

**Project 2 - Milestone 3- Improve Call Center Effectiveness Using Machine Learning**

Guruprasad Velikadu Krishnamoorthy

College of Science and Technology, Bellevue University

DSC680-T301: Applied Data Science

Professor. Amirfarrokh Iranitalab

July 21, 2024

## **Improve Call Center Effectiveness Using Machine Learning**

### **1. Project Description:**

The project aims to build a machine-learning model for a Call center to understand the call's sentiment by analyzing the employees' facial expressions and the Customer's tone, thus providing real-time suggestions on the screen to improve the overall call efficiency and offer better customer experience. Two Convolutional Neural network models were built using Image and Speech emotion recognition algorithms. The results from these models were combined to make the appropriate choice of providing suggestions to improve the overall call effectiveness.

### **2. Background:**

When customers contact the help desk for support, they are forced to navigate through a series of options to collect the details through the Interactive Voice Response (IVR) systems before talking to the live agent. Sometimes, the wait times can be too long, which can annoy the already frustrated customer. When the agent cannot help the customer or keeps the customer on hold for too long while trying to find answers, it can lead to more frustration and, in some cases, losing the customer.

As many corporations have outsourced their call center jobs overseas, understanding the foreign accent on the phone can be a huge barrier for some, which can quickly escalate the situation, especially after waiting on the phone for long to talk to a person.

Conversely, staffing issues can pose a considerable challenge for the call centers. Employee turnover is much higher in the call center industry compared to others due to demanding workloads, odd hours, and repetitive and fast-paced workstyles. Also, it is worth mentioning that

understaffed call centers only worsen the situation by increasing employee burnout, reducing productivity, and thus reducing the quality of service overall. Training employees can be expensive and time-consuming; hence, losing them comes with a hefty price tag, including losing all the knowledge and talent and rehiring and training new employees. (Solutions, 2023)

### **3. Business Problem:**

One of the significant challenges that most corporations face nowadays is maintaining a good balance between employee satisfaction and providing a better customer experience. This is particularly evident in customer-facing industries such as call centers. Though this can be handled by effectively training the employees, aiding on the field with real-time suggestions on the screen can help de-escalate the situation and reduce friction, thus improving the overall customer experience. If the employee appears to be stressed out, the Model can provide real-time suggestions on the screen, such as the “*phrases to use*,” “*words to avoid*,” “*Need additional help?*” etc.

### **Why is sentiment analysis based on text classification not the best solution for call centers?**

The Sentiment analysis is usually performed by extracting the text and building a Natural Language model to understand the underlying sentiments. These models can generally predict if the overall sentiment is Positive, Negative, or Neutral. This can be a perfect solution for Chatbots or the customer service provided through chat, such as Live chats. However, the Call centers operate entirely differently, where they interact with the customers on the phone. So, it is essential to understand the emotions expressed by the customer to provide better customer service. Hence, sentiment analysis based on text classification methods may not be an optimal solution for call centers, although it can supplement emotion recognition models.

Two Neural Network models were built in this project. One was used to analyze the employees' sentiments by examining the pictures using Image Classification algorithms. The other algorithm was to investigate the tone of speech to understand the emotion using Speech emotion recognition algorithms.

#### **4. Data Sources:**

For the image recognition part of the project, the data was downloaded from the web using the Python library “bing-image-downloader.” About 100 Images of happy and stressed employees were downloaded to study their facial emotions and train the Model.

The RAVDESS dataset from Kaggle was used for the speech emotion recognition part of the project. This dataset had 1440 files recorded by 12 Male and 12 female actors displaying emotions such as Fear, pleasant surprise, sadness, Anger, etc. These samples were available in the .wav file format. (Livingstone, 2018)

#### **5. Methods and Analysis:**

##### **Employee Sentiment Analysis Using Image Classification Model**

##### **5.1. Data Cleansing and Exploration:**

The images downloaded from the web may be of varying sizes and file types. In this project, the jpeg, jpg, png, and bmp files were processed, and the rest were ignored. Also, only files larger than 10KB were processed, and smaller files were removed. After data cleansing, sample images were validated using the “skimage” Python library as shown in the Figure below.



*Figure 1: Sample Images after Data Cleansing*

## 5.2. Data Preparation and Model Building:

The images were converted into Numpy arrays using the skimage library as Machine Learning requires numeric inputs. During this step, all images were resized to a standard size of 256x256 and rescaled between 0 and 1 for further processing by Machine learning models. There were about 100 images for each Target class, so a batch size of 20 Images was chosen. The dataset was split into Train, Validation, and Test sets in the 60, 20, and 20 ratios.

Two models were created for the Image classification. The first was a Convolutional Neural Network model using TensorFlow and Keras Sequential API. As shown in the Figure below, several Hidden layers were added to the Neural Network using Conv2D and MaxPooling2D. Rectified Linear Activation function, commonly known as "ReLU," was used in the hidden layers. In contrast, the "sigmoid" function was used in the final output layer, as the Target variable had only two classes- A Happy employee and a Disappointed Employee. The summary of the Model is shown below.

```
[ ] image_model1 = Sequential()

time: 18.3 ms (started: 2024-07-11 00:17:20 +00:00)

[ ] image_model1.add(Conv2D(16, (3,3), 1, activation='relu', input_shape=(256,256,3)))
image_model1.add(MaxPooling2D())
image_model1.add(Conv2D(32, (3,3), 1, activation='relu'))
image_model1.add(MaxPooling2D())
image_model1.add(Conv2D(16, (3,3), 1, activation='relu'))
image_model1.add(MaxPooling2D())
image_model1.add(Flatten())
image_model1.add(Dense(256, activation='relu'))
image_model1.add(Dense(1, activation='sigmoid'))
```

```
image_model1.summary()

Model: "sequential"
```

