Speech-Based Techniques for Emotion Detection in Natural Arabic Audio Files

Ashraf Kaloub
Department of Multimedia and Information Technology
Al-Aqsa University, Palestine
ai.kaloub@alaqsa.edu.ps

Eltyeb Abed Elgabar
Department of Computer Science
Al-Neelain University, Sudan
tayebsamani@neelain.edu.sa

Abstract: *Emotion detection is one of the greatest challenges of Natural Language Processing (NLP). Often referred to as emotion recognition, it is the process of identifying a person's various feelings or emotions such as: happiness, sadness, or anger. Emotions are a strong feeling regarding a human's situation or relation with others. They are the mental states that affect human behavior and interactions. In this paper, we propose an approach for emotion detection in audio files, focusing on a natural Arabic audio dataset and applying several Machine Learning (ML) classifiers: Sequential Minimal Optimization (SMO), Random Forest (RF), K-Nearest Neighbours (KNN), and Simple Logistic (SL). The classification experiments were conducted using sixteen acoustic feature sets. Many acoustic features were explored including Mel Frequency Cepstral Coefficient (MFCC), Mel spectrogram, spectral contrast, Zero Crossing Rate (ZCR), and Intensity. The experimental results show that SMO and SL classifiers achieved the highest overall accuracy 83.82% when using combinations of all acoustic features (MFCC, Mel spectrogram, Spectral contrast, ZCR and intensity). Additionally, The RF and KNN classifiers yielded Competitive results, with accuracies of 81.71% and 77.34%, respectively. These results suggest that combining multiple acoustic features significantly enhances the performance of emotion detection models, especially for complex emotions in natural Arabic audio datasets.*

Keywords: *Emotion detection, natural language processing, machine learning, Arabic language, acoustic features.*

Received June 27, 2024; accepted October 14, 2024
https://doi.org/10.34028/iajit/22/1/11

1. Introduction

Speech is the vocalized form of language that humans use to communicate and express thoughts, ideas, and emotions. Many studies have been conducted on speech production and perception of sounds used in vocal languages. Speech production refers to how speech organs involved in making a sound whereas speech perception refers to the processes by which humans are able to interpret and understand the sounds used in language. In speech, each word is formed from a phonetic combination of a limited set of vowel and consonants speech units. These speech units can be digitally represented as speech signals [26]. One of the challenging problems in speech processing is identification the speaker's emotional state. Emotion detection, often referred to as emotion recognition, is the process of identifying a person's various feelings or emotions such as: happiness, sadness, or anger [49]. In psychology, human emotion has always been a core interest of study. It is defined by many professors and specialists as involving "...physiological arousal, expressive behaviors, and conscious experience" [22]. Another definition, that describes emotion as "emotion is defined as an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism" [69].

Emotion detection is regarded as a kind of higher, evolved form of Sentiment Analysis (SA). The purpose of SA is to categorize texts, posts, sentences, or documents as negative, positive or neutral. Emotion detection, on the other hand, is a more details, that tries to check the psychology of diverse user behaviors revealing deeper human emotional connotations such as anger, disgust, sadness, joy, surprise, etc. [23]. Emotions are the mental states that effect on the human behavior. It is easy for human using available senses to detect the emotional states from a speaker's speech, but this is a very difficult task for machines [61]. However, detecting the emotional content of an audio signal presents several challenges. One of the main difficulties results from the fact that it is difficult to define what emotion means in a precise way [40]. In addition, Audio signals transfer affective information through explicit (linguistic) messages and implicit (acoustic) messages that reflect the way how the words are spoken [28].

Regarding to languages, most Speech emotion databases are executed in Japanese, German, English, Spanish, Danish, Swedish, Russian, Chinese, and Greek [60]. However, for the Arabic language, there is a shortage of speech emotions datasets although that it is spoken by more than 450 million people [68]. Additionally, most Arabic audio datasets are divided into acted, elicited and semi-natural audio datasets [36, 1]. Modern Standard Arabic (MSA) is the official language over Arab nations, despite different Dialectal Arabic (DA) are used as well.

In this paper, we create a natural Arabic audio dataset. We are extracted an acoustics features from Arabic audio files to detect four emotions (anger, sadness, happiness, and neutral), our dataset was collected from numerous YouTube channels, it includes a total of 2083 audio files (522 are anger, 518 are happiness, 506 are sadness and 537 are neutral). Four different supervised classification methods are applied: Sequential Minimal Optimization (SMO), Random Forest (RF), K-Nearest Neighbours (KNN) and Simple Logistic (SL), which are widely used in Speech Emotion Recognition (SER) systems.

