Capstone Project-5

Speech Emotion Recognition (Deep Learning)

- Soma Pavan Kumar(Pro Flex)
- spkumar1998@gmail.com

Contents

1. Problem Statement
2. Introduction
3. Understanding Data
4. Data Modelling and Model implementation
5. Post Deployment on AWS
6. Conclusion

Problem Statement

- Verbal Communication is valuable and sought after in workplace and classroom environments alike. There is no denying the notion that Indians lack verbal communication and consequently lag behind in the workplace or classroom environments. This happens despite them having strong technical competencies.

- Clear and comprehensive speech is the vital backbone of strong communication and presentation skills. Millions of people are affected by stuttering and other speech disfluencies, with the majority of the world having experienced mild stutters while communicating under stressful conditions.

- Research shows that mild disfluencies can be cured without medical help, just practicing speech regularly and constructive feedbacks are effective ways to improve. We will solve the above-mentioned problem by applying deep learning algorithms to audio/speech data. The solution will be to identify emotions in speech

- We will solve the above-mentioned challenge by applying deep learning algorithms to audio speech data. The solution to this problem is by recognizing emotions in the speech.

Introduction

- Speech is the most natural form of communication between humans and computers speech is a complex signal. It contains information regarding message speaker language and emotion. so, emotion makes speech more attractive more effective more expressive. Speech emotion recognition means understanding the emotional state of a human by extracting or detecting the feature extracted by his/her voice.

- There are some universal emotions including Neutral, anger, joy, sadness in which any intelligent system with limited computer resources can be trained to recognize or synthesize as needed.

- While digital platforms have limitations in terms of physical surveillance but it comes with the power of data and machines which can work for you. It provides data in the form of video, audio, and texts which can be analyzed using deep learning algorithms.

- Deep learning backed system not only solves the speech emotion recognition issue, but it also removes the human bias from the system, and all information is no longer in a human brain rather translated in numbers that can be analyzed and tracked.

Dataset Information

- The data used here are audio files each of them about 2 to 3 seconds long and the some of them are 4 seconds long. All these audio files are publicly available in the TESS Toronto emotional speech set on Kaggle.

- Some information on the dataset: A dataset for training emotion (7 cardinal emotions) classification in audio. What's interesting is that this dataset is female only and is of very high quality audio. Most of the other dataset is skewed towards male speakers and thus brings about a slightly imbalance representation. So because of that, this dataset would serve a very good training dataset for the emotion classifier in terms of generalisation (not overfitting).

There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total.

The dataset is organised such that each of the two female actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.

Data Augmentation:

- Data augmentation is that the method by which we make unused engineered information tests by including little annoyances to our introductory preparing set.

- To create syntactic data for sound, we are going to apply commotion infusion, moving time, changing pitch, and speed.

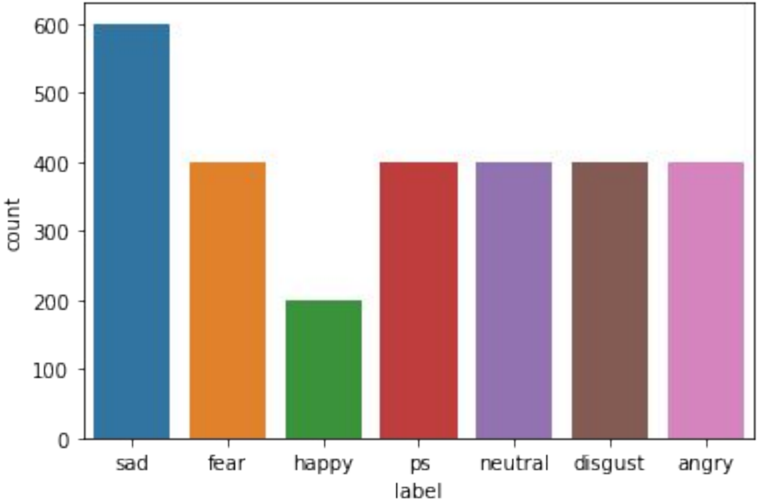
- Data augmentation is a technique that it reduces the overfitting of model and act as a regularizer.