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 254, 254, 16)	448
max_pooling2d (MaxPooling2D)	(None, 127, 127, 16)	0
conv2d_1 (Conv2D)	(None, 125, 125, 32)	4640
max_pooling2d_1 (MaxPooling2D)	(None, 62, 62, 32)	0
conv2d_2 (Conv2D)	(None, 60, 60, 16)	4624
max_pooling2d_2 (MaxPooling2D)	(None, 30, 30, 16)	0
flatten (Flatten)	(None, 14400)	0
dense (Dense)	(None, 256)	3686656
dense_1 (Dense)	(None, 1)	257

```

=====
Total params: 3696625 (14.10 MB)
Trainable params: 3696625 (14.10 MB)
Non-trainable params: 0 (0.00 Byte)
=====
time: 44 ms (started: 2024-07-11 00:17:21 +00:00)
```

*Figure 2: CNN Image Classification Model and Summary*

The second Support Vector Machine Classification model was built using the Scikit Learn Library. For this Model, hyperparameter tuning was done by evaluating various parameters before training the Model. The outcome is shown in the Figure below.

```
from sklearn.model_selection import GridSearchCV
from sklearn import svm
param_grid_svc=[
    {"C": [1,10,100,1000], "kernel": ["linear"]},
    {"C": [1,10,100,1000], "gamma": [0,0.001,0.0001], "kernel": ["rbf"]}
]
svc=svm.SVC(probability=True)
image_model2=GridSearchCV(svc,param_grid_svc)
image_model2.fit(x_train, y_train)
```

GridSearchCV

GridSearchCV(estimator=SVC(probability=True),  
param\_grid=[{'C': [1, 10, 100, 1000], 'kernel': ['linear']},  
{'C': [1, 10, 100, 1000], 'gamma': [0, 0.001, 0.0001],  
'kernel': ['rbf']}])

estimator: SVC

SVC(probability=True)

SVC

SVC(probability=True)

time: 21min 37s (started: 2024-07-11 00:42:06 +00:00)

Figure 3: SVM Image Classification Model

### 5.3. Evaluation of the Image Classification models:

The model performances were measured by calculating and plotting the Accuracy and Loss functions as shown in the figures. Accuracy represents the percentage of predictions where the predicted value is the same as the actual value. From the plot, the Accuracy appears to increase as the number of Epochs increases and then starts to flatten. The Loss, which represents the sum of errors made in each sample, decreases as the number of epochs increases.

The confusion matrix that compares the actual values versus predicted values shows that the SVM model could predict about 80% of the values correctly.