The reminder of this paper is organized as follows: Section 2 presents the related works. Section 3 offers the proposed approach for emotion detection in natural Arabic audio files. Finally, section 4 presents our experiments and discusses the obtained results.

2. Related Works

In this section, several previous works are studied and investigated. The previous works are introduced and analyzed for emotion detection on audio files. all of the presented previous works is based on acted, semi-natural and elicited audio datasets for Arabic language except two represented natural Arabic audio datasets. Additionally, there is a lack of published work related to Arabic language that utilizes natural audio datasets.

Mohammad and Elhadef [46] presented a method for Arabic SER. They used an Audio dataset that containing four emotions (happy, surprised, sad, questioning). Emotion speech audio files were gathered and recorded by humans (5 males and 5 females), every one of them recorded 20 sentences for each type of emotion and the results were 200 collected records. The recorded audio files were in the extended WAV format. In this method, five classification algorithms (MultiLayer Perceptron (MLP), KNN, decision tree, Support Vector Machine (SVM) and logistic regression) were applied. The experiments done using the same extracted features for all of them and the results were 66.7%, 66.7%, 91%, 75%, 91.7%, respectively for these classification algorithms.

Advantages:

- Multiple classification classifiers were used.
- Achieving high accuracy for some classifiers up to 91.7%.

Disadvantages:

- The dataset size is limited, they used only 200 collected records for the experiments.
- There is no information about the number of instances for each emotion in the dataset.
- The dataset is acted, this may not accurately reflect natural emotional expressions.
- The obtained accuracy for some of classifiers is an indication of the difficulty of the task.

Khalil *et al*. [36] introduced a framework to detect anger from natural Arabic conversations. The corpus was gathered from a TV Debate Show and an angry Customers Calls. The total number of gathered audio files were as follows: more than 400 utterances from TV Debate Show represents anger emotion state, varied from 1 to 9 seconds and 45 utterances extracted from an Angry Customers Calls represent anger state with total number of utterances 484. They conducted emotion survey to check the accuracy of labeling, where a listening test of the initial emotional utterances was carried out with the help of groups of an odd number of randomly-selected volunteer human judges. Every one asked to describe each audio clip into one of four options: “Neutral”, “Anger”, “Other” and “Unclear” as a result the experiments were conducted in two sets of data “context-aware” and “context-free”, the first set “context-aware” includes the initial division from the researches for the audio files where 240 utterances represent anger state and 244 utterances represent neutral state, the second set “context-free” includes 185 utterances of audio files represent neutral state and 152 utterances represent anger state. This set exclude all clips that have high score disagreement. Many acoustic features extracted includes fundamental frequency (pitch), formants, energy (intensity) and Mel-Frequency Cepstral Coefficients (MFCCs). They used Three classification algorithms SVM, Probabilistic Neural Network (PNN), Decision Tree Forest used for the experiments. The results showed that SVM applied the highest accuracy at 77.2% for anger detection in real-time.

Advantages:

- A natural dataset was gathered from TV debate shows and customer service calls.
- Using a comprehensive labeling approach by conducting an emotion survey, utilizing listening test with multiple human judges.
- A variety of Acoustic Features were used in experiments.

Disadvantages:

- The emotional scope is limited, concentrating on only two emotional states (anger and neutral).
- The dataset size is limited, they used only 484 collected records for the experiments.
- Limited classification classifiers, the study used only three classification classifiers (SVM, PNN and decision tree).

Aljuhani *et al*. [19] presented a new approach for Arabic SER from Saudi Dialect Corpus, they created semi-natural audio dataset from YouTube videos taken from the popular Saudi YouTube channel (Telfaz11), a group of videos was checked and viewed to choose the scenes that represent the best emotion references for ML classifiers. The final result of dataset was included of

175 records, with male and female actors divided into 113 chunks for males and 62 for females with total duration 11 minutes. The three emotional states used from the dataset for anger, happiness, neutral and sadness included 69 chunks, 31 chunks, 37 chunks and 38 chunks respectively. They used three classifiers SVM, MLP and KNN to predict the four in audio dataset. For the classification, spectral features used where MFCC and spectral contrast showed the best accuracy for KNN at 68.57%, by adding the Mel spectrogram features to the previous features the prediction enhance for SVM and MLP with accuracy of 77.14% and 71.43, respectively. The Results also showed that anger was the best predicted emotion by all classifiers.

Advantages:

- Effective features utilization achieved by combining multiple features this include MFCC, spectral contrast and Mel spectrogram.

