ISTANBUL COMMERCE UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

SENTIMENT ANALYSIS OF MEETING ROOM

Mert İLERİ

Supervisor
Asst. Prof. Dr. Metin TURAN

MSc. THESIS
DEPARTMENT OF COMPUTER ENGINEERING
ISTANBUL – 2022

ACCEPTANCE AND APPROVAL PAGE

On 28/01/2022 **Mert İLERİ** successfully defended the thesis, entitled "**Sentiment Analysis of Meeting Room**", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury members whose signatures are listed below. This thesis is accepted as a **MASTER'S THESIS** by Istanbul Commerce University, Graduate School of Natural and Applied Sciences **Department of Computer Engineering**.

Supervisor	**Asst. Prof. Dr. Metin TURAN** Istanbul Commerce University
Jury Member	**Asst. Prof. Dr. Feyza Merve HAFIZOĞLU** Istanbul Commerce University
Jury Member	**Asst. Prof. Dr. Zeynep Turgut AKGÜN** Istanbul Medeniyet University

Approval Date: 10.02.2022

Istanbul Commerce University, Graduate School of Natural and Applied Sciences, accordance with the 2nd article of the Board of Directors Decision dated 10/02/2022 and numbered 2022/339, "Mert İLERİ" who has determined too fulfill the course load and thesis obligation was unanimously decided to graduated.

Prof. Dr. Necip ŞİMŞEK
Head of Graduate School of Natural and Applied Sciences

ACADEMIC AND ETHICAL RULES DECLARATION OF CONFORMITY

In this project I prepared in accordance with the rules of thesis writing, Istanbul Commerce University, Institute of Science,

- I obtained all the information and documents in the project within the framework of academic rules.
- I present all visual, audio and written information and results in accordance with scientific moral rules.
- I refer to the related works in accordance with scientific norms in case of using others' works.
- I cited all the works I cited as a source.
- I did not make any distortions in the data used.
- and that I do not present any part of this project as another thesis study at this university or another university.

I declare.

Mert İLERİ
10.02.2022

CONTENTS

ABSTRACT

M.Sc. Thesis

SENTIMENT ANALYSIS OF MEETING ROOM

Mert İLERİ

Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Computer Engineering

Supervisor: Asst. Prof. Dr. Metin TURAN

2022, 37 pages

In the last decade, enormous data has been shared throughout the world. With analysis applications in today's big data world, companies try to use sentiment analysis techniques to analyze their customers' moods and improve their efficiency according to their sensitivity. In this research, as a different application of emotion analysis, speech analysis in closed places was focused on the detection of emotion analysis in the meeting. The research needs low-noise environments. Otherwise, it may be affected by noises (other sounds) and conflicting situations may occur for more than one emotion (for example, noise will generally create negative emotion). As a solution, an artificial neural network that using meaningful sound features is proposed. Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS) data was used in this study. After the sound features were extracted, normalization (Z-score standardization) was applied to the data. The artificial neural network is fed with training data and a classifier machine learning model is created. In the performance measurements made using the test data, the average success rate was approximately 88%.

Keywords: Artificial neural network, sentiment analysis, voice features.

ÖZET

Yüksek Lisans Tezi

TOPLANTI ODALARININ DUYGU ANALİZİ

Mert İLERİ

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Asst. Prof. Dr. Metin TURAN

2022, 37 sayfa

Son on yılda, dünya çapında muazzam veriler paylaşılmaktadır. Günümüzün büyük veri dünyasında analiz uygulamaları ile şirketler, müşterilerinin ruh hallerini analiz etmek ve duyarlılıklarına göre verimlerini artırmak üzere duygu analizi tekniklerini kullanmayı denemektedirler. Bu araştırmada duygu analizinin farklı bir uygulaması olarak, kapalı mekanlarda konuşma analizi yapılarak toplantıda duygu analizi tespitine odaklanılmıştır. Araştırma düşük gürültülü ortamlara ihtiyaç duymaktadır. Aksi takdirde gürültülerden (diğer seslerden) etkilenebilir ve birden fazla duygu için çelişkili durumlar oluşabilir (örneğin gürültü genelde olumsuz duygu oluşturacaktır). Çözüm olarak anlamlı ses özelliklerini kullanan bir yapay sinir ağı önerilmiştir. Bu araştırmada Ryerson Duygusal Konuşma-Şarkı Görsel İşitsel Veritabanı (RAVDESS) verileri kullanılmıştır. Ses özellikleri çıkarıldıktan sonra verilere normalizasyon (Z-score standardizasyonu) uygulanmıştır. Yapay sinir ağı, eğitim verileriyle beslenmiş ve bir sınıflayıcı makine öğrenmesi modeli oluşturulmuştur. Test verileri kullanılarak yapılan başarım ölçümlerinde ortalama başarı olarak yaklaşık %88 değerine ulaşılmıştır.

Anahtar Kelimeler: Duygu analizi, ses özellikleri, yapay sinir ağı.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my esteemed advisor, Asst. Prof. Dr. Metin TURAN, who guided me for this research and helped me to overcome the difficulties I faced with his knowledge and experience. I would like to show my gratitude and my love to my precious family for their love, kindness and endless belief in me.

Mert İLERİ
ISTANBUL, 2022

LIST OF FIGURES

Page

LIST OF TABLES

SYMBOLS AND ABBREVIATIONS

ANN	Artificial Neural Network
b	Bias for Activation Function
CNN	Convolutional Neural Network
DNN	Deep Neural Network
f	Activation Function
HMM	Hidden Markov Model
MEDC	Mel Energy Spectrum Dynamic Coefficients
MFCC	Mel Frequency Cepstrum Coefficient
MLP	Multilayer Perceptron
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SER	Speech Emotion Recognition
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
WEKA	Waikato Environment for Knowledge Analysis
x_i	Output

1. INTRODUCTION

Companies and individuals are trying to specify other people's tendency or moods according to situations. So, determining general mood or sentiment of a customer or a person has been gaining more importance. Sentiment analysis is a set of methods determining people's emotions and reactions. It can be named as emotion recognition or mood analysis. With increase of sentimental analysis technologies, companies started to examine people's tendency. Because it saves time and the process is generally automated. This area is getting more important and popular such as a result of the development of machine learning, natural language processing technologies and getting more data accessible day to day. We can see an image about the increase of searching "sentiment analysis" string in a search engine (Figure 1.1). Mäntylä et al. (2018) said that this information can show us that concept of emotion analysis has recently started to increase. We sure about that sentiment analysis technology will become more accessible and more accurate in future. Because every type of data is growing up tremendously in every platform continuously. Not only companies, but also medical services used sentiment analysis according to Kim et al. (2004). In addition to that, in music genre classification or gender classification systems emotion analysis are being used (Darji, 2017). With these information, companies or peoples are trying to find opportunities for themselves. Because decreasing of time and increasing of customer assessment may bring speedy information and they can make a prediction of future trends. Therefore, sentiment analysis provides these organizations or persons useful knowledge to obtain people's tendency. Humans determine an idea or react to a situation always with emotions. These emotions are kept in voices or text-based systems. First, these companies or individuals must use comprehensive data for this task. Data can be found in any kind of platform. For example, social media platforms like Twitter etc. are supplying emotional data (Gebremeskel, 2011). Tweets and replies to these tweets contain many emotions about persons (Gebremeskel, 2011). Because in social media, millions of people are commenting and reacting. In addition to social media platforms, some specific datasets can be used for data source. But there are some problems in order to make text ready for usage of sentiment analysis

(Mohammad, 2017). Cleaning of unstructured parts of text is the initial work according to Saglani et al. (2020). Furthermore, words may have ambiguity or some words may not reflect the emotion being described. For example, some texts may contain disappointment, if the program extracts the emotion as anger, then this is not a useful knowledge for classification.