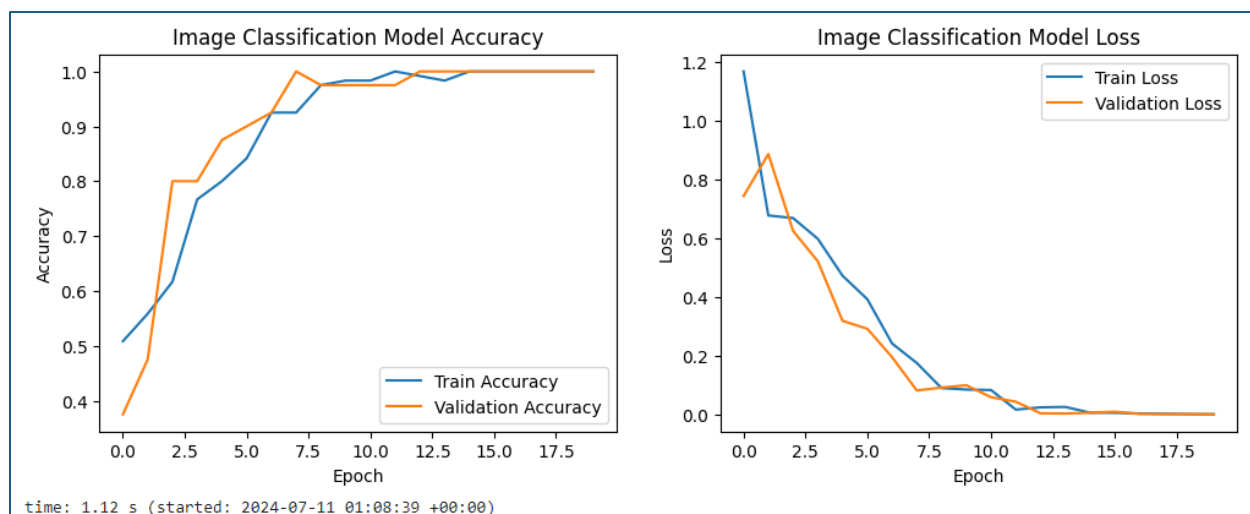


Figure 4: Accuracy and Loss function plots of Image Classification Model

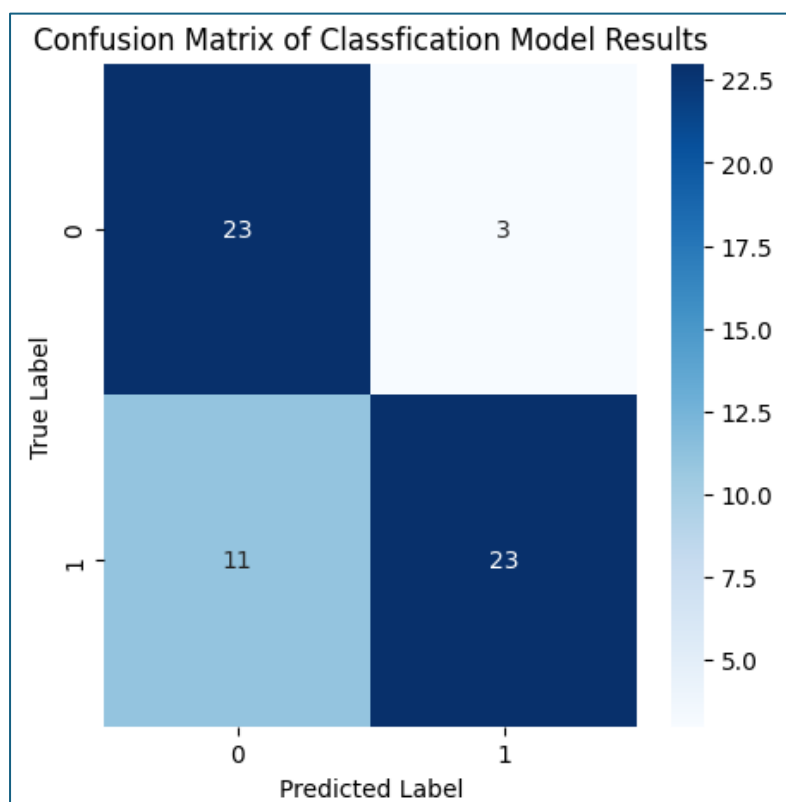


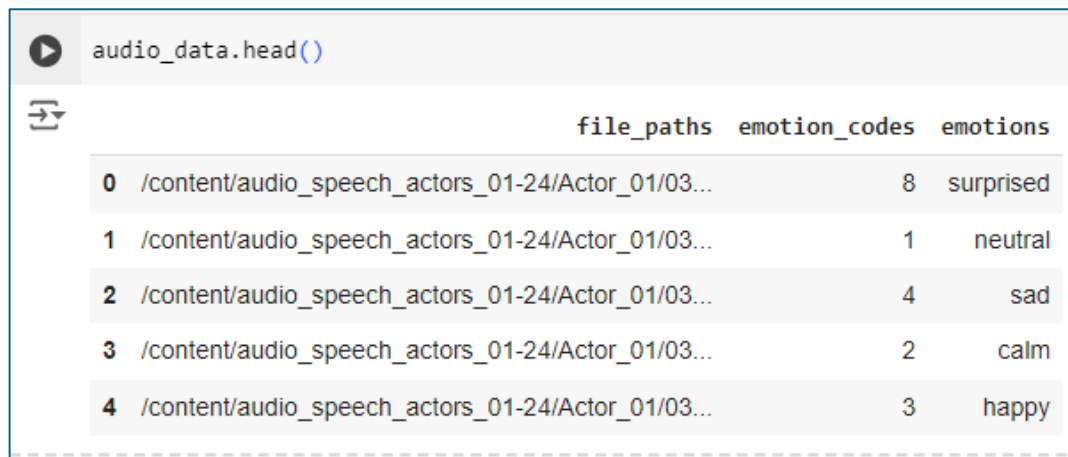
Figure 5: Confusion Matrix of Image Classification Model



## Speech Emotion Recognition Model to Analyze the Customers' Emotions

### 5.4. Data Exploration and Preparation:

The audio files taken from the RAVDESS dataset did not require any cleaning as the emotions were recorded by various voice actors displaying emotions such as fear, anger, happiness, disgust, sadness, etc. The emotion and other details of the Audio were encoded in the file name, so they were extracted and stored in a data frame with each row representing the file's location on the drive, emotion code, and emotion label, as shown in the Figure below.



The screenshot shows a Jupyter Notebook cell with the command `audio_data.head()` executed. Below the command, a table displays the first five rows of the data frame. The table has three columns: `file_paths`, `emotion_codes`, and `emotions`. The rows are indexed from 0 to 4.

	file_paths	emotion_codes	emotions
0	/content/audio_speech_actors_01-24/Actor_01/03...	8	surprised
1	/content/audio_speech_actors_01-24/Actor_01/03...	1	neutral
2	/content/audio_speech_actors_01-24/Actor_01/03...	4	sad
3	/content/audio_speech_actors_01-24/Actor_01/03...	2	calm
4	/content/audio_speech_actors_01-24/Actor_01/03...	3	happy

Figure 6: Data Frame showing the top 5 rows of extracted Audio data

The bar plot in the Figure below indicates the dataset's emotions distribution.