Disadvantages:

- The dataset is limited, containing only 175 records.
- Imbalanced gender representation 113 chunks for males and 62 for females.
- The Dataset scope is limited by concentrating on specific regional dialect (Saudi gulf dialect).
- Semi-Natural Audio dataset collected from popular Saudi YouTube channel Telfaz11.

Meftah and Zakariah [44] proposed an audio dataset called King Saud University (KSU) Emotions for MSA using 23 speakers (10 males and 13 females) from three Arabic countries: Yemen, Saudi Arabia, and Syria. 16 sentences spoken by speakers are selected from the original corpus, King Abdulaziz City for Science and Technology Text-To-Speech Database (KTD) for six emotions: Neutral, Happiness, Sadness, Surprise, Anger and questioning. The experiments were conducted in two Phases: phase 1 and phase 2, each phase represents a group of selected speakers to read 16 sentences. The final audio files recorded for phase 1 were 1600 audio files and 1680 audio files were recorded for Phase 2 with total numbers of 3280 audio files for two phases. To evaluate the two phases' recordings, a blind human perceptual test was performed, where nine listeners (6 males and 3 females) were involved to listen to the recorded files to determine whether they are able to detect the recorded emotions. Five experiments were conducted using both phase 1 and phase 2, either alone or together. various feature-extraction techniques were applied to the audio dataset, including the Zero-Crossing Rate (ZCR), short-term energy, MFCCs, and delta feature. Two classification algorithms were applied KNN and SVM. The final results showed that KNN has attained better accuracy nearly 87.04% compare with SVM which attained accuracy of 78.96%. The results showed that phase 2 of the corpus is better than phase 1.

Advantages:

- Diverse geographic representation, this include three different Arabic countries (Yemen, Saudi Arabia, Syria).
- Comprehensive emotion range, this include (neutral, happiness, sadness, surprise, anger and questioning).
- Large dataset collection, it includes 3280 audio files.

Disadvantages:

- Limited classification classifiers, the study used only two classification classifiers (KNN and SVM).
- Acted dataset, this may not accurately reflect natural emotional expressions.
- Lack of emotion distribution details.

Klaylat *et al.* [37] presented a method to detect emotions from natural audio files, the first a realistic corpus from Arabic TV shows were gathered, eight videos were downloaded from different Arabic online live talk shows, these videos were live calls between the presenter and a human outside the studio. The videos contain Egyptian, Gulf and Lebanese speakers, the videos were diverse in length, comprising both male and female speakers. Listening test was done to label each video, where 18 listeners were asked to listen to each video to distinguish one of the three emotion states: happy, angry or surprised. Each video was divided into smaller segments based on who is speaking the representer or the caller, some pre-processing operations were done to eliminate Silence, laughs and noisy segments. Every segment was automatically split into 1 sec speech units, the final result of audio dataset was involved of 1384 records with 505 happy, 137 surprised and 741 angry segments. Thirty-five classification algorithms were used to classify these audio files based on 845 audio features extracted using 19 statistical functions applied to each of 25 initial features. These features include: intensity, ZCR, MFCC 1-12, F0 (Fundamental frequency) and F0 envelope, probability of voicing and, LSP frequency 0-7. Additionally, delta coefficients for each LLD were computed, leading to a total of 950 features. The Kruskal-Wallis test was then applied to reduce dimensionality, removing features with p-values>0.05. The best result was 95.52% using SMO algorithm and the worst result was 53.58% by five algorithms, thirteen algorithms gave more than 90% accuracy, nine algorithms between 89 and 80, four between 79 and 70; three algorithms in the 60's and sex algorithms in 50's.

Advantages:

- Natural audio dataset collected from various Arabic TV shows.
- Extensive listening test for labeling by using 18 listeners for labeling emotions.
- Achieves high accuracy rates, with best results using

the SMO classifier.

Disadvantages:

- Imbalanced audio dataset, the distribution of emotional states is (505 happy, 137 surprised, 741 angry), which might bias the classification performance towards more frequently represented emotions.
- Limited emotion range, only recognizes three emotions (happiness, anger, surprise)
- Short segment duration, using 1 second segments not completely enough to detect some other emotions.
- The audio dataset contains speakers of limited dialects, specifically Egyptian, Gulf, and Lebanese.
- Different classifier performance, while some classifiers performed well, others yielded lower accuracies.
- The dataset consists of 1384 chunks, which might be consider small for developing a robust model.