- More data is generated using the training set by applying transformations. It is required if the training set is not sufficient enough to learn representation.

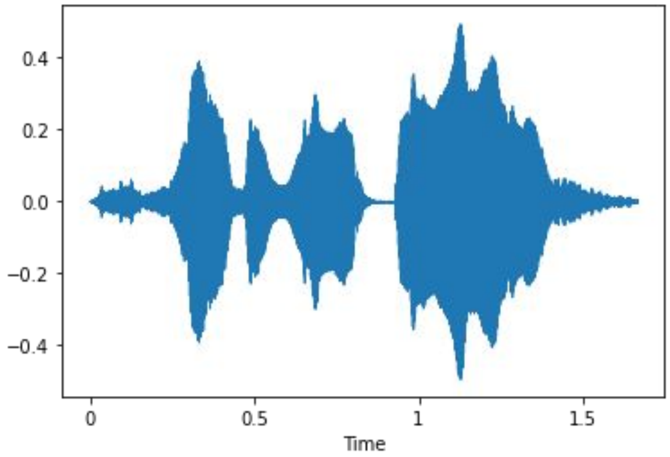
Exploratory Data Analysis:

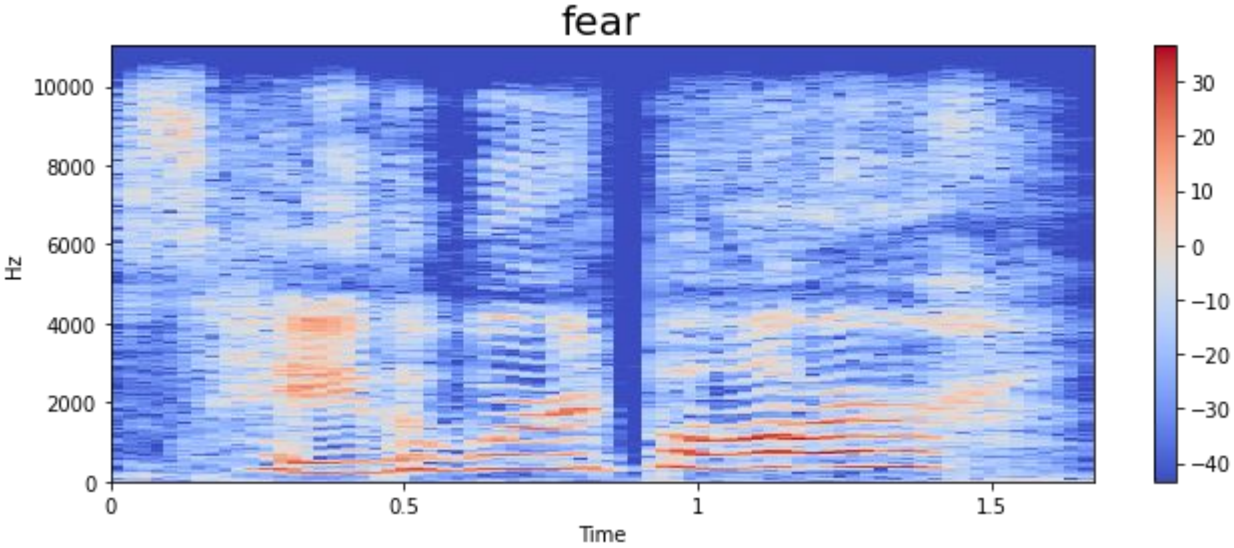
The First Thing First EDA using this we get intuition about how the data is structured.

After saw the data set first we analysis how many emotion are present in the dataset so after that we got most common seven emotion. It helped us understand which features behave in which ways in relation to the target variable.



A **spectrogram** is a visual representation of the spectrum of frequencies of a signal as it varies with time. When applied to an audio signal, spectrograms are sometimes called **sonographs**, **voiceprints**, or **voicegrams**.





