EMOTION ANALYSIS IN AUDIO CONVERSATION

MECE-6397 PROJECT

INTRODUCTION

The goal of this research work is to use machine learning algorithms to predict the emotion of a person who is contacting customer care for their product queries. In this project, the audio is divided into chunks based on customer and caller. Now the audio is converted into text using Google API. This text sentiment is trained using logistic regression With this the converted text from the audio is tested. Also, the audio chunks are trained used models like KNN, Random Forest and SVM and the accuracies are compared.

OBJECTIVE

- The goal of this research work is to use machine learning algorithms to predict the emotion of a person who is contacting customer care for their product queries.

- Our work proposes a novel method of predicting the emotions of the customer, both pitch-wise and the words used.

- Audio is divided into chunks using audio segmentation and audio is converted into text to classify both text and audio chunks.

PREVIOUS WORKS

In the past, researchers have shown great interest in the text conversion of audio and analyzing the text intensity based on the words used and classifying the emotions with the words, with this intensity of customer pitch is not considered which is a major component to predict the emotion.

Researchers evolved the project by analyzing the intensity in the voice with this soft-spoken people are always considered calm and loud spoken as harsh irrespective of the language they use. To solve these problems, I tried to create an hybrid model to satisfy the problem statement

RESEARCH QUESTION

Usually, in emotion classification, researchers consider the acoustic features alone. For strong emotions like anger and surprise, the pitch and energy of the acoustic features are both high. In such cases, it is very difficult to predict emotions correctly using acoustic features alone. But, if we classify speech solely on its textual component, we will not obtain a clear picture of the emotional content. In the proposed hybrid approach we consider both text and audio features. Fig shows the framework for the hybrid approach.

DATA COLLECTION

I am using the dataset from RAVDESS for classification purpose of audio chunks obtained. It consists of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

Dataset consists of different emotion like - neutral, calm, happy, sad, angry, fearful, disgust, surprised

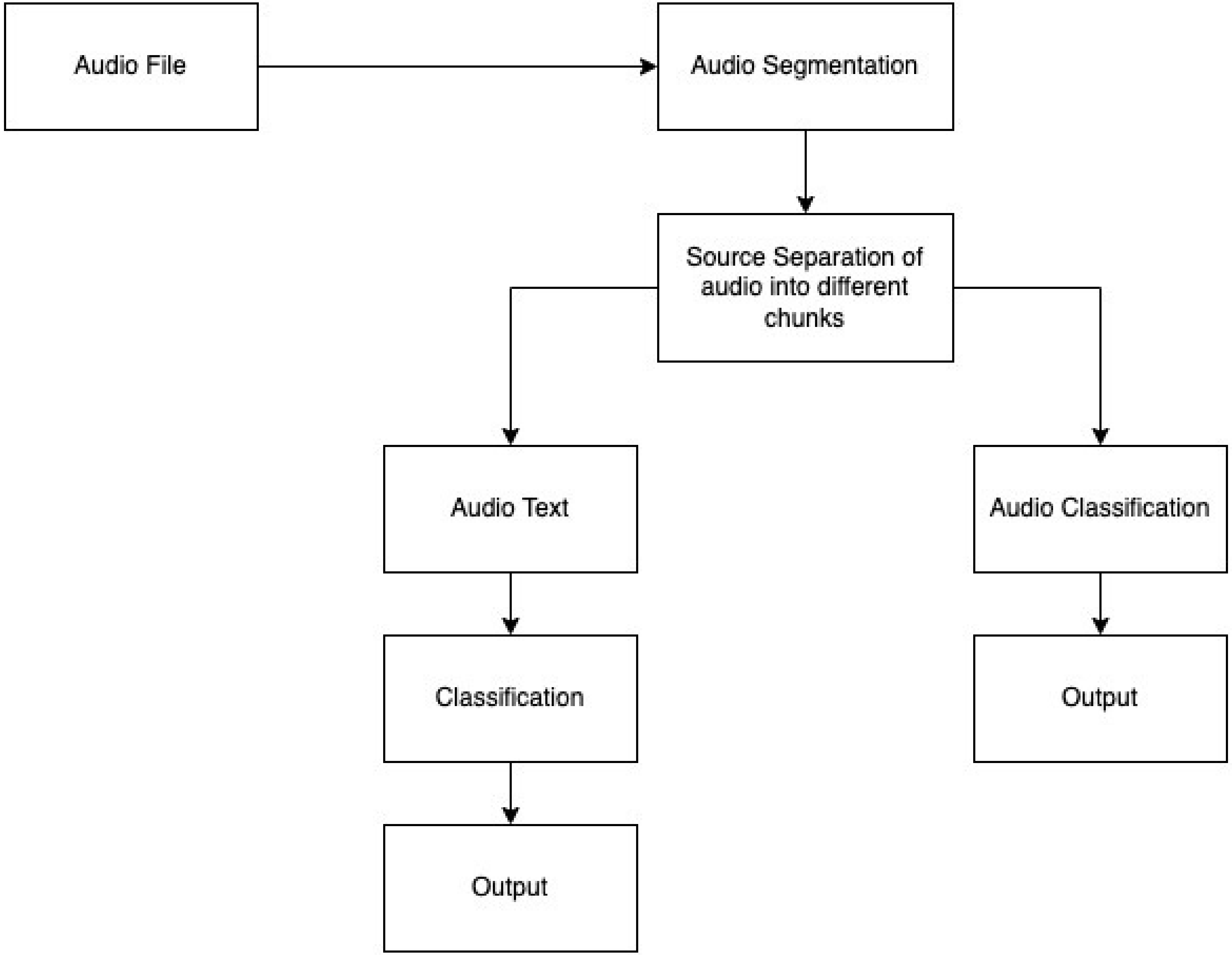
the labeled descriptions for each dataset in the same order as the label categories in the datasets. For instance, if an observation in SCv1 has label 1, then that observation is in the sarcasm category. Mutiple emotions like happy calm surprised were classified as positive and angry fearful disguist were classified as negative emotion.