Abdel-Hamid [1] introduced Egyptian Arabic Speech Emotion (EYASE) database. It is a semi-natural dataset that was created from award winning Egyptian drama series 'Hatha Al-Masaa'. The EYASE dataset contains 579 utterances from six professional actors (3 females and 3 males) for four emotions: sad, angry, happy, and neutral. Prosodic (pitch-intensity), Spectral (formants, MFCC, Long-Term Average Spectrum (LTAS) and wavelet features are extracted from the utterances. Two experiments were performed speaker-dependent and speaker-independent for three cases:

1) Multi-emotion classifications.
2) Neutral versus emotion classifications.
3) Valence and arousal classifications.

The results showed anger emotion was found mostly detected and happiness was the most challenging, also Arousal (angry/sad) recognition rates were shown to be superior to valence (angry/happy) recognition rates. Furthermore, higher accuracies achieved for male subjects than for female subjects in all performed experiments by applying two classification algorithms KNN and SVM.

Advantages:

- Effective features utilization achieved by using prosodic, spectral and wavelet features.
- Gender analysis, where the analysis included gender-based performance, providing insights into how the system performs across male and female subjects.

Disadvantages:

- Limited classification classifiers, the study used only two classification classifiers (KNN and SVM).
- The dataset is limited, it contains only 579 utterances.
- The dataset is limited in scope, concentrating on specific regional dialect (Egyptian Arabic dialect).
- Semi-Natural Audio dataset collected from popular Egyptian drama series.

Horkous and Guerti [30] presented the Algerian Dialect Emotional Database (ADED). This dataset created from six famous movies in Algerian dialect. These movies describe the civil war between (1992-2000) in addition to the period that followed. The dataset includes 32 actors (16 males and 16 females) with different ages (from 18 to 60 years). The dataset contains 200 semi-natural emotion utterances of duration ranging from 0.2 s to 3 s. ADED includes four emotions: fear, anger, sadness and neutral. Different features were extracted from the speech utterances of the ADED, including pitch, intensity, duration, unvoiced frames, jitter, shimmer, HNR, formants and MFCCs. The KNN technique was used as a classifier in all the experiments. Multiple experiments were performed to evaluate the performance of features extracted on the recognition systems by using different features in every experiment.

The results showed that the use of MFCCs features with other features gave recognition rate very important and the formants features gave a weak performance on the experiments. Fear and neutral emotions gave the higher recognition rate 87.50%, when the anger emotion was added to the system the recognition rate decreased to 84.02%, and when the sadness emotion was added to the system the recognition rate decreased to 82.29%.

Advantages:

- Diverse speaker representation, this include 32 actors of different ages (from 18 to 60 years).
- Comprehensive feature extraction, including pitch, intensity , duration , unvoiced frames, jitter , shimmer, HNR, formants and MFCCs.

Disadvantages:

- Limited classification classifiers, the study used only one classification classifiers KNN.
- Small dataset size, with only 200 utterances.
- Acted dataset, this may not accurately reflect natural emotional expressions.
- The Dataset is limited in scope by concentrating on specific regional dialect (Algerian Arabic dialect).
- Semi-natural audio dataset collected from six famous movies in Algerian dialect.

3. The Proposed Approach

This section presents our proposed method for emotion detection in a Natural Arabic audio dataset. Multiple steps have to be implemented to achieve emotion detection, including audio dataset collection, audio files pre-processing, annotation, audio files normalization, features extraction and normalization, features selection, supervised learning classification and evaluation. Figure 1 depicts the multiple steps of our proposed approach. First, the natural Arabic audio files are pre-processed after downloaded from YouTube and this include: segmentation, noise removal from audio files. Then, the annotation process is done. Next, a set of acoustics features are extracted and normalized such as: MFCC, Mel Spectrogram, Spectral Contrast, Chroma, ZCR, Pitch and Intensity. Four classifiers are applied including SMO, RF, KNN, and SL, to judge whether the given audio files is one from these emotional states (Anger, Happiness, Sadness or Neutral), these classifiers are trained from the annotated audio files. Finally, the classification results have assessed using various classification metrics such as accuracy, precision and recall.