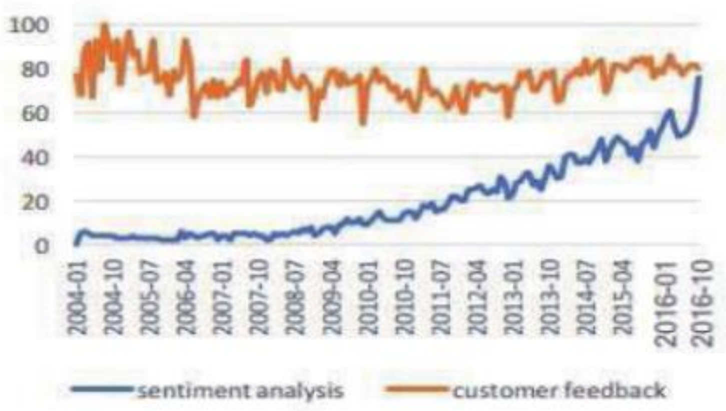

Figure 1.1. Sentiment analysis by years

Additionally, idioms and proverbs are making analysis difficult for researchers. As a different application of sentiment analysis, we focused on voice instead of text and restricted research with closed places possess to low distortion. We need low-noise environment because distortion or noise can affect the sentiment. Meetings are held on many different topics in many different companies. In addition to that, meetings may be critically important in terms of future of company and decisions to be made. Therefore, the current mood can also change depending on subject in discussion. The purpose of the idea is to make meetings under control in companies. If the general situation is anger or anxiety, soothing music can be played or meeting can be paused. So, meeting efficiency can be increased or prevent discussions being harmful for employee relationships. This system can also be used in case of need for security issues or medical services. One room may be listened by a police officer (possibly the language of the people in the room is not main language of country). So, general mood may trigger the actions that can be taken. To develop a successful model in this research, various voice recordings for various emotions were collected and meaningful features were extracted (Figure 1.2). Then, we split data into train and test sets. A classifier was trained according to these features. After that, test data set were used to evaluate the accuracy of the classifier. Considering all these issues, “How could we achieve a high success rate?” and “what should we do differently?” are the questions asked ourselves. Because every feature is not useful in the classification. For example, some of them is being correlated with others and decreases learning (or over-fitting). Furthermore, this may cause high dimensionality that prevents data to be processed in a reasonable time. Even only audio features have various types according to its return value.

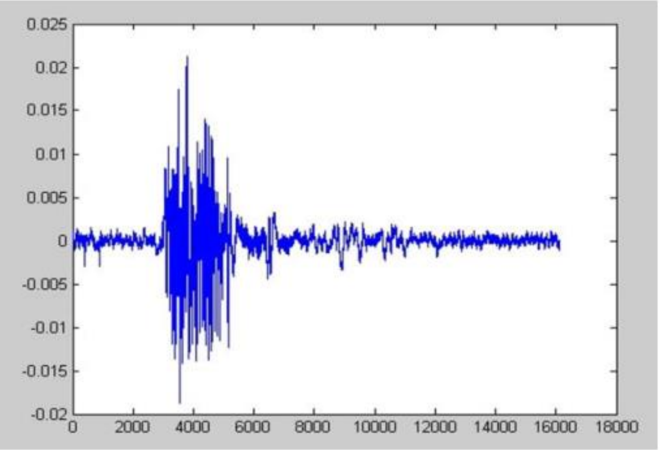

Figure 1.2. Voice signal of a word (Szabo, 2013)

A multilayer perceptron (MLP) is a neural network connecting multiple layers in a directed graph, which means that the signal path through the nodes only goes one way. Each node, apart from the input nodes, has a nonlinear activation function. An MLP uses backpropagation as a supervised learning technique. Since there are multiple layers of neurons, MLP is a deep learning technique. MLP is widely used for solving problems that require supervised learning as well as research into computational neuroscience and parallel distributed processing (Bhattacharjee et al., 2009; Nazzal et al., 2008). In section 2, we have a look at related works with different methods. Section 3 includes technical information about libraries, features, classifier that are used in solution and includes methodology that we chose.

2. LITERATURE REVIEW

As mentioned in the introduction section, there are several voice features to classify emotions. One of these features is MFCC. In one of the MFCC work, Davletcharova et al. (2015) studied sound waves of the people. They applied emotion classification via analyzing vocalic differences about different circumstances. These sounds separated word by word and when they were graphed, it was revealed emotions according to the peaks of the MFCC feature. People experimented are Russian and the result was Russian words under different emotions. Voice records were kept in same environment and MATLAB programming language was chosen. They mainly focused on MFCC feature for classification and classified test data for 3 emotions. The accuracy for three emotions was approximately 70%.

In another research, Mihalcea et al. (2011) said that sentiment analysis can be done with a different approach, a multimodal system. They did not use only audio features; they interested in textual and visual features too. For this purpose, they used videos where people talk about any kind of news. For textual features they used bag of words from videos to generate unigram counts. For audio features, they used video clips as well and used OpenEAR tool to calculate some numerical values from voice. They calculated pitch, intensity, loudness for each video clip. For visual features, they used OKAO Vision software. They detected faces and eyes with this software and determined 2 categories for visual features: Smile duration and time of looking at the camera. Thanks to all the features, they experimented this multimodal system and they reached 75% accuracy approximately.

In a different study, Chavhan et al. (2010) developed a model that contains SVM classifier. They used MFCC, MEDC as features and windowing procedure for minimizing the signal fluctuations between the beginning and end of each frame. They used SVM for training and classify the emotions. To data acquisition, they gathered wav formatted files from Berlin Emotional Database. The target emotions was anger, happiness, sadness, neutral, fear. At the end of the research, classification accuracy was 93%.

In another research developed by Franti et al. (2017), CNN classifier was used. In this research, researchers implemented emotion recognition using deep learning technique CNN. They used MFCC feature and PRAAT free software developed for speech analysis. PRAAT software helped them to extract MFCC coefficients and preprocess input data. They used wav formatted voice files. CNN classifier has ReLU activation function and has 400 x 12 neurons for input. The CNN classifier was developed with Python language and TensorFlow library. They gained an accuracy of 71.33% for 6 emotions (happiness, fear, sadness, disgust, anger, surprise).

In another CNN based research, Tripathi et al. (2019) combined speech features and textual features. They used MFCC and spectrogram beside the texts. They used USC-IEMOCAP stands for University of Southern California's Interactive Emotional Motion Capture database for input. First, they developed a model that contains MFCC and spectrogram speech features to classify emotions then, developed a model that contains both speech features and textual features. They tried different combinations of features that can affect the classification result. According to these combinations when they take MFCC and text as inputs. Accuracy reached to 76% approximately.

In a survey generated by Abbaschian et al. (2021), there are several datasets for emotion recognition from voice studies. Some of those databases are IEMOCAP (Interactive Emotional Dyadic Motion Capture Database), TESS (Toronto Emotional Speech Set), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) and EMO-DB (Berlin Database of Emotional Speech). These datasets which differ in the number of emotions and number of samples have been used in different studies. Also, it gives information about the different methodologies for speech emotion recognition task. Some of these, traditional machine learning methods and signal processing methods, HMM and SVM. Artificial neural networks, convolutional neural networks and deep neural networks are classifiers more preferred nowadays. Former studies mostly used traditional methods like SVM, however researchers have been interested in deep neural network implementations recently because of their ability of extracting complex features.

In a deep neural networks-based study developed by Han et al. (2014), the input voice signals have been converted into the segments and a feature vector was generated for each segment. These feature vectors includes MFCC and pitch-based features were used to train the DNN. In the second phase of study after the training of DNN, researchers created utterance-level features from segment-based features and fed an ELM (Extreme Learning Machine) to classify the whole utterance. With ELM they gained higher accuracy contrary to DNN based system. The principle of ELM is to assign weights between the input layer and hidden layer randomly. Then these weights are fixed. Therefore, random weights chosen independent from the training set and it has an advantage of faster training process then standard neural network structure for a small dataset.

One of the other researches aimed to detect speaker and then apply sentiment analysis to classify sentiments from recorded voice dataset. They collected speeches and separate voices from signals and stored in a database. They passed voices through the speech recognition and speaker discrimination systems. While speaker recognition system helps to identify speakers individually, speech recognition system converted these voice packages into texts. Then sentiment analysis from texts have made by researchers positive, neutral and negative results using different type of classifiers like Naïve Bayes, linear SVM.

In another research, Parlak et al. (2013) created a new emotional database that named EmoSTAR for emotion classification task. They gathered voices from TV and internet recordings in order to create the database. It contains 4 emotions (angry, sad, happy, neutral). EmoSTAR database has been used as training dataset. The test set was chosen from EmoDB (Berlin Emotional Database). They used WEKA and openSMILE tools to extract features. Weka tool is used for data mining applications and has its own classifiers like Naïve Bayes, SMO and Bagging

(Bag). They used LPCMCC (Linear Prediction Coding Mel Cepstrum Coefficients) feature to classify. They reached different success rates for different classifiers like 83% for Naïve Bayes, 84% for Bag and the highest success rate of 96% for SMO classifier for 4 emotions. Now, they have been trying to expand capacity of EmoSTAR database.

One of the researches that includes RAVDESS, researchers used CNN model to implement SER (speech emotion recognition). They used spectrograms and MFCC feature to classify emotions from audio signals. According to Mustaqeem et al. (2019), spectrograms were useful in SER because of their rich contents about frequency and time aspects of audio data. Therefore, they did not prefer textual features and choose Softmax as activation function. Generally, they reached 79,5% of accuracy with RAVDESS dataset.

In another research that involves multimodality, researchers chose both visual and audio features in order to classify emotions. Jimenez et al. (2021) used RAVDESS dataset and used several classifiers like CNN, SVM and logistic regression. Because of 2 types of input data, they trained CNN with audio features and they trained the network that called AlexNet with images. They reached maximum 80.08% of accuracy with SVM classifier and RAVDESS dataset.