Feature Extraction :

- MFCC (Mel Frequency Cepstral Coefficients)- MFCC can be used to extract the distinctive properties of human voice, and this MFCC also represents the short-term power spectrum of human voice. MFCC is used to produce the coefficients that describe the frequency Cepstral; these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale frequency.

- Mel Spectrogram- A Fast Fourier Transform is computed on overlapping windowed segments of the signal and that we get what's called the spectrogram A spectrogram may be a visual way of speaking to the flag quality, or “loudness”, of a flag over time at different frequencies display in a specific waveform.

Algorithms Used:

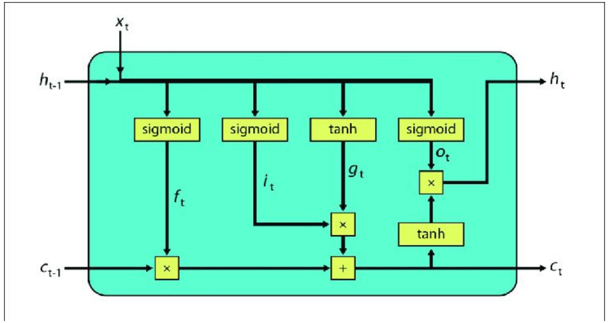
1. MLP (Multi Layer Perceptron)

- Multilayer perceptron (MLP) classifier is a supervised classification technique that uses backpropagation for training. It is one of the feed-forward artificial neural networks (ANN) classes. It consists of more than one perceptron. It consists of one output layer, one input layer, and in between these input and output layers, there may be an arbitrary number of hidden layers based on the user's choice.

- Its nonlinear activation function, multiple layers distinguish this from a single layer feed-forward neural network. Since it has nonlinear activation, it can be able to distinguish the data that is not linearly separable.

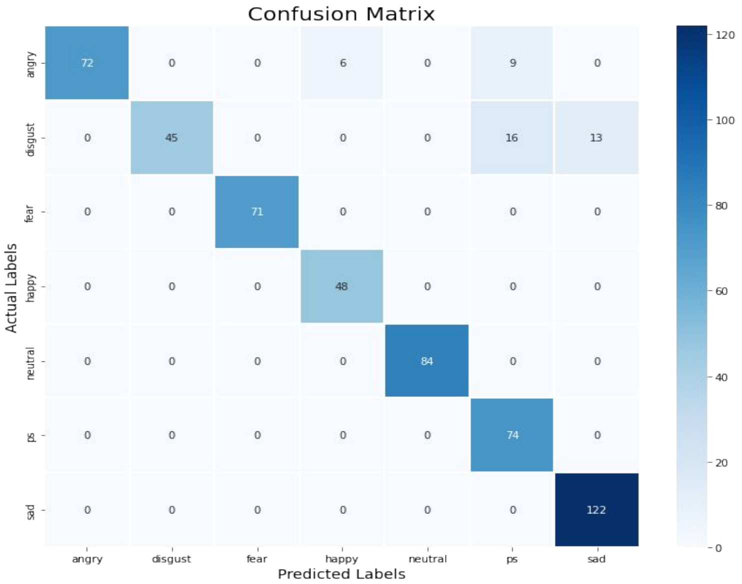
2. LSTM (Long Short Term Memory)

Long Short-Term Memory systems are the special mode of RNN that have the capabilities to memorize long-term conditions and rectify choice to work in a profound assortment of problems. These are specially planned to manage with long-term conditions issues, by default behaviour, they can keep in mind data for a colossal span to time. Like the chain-like structure of rehashing modules of RNN, LSTM incorporates a distinctive structure of the rehashing modules, it has the set of four neural organize layer that connected with each other in a special way.



Evaluation Metrics :

- The confusion matrix is a table that summarizes how successful the classification model is in estimating examples related to different classes. One axis of the confusion matrix is the label predicted by model , and the other axis is the actual label.