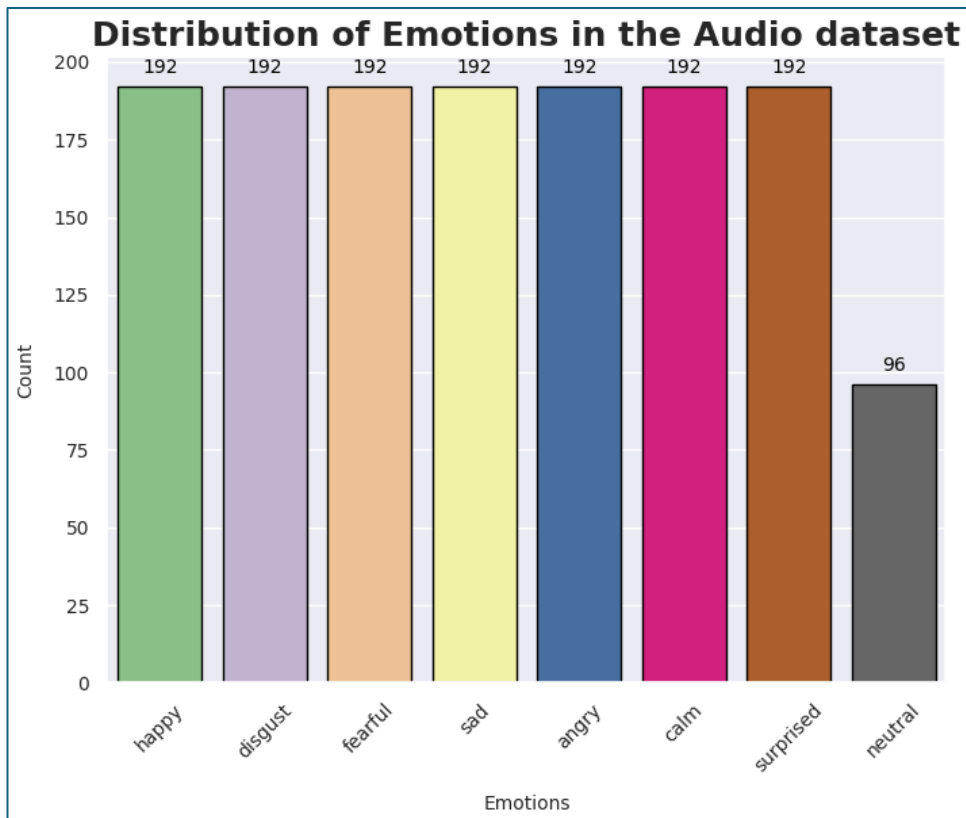
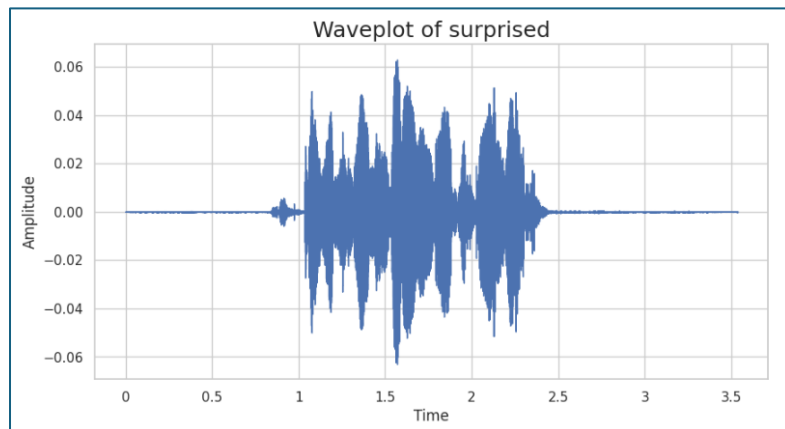
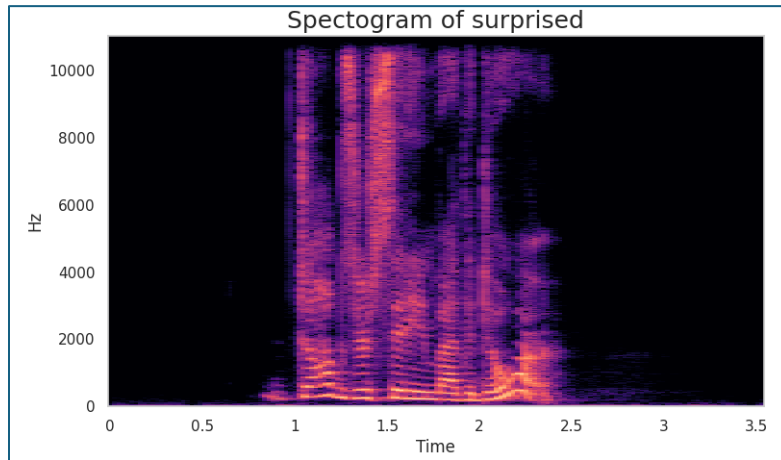


Figure 7: Bar Plot showing the Distribution of Emotions data

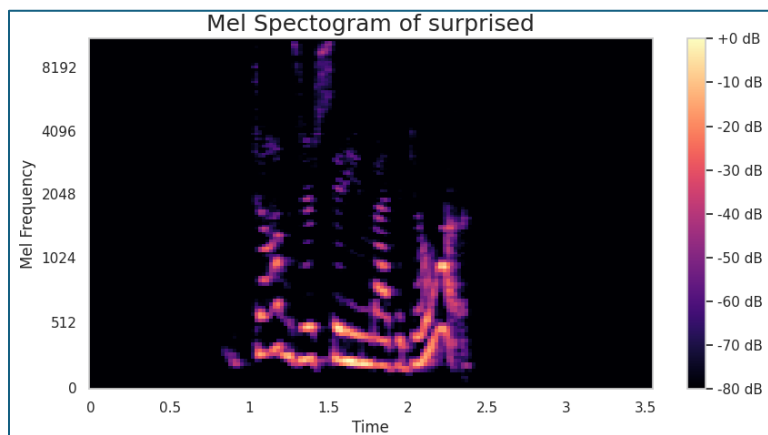
Librosa Python Library is widely used in the project to convert the audio files into the Numpy array required for Machine Learning model building. Further analysis of the speech files was performed by plotting the Wave plot, Spectrogram, and Mel Spectrogram using methods available in the Librosa library.



*Figure 8: Sample Wave plot*



*Figure 9: Sample Spectrogram*



*Figure 10: Sample Mel Spectrogram*

### **5.5. Feature Extraction and Model Building:**

Various features such as Zero Crossing Rate, Root Mean Square, Mel-Frequency cepstral coefficients, short-term Fourier transforms of Chroma Variants, Spectral contrast, and Mel-Spectrogram were extracted for each speech file, and the resulting NumPy array of features were flattened and combined to form the final array of features. Further Noise and Pitch were added for each speech file, and the resulting NumPy arrays were combined to form a final input for Machine

Learning Model building. The combined array was then standardized using a standard scaler and label encoded before being split into Train and Test sets.