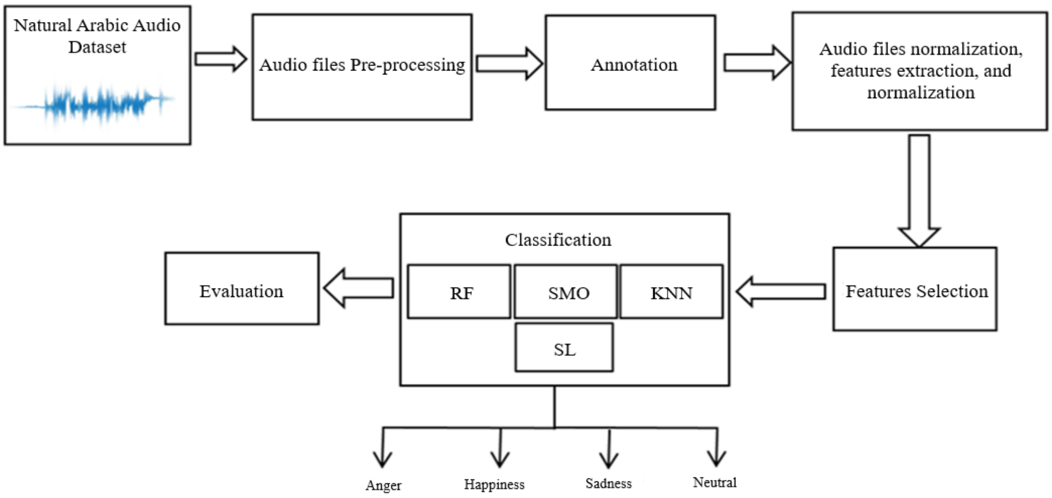

Natural Arabic Audio Dataset
Audio files Pre-processing
Annotation
Audio files normalization, features extraction, and normalization
Features Selection
Classification
RF
SMO
KNN
SL
Evaluation
Anger
Happiness
Sadness
Neutral

Figure 1. The multiple steps of our proposed approach.

3.1. Dataset Collection

In this work, a natural Arabic audio dataset was constructed using freely accessible YouTube videos on the internet. In order to include a wider range of emotional content in the collected dataset, 1103 videos ranging from 1 to 50 minutes were downloaded from various YouTube channels. These videos, representing four emotional states (Anger, Happiness, Sadness and Neutral), were considered raw videos. Initially, each video was carefully listened to and selected based on its potential to contain the suitable emotional content, including both the sounds and words expressed in the audio files, corresponding to each emotion. The collection process spanned 10 months. The videos include talk shows and meetings with different guests contain discussions on interesting topics which can induction multiple emotions from the speakers. The collected dataset consists of audio files spoken in MSA and different DA or a mix of both, providing a diverse linguistic range. The following is a detailed description of our audio dataset sources:

1) The opposite direction program on al-jazeera YouTube channel to represent the anger emotion: This source was accessible on YouTube. It consists of a set of recorded episodes of Arabic political issues. We chose this TV program for multiple reasons. First, the Anger emotion state found in most of episodes due to the nature of the program, which includes dialogical arguments. Second, all speakers in this program are guests from various Arab countries, providing us with diversity of political topics, the source of these audio files is [17].
2) Topics like winning in sports and achieving success in Tawjihi exams (secondary school leaving examinations in most Arab countries) were chosen to represent the emotion of happiness: the audio files were collected from meetings with individuals in public places such as: the streets, homes, stadiums, and secondary schools. Here, they express their happiness about both success in tawjihi or winning in sports according to the context of the meeting. To effectively narrow down the search for these types of audio files on YouTube, some keywords such as: 'فرحة الفوز في المباراه', 'فرحة النجاح في توجيهي', and ' فرحة النجاح في الثانوية العامة' were used. The sources of these audio files are [2, 3, 10, 11, 13, 14, 20, 21, 24, 25, 27, 33, 35, 47, 48, 51, 53, 55, 56, 58, 65].
3) Topics related to loss, to represent Sadness emotion: The sources of these audio files are meetings with individuals who lost some of their families in wars, accidents or natural deaths. In these cases, people express their feelings towards this with emotional content convey sadness. To effectively narrow down the search for these types of audio files on YouTube, some keywords such as: 'الحزن على وفاه' and 'فقدان' were used. The sources of these audio files are [5, 6, 9, 12, 16, 18, 27, 33, 41, 50, 54, 55, 59, 64, 65].
4) Neutral speech topics, to represent neutral emotion: The audio files were sourced from podcasts, news, and documentary programs, where all the speech conveys neutral emotions. The sources of these audio files are [4, 7, 8, 16].