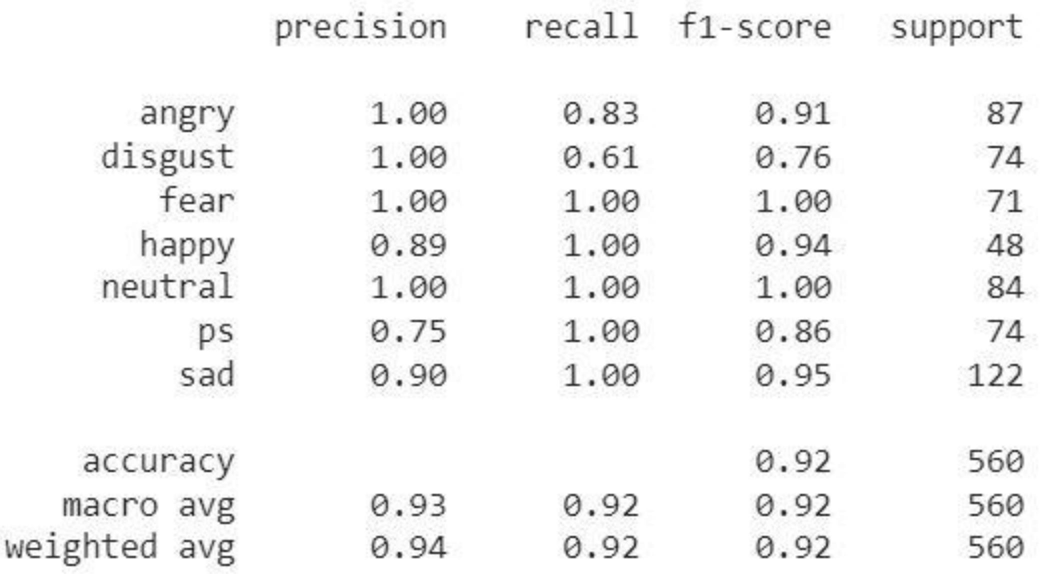


Confusion Matrix

Actual Labels \ Predicted Labels	angry	disgust	fear	happy	neutral	ps	sad
angry	72	0	0	6	0	9	0
disgust	0	45	0	0	0	16	13
fear	0	0	71	0	0	0	0
happy	0	0	0	48	0	0	0
neutral	0	0	0	0	84	0	0
ps	0	0	0	0	0	74	0
sad	0	0	0	0	0	0	122

Classification report

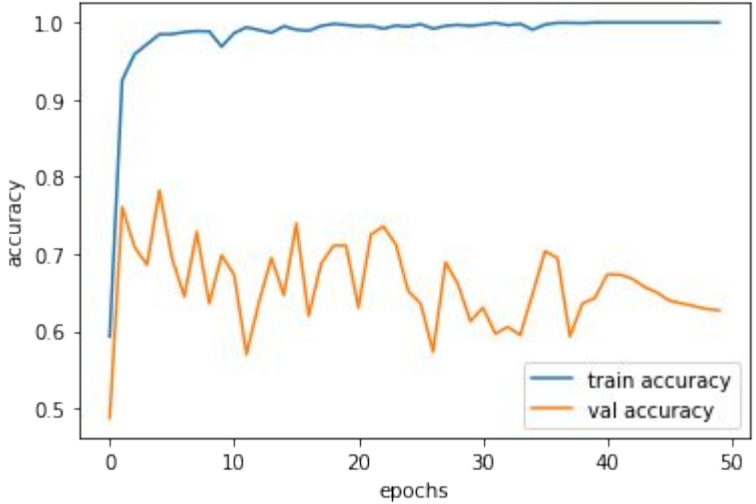
A classification report is **a performance evaluation metric in machine learning**. It is used to show the precision, recall, F1 Score, and support of your trained classification model.