A Convolutional Neural Network Model was built using TensorFlow, which uses Keras Sequential API. Like the Image classification model, several Hidden layers were added to the Neural Network using Conv2D and MaxPooling2D, as shown in the Figure below. The Rectified Linear Activation function, commonly known as "ReLU," was used in the hidden layers. In contrast, the "softmax" function was used in the final output layer as there were multiple classes in the Target variable. The Figure below shows the Model definition and the output of Training the Model.

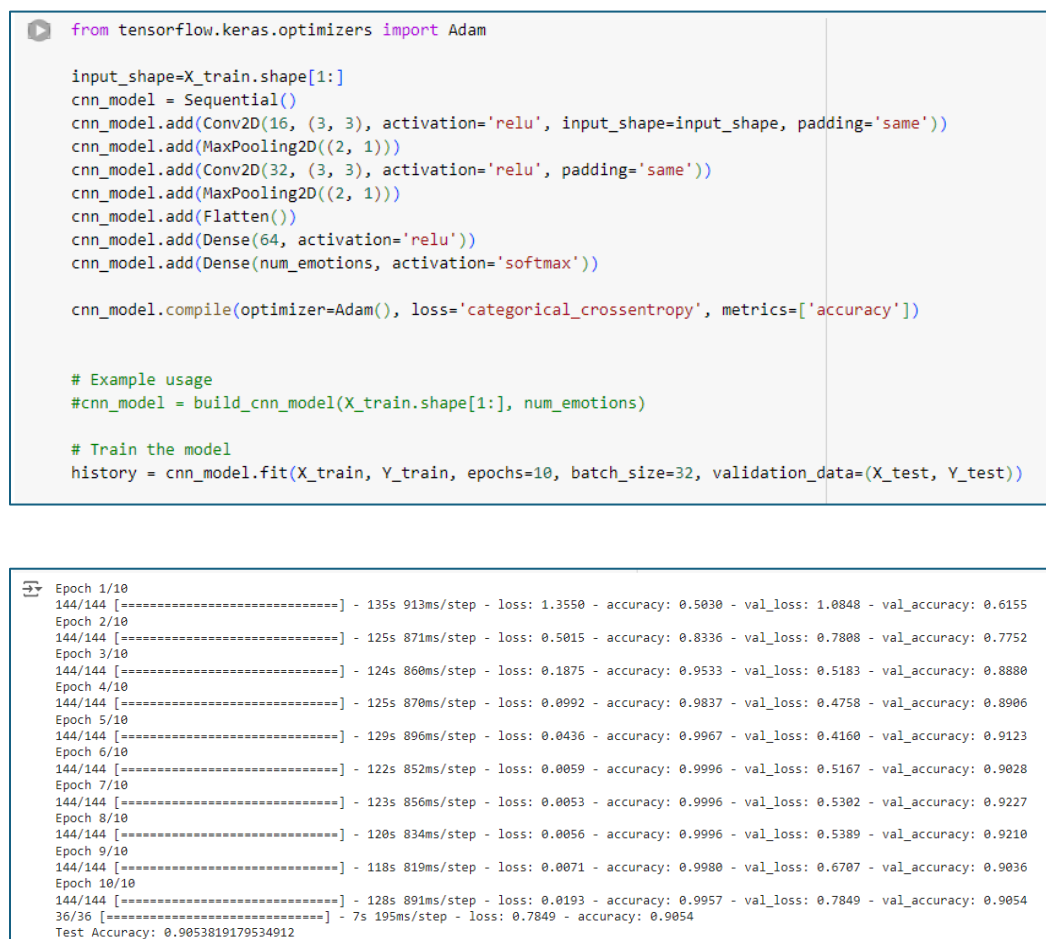


Figure 11: CNN Speech Emotion Classification Model

### 5.6. Evaluation of the Speech Emotion Classification models:

The Speech Emotion model was evaluated by plotting the Loss and Accuracy models and the Confusion Matrix. Like the Image models, in this Model, the Accuracy appears to increase as the number of Epochs increases and then starts to flatten. The Loss, which represents the sum of errors made in each sample, decreases as the number of epochs increases.

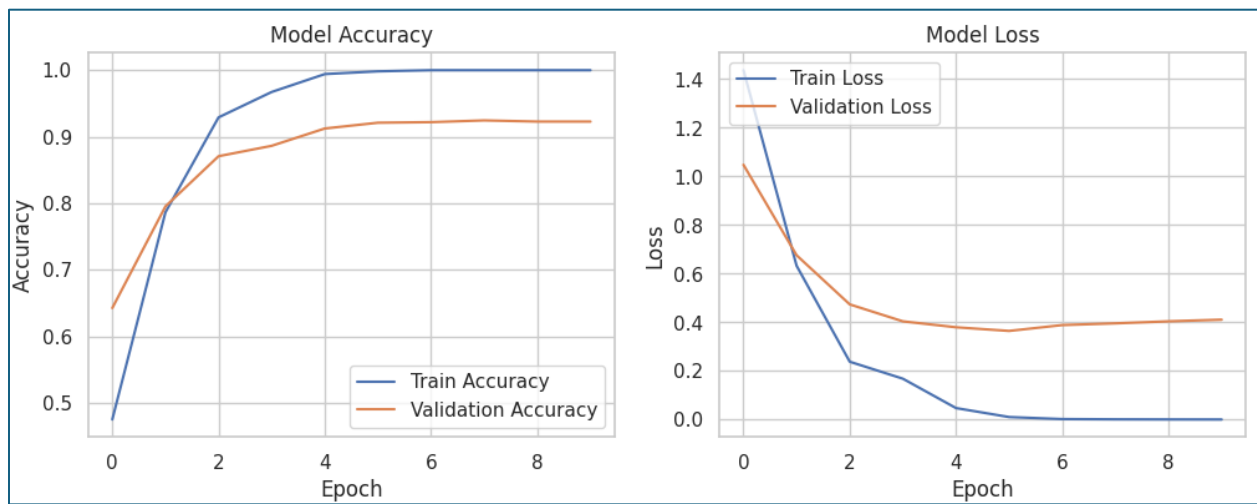


Figure 12: Accuracy and Loss function plots of Speech Emotions Classification Model

The Confusion Matrix, commonly used to describe the Classification model performance, suggests that the Model predicted most values correctly.