3.2. Audio Files Pre-Processing

- Segmentation and Noise Removal

For every video, we striped the audio files alone as we are now concentrating on speech data. All audio files were segmented into smaller chunks based on the emotional content of every single audio file. During the segmentation process, we check that the emotional content for every segment is consistent and not change in the same chunk. The process of segmentation was done manually to ensure consistency of the speech chunks. As a result, the chunks ranged from 1 to 9 seconds in length, each selected based on its clear emotional content. We removed noise from each chunk, including background music, to maintain the appropriate quality for every audio file. For segmenting the audio files and removing all noise, we used two adobe speech processing tools ([1]Adobe Premiere and [2]Adobe Audition). The overall audio dataset consisted of 2160 records from both male and female speakers, with a total of 540 chunks for every emotion (anger, happiness, sadness, and neutral). The total duration of the dataset was about 119 minutes, recorded at 48kHz sampling rate and saved as '.wav' files.

3.3. Annotation

In this step, we obtained 540 chunks for emotional states (anger, happiness, sadness, or neutral) during the segmentation process. We then asked three human listeners (2 females and 1 male) to evaluate these chunks. They listened to the emotional content of the chunks labeled as anger. If two or more listeners agreed on the emotional state, then we labeled it as anger.

If there was disagreement between two or more listeners, about this emotional state then we will discard it, this process was repeated for each emotion.

Table 1. Distribution of emotions labels in the audio dataset.

Emotion	Number of chunks	Duration in minutes
Anger	522	20.29
Happiness	518	20.44
Sadness	506	28.05
Neutral	537	46.50

The details of the audio dataset and emotions distributions can be observed in Table 1, the dataset comprises of 2083 audio files, classified by emotion (522 for anger, 518 for happiness, 506 for sadness, 537 for neutral), this distribution ensures a balanced representation for every emotional state, which is important for training robust emotion detection models.

The table also shows the total duration in minutes for every emotion, with the overall duration of the dataset approximately 115 minutes.

3.4. Audio Normalization, Features Extraction and Features Normalization

The process of features extraction is the initial step to extract important information from audio files. First, we normalize audio files to ensure consistent loudness levels for all audio files, this a very important step for accurate features extraction, especially in different recoding conditions. This normalization is done for each audio file by scaling its amplitude to a range between -1 and 1, ensuring that no parts of the audio files exceed this range, which helps in keeping the integrity of the audio signal avoiding any distortion that can occur if the signals amplitude is too high. Second, after features extraction, we execute another normalization process to ensure uniform scaling of these features. This normalization is done by calculating the mean and standard deviation of each raw features and then adjusting these features to have zero mean and unit variance. This step involves subtracting the mean from each feature value and dividing by standard deviation, which standardized the features. This standardization is important for effective ML, as it prevents features with larger scales from dominating the learning process. In this work, we have extracted seven types of raw speech features: MFCC, Mel Spectrogram, Spectral Contrast, Chroma, ZCR, Pitch and Intensity. Each of these features was normalized using the method described above to ensure consistency and improve the effectiveness of ML classifiers. These audio features can be broadly classified into two categories (spectral and prosodic), these seven features were chosen because they are widely recognized in the field of speech emotion detection for their effectiveness in capturing the acoustic properties relevant to various emotional states, as supported by previous works.

1) Mel Frequency Cepstral Coefficient (MFCC).

MFCC is one of the most broadly used spectral features. It has multiple advantages, including simplicity in computation, enhanced diversity capabilities, and high noise resistance [45].

2) Mel spectrogram.

Its representation of the sound signal on a Mel Scale. The logarithmic form of Mel-spectrogram assistances to better understand emotions, as humans perceive sound logarithmically [62].

3) Spectral contrast.

Spectral contrast analyzed the strength of peak and valley of the spectral and the variance between them, also, spectral contrast features represent the relative spectral characteristics, in addition it has more spectral information compared to MFCC feature [34].

4) Chroma.

Chroma feature is usually comprising a 12-element feature vector that representing how much the energy level of every pitch class within the signal, based on a

[1] https://www.adobe.com/mena_en/products/premiere.html
(Premiere Pro is designed for a video editors post-production workflow)

[2] https://www.adobe.com/mena_en/products/audition.html?promoid=2XBSC386&mv=other
Audition is designed for creating, mixing, editing, and restoring audio content.

standard chromatic scale [63].

5) Zero rossing Rate (ZCR).

The ZCR defined as the rate at which a signal changes from positive to zero to negative or from negative to zero to positive referred to the ZCR. Its importance has been extensively recognized in sound detection and music information retrieval, and it is an essential element in classifying rhythmic sounds [15].

6) Pitch.