Twitter dataset of sentiment analysis of text documents is chosen.Assigning a sentiment score to each phrase and component (-1 to +1)

METHODLOGY

Our methodology is divided in 4 parts-

1. Source Separation of audio into different chunks [Separating customer audio]
2. Conversion of audio to text
3. Classification based on audio
4. Classification based on Text

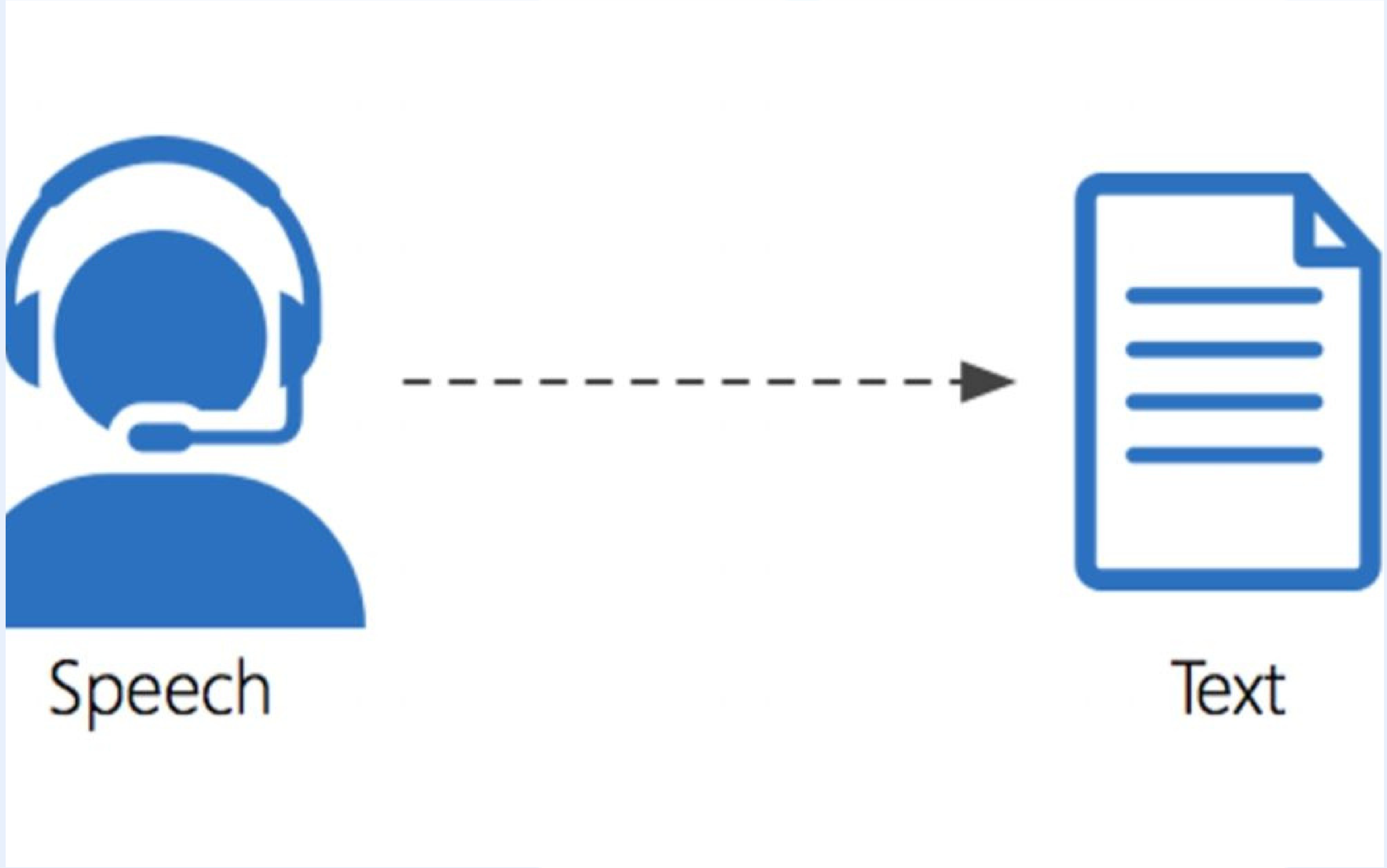

Audio File
Audio Segmentation
Source Separation of audio into different chunks
Audio Text
Classification
Output
Audio Classification
Output