Another study that includes audio signals, RAVDESS dataset were used. Byun et al. (2021) classified emotions with different input types like audio and image. They used MFCC, spectral features, chroma and harmonic features. CNN was used for visual features, whereas RNN classifier was used to classify emotions from speech files. Consequently, they reached 87% of accuracy approximately.

3. METHODOLOGY

In this section, the methodology that has been followed to implement system will be mentioned. But before that, an informative summary will be given about the system ingredients. In machine learning problems, there are several issues to consider. One of them is data collection. Data collection is crucial for most of the experimental researches. If the data that gathered is not enough, classifier cannot train itself good (the collected data is not enough to learn the details in the all space). So, high accuracy cannot be achieved on test data (or real examples). Otherwise, if there is too much data than it should be, it may cause overfitting (memorizing the given examples). So that, enough data should be supplied in equal size for each class. Data can be text, image, voice or it can be all of them. Any social media platform, television or internet videos, images, voice recordings, plenty of texts can be used as data source. If it is possible, a dataset can be created by researchers.

3.1. Data Acquisition

Data was collected from RAVDESS dataset that stands for Ryerson Audio-Visual Database of Emotional Speech and Song named by Livingstone et al. (2018). This dataset consists of voice recordings of several emotions (angry, sad, happy, disgust, fear, neutral, surprised, calm) and video recordings to assist these emotions with visual materials. So, it can be said that this is a multimodal database. These are 3 different environment that contains these emotions. These are audio-only, visual only and audio-visual environments. 24 professional actors (12 male, 12 female) contributed the database with their own voices or faces. RAVDESS contains thousands of voice recordings totally. All voice data are kept as wav formatted and all files are determined according to pre-defined indicators like modality, vocal channel, emotion, intensity and actor number. For example, emotions were determined from 01 to 08. Whether a file is a speech file or video, it is also indicated by vocal channel. Therefore, any file in this database has unique name. For this study speech files and song files were used for train and test data sets. And every voice recording was used to extract features. These features feed a multilayer perceptron. Details about this MLP classifier is given at the following sections.

3.2. Mlp Classifier

Artificial neural networks or commonly known as ANN has been used in many applications. An ANN is a system that facilitates the machine learning from input data. Its main purpose is to take input and process the input according to a certain activation function and produce expected output. However, the output produced may not be equal to the required output in the training phase. In this case, the weight values of the input neurons are updated and the input neurons are included in the calculation again (Nazzal, 2008). The difference between the desired output and the obtained output is an indicator of the learning success of the system. This cycle is called backpropagation. The purpose is to reduce the difference between the desired output and the obtained output. Input weights change randomly, so the outputs are different each time. An artificial neural network basically consists of 3 main parts. Input layer, hidden layer(s), and output layer. When certain inputs are given to the system, certain outputs are expected to be produced. And outputs are compared with predefined output labels in the training phase. For this reason, artificial neural network system is one of the supervised learning methods. In Figure 3.1 a simple single layer input neuron can be seen.

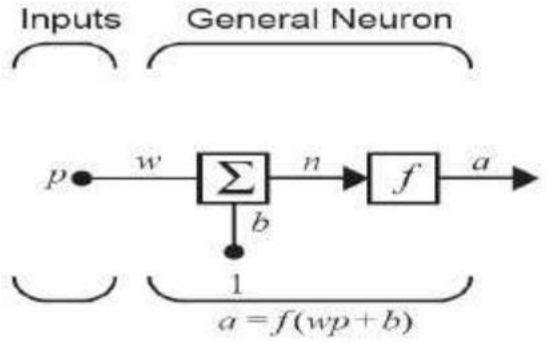

Figure 3.1. Single layer input neuron (Nazzal, 2008)

Where input p is multiplied by w (weight) value and added to the bias value b to generate neuron value n. Finally, a is neuron output according to generated f(n) value (activation function). Different activation functions can be used according to the problem type. Sigmoid function is usually used in machine learning applications. Sigmoid function is a function that takes input between – infinity and + infinity and align the output within the range 0-1. It is used in multi-layer perceptrons generally that are trained with the backpropagation algorithm (Nazzal, 2008). For example, in Figure 3.2, a multi-layer perceptron can be seen.

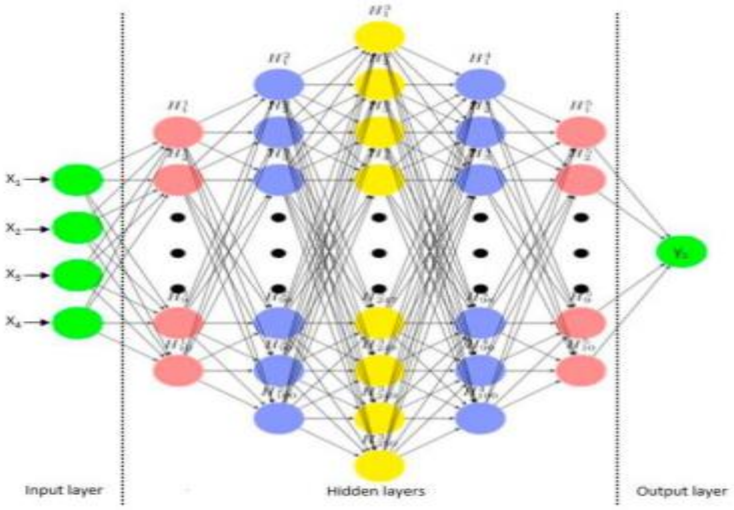

Figure 3.2. Multi-layer Perceptron Architecture (Serin, 2021)

An MLP, consists of several layers and each layer has its own weights and bias values. A multi-layer perceptron or an artificial neural network is generally used for training a certain proportion of input data and expected to fit model. Then test data is applied to the trained model and prediction is evaluated for accuracy of the model. Prediction varies according to research topic, features to be extracted and input data. It can be an object that recognized by image processing, a music that classified in terms of genre can be criterion of accuracy. However, it is the emotional state that will be measured in this study. If it is desired to be examined technically, multi-layer perceptrons have a more complex structure than single-layer ones. The output of neurons in the input layer is the same as their input and those outputs form the inputs of the middle layer (hidden layer). According to Hounmenou et al. (2021) hidden layer's net value can be calculated by summation of multiplying related input values and weight values within the same direction. Then this value is added to the bias.

According to (Formula 3.1) j indicates the neuron number in hidden layer and k indicates the neuron number in the input layer.

$$Net_j = \sum \mathrm{I_k} * \mathrm{w_k} + b \quad (3.1)$$

$\mathrm{w_k}$ symbolizes the weight value of the kth neuron in the hidden layer path. $\mathrm{I_k}$ symbolizes the input value of the kth input neuron. Net input value of jth hidden layer neuron can be found by this summation of the k multiplication operations. This process is needed for every hidden layer neuron. The output value is calculated by giving the Net_j value to a predetermined activation function. Activation functions can be divided into 2 sections according to their linearity.

3.2.1. Activation functions

In order to calculate the output value of a neuron in hidden layer, it is necessary to have knowledge about the activation functions. Activation functions differ from each other according to mathematical formula and selected in terms of subject of the study. Its purpose bring the output of a neuron closer to non-linearity. Basically, these functions are divided into 2, linear functions and non-linear functions.

3.2.1.1. Linear activation functions

Linear functions do not help with the future complexity. Also, linearity does not support the model among various data. Because output value of a neuron is not confined between a range. So, output value can vary between -infinity and +infinity. As you can see, in Figure 3.3, there is a linear function that is plotted f(x) = x and its borders are extended from -infinity to +infinity. In Figure 3.3 a linear activation function (linear equation) can be seen.

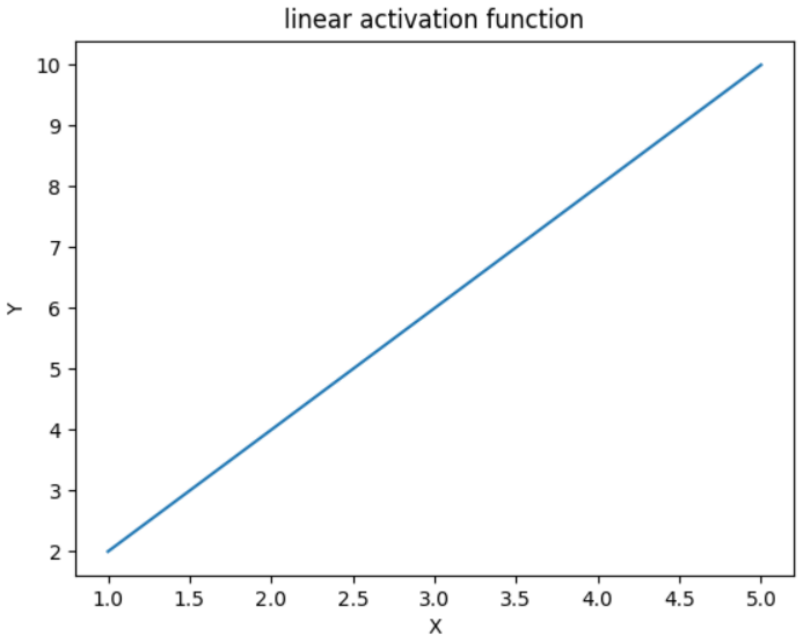

Figure 3.3. Linear activation function

3.2.1.2. Non-linear activation functions

Non-linear activation functions are the most used functions in machine learning and neural networks applications. This type of functions has ability to adapt when the data have more dimensions. In addition, since there is a limitation between specific borders, the outputs can vary between reasonable values. Like in the Figure 3.4, sigmoid function returns output values between [0,1]. Non-linearity provides smoothness to the graphics. The purpose of nonlinear functions is to approximate linear input values to nonlinear outputs. According to Nwankpa et al. (2018) this conversion is done for the next steps and calculations. Several commonly used linear and non-linear activation functions will be mentioned briefly in the following sections.