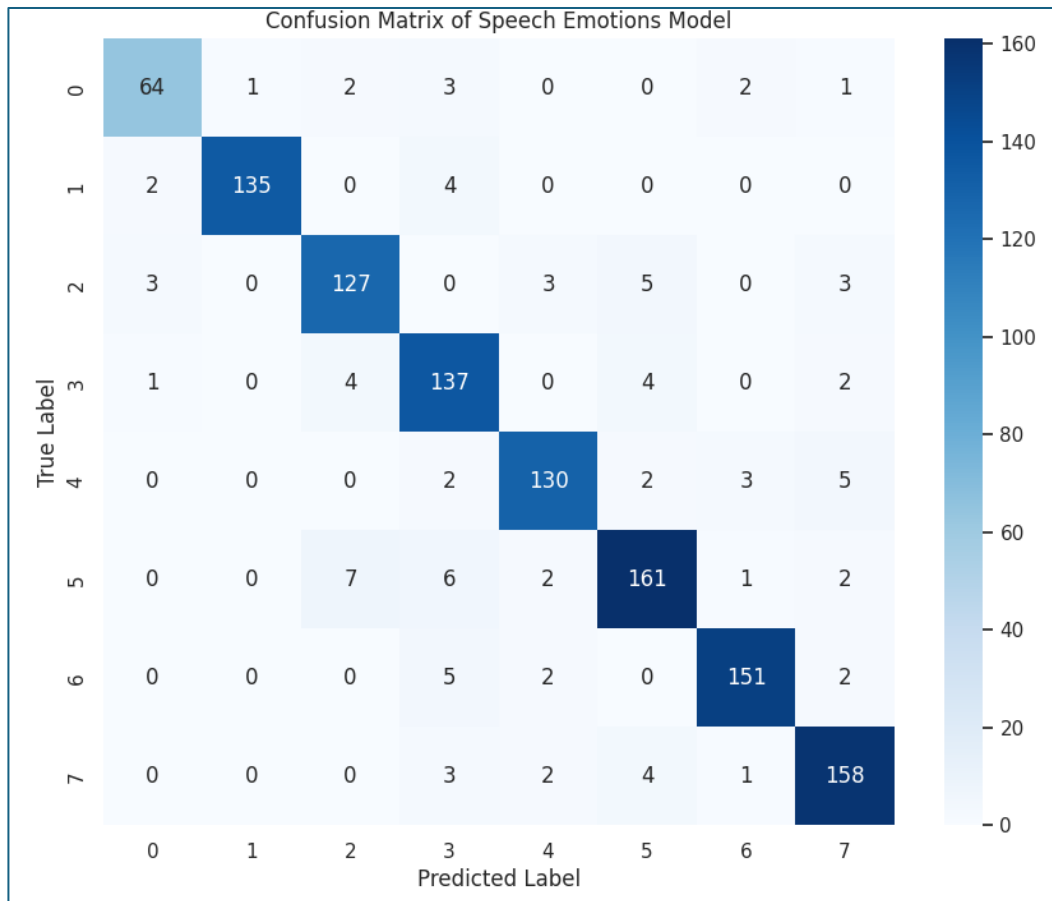


Figure 13: Confusion Matrix of Speech Emotions Classification Model

### 5.7. Validation of Model Results:

To validate the model results, the text from the audio files was extracted, and the sentiments in the text were analyzed using a BERT (Bidirectional Encoder Representations from Transformers) Natural Language Processing model. The “Whisper” library in Python was used to extract the audio from the Speech files and was fed to a pre-trained Neural network model (nlptown/bert-base-multilingual-uncased-sentiment). The encoded text was fed to the Model that returns the sentiment as an Integer value ranging from 1 to 5, with 1 being extraordinarily Unpleasant and 5 being very pleased/ happy.

```
# Downloading the Pre trained BERT Model for Validations
tokenizer = AutoTokenizer.from_pretrained('nlpTown/bert-base-multilingual-uncased-sentiment')
bert_model = AutoModelForSequenceClassification.from_pretrained('nlpTown/bert-base-multilingual-uncased-sentiment')
```

/usr/local/lib/python3.10/dist-packages/huggingface\_hub/utils/\_token.py:89: UserWarning:  
The secret `HF\_TOKEN` does not exist in your Colab secrets.  
To authenticate with the Hugging Face Hub, create a token in your settings tab (<https://huggingface.co/settings/tokens>).  
You will be able to reuse this secret in all of your notebooks.  
Please note that authentication is recommended but still optional to access public models or datasets.

warnings.warn(  
tokenizer\_config.json: 100% ██████████ 39.0/39.0 [00:00<00:00, 1.39kB/s]  
config.json: 100% ██████████ 953/953 [00:00<00:00, 45.5kB/s]  
vocab.txt: 100% ██████████ 872k/872k [00:00<00:00, 5.45MB/s]  
special\_tokens\_map.json: 100% ██████████ 112/112 [00:00<00:00, 3.69kB/s]  
pytorch\_model.bin: 100% ██████████ 669M/669M [00:04<00:00, 157MB/s]

```
[15] emotions_dict={1:'Extremely Unpleasant',2:'Unpleasant',3:'neutral',4:'Happy',5:'Very Happy'}

[16] # Extracting the text from the Audio File using Whisper Library
whisper_result1=whisper_model.transcribe("/content/03-01-01-01-01-01.wav", fp16=False)
print(f" The Extracted text from the Audio: {whisper_result1['text']}")

The Extracted text from the Audio: Kids are talking by the door.

[17] # Encoding the Tokens
tokens_1 = tokenizer.encode(whisper_result1['text'], return_tensors='pt')
```