METHODLOGY

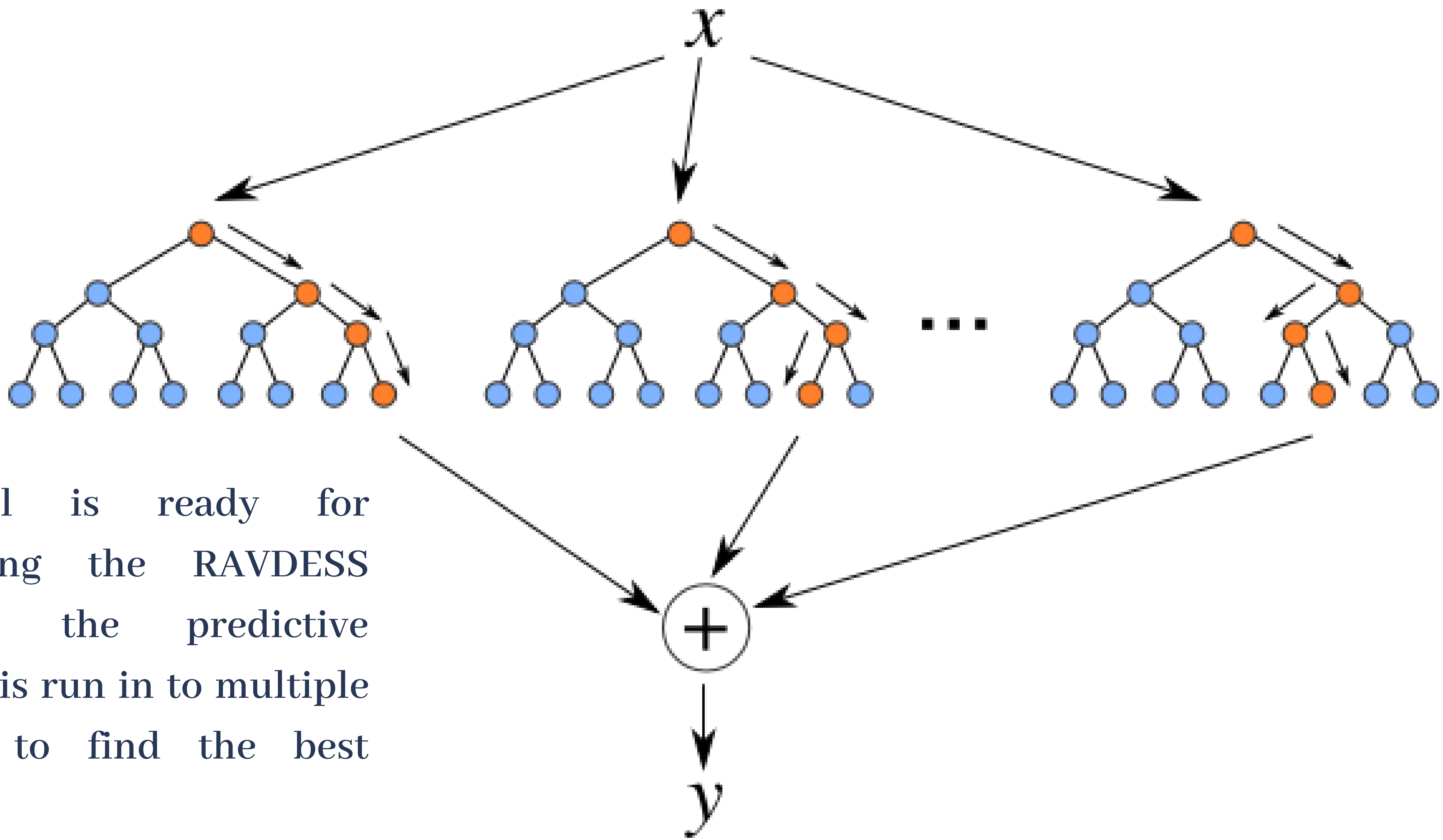
The first phase of our system is segmentation of the audio into parts using Google's webrtc Voice Activation Detection . The splitted parts of audio are run through Adaptive MAP estimation to find the customer and remove the audio chunk of company caller using the generated super vectors.

METHODLOGY

Once the audio file of conversation between the customer and call center agent is divided into chunks, we pass it through Google API to convert the customer chunk audio to text which divides our project into two arena.



METHODLOGY



Now the model is ready for classification, Using the RAVDESS emotion dataset the predictive sentiment analysis is run in to multiple classifier models to find the best accuracy of all.