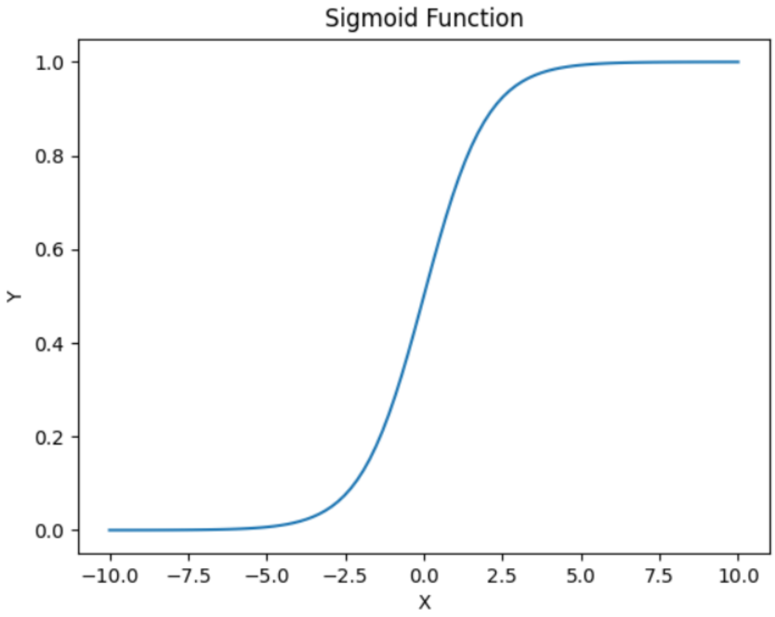

Figure 3.4. Non-linear activation function

3.2.1.2.1. Sigmoid function

Sigmoid functions are commonly referred as logistic functions or squashing functions according to Turian et al. (2009). This type of non-linear activation functions are used in probabilistic output values and can be applied to the binary classification problems generally. It shapes as 'S' and output values are between 0 and 1 using the Formula 3.2.

$$f(x) = 1/(1 + e^{-x}) \tag{3.2}$$

In this equation sigmoid activation function can be seen. Value e is a constant and x is the input.

3.2.1.2.2. Hyperbolic tangent function (Tanh)

Hyperbolic tangent function is another type of activation function similar to the sigmoid activation function (Figure 3.5). Tanh activation function generally has better training performance for multi-layer neural networks than the sigmoid function according to Nwankpa et al. (2018). Output range is between -1 and 1.

Therefore, the mean of the hidden layer approximates to 0 and this helps in data centralization. Additionally, it is a shifted version of sigmoid function and smoother as well. So, they can be produced from each other. The formulization of hyperbolic tangent function can be shown as follows. According to the (Formula 3.3), hyperbolic tangent function can be given as follows.

$$f(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x}) \tag{3.3}$$

The x value represents the axis. Constant e represents the exponentiality.

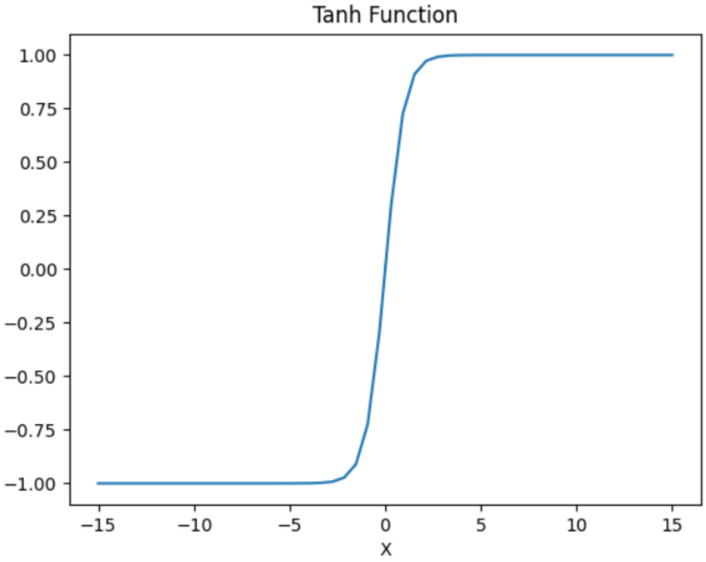

Figure 3.5. Tanh activation function

3.2.1.2.3. Rectified linear unit function (Relu)

Rectified linear unit function (Figure 3.6) is the mostly used and successful activation function in nowadays neural network researches according to Nwankpa et al. (2018). Among the reasons for this intensive use is that it is suitable for most of the different problems and not too complex mathematically. The function (3.4) mathematically returns 0 if the input value is 0 or less. For values greater than 0, the input value becomes output directly. This equation reduces complexity and processing time in neurons. Consequently, it makes ReLU faster than the sigmoid and tanh function. The formulation resemble a linear activation function, however it is basically a nonlinear activation function. The output range is from 0 to +infinity. Its function is defined as follows.

$$f(x) = \max(0, x) = \begin{cases} x_i, & if \ \ x_i \geq 0 \\ 0, & if \ \ x_i < 0 \end{cases} \tag{3.4}$$

In this study ReLU activation function has been used as well because of the complexity minimalization and fast processing in hidden layer neurons. Variable x_i symbolizes input.

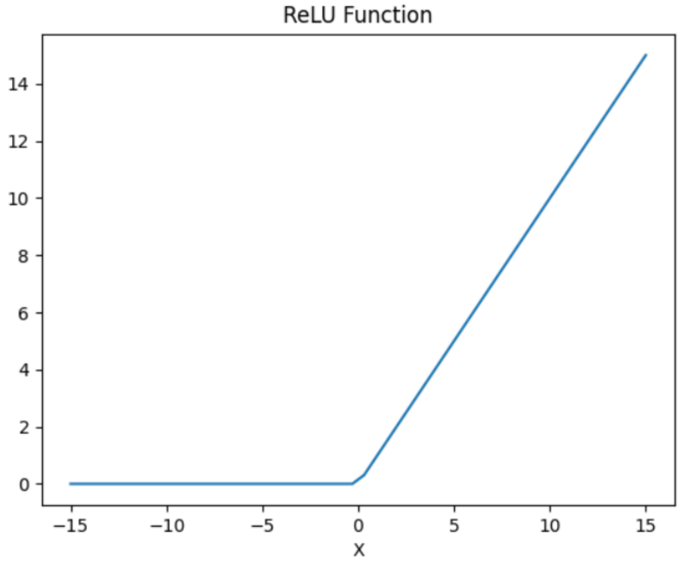

Figure 3.6. ReLU activation function

3.2.1.2.4. Softmax function

Softmax function (Figure 3.7) is another type of non-linear activation function similar to the sigmoid function. Hence the outputs are approximated to interval 0 and 1 included. But this function is more useful for multi-classification problems and where outputs are divided by the sum of the outputs. First, exponential of all values in input array is done and each individual result is divided by sum of these results as given in Formula 3.5. The final values are the probability values of each class.

$$f(x) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_i}} \tag{3.5}$$

K is the number of classes. Value of x_i symbolizes the input array's elements. The numerator section denotes exponential function for each input values. The denominator section denotes summation operation of all the values that has been calculated in numerator. The primary difference between sigmoid function and softmax function is that the sigmoid function is more suitable for binary classification while the softmax function is used for multiple classification studies.