It is referred to it as the fundamental frequency of a signal. It reflects the vibration of the vocal cords. Typically, females show a higher pitch than males and Children pitch are similar to female pitch. Pitch can be calculated through time domain analysis using short-time average magnitude difference function. Cepstrum analysis is a frequency domain method to calculate pitch based on harmonics improvement [52].

7) Intensity.

Indicates the force with which the sound is produced and is measured in decibels. Unlike the fundamental frequency, this parameter is not related to the physiology of the speaker. In addition, it varies significantly and one of it challenging is hard to normalize. The measurement time may greatly different throughout a telephone conversation, for instance, based on caller distance from the microphone. Additionally, it is a very significant parameter: emotions for example anger usually have a high intensity, unlike emotional states like neutral or sadness. So, intensity plays avital role to determining the emotional state of human [43].

Finally, a large set of 326 acoustic features are generated from every speech chunk. Table 2 illustrate the distribution of acoustic features for every chunk derived from each raw features along with their corresponding categories. For audio normalization, features extraction, and features normalization, we used librosa library [42], which consider a Python package for music and audio analysis, it provides the building blocks necessary to create music information retrieval systems. Librosa design involves decomposing complicated functions into simpler, making it appropriate for academic research and practical applications. The library utilizes NumPy for numerical computation and integrates well with the broader Python scientific computing framework, enabling rapid prototyping and efficient experimentation. Librosa offers several features for audio and music analysis, this includes time-series analysis, spectral representation, and features extraction. Librosa aims to make a balance between flexibility for skilled users and simplicity for beginners by providing detailed documentation and maintaining the conventional python coding style.

Table 2. Distribution of acoustic features and their corresponding categories.

Raw feature	Number of derived features	Category	Total number of features per category
MFCC	26	Spectral	320
Mel spectrogram	256		
Spectral contrast	14		
Chroma	24		
ZCR	2	Prosodic	6
Pitch	2		
Intensity	2		

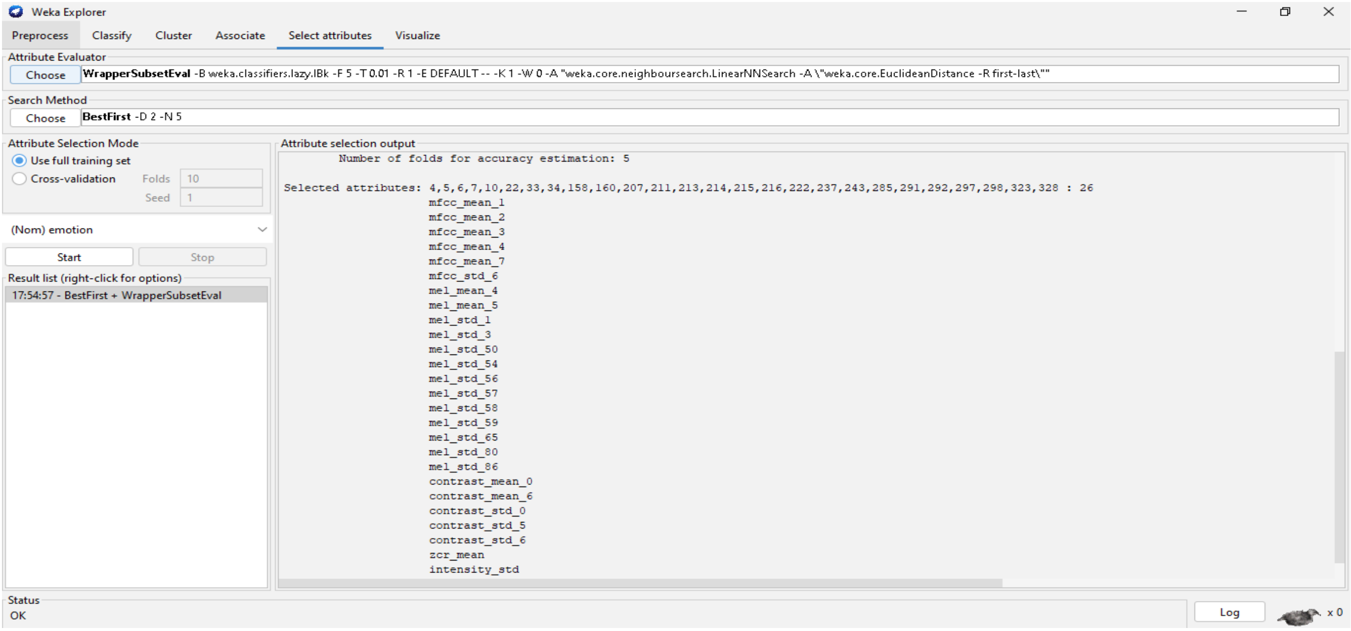
3.5. Features Selection

After extracting 326 acoustic features from the speech signals, we applied features selection techniques to select the best appropriate features that have information to obtain better performance of the classifiers. For this we used [3]Weka environment, which is freely and open-source software. Features selection is separated into two parts: Attribute evaluator and search method and each part have several techniques from which to select. Figure 2 Illustrates screenshot of using Weka during features selection, at the end we obtained 26 acoustic features as shown in Table 3.