METHODLOGY

And the text is passed into text sentiment analysis model built using logistic regression trained with nplk twitter dataset to predict the emotion of the customer

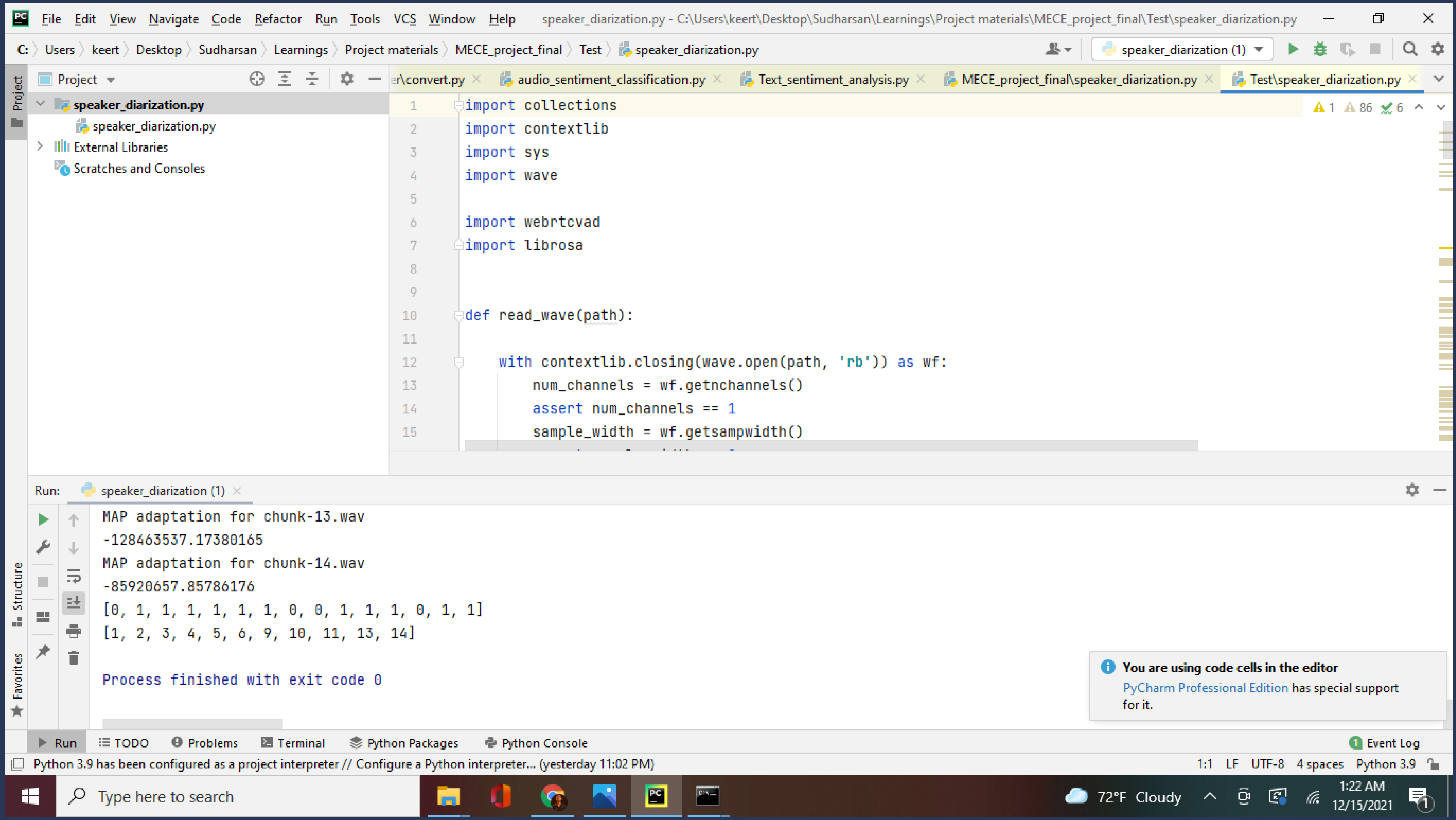
SIMULATIONS & RESULTS

COMPARISON AND INFERENCE

SIMLUATION & RESULTS-1

Through VAD detection, segments of an audio file are separated. These segments contain voiced data, which are then passed on to the Adaptive MAP estimation process to obtain the super vector.