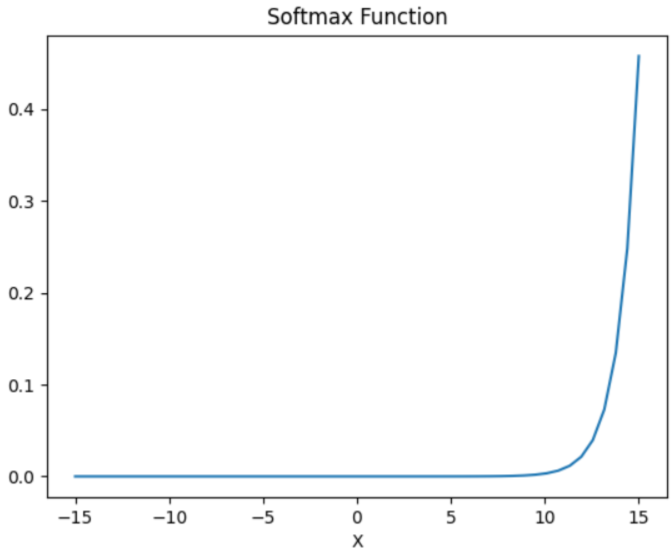

Figure 3.7. Softmax activation function

Output of the function (Formula 3.6) is become the output of specific hidden layer neuron. Net_j is the final input value of specific hidden layer neuron. O_1 is the output value that have calculated recently, is evaluated as the input value of the next neuron.

$$O_1 = f(Net_j) \tag{3.6}$$

While calculations continue in this way, the neurons in the output layer are reached to and output value of the model is obtained. As discussed above, output value may be far from the deserved output. Backpropagation method can be used to reduce this inequality (error) between desired output and the actual output. There are several parameters need to be discussed before creating a model. ReLU activation function was used in this research because of its non-complexity and faster processing compared to other activation functions. But these activation functions were not the only parameters that we needed to take into consideration. In a machine learning research, features are so important for successful model. Which features that are used and how we use them are also important for the research. Since most of the sentiment analysis approaches that use machine learning techniques, the salient features of texts, images or voices are represented as feature vector (Gebremeskel, 2011). For this study, every voice recording in dataset was examined in terms of each voice feature that we selected.

3.3. Feature Extraction

Several features are required to be used in every research or for data types such as image, voice or text specifically. An image has specific features like intensity, sharpness, resolution etc. Text-based data has its own features like word types, punctuations, emoticons etc. But according to (Mohammad, 2017) textual features and visual features have their own disadvantages. For an image, noise, deterioration or resolution quality may be a problem. For text document, idioms, emojis and sarcastic statements may be a problem. Because of such problems, moods may not be properly classified. So, accuracy may remain low. Audial features have some obligations like noisy data and complex voices in terms of classification accuracy. These features can be used not only for sentiment analysis researches, but also for other machine learning studies. For our study, because of voice recordings are used for sentiment analysis, audio features were used and these features will be discussed. MFCC, zero crossing rate, spectral centroid, spectral flux, spectral roll-off, chroma and contrast are among the most used audio features in machine learning studies.

3.3.1. Mfcc (Mel frequency cepstrum coefficients)

MFCC is abbreviation of Mel Frequency Cepstrum Coefficients which is the mostly used feature in voice researches. Because MFCC is a set of features that facilitates to determine the shape of sound. It contains approximately 40 individual features and MFCC is a criterion showing the perceptron of the human ear's hearing in specific frequencies according to Turan et al. (2021). A sample MFCC feature spectrogram of a word can be seen in Figure 3.8.

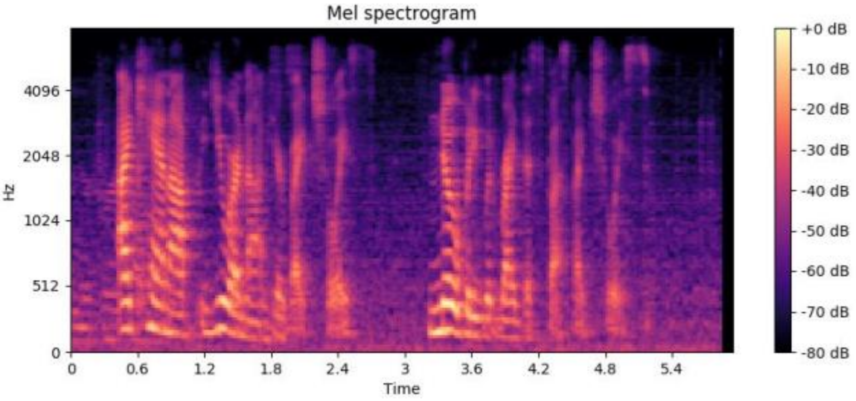

Figure 3.8. Sample spectrogram of a word (Tripathi, 2019)

For such a graph, researchers can obtain the emotion from the distance between peak and bottom values. For example, happiness emotion has smaller distance between peak and bottom. MFCC has two types of filters that are spaced linearly at lower than 1000 Hz and spaced logarithmic upper than 1000 Hz according to Muda et al. (2010).

3.3.2. Zero crossing rate

Zero Crossing Rate is another audio feature that can be used in vocal researches. Generally, it measures that how many times the amplitude of voice signal passes through 0 level in a specific time. This feature is used generally in speech recognition and music information retrieval. Figure 3.9 presents shape of zero crossing rate of an example word.

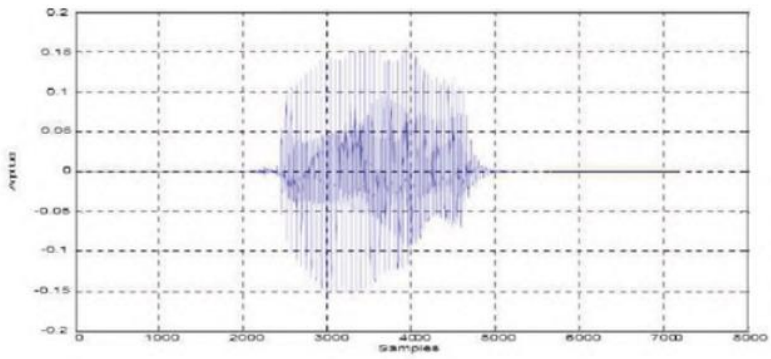

Figure 3.9. Zero crossing rate of a word (Barkana, 2015)

Zero crossing rate can be used in differentiate the voiced and unvoiced signal and can be calculated as follows (Darji, 2017). According to Formula 3.4, it can be found how many times a voice signal crosses the horizontal axis.

$$F = \frac{1}{T-1}\sum_{t=1}^{T-1} F(S * S - 1 < 0) \qquad (3.4)$$

Where S is the signal of T time length voice, this formulation can help us to calculate zero crossing rate.

3.3.3. Spectral flux

Spectral flux feature helps to calculate the spectral rate of change. Therefore, it is given by frame-to-frame difference of the spectral vector. In addition to this, spectral flux is a part of spectral features. According to Sadjadi et al. (2013) this feature is extracted in frequency domain and if there is a high flux value, it can be said that there has been a sharp change. This calculation is given in Formula 3.5.

$$F_k = \sum_{r=1}^{\frac{N}{2}}(|X_k[r]| - |X_{k\ \_\ 1}[r]|)^2 \qquad (3.5)$$

X_k symbolizes the sample rate of voice and spectral rate of change can be found according to this formula.

3.3.4. Contrast

Contrast examines the spectral peak, spectral valley and their difference in each sub-band. According to Jiang et al. (2002) contrast feature may provide better separation than MFCC in terms of music type classification. It is mostly used in music type classification. Frequency sub-bands are determined according to some specific intervals like 0hz-200hz, 200hz-400hz etc.

3.3.5. Chroma

Chroma feature is a set of features like MFCC. However, chroma is a feature that represents the tonal content of an audio signal and it is referred as color of the sound. Chroma feature captures melodical and harmonic information of voice. Chroma examines this information in terms of twelve different pitch classes.

Two of the primary chroma features as follows:

- Chroma vector is a feature vector that indicates how much energy of each pitch class has (Shah et al., 2019).
- Chroma deviation is another primary feature of chroma. It denotes deviation of each 12 pitch classes.

In the Figure 3.10, it can be seen a chromagram produced by an audio recording.

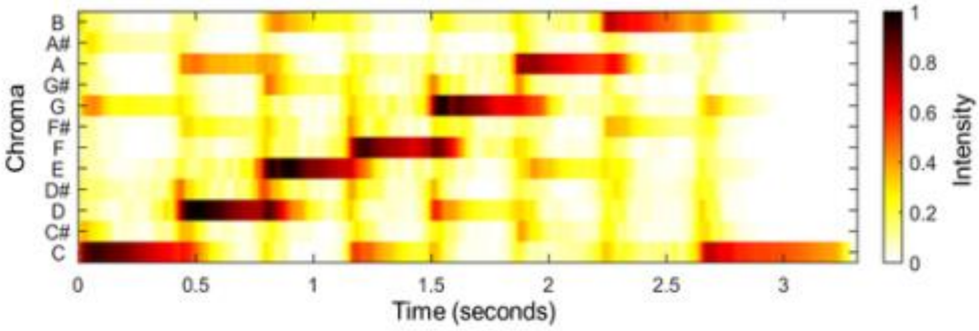

Figure 3.10. Chromagram of an audio recording (Shah et al., 2019)

Chromagram means that time-chroma representation of a voice in terms of 12 pitch classes.