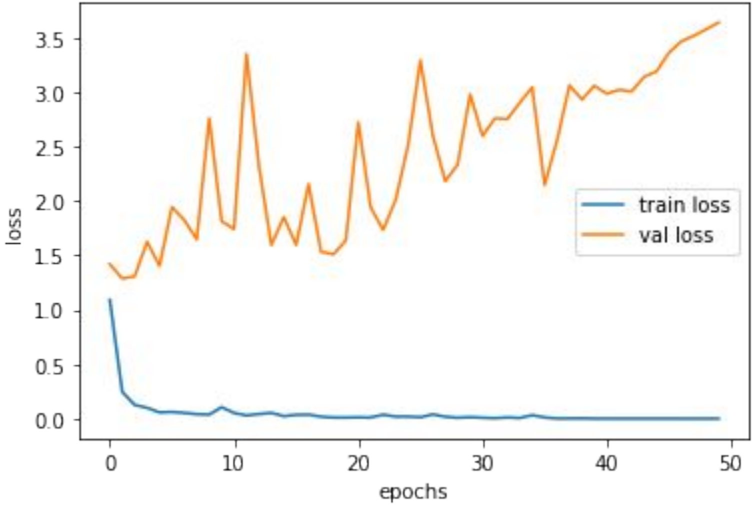


```
              precision    recall  f1-score   support

       angry       1.00      0.83      0.91        87
     disgust       1.00      0.61      0.76        74
        fear       1.00      1.00      1.00        71
       happy       0.89      1.00      0.94        48
     neutral       1.00      1.00      1.00        84
          ps       0.75      1.00      0.86        74
         sad       0.90      1.00      0.95       122

    accuracy                           0.92       560
   macro avg       0.93      0.92      0.92       560
weighted avg       0.94      0.92      0.92       560
```

Results from training the model

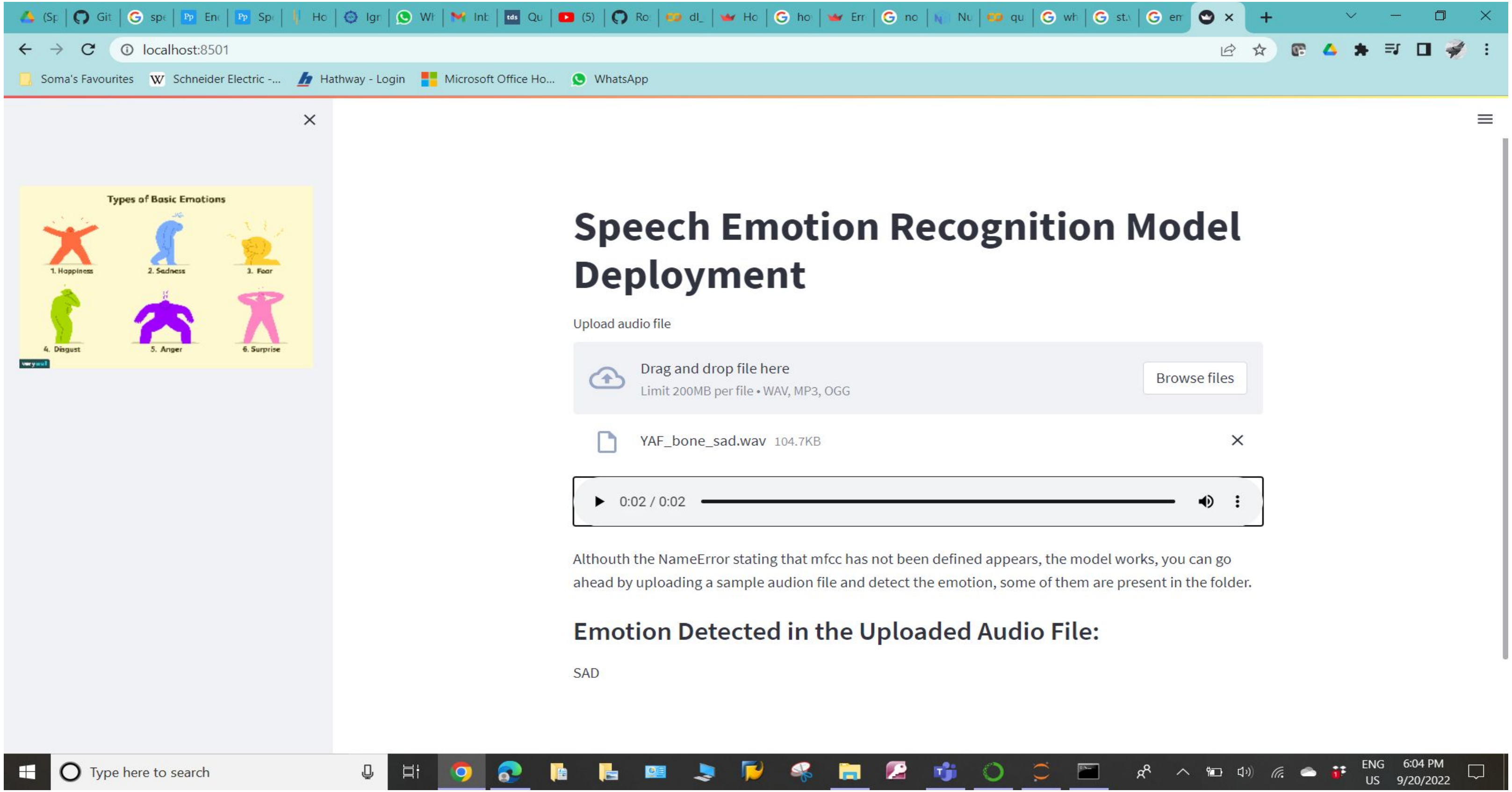
- The last few epoch results from the training are shown below