# Applying BERT model on the extracted tokens  
text\_result1 = bert\_model(tokens\_1)  
# Getting the Integer value of the predicted results  
int\_result1=int(torch.argmax(text\_result1.logits))+1  
# Extracting Emotion label  
emotion\_result1=emotions\_dict[int\_result1]  
print(f" The Emotion of the Audio: {emotion\_result1}")

The Emotion of the Audio: Happy

Figure 14: BERT NLP model and its Predictions

## 5.8. Audience Questions:

Some questions that the Audience may ask about the project are:

### 1. ***How can the Model help the employee during a heated conversation with a customer?***

**Solution:** During the call with the customer, the model captures the images of the employees periodically and analyzes their emotions. If the Employee appears to be stressed during the entire call, the Model provides suggestions such as *"Do you need Additional Help?" "Do you want to redirect the call to the Supervisor?" etc.*

### 2. ***How can the Model help to handle a frustrated customer on the phone?***

**Solution:** The Model evaluates the customer's tone on the line, and if it detects anger and frustration, it will provide suggestions to the employee on how to use Pleasurable words or words to avoid, etc. Suppose the heated exchange continues between the employee and the customer. In that case, the Model can alert the Supervisor to see if they can take over the situation to assist the customer on the phone, thus improving the customer experience.

### 3. ***How often does the Model provide Predictions during a call?***

**Solution:** This can be customized such that the Model waits for a few iterations before making a final decision. For instance, if the call lasts 3 minutes, the Model can be invoked every 10 seconds. However, it can start to make recommendations to the Employee after 60 seconds, thus avoiding false or misleading recommendations.

### 4. ***What emotions or sentiments can the Model Predict?***



**Solution:** The Emotions model can predict if a customer on the phone is Angry, Disgusted, Happy, Surprised, Fearful, or Calm. The Image recognition model can indicate if the Employee is Stressed or acting Normally.

**5. *What Additional enhancements can be made to the Model?***

**Solution:** The Model's scope can be expanded by including a Natural Language Processing Model that can translate the speech into Text and process the Text to predict Emotions. This can supplement the Audio Speech Processing Model.

Though this functionality is included in the project for validation purposes, it can be implemented in Production to supplement the original predictions. However, the downside of this approach is that the sentiment classification can only predict Positive, Negative, or Neutral sentiments, and it may not be as exhaustive as the Emotions Recognition model.

**6. *Can the Model be trained for the local call center requirements?***

**Solution:** The Model can be customized so that it can be trained to meet the needs of the local call center. Both the Images and Audio taken from the local call center can be used to Train the Model and deploy the Custom Model for Implementation.

**7. *Can the model record the conversations or video them for further analysis and training?***

**Solution:** Though the Model does not record the conversations and Videos, it is necessary for Training purposes. So, the recording and storage functionality can be achieved through different means and used to train the models.

**8. *What if the Model predictions are wrong? Will the Model learn from its mistakes?***

**Solution:** The Model used forward and Backward propagation methods of neural networks during the training process, which adjusts the weights and activation functions to achieve the desired output. By Implementing Reinforcement Learning, the Model can learn from its mistakes.

**9. *What emotions/sentiments are not captured by the Model?***

**Solution:** The Model can predict if an employee is under stress or if they are Happy. It can be enhanced further to predict additional Employee emotions per the needs.

**10. *Can the Model generate Monthly and weekly reports based on the overall trend?***

**Solution:** Currently, the Model does not have this capability, but it can be an excellent enhancement to integrate with visualization tools, such as Tableau, to analyze the overall sentiments and send automated reports.

**6. Conclusion:**

The Accuracy of both the Image and Speech emotion classification models was around 90%, indicating that the Model may be a good fit for the dataset. To make predictions and final recommendations, the results from both models were analyzed together before a conclusion was made. For instance, if the Image and Audio speech model predicts "Anger," it can indicate a heated exchange between the employee and the Customer. If the Employee and the customer appear stressed, the Model will provide additional resources or suggestions to calm the situation. It will also prompt if the employee needs extra assistance. If everything seems normal, the Model will prompt a message stating the same. This is shown in Figures 15a and 15b.

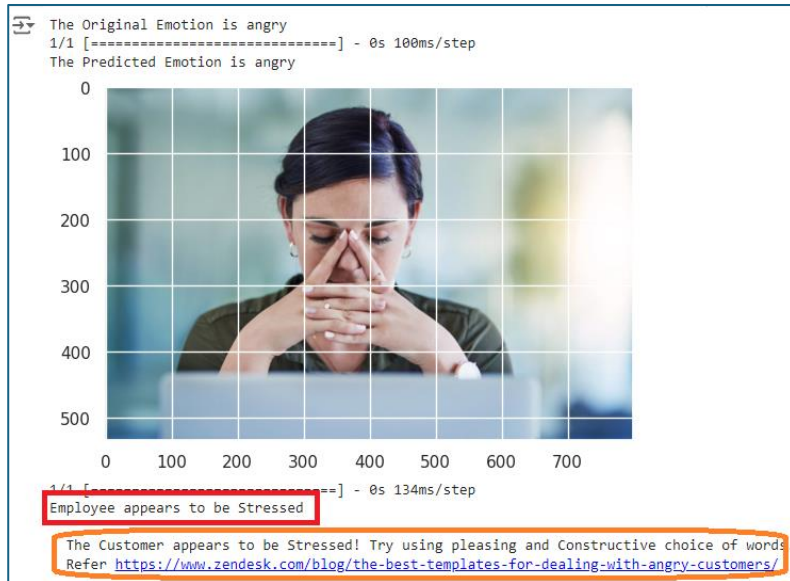


Figure 15a: Recommendations by the Model if the Employee is Stressed



Figure 15b: Recommendations by the Model for a Normal Conversation

## **7. Assumptions:**

- The Image and Speech emotion classification models were trained assuming that the samples used in the project represent the general population.
- The classification models were built assuming that the features portrayed by the images belong to one of the Target Groups and are independent of each other.
- Also, the data was assumed to have no Multicollinearity between the independent features.
- The audio dataset was trained assuming it represented the Voice of the Customer and not of the Employee.

