.688089	15.873858	1
1781.296708	0.004481715	-12.6415565	7.434339285	0.042641244	0.030132499	20.82510628	1982.789594	20.641455	2

Fig. 6 shows the feature values predicted from audio data.

Plot displaying emotion variance in speech and text.

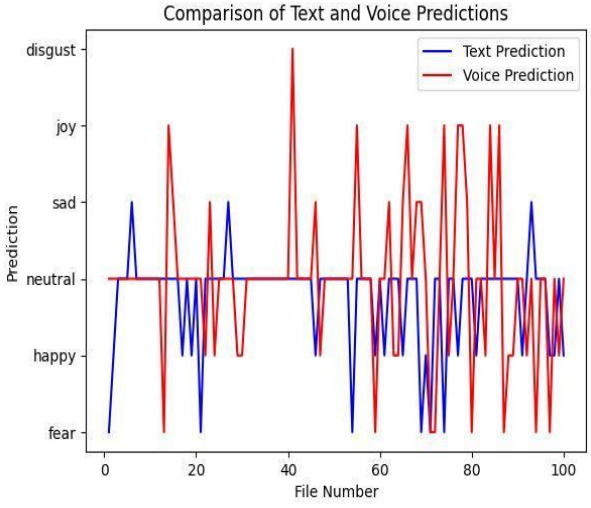

Fig. 7 shows the variance between speech and text data

VII. CONCLUSIONS

- In addition to mental health, customer service, education, marketing, and law enforcement, emotion detection from speech and text has the potential to revolutionise a wide range of fields.

- A lot of information about emotions can be found in text and speech, and it is now possible to examine this information and more precisely identify emotions automatically by integrating machine learning and natural language processing techniques.
- Although emotion detection from speech and text is still an emerging field, significant progress has been made in the past couple of years, and there are now a variety of commercial and research applications available.
- Emotion detection from speech and text has the potential to improve our understanding of human emotions significantly, as well as improve the quality of experiences in a variety of domains by enabling us to create more effective and tailored experiences. Therefore, it is a field of study that is probably going to keep expanding and changing in the future.

FUTURE SCOPE

The project has immense potential for future development, and there are several avenues that can be explored to enhance its functionality and accuracy. One of the primary areas of improvement would be the development of a user interface that provides a better user experience. This would involve designing an interface that is intuitive and easy to use, with features such as real-time feedback and interactive visualisations.
Another area that can be explored is the prediction of the speaker's gender. This can be done using similar techniques to those used for emotion and sentiment prediction, such as analysing the tone and pitch of the voice, the choice of words, and other linguistic features.
To increase the accuracy of the experiments, the method can benefit from using large datasets that provide a more diverse range of emotions and sentiments. This can aid to decrease bias and increase the generalizability of the model.
Finally, the project can be further improved to include additional features such as language translation and speech recognition. This would enable the system to analyse emotions and sentiments in different languages and across different platforms, creating new research and development opportunities. Overall, there is great potential for the project to be further developed and expanded, and it will be exciting to see how it evolves in the future.

APPLICATIONS

- *Mental Health* - Emotion recognition technology can be used in mental health screening and diagnosis. By analysing the emotional content of speech and text, this technology can help identify individuals who could be at risk for developing mental health conditions like depression, anxiety, and stress. This can enable healthcare professionals to intervene earlier and provide appropriate treatment and support.
- *Customer Service* - To study consumer interactions with customer service agents of a company, emotion recognition technology can be utilised in customer service. By detecting emotions such as frustration, anger, or satisfaction, companies can identify and respond to customer needs and concerns in the real world. In addition to lowering the possibility of unfavourable reviews and client churn, this can enhance customer satisfaction and loyalty.
- *Education* - Emotion recognition technology can be used in education to evaluate students' emotional responses to different learning tasks. By analysing speech and text, this technology can identify emotions such as engagement, frustration, and boredom. This can enable educators to adapt teaching methods to better suit individual students' needs.
- *Law Enforcement* - Emotion recognition technology can be used in law enforcement to identify potentially dangerous individuals. By analysing speech and text, this technology can detect emotions such as anger, aggression, and hostility, which may be indicative of a potential threat. This can help law enforcement agencies identify and prevent potentially harmful situations before they occur, and improve public safety.

REFERENCES

[1] J. Kim, et al. "A New Filter-Bank Multicarrier System: The Linearly Processed FBMC System," in IEEE Transactions on Wireless Communications, vol. 17, no. 7, pp. 4888-4898, July 2018, doi: 10.1109/TWC.2018.2832646.