- The training Loss value is close to “0.01” and accuracy is “0.99” at 50 epochs.
- The Validation/Testing loss is around“3.5” and accuracy is “0.60”.

Deploying the model on Streamlit :



Speech Emotion Recognition Model Deployment

Upload audio file

Drag and drop file here
Limit 200MB per file • WAV, MP3, OGG

Browse files

YAF_bone_sad.wav 104.7KB

0:02 / 0:02

Althouth the NameError stating that mfcc has not been defined appears, the model works, you can go ahead by uploading a sample audion file and detect the emotion, some of them are present in the folder.

Emotion Detected in the Uploaded Audio File:

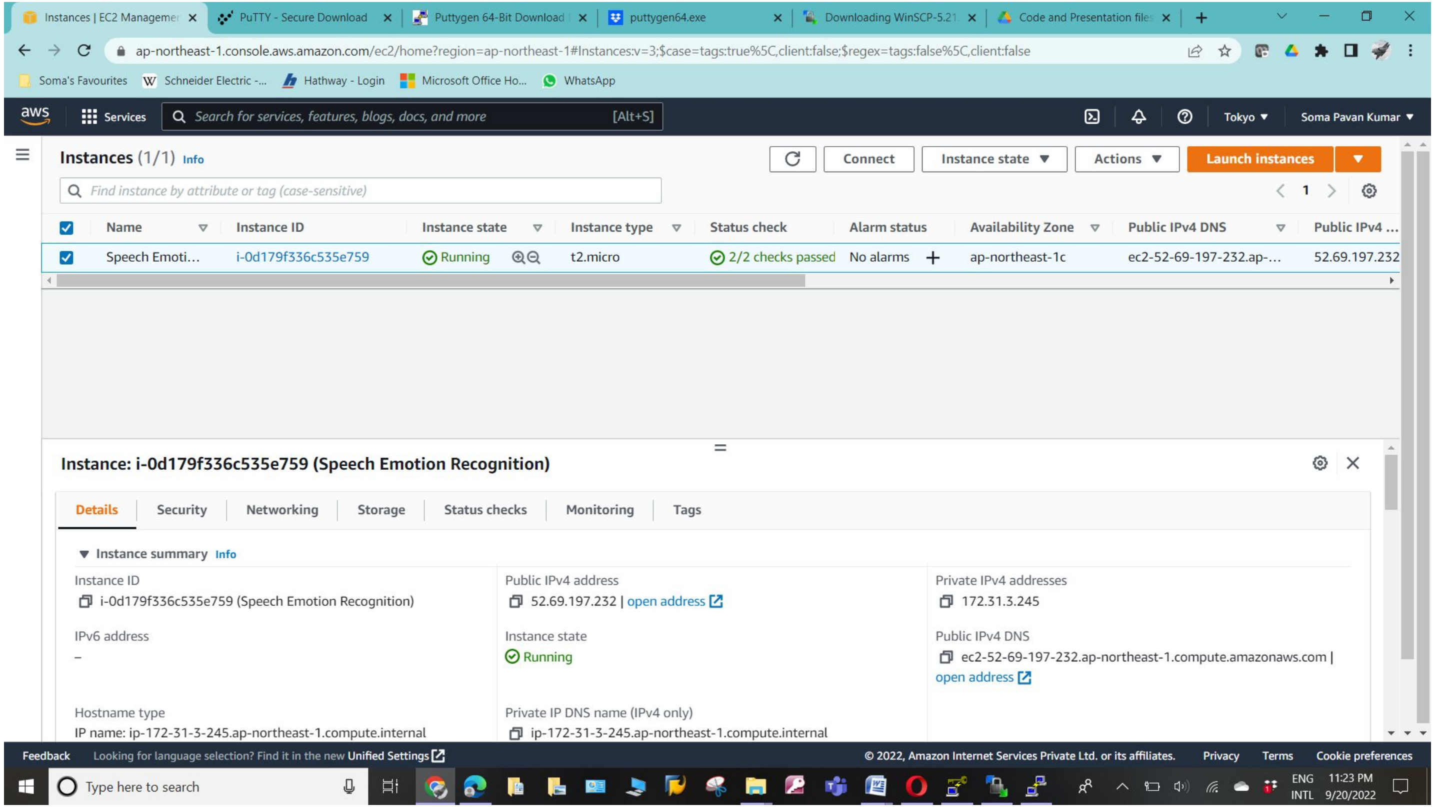
SAD

Deploying on AWS (Amazon Web Service) servers:

- AWS is The leading cloud provider in the marketplace is Amazon Web Services. It provides over 170 AWS services to the developers so they can access them from anywhere at the time of need.

- AWS has customers in over 190 countries worldwide, including 5000 ed-tech institutions and 2000 government organizations. Many companies like ESPN, Adobe, Twitter, Netflix, Facebook, BBC, etc., use AWS services.

Snippet of EC2 instance deployed on aws:


Instances | EC2 Managemer
PuTTY - Secure Download
Puttygen 64-Bit Download