SIMLUATION & RESULTS-1



SIMLUATION & RESULTS-2

Sample Output

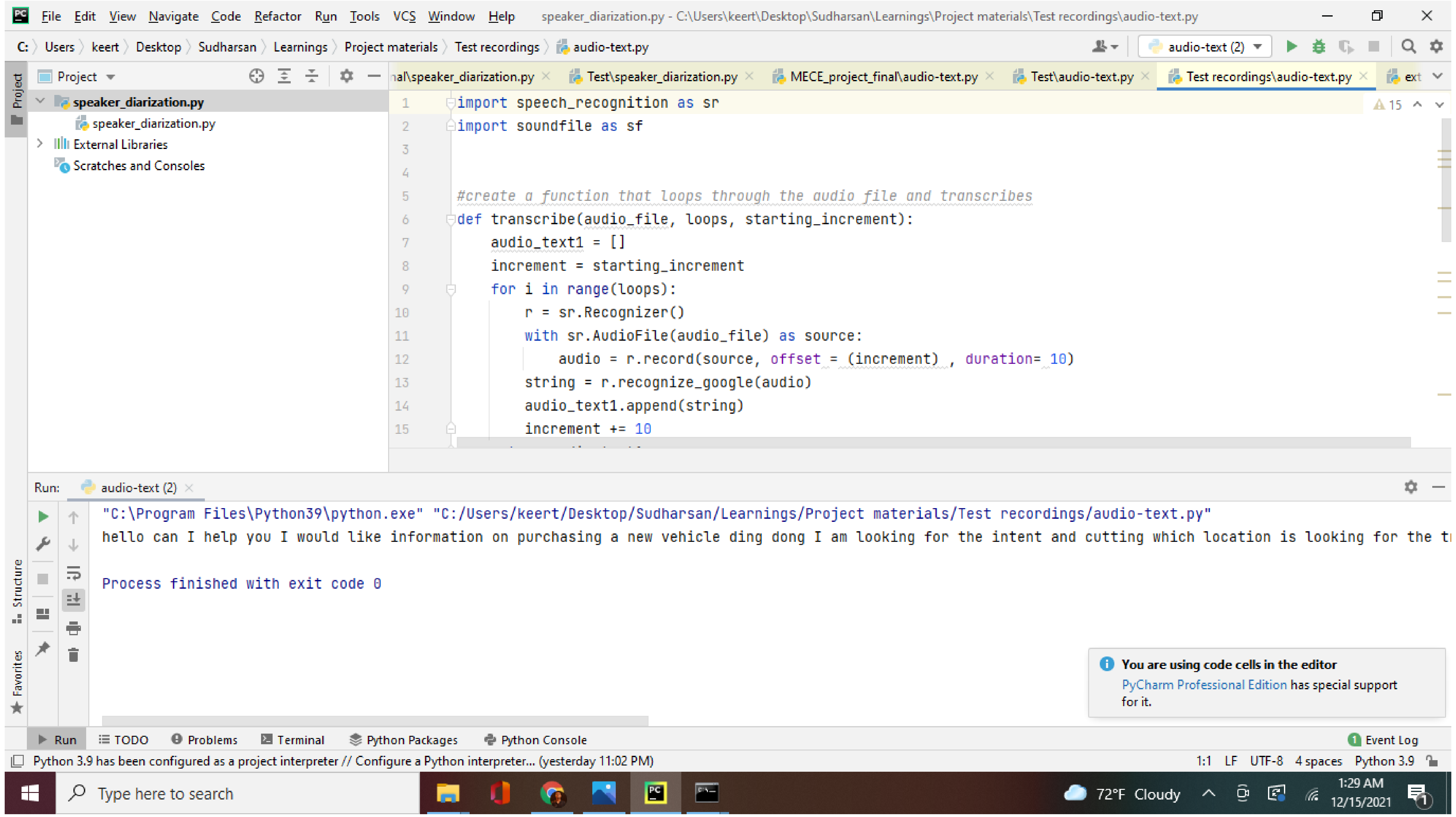
chunk-01.wav I am falling from market I am falling from market

chunk-06.wav Hanuman Mahima Mama ji great

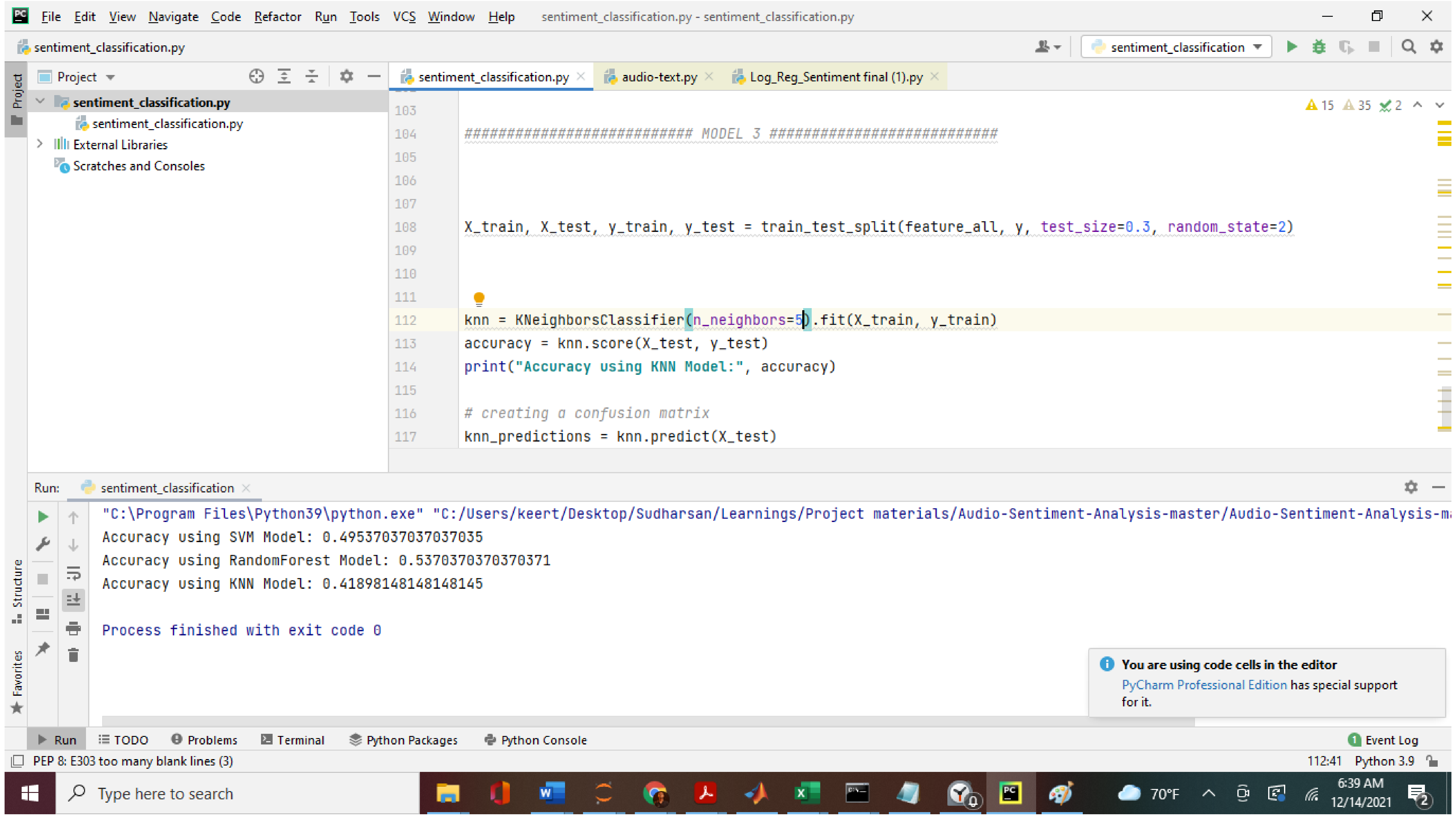
chunk-09.wav Anjali Sharma spoken to nahin number in about drawing about school number about