## **8. Limitations:**

The datasets used for Training the models contained the voices of professionally trained actors expressing their emotions. However, in the real world, people's emotions may not be as clear as in the training dataset.

The Model also considers the voice pitch as one of the inputs to understand the emotion of the speech. So, if the customer does not speak louder or clearer or if there is a problem with collecting the audio signals, it can significantly affect the Model's outcome.

## **9. Challenges:**

The data used in the model was acquired from the actors who speak English with a North American accent. There may be challenges while analyzing the speech and emotions of a customer with a different accent.

The models were trained using audio files that were only a few seconds long. The calls usually last for a few minutes or longer. This can be particularly challenging for the Model to predict accurate outcomes if there are a variety of emotions during the duration of the call.

The project aims to provide suggestions by studying the photos of the employees taken at the call centers from their web cameras. If the employee is not in the focus of the camera or if they wear a face mask, the Model may not be able to make accurate predictions. Also, if the employee turns the camera off, the Model may be unable to predict anything.

Apart from these challenges, if the hackers were to hack the facial data, it may lead to data breaches, privacy concerns, and expensive lawsuits.

#### **10. Future uses/ Additional Applications:**

The project aims to analyze facial expressions and understand speech emotions by processing the audio signals. However, it can be expanded further by extracting the Audio, converting it to text, and analyzing the emotions from the text. This can primarily supplement the sentiment analysis the project is trying to achieve.

The project scope can also be expanded so that the model can be invoked multiple times during the call. However, the outcomes must be compared with the combination of previous outcomes, and a final decision must be made based on that. For instance, the Customer may be “really” angry if the Model predicts “anger” in five consecutive invocations. Otherwise, it must be ignored as a False Positive.

#### **11. Recommendations:**

The Model for this project was trained based on the data available on the web. However, it is recommended that the Model be further trained using the real-world data acquired from the call centers.

The image classification model can predict if an employee is happy or stressed out and can be expanded further by including other emotions such as being nervous, angry, etc.

The speech classification model can be trained further by using the voices of the local population so the Model can understand the emotions better. For instance, the American and British accents sound different, and some words are spoken differently, which may confuse the Model and lead to incorrect results due to how the words are pronounced and stressed.

## **12. Implementation plan:**

Before deploying the Model in Production, it must be tested in production-like simulation environments to assess its performance. To test the Model, an app using Streamlit or a Web User interface can be built to take both the Audio and Image inputs and produce the results.

For the actual production implementation, the Model must be integrated with the cameras and the phone systems to capture the image and the Audio and provide real-time suggestions on the employee's screen. Also, the Model must be able to access the emotions for small periods, such as every 15 seconds of the call, and compare the results with the previous outcomes before making the final recommendations, thus avoiding False positives.

## **13. Ethical considerations:**

Though the project only intends to capture the employees' facial expressions to analyze their moods, it may be mistaken for an employee monitoring tool. Hence, it is essential to be transparent to the employees about the tool's purpose to avoid confusion. Employees must not feel that they are being eavesdropped on but that the tool only tries to assist them. (Ethical Considerations for Employee Monitoring, n.d.).

The Model was trained using voices recorded by actors speaking with a North American English accent. This may raise ethical concerns for non-English speakers, such as customers who prefer to speak Spanish.

Also, there may be privacy concerns about the pictures of the employees being taken. Hence, getting legal counsel and not violating local laws is necessary.

As the Customer's Voice is being recorded for further analysis by the Model, informing them that the call may be recorded for training purposes to avoid legal disputes is essential.

#### **14. References:**

Solutions, R. C. C. (2023, June 20). Common Call Center Staffing Issues (and How to Avoid Them). ROI CX Solutions. <https://roicallcentersolutions.com/blog/common-call-center-staffing-problems-avoid-them/#:~:text=Inadequate%20training%20not%20only%20leads>

Ethical Considerations for Employee Monitoring. (n.d.). ActivTrak. <https://www.activtrak.com/solutions/employee-monitoring/ethical-considerations/#:~:text=Be%20transparent%3A%20When%20it%20comes>

Livingstone, S. R. (2018). RAVDESS Emotional speech audio. Kaggle.com. <https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio/data>

## **Appendix:**

### **Differences between Sentiment and Emotional Analysis:**

Though both Sentiment and Emotion may be relatable, they are vastly different and are usually used interchangeably. Sentiment Analysis, often known as "Mining of Opinions," is used to understand the feelings or Opinions about various fields such as Marketing Reviews, social media, etc. The Outcomes are usually values such as Positive, Negative, and Neutral.

Sometimes, the sentiment analysis cannot fully understand the customer's emotion, leading to ineffective customer experience. Emotion analysis extends beyond the basic sentiment and can study a variety of human emotions such as Sadness, Happiness, Fear, Disgust, Surprise, etc. This is achieved by studying facial features, or Voice signals to understand the feelings expressed.

In this Project, Sentiment analysis of the employees was performed using Image classification algorithms, and Speech Audio data was used to build an Emotion Recognition Model.