puttygen64.exe
Downloading WinSCP-5.21
Code and Presentation files
ap-northeast-1.console.aws.amazon.com/ec2/home?region=ap-northeast-1#Instances:v=3;$case=tags:true%5C,client:false;$regex=tags:false%5C,client:false
Soma's Favourites
Schneider Electric -...
Hathway - Login
Microsoft Office Ho...
WhatsApp
aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Tokyo
Soma Pavan Kumar
Connect
Instance state
Actions
Launch instances
Find instance by attribute or tag (case-sensitive)
1
Details
Security
Networking
Storage
Status checks
Monitoring
Tags
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
ENG
INTL
11:23 PM
9/20/2022

Instances (1/1) Info

☑	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS	Public IPv4 ...
☑	Speech Emoti...	i-0d179f336c535e759	✓ Running	t2.micro	✓ 2/2 checks passed	No alarms +	ap-northeast-1c	ec2-52-69-197-232.ap-...	52.69.197.232

Instance: i-0d179f336c535e759 (Speech Emotion Recognition)

▼ Instance summary Info

Instance ID
i-0d179f336c535e759 (Speech Emotion Recognition)

Public IPv4 address
52.69.197.232 | open address

Private IPv4 addresses
172.31.3.245

IPv6 address
–

Instance state
✓ Running

Public IPv4 DNS
ec2-52-69-197-232.ap-northeast-1.compute.amazonaws.com | open address