3.3.6. Spectral roll-off

Spectral roll-off is a feature that measures the shape of the voice signal. It indicates proportion of spectral energy and it can be extracted within a sample rate of data file. Spectral roll-off can be formulated as given Formula 3.6.

$$\sum_{k=1}^{K} X[k] \geq 0.85 \sum_{k=1}^{N} X[k] \qquad (3.6)$$

According to Kos et al. (2000), we can see that the K variable is the spectral roll-off and the minimum index of the spectral energy of voice signal satisfying the inequality (3.6), where X[k] is the energy of voice signal and N is the length of the spectrum.

3.3.7. Tonnetz

Tonnetz feature is an exposition of harmonic links in music. According to Humphrey et al. (2012) it is a geometric representation of equal tempered pitch intervals in musical tuning. Tonnetz feature is a diagram that represents tonal features of 12 pitches. Therefore, it can be used in musical research that contains harmony. It can be used in automatic chord transcription, song segmentation studies (Humphrey et al., 2020). In the Figure 3.11 it can be seen the modern rendering of the tonnetz feature.

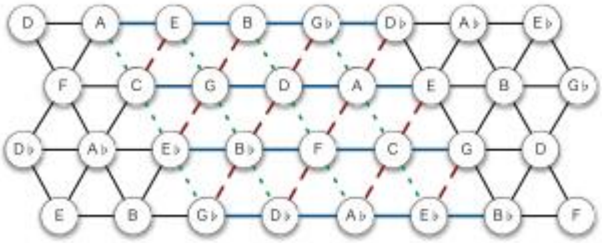

Figure 3.11. Tonnetz feature representation of an audio (Humphrey et al., 2020)

3.3.8. Spectral centroid

Spectral centroid indicates center of speech signal mass. Therefore, it can be obtained where the most of voice energy is intensified. It is a measurement of brightness of audio signal (Darji, 2017). So, it can be said that if there is a high energy placed, then centroid will approach that place. According to Kalamani et al. (2014) high values belongs to brighter sounds and spectral centroid feature can be calculated using Formula 3.7.

$$C = \frac{\sum_{m=0}^{N-1} f(m) X_i(m)}{\sum_{m=0}^{N-1} X_i(m)}$$

(3.7)

In the formula, C denotes result centroid and f(m) and $X_i(m)$ are the center frequency and the amplitude of voice signal respectively.

3.3.9. Pitch

Pitch is a feature that declaration of each person's subjective perception of audio signal. It is auditory sensation for assigning musical tones into a musical scale. Therefore, it cannot be directly measured. Pitch is like frequency but frequency is a scientific attribute. So, it can be calculated.

If we want to classify sound waves within the scope of the signal area, we can classify them as in the Table 3.1 below. Time domain features are derived from unprocessed (raw) sound waves. Frequency domain contains features that includes the numerical values of the sound, such as frequency. Time-frequency domain features, on the other hand, contain features that use these 2 types.

Table 3.1. Signal domain and feature relation

Signal Domain	Feature
Time Domain	Zero crossing rate
Frequency Domain	Spectral bandwith, spectral centroid, spectral flux
Time-Frequency Domain	MFCC, chroma, contrast, spectral roll-off, spectral centroid

3.4. Methodology of System

We used the functions of the librosa library in python for development in this research. Thanks to the librosa library, we had the opportunity to extract and examine the characteristics of the sound files in a short time and accurately. Therefore, this library has a great influence on the creation of feature vectors. Librosa is a built-in library that was written in Python programming language in order to examine audio signals, generate spectrograms and extract features. Python 3.6.5 version was used for this study and sklearn library was required to design multilayer perceptron model and for training and test operations.

3.4.1. Data gathering and feature extraction

As mentioned in the methodology section before, data was collected from RAVDESS database. Wav formatted speech files and song files were used. Once data were complete, whole arranged for file reading in order to become input for MLP neural network. So that, all data files were located in same directory and their names were split into according to the emotion numbers specified by RAVDESS. For arranging and splitting operations in program, “glob” and “os” libraries in Python were used. Then, feature extraction was made on the current file. Feature extraction includes examining the features of the audio signal and using it for emotion classification. Basically, there are two processes, examining the signal in an audio file according to predetermined features and transferring it to an array. The algorithm of this process extracts the features and stack them into an array, finally array is returned back. Since the features to be examined will change according to the experiment planned, the length of the array may also change. After extracting the features and determining the mood for a particular audio file, the extracted features and their related emotions are stored to be processed later. When this process is repeated for all audio files, we have the features of all audio files. Inputs with arrays of features and arrays of emotions are given to the “train test split” function. As output, training set and test set are obtained. The density of training and test sets can be changed as any proportion. Finding the optimal proportion is important in emotion classification. In this research, the optimal rate was determined as 75% training and 25% testing. For example, if wanted to have the MFCC feature in a signal, the following code snippet can be used:

```
mfcc_feature = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40)
.T, axis=0)

result = np.hstack((result, mfcc_feature))
```

In this code snippet, **np** means NumPy library that facilitates the numerical operations in Python. **Mean** returns the average of the array elements, librosa is the audio feature examination library, **X** is the sound file that is going to be read, **sample_rate** is the sample rate of sound file and **n_mfcc** denotes the number of MFCC features to return. The result variable adds the currently examined feature to the feature vector using the numpy hstack method. **Hstack** method is used to concatenate array elements one after another.

3.4.2. Data normalization (Z-score standardization)

Data normalization is an important step for machine learning. Data normalization or standardization are 2 methods which prepare dataset for train. Main purpose of these methods is scaling the data. Because the data set itself may not be suitable to enter directly into the classifier. By applying normalization, the dataset can be scaled. Normalization is mostly done to fit the dataset between 0 and 1. However, z-score standardization was applied in this research. As a result of this process arithmetic mean of new dataset formed is 0 and its standard deviation is 1(Ali et al., 2014). To apply z-score standardization to the dataset, it is necessary to subtract the arithmetic mean of the dataset from each element and divide the result by the standard deviation as given in Formula 3.8.

$$Z = \frac{x-mean}{SD} \tag{3.8}$$

In the Formula 3.8, x represents each element in dataset, mean is arithmetic mean and SD denotes standard deviation. Z is the result that standardized dataset. After data normalization, data is input to a machine learning model like support vector machines, artificial neural networks etc. classifiers to train model.

```
Mean = np.mean(train_set, axis=0)

standard_deviation = np.std(train_set, axis=0)

train_set = (train_set – mean) / standard_deviation
test_set = (test_set – mean) / standard_deviation
```

This code snippet is used for data standardization in the research. Once the mean and standard deviation is calculated, dataset is set to their new values according to z-score standardization approach.

3.4.3. Classifier training and model initialization

So far, all the tasks such as data gathering, feature extraction, split into training and test sets have been implemented. The structure of machine learning model has been designed. Later, the model trained with train data set and prediction accuracy was evaluated with test data set. There are some parameters should be considered while model design. As discussed before, hidden layer size, activation function, learning rate and maximum iteration number are the parameters important in multilayer perceptron. Hidden layer size represents the number of neurons in the specific hidden layer (default is 100). Activation parameter is required for that which activation function will be chosen for hidden layer (default is ReLU). Learning rate determines the update pattern of weights. Its options are constant, adaptive, invscaling but default option is constant. Finally, max_iter parameter determines maximum number of iterations. In this research, max_iter was chosen as 700, learning rate was chosen adaptive and hidden layer size was chosen as 200 neurons. Activation function remained as default.

4. EXPERIMENTAL SETUP AND RESULTS

This section includes all the experiments executed for the system. According to Huang et al. (2019), to receive higher accuracy, which features are selected and what data pre-processing approximation is applied to is important. Experiments are important for machine learning studies to be evaluated accurately. Each experiment was prepared according to the combinations of features and existence of data pre-processing. The four emotions are angry, sad, neutral and happy and features are MFCC, chroma, contrast, zero crossing rate, spectral roll-off and spectral flux. MFCC feature discarded and the other features remain, accuracy is approximately 77.64% in experiment I, where data standardization is applied to. The confusion matrix of the experiment I is given at Figure 4.1. As MFCC plays an important role in modeling sound, the accuracy remained less than the best obtained. The confusion matrix in the Figure 4.1 can show the success of the emotion classification. Rows and columns designates the emotions. Therefore, angry, sad, neutral, happy emotions are shown in order (0-1-2-3). For experiment I, it can be said that 78 angry voices, 67 sad voices, 35 neutral voices and 70 happy voices are classified correctly.