chunk-14.wav Vikram 951 number phone number 951 number phone number

SIMLUATION & RESULTS-2



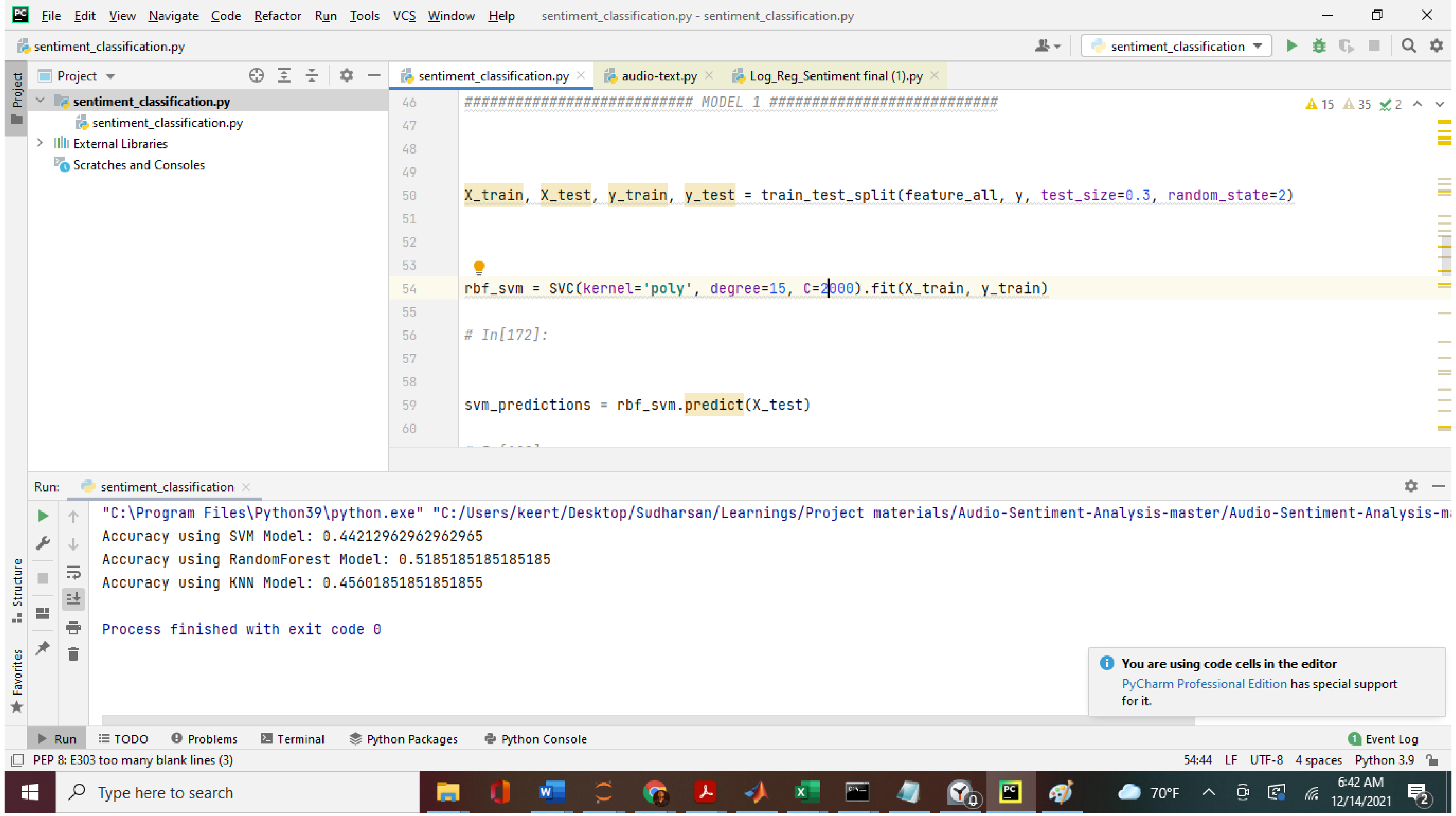
SIMLUATION & RESULTS-3



1. SVM- kernel= “poly”; degree=10; c=1000
2. Random forest n_estimators=100, criterion='entropy', random_state=58
3. KNN, nearest neighbour=5