Hostname type
IP name: ip-172-31-3-245.ap-northeast-1.compute.internal

Private IP DNS name (IPv4 only)
ip-172-31-3-245.ap-northeast-1.compute.internal

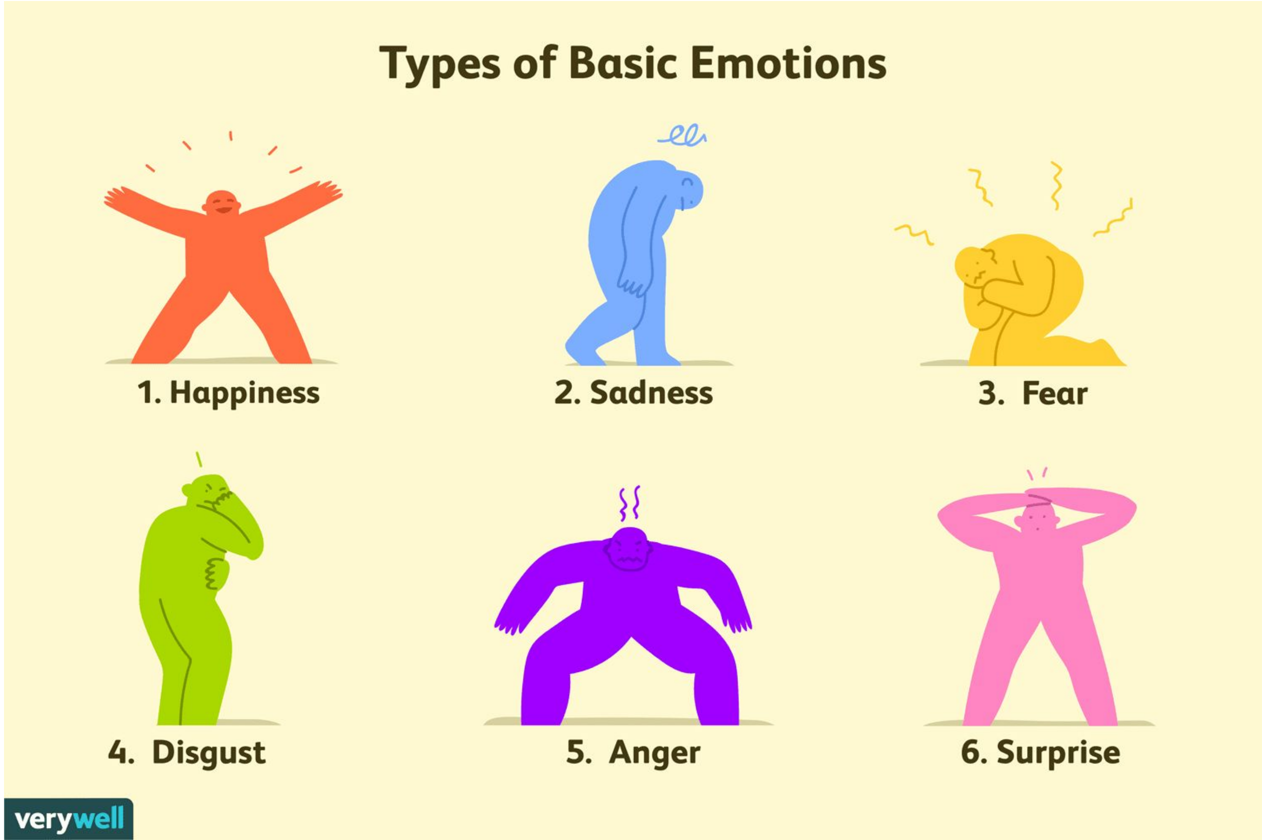
Conclusion:

- The Experiment I chose is to an end to end deployable model of Speech Emotion Recognition based on Deep Learning.

- In this Project, we utilize a few excellent procedures like LSTM, GRU. After utilizing all demonstrate LSTM gave a good accuracy. After that we also use Data Augmentation could be a strategy utilized to extend the sum of information by including marginally adjusted duplicates of as of now existing information or recently made engineered information from existing information. It acts as a regularizer and makes a difference decrease overfitting when preparing a machine learning demonstrate.

- The model was further deployed on Azure from Amazon Web Service by creating an EC2 instance.

THANK YOU


Types of Basic Emotions
1. Happiness
2. Sadness
3. Fear
4. Disgust
5. Anger
6. Surprise
verywell