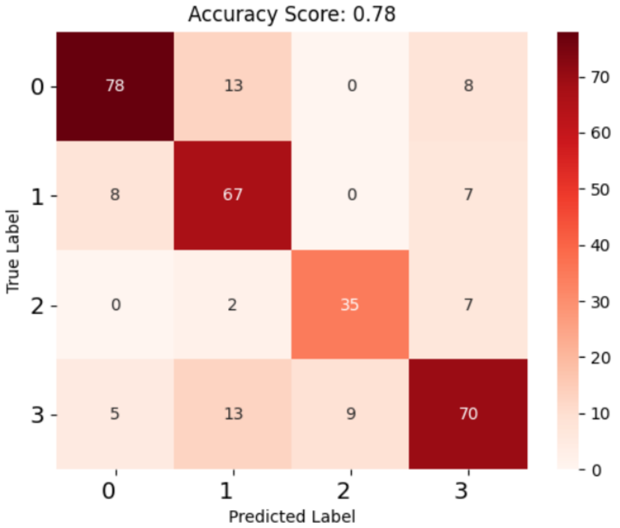

Figure 4.1. Confusion matrix of experiment I

With similar features, but data standardization is not applied to executed as experiment II (Figure 4.2). As a result of the removal of both MFCC feature and data standardization, accuracy decreased dramatically. Accuracy is very low being 34.16%.

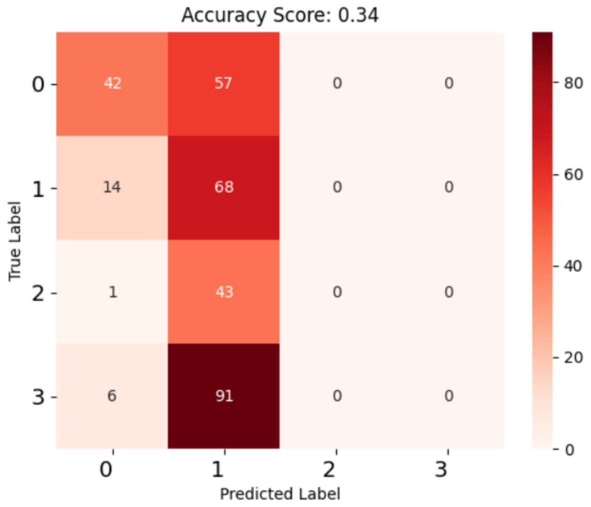

Figure 4.2. Confusion matrix of experiment II

In experiment III (Figure 4.3), only MFCC feature is extracted. Data standardization is not applied. Therefore, execution time decreased. As can be seen in experiment III, 80 angry voices, 66 sad voices, 34 neutral voices and 68 happy voices classified correctly. Accuracy reached 77.02%.

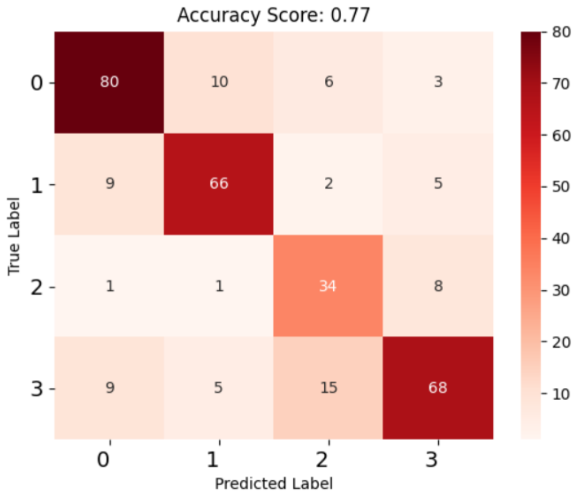

Figure 4.3. Confusion matrix of experiment III

For experiment IV (Figure 4.4), only MFCC feature was extracted and data standardization is enabled. As a result of the change, accuracy reached 86.34%. These 4 emotions are better classified in this experiment.

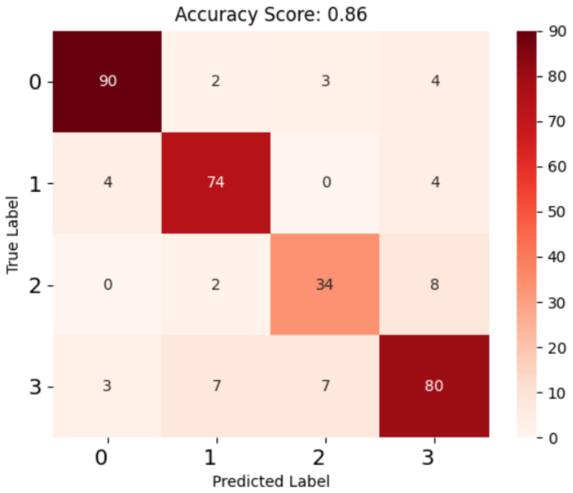

Figure 4.4. Confusion matrix of experiment IV

In experiment V, every feature that mentioned in the beginning of experimental setup section (MFCC, chroma, contrast, zero crossing rate, spectral roll-off and spectral flux) are extracted (Figure 4.5). In addition to that data standardization enabled. Accuracy reached 87.58%. This is the best accuracy we have had of all the experiments we have done. Most successful classification of the voice recordings have been made according to the 4 emotions.

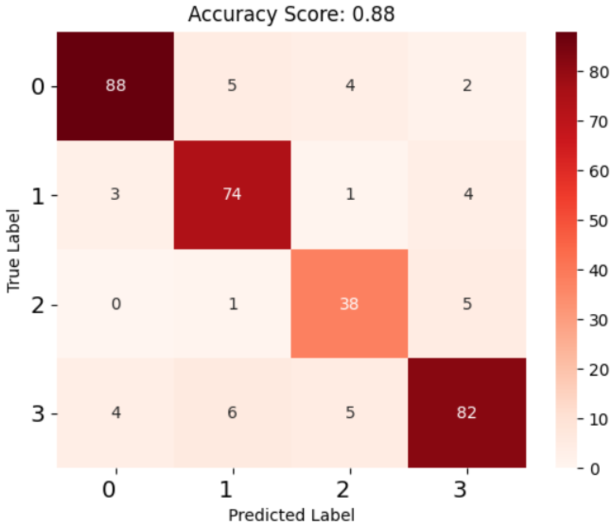

Figure 4.5. Confusion matrix of experiment V

When tonnetz feature is added to the experiment that gave the best accuracy and experiment is repeated, the accuracy reached 86.34%. Its confusion matrix can be seen in Figure 4.6. As mentioned in the methodology section, tonnetz is a feature that includes harmonical information in a voice signal. Therefore, it needs more time to extract.

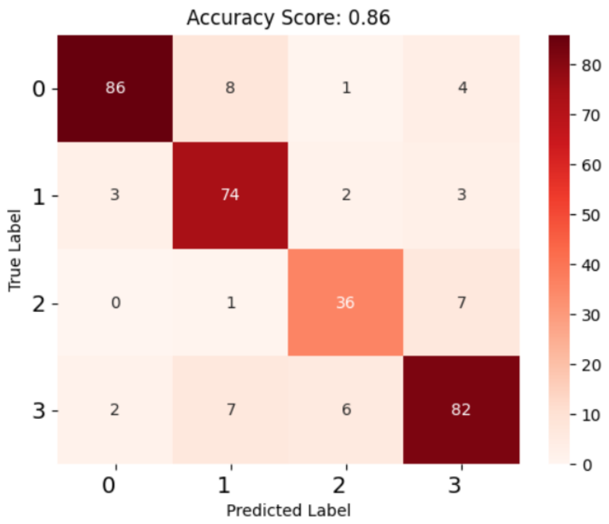

Figure 4.6. Confusion matrix of experiment VI

When spectral centroid feature is used in addition to the experiment that gave the best result, the accuracy reached 86.02%. According to Figure 4.7 new confusion matrix can be seen.

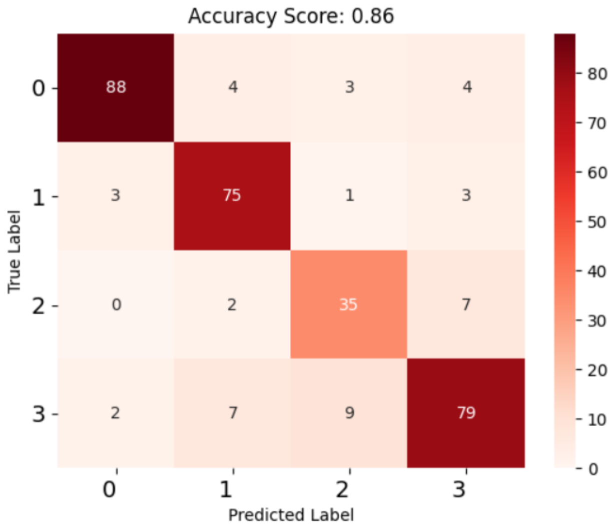

Figure 4.7. Confusion matrix of experiment VII

According to the some researches referred on the literature which uses RAVDESS dataset and our proposed different trained systems are compared in terms of features and gained accuracies in Table 4.1. Consequently, it can be seen that some features have more effect on the accuracy than others.