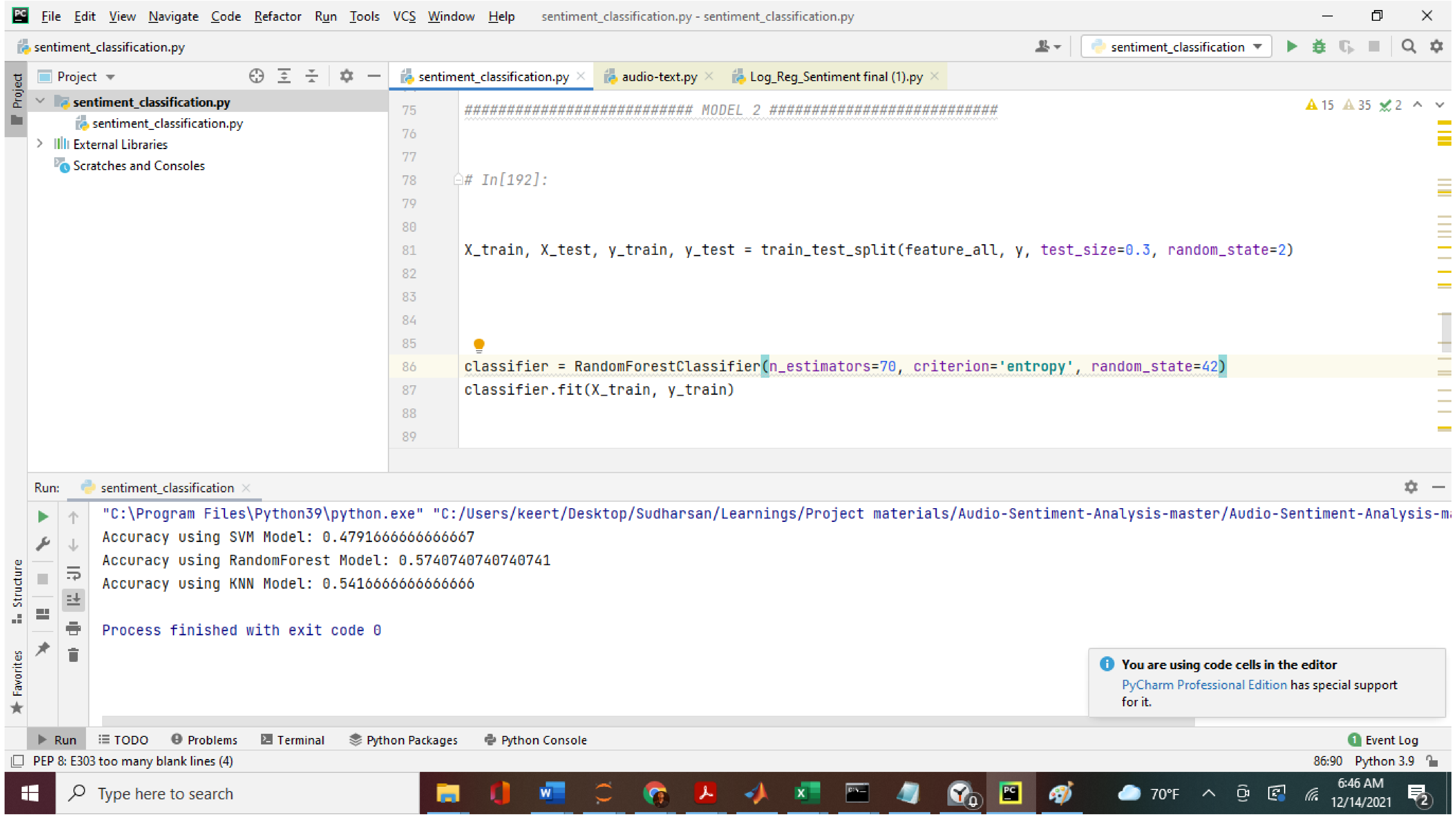
SIMLUATION & RESULTS-3

1. SVM- kernel= “poly”; degree=15; c=2000
2. Random forest n_estimators=50, criterion='entropy', random_state=46
3. KNN, nearest neighbour=3



SIMLUATION & RESULTS-3

1. SVM- kernel= “poly”; degree=12; c=5000
2. Random forest n_estimators=70, criterion='entropy', random_state=42
3. KNN, nearest neighbour=1



SIMLUATION & RESULTS-4

Given the test data and the weights of trained model, we calculated the accuracy of our logistic regression model.