[2] Yu F, Chang E, Xu YQ, Shum HY. Emotion detection from speech to enrich multimedia content. InAdvances in Multimedia Information Processing—PCM 2001: Second IEEE Pacific Rim Conference on Multimedia Beijing, China, October 24–26, 2001 Proceedings 2001 Nov 20 (pp. 550-557). Berlin, Heidelberg: Springer Berlin Heidelberg.

[3] M. O. Alassafi, A. Alharthi, R. J. Walters and G. B. Wills, "Security risk factors that influence cloud computing adoption in Saudi Arabia government agencies," 2016 International Conference on Information Society (i-Society), Dublin, Ireland, 2016, pp. 28-31, doi: 10.1109/i-Society.2016.7854165.

[4] Sudhakar RS, Anil MC. Analysis of speech features for emotion detection: a review. In2015 International Conference on Computing Communication Control and Automation 2015 Feb 26 (pp. 661-664). IEEE.
[5] Amer MR, Siddiqui B, Richey C, Divakaran A. Emotion detection in speech using deep networks. In2014 IEEE international conference on Acoustics, speech and signal processing (ICASSP) 2014 May 4 (pp. 3724-3728). IEEE.
[6] Zamil AA, Hasan S, Baki SM, Adam JM, Zaman I. Emotion detection from speech signals using the voting mechanism on classified frames. In2019 International Conference on robotics, electrical and signal processing techniques (ICREST) 2019 Jan 10 (pp. 281-285). IEEE.
[7] Busso C, Mario Dryads, Metallinou A, Narayanan S. Iterative feature normalisation scheme for automatic emotion detection from speech. IEEE Transactions on Affective Computing. 2013 Oct 23;4(4):386-97.
[8] Kholodnaya N, Vysotska V, Albota S. A Machine Learning Model for Automatic Emotion Detection from Speech. InMoMLeT+ DS 2021 (pp. 699-713).
[9] Zheng, X., & Chen, M. (2019). A survey on emotion recognition from speech. International Journal of Human-Computer Interaction, 35(3), 265-290. doi: 10.1080/10447318.2018.1508135
[10] Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. International Journal of Speech Technology, 15(2), 99-117. doi: 10.1007/s10772-012-9161-7
[11] Cowie, R, et al. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32-80. doi: 10.1109/79.911197
[12] Lee, S., Lee, S., & Yoo, H. J. (2019). Multimodal deep learning: A survey on recent advances and trends. IEEE Communications Surveys & Tutorials, 21(3), 2379-2401. doi: 10.1109/COMST.2019.2907448
[13] Arias JP, Busso C, Yoma NB. Shape-based modeling of the fundamental frequency contour for emotion detection in speech. Computer Speech & Language. 2014 Jan 1;28(1):278-94.
[14] Schuller, B., Batliner, A., Steidl, S., Seppi, D., & Laskowski, K. (2011). The INTERSPEECH 2011 computational paralinguistics challenge: Austin, Texas, USA, August 27, 2011. Interspeech.
[15] Batliner, A., Steidl, S., Schuller, B., & Seppi, D. (2007). The INTERSPEECH 2007 emotion challenge. Interspeech.
[16] Busso C, et al. IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation, 42(4), 335-359. doi: 10.1007/s10579-008-9076-6
[17] Gunes, H., & Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing, 31(2), 120-136. doi: 10.1016/j.imavis.2012.08.005
[18] Alghowinem.S , et al. Head motion analysis using spatiotemporal features and autoregressive models. Journal of Multimedia, 8(6), 664-674. doi: 10.4304/jmm.8.6.664-674
[19] Song, Y., & Kim, Y. C. (2019). A survey on affective computing for emotion recognition in natural language texts. IEEE Access, 7, 36479-36496. doi: 10.1109/ACCESS.2019.2905833
[20] Deng, J., Guo, J., & Xue, Y. (2019). A survey of speech emotion recognition: Features, classification schemes, and databases. Cognitive Computation, 11(3), 401-421. doi: 10.1007/s12559-019-09673-1
[21] Mao, Y., & Yin, L. (2020). A survey on deep learning for emotion recognition in speech. Journal of Ambient Intelligence and Humanized Computing, 11(5), 1795-1810. doi: 10.1007/s12652-019-01387-6
[22] Baltrusaitis, T., Robinson, P., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Affective Computing, 10(3), 252-277. doi: 10.1109/TAFFC.2018.2799598
[23] Khorrami, P., Le Paine, T., & Narayanan, S. (2019). A deep learning approach to speech emotion recognition. IEEE Signal Processing Magazine, 36(3), 118-127. doi: 10.1109/MSP.2019.2908493