Table 4.1. Comparison of 4 research

Research	Features	Accuracy
Mustaqeem (2019)	Spectrograms, MFCC	79,5%
Jimenez (2021)	Images, MFCC	80,08%
Byun (2021)	Images, MFCC, spectral features, chroma and harmonic features	87%
Proposed (Without z-score normalization and MFCC)	Zero crossing rate, chroma, contrast, spectral flux, spectral roll-off	34,16%
Proposed (Without MFCC, with z-score normalization)	Zero crossing rate, chroma, contrast, spectral flux, spectral roll-off	77,64%
Proposed (With z-score normalization and MFCC)	MFCC, zero crossing rate, chroma, contrast, spectral flux, spectral roll-off	88%

5. CONCLUSIONS AND IMPLICATIONS

The emotions classified are sadness, happiness, neutral and anger. The reason for choosing these 4 emotions is that they are common emotions being in interest for a meeting room. There are also some audio features that we didn't consider for evaluation. Examples of these unused features are spectral centroid and tonnetz. When using the Tonnetz feature, the training time increases too much and there was no significant increase in accuracy as expected. For this reason, this feature was not considered. When the spectral centroid feature was also used, there was no significant increase in accuracy as expected, so spectral centroid was not used too. Therefore, in this research, it is seen that which features affect such system more significantly and the importance of the data preprocessing. Due to the lack of data standardization, the classification accuracy for some experiments could not achieve the desired level. According to Swamy et al. (2013), data standardization is critical for training. Additionally, feature extraction is highly important for gaining high accuracy according to Parikh et al. (2020). It was seen that using all features at once, can reduce the accuracy. Especially for the last two experiments, it can be seen that the additional features used did not increase the accuracy at the desired rate and were not included in the combination that gave the best result.

REFERENCES

Abbaschian, B., Sosa, D., S., Elmaghraby, A., S., 2021. Deep Learning Techniques for Speech Emotion Recognition, from Database to Models. Sensors, 21(4), 1-27.

Barkana, B.D., Bachu, R.G., Adapa, B., Kopparthi, S., 2008. Separation of Voiced and Unvoiced Using Zero Crossing Rate and Energy of the Speech Signal. American Society for Engineering Education, March 2008, America.

Bhattacharjee, D., Bhowmik, M. K., Nasipuri, M., Basu, D. K., Kundu, M., 2010. Classification of Fused Face Images Using Multilayer Perceptron Neural Network. Department of Computer Science and Engineering, 27 January, India, 1-6.

Byun, S. W., Lee, S. P., 2021. A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms. Applied Sciences, 11(4), 1890-1905.

Chavhan, Y., Dhore, M., L., Yesaware, P., 2010. Speech Emotion Recognition Using Support Vector Machines. International Journal of Computer Applications, 1(20), 6-9.

Darji, M., 2017. Audio Signal Processing: A Review of Audio Signal Classification Features. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(2), 227-230.

Davletcharova, A., Sugathan, S., Abraham, B., James, A., P., 2015. Detection and Analysis of Emotion from Speech Signals. Procedia Computer Science, In Press.

Franti, E., Ispas, I., Dragomir, V., Dascälu, M., Zoltan, E., Stoica, I., C., 2017. Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots. Romanian Journal of Information Science and Technology, 20(3), 222-240.

Gebremeskel, G., 2011. A Sentiment Analysis of Twitter Posts About News. University of Malta, Department of Computer Science and Artificial Intelligence, M.Sc. Thesis, 123, Malta.

Han, K., Yu, D., Tashev, I., 2014. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. International Speech Communication Association, 14-18 September 2014, Singapore, 223-227.

Hounmenou, C., G., Gneyou, K., E., Kakai, R., G., 2021. A Formalism of the General Mathematical Expression of Multilayer Perceptron Neural Networks. Preprint, 1(5), 1-12.

Huang, A., Bao, P., 2019. Human Vocal Sentiment Analysis. NYU Shanghai Computer Science Symposium, 19 May, Shanghai, 1-16.

Humphrey, E. J., Cho, T., Bello, J. P., 2012. Learning a Robust Tonnetz-Space Transform for Automatic Chord Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, 25-30 March, Kyoto, 453-456.

Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H., Cai, L. H., 2002. Music Type Classification by Spectral Contrast Feature. IEEE International Conference on Multimedia and Expo, 26-29 August, Lausanne, 0-3.

Jimenez, C. L., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M., Martinez, F. F., 2021. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. Sensors, 21(22), 7665-7694.

Kalamani, M., Valarmathy, S., Anitha, S., Mohan, R., 2014. Review of Speech Segmentation Algorithms for Speech Recognition. International Journal of Advanced Research in Electronics and Communication Engineering, 3(11), 1572-1574.

Kim, K. H., Bang, S. W., Kim, S. R., 2004. Emotion Recognition System Using Short-Term Monitoring of Physiological Signals, Medical and Biological Engineering and Computing, 42(3), 419-427.

Kos, M., Kacic, Z., Vlaj, D., 2000. Speech Bandwidth Classification Using General Acoustic Features, Modified Spectral Roll-Off and Artificial Neural Network. Mathematical Models and Methods in Modern Science, 212-217.

Livingstone, S., R., Russo, F., A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, 13(5), e0196391.

Mäntylä, M. V., Graziotin, D., Kuutila, M., 2018. The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. Computer Science Review, 27(2), 16-32.

Mihalcea, R., Rosas, V., P., Morency, L., P., 2013. Multimodal Sentiment Analysis of Spanish Online Videos. IEEE Intelligent Systems, 28(3), 38-45.

Mohammad, M., 2017. Challenges in Sentiment Analysis, National Research Council Canada, 61-83.

Muda, L., Begam, M., Elamvazuthi, I., 2010. Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Wrapping (DTW) Techniques. Journal of Computing, 2(3), 138-143.

Mustaqeem, I., Kwon, S., 2019. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. Sensors, 20(1), 183-198.

Nazzal, J., M., Emary, I., M., Najim, S., A., 2008. Multilayer Perceptron Neural Network (MLPs) For Analyzing the Properties of Jordan Oil Shale. World Applied Sciences Journal, 5(5), 546-552.

Nwankpa, C., E., Ijomah, W., Gachagan, A., Marshall, S., 2020. Activation Functions: Comparison of trends in Practice and Research for Deep Learning. 2nd International Conference on Computational Sciences and Technologies, 13-19 December, Jamshoro, 124-133.

Parikh, D., Sachdev, S., 2020. Improving the Efficiency of Spectral Features Extraction by Structuring Audio Files. IEEE-HYDCON, 11-12 September, Hyderabad, 1-5.

Parlak, C., Diri, B., Gürgen, F., 2014. A Cross Corpus Experiment in Speech Emotion Recognition. Workshop on Speech, Language and Audio in Multimedia, 11 September 2014, Peneng, Malaysia, 1-5.

Sadjadi, S. O., Hansen, J. H., 2013. Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux. IEEE Signal Processing Letters, 20(3), 197-200.

Saglani, K., Janwe, N., 2020. Machine Learning Based Sentiment Analysis for Text Messages. International Journal of Computing and Technology, 7(4), 425-428.

Serin, G., Sener, B., Gudelek, M., U., Ozbayoglu, A., M., Unver, H., O., 2020. Deep Multi-Layered Perceptron based Prediction of Energy Efficiency and Surface Quality for Milling in the Era of Sustainability and Big Data. 30th International Conference on Flexible Automation and Intelligent Manufacturing, 15-18 June 2020, Athens, 1166-1177.

Shah, A. K., Kattel, M., Nepal, A., Shrestha, D., 2019. Chroma Feature Extraction. Chroma Feature Extraction Using Fourier Transform, Nepal, January, 1-13.

Swamy, M.N., Du, K.L., 2013. Neural Networks and Statistical Learning. Springer London, 856, London.

Szabo, R., Gontean, A., 2013. Human Voice Signal Synthesis and Coding. 12th IFAC Conference on Programmable Devices and Embedded Systems, 25-27 September 2013, Czech Republic, 336-341.

Tripathi, S., Kumar, A., Ramesh, A., Singh, C., Yenigalla, P., 2019. Deep Learning Based Emotion Recognition System Using Speech Features and Transcriptions. International Conference on Computational Linguistics and Intelligent Text Processing, 7-13 April 2019, La Rochelle, France, 1-12.

Turan, M., İleri, M., 2021. Sentiment Analysis of Meeting Room. 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 11-13 June, Ankara, 1-5.

Turian, J., Bergstra, J., Bengio, Y., 2009. Quadratic Features and Deep Architectures for Chunking. Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, 31 May-5 June, Colorado, 245-248.

BIBLIOGRAPHY

Name and Surname	: Mert İLERİ

Educational Status

Bachelor’s Degree	: Ege University Computer Engineering, 2017
Master’s Degree	: Istanbul Commerce University Computer Engineering, 2021

Publications

Turan, M., İleri, M., 2021. Sentiment Analysis of Meeting Room. 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 11-13 June, Ankara, 1-5.