- Using predict_tweet() function to make predictions on each tweet in the test set.
- If the prediction is > 0.5, set the model's classification y_hat to 1, otherwise set the model's classification y_hat to 0.
- A prediction is accurate when y_hat equals test_y. Sum up all the instances when they are equal and divide by m.

SIMLUATION & RESULTS-4

Accuracy for Logistic regression= 0.9834

Output for chunk 09 and chunk 06

['anjali', 'sharma', 'spoken', 'nahin', 'number', 'draw', 'school', 'number']

[[0.50014869]]

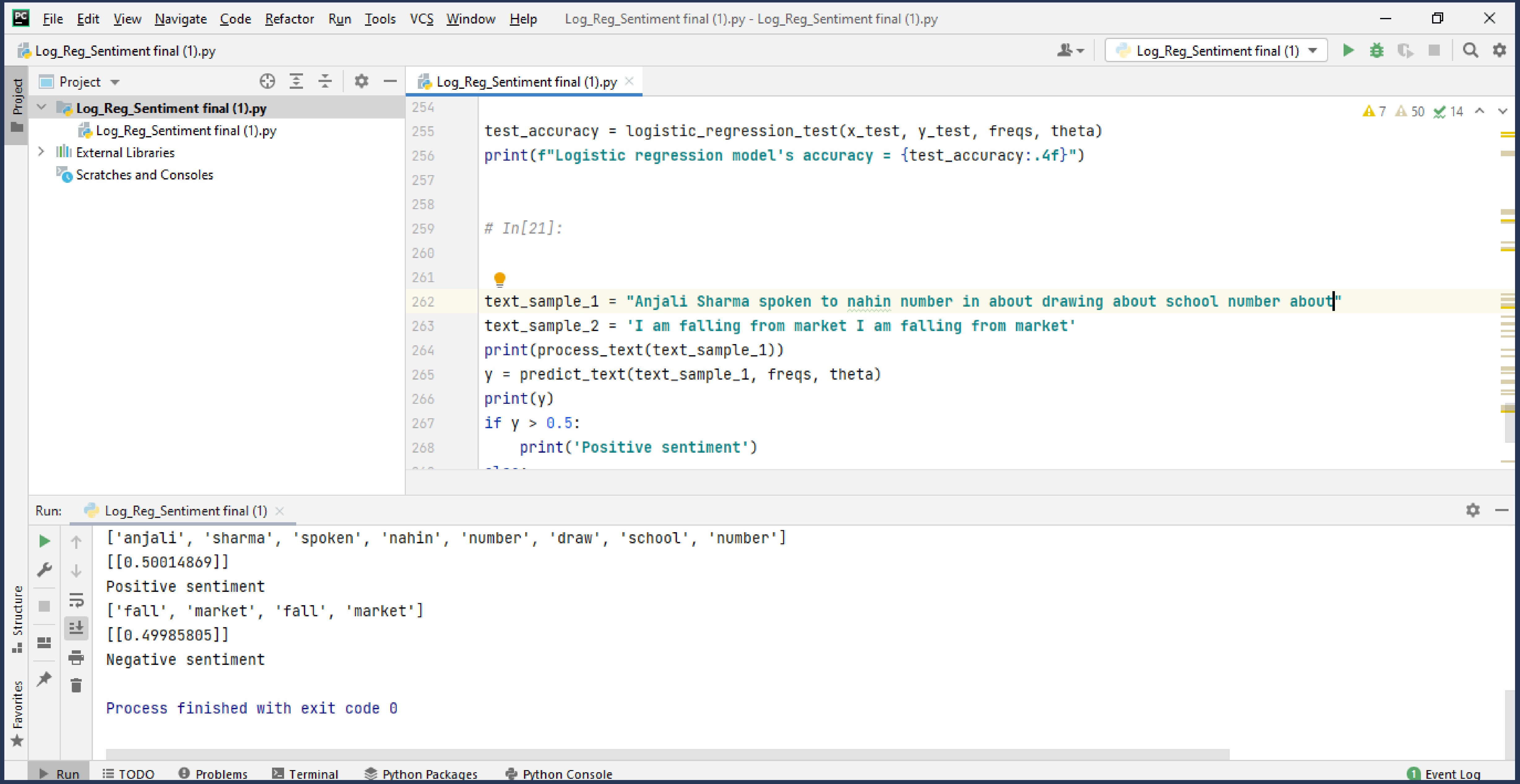
Positive sentiment

['fall', 'market', 'fall', 'market']

[[0.49985805]]

Negative sentiment

SIMLUATION & RESULTS-4



FUTURE ASPECTS

- A model is to be developed to get the weighted average of the accuracies calculated from both the classifications of audio chunks and texts derived from the audio which is now limited.
- Accuracy of audio classification must be improved as the major part emotion will be displayed in the pitch of the customer.
- Audio can be further divided into smaller chunks where the conversation of customers can be divided into smaller parts to run through reinforcement learning and train the initial part to test the latter parts.
- In this way pitch of the person can be identified as the pitch is a personal trait and a louder mouth can be classified and not directly noted as anger person.
- This will help the model to improve the accuracy invariably.
- This will be a stepping stone for machines to take over customer care which will reduce the manpower in the industry.

THANK YOU