Table 3. Distribution of acoustic features after features selection process.

Derived features	Raw feature	Category	Derived features	Raw feature	Category
Mfcc_mean_1 Mfcc_mean_2 Mfcc_mean_3 Mfcc_mean_4 Mfcc_mean_7 Mfcc_std_6	MFCC	Spectral	Contrast_mean_0 Contrast_mean_6 Contrast_std_0 Contrast_std_5 Contrast_std_6	Spectral contrast	Spectral
Mel_mean_4 Mel_mean_5 Mel_std_1 Mel_std_3 Mel_std_50 Mel_std_54 Mel_std_56 Mel_std_57 Mel_std_58 Mel_std_59 Mel_std_65 Mel_std_80 Mel_std_86	Mel spectrogram		Zcr_mean	ZCR	Prosodic
Intensity_std	Intensity	Prosodic			

[3] http://www.weka.io



Attribute Evaluator: **WrapperSubsetEval** -B weka.classifiers.lazy.IBk -F 5 -T 0.01 -R 1 -E DEFAULT -- -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Search Method: **BestFirst** -D 2 -N 5

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) emotion

Result list (right-click for options): 17:54:57 - BestFirst + WrapperSubsetEval

Attribute selection output

```
        Number of folds for accuracy estimation: 5

Selected attributes: 4,5,6,7,10,22,33,34,158,160,207,211,213,214,215,216,222,237,243,285,291,292,297,298,323,328 : 26
                     mfcc_mean_1
                     mfcc_mean_2
                     mfcc_mean_3
                     mfcc_mean_4
                     mfcc_mean_7
                     mfcc_std_6
                     mel_mean_4
                     mel_mean_5
                     mel_std_1
                     mel_std_3
                     mel_std_50
                     mel_std_54
                     mel_std_56
                     mel_std_57
                     mel_std_58
                     mel_std_59
                     mel_std_65
                     mel_std_80
                     mel_std_86
                     contrast_mean_0
                     contrast_mean_6
                     contrast_std_0
                     contrast_std_5
                     contrast_std_6
                     zcr_mean
                     intensity_std
```

Status: OK

Figure 2. Screenshot of the WEKA during feature selection.

3.6. Classification

After features selection, features are combined into a single vector for each audio file. This combination includes combining all normalized features (MFCC, Mel Spectrogram, Spectral Contrast, ZCR, and Intensity) into comprehensive feature vector that represents the entire audio file. These features are then used as inputs to the classifiers. The classifiers are trained to detect patterns in these vectors that correspond to different emotional states. For this purpose, we have to choose the appropriate ML classifiers that achieve the higher classification results. We have selected four ML classifiers within [4]RapidMiner Platform to judge whether the audio files are (anger, happiness, sadness or neutral) emotion, including: SMO, RF, KNN and SL.

While RapidMiner supports KNN classifier, we employed the WEKA extension within RapidMiner to access the additional three classifiers SMO, RF and SL. These ML classifiers were selected because they have shown the best results in many emotion detection tasks in audio files. The training process includes feeding these classifiers with acoustic features vectors generated from the audio files to classify every audio file into one of the emotional states: anger, happiness, sadness, or neutral. Although these classifiers are mentioned, the performance and results from using them will be detailed in section 4 experiments and results. This section will demonstrate the significance of the results, accompanied by visual representations such as confusion matrices and a comparative overview of each classifier efficiency.

1) Sequential Minimal Optimization (SMO).

SMO is a classifier for solving the quadratic programing problem that occurs during the training of SVM. It is broadly used for training SVMs and is executed by the popular LIBSVM tool. SMO breaks the problem into groups of smallest possible problems, which are then solved analytically [66].

2) Random Forest (RF).

The RF classifier is built on the principle of ensemble learning, which is a process of merging multiple classifiers to solve complex problems and to improve the performance of the model. It includes a number of decision trees on different subsets of the given dataset and takes the average to enhance the predictive accuracy of that dataset. Instead of depend on one decision tree. From each tree, the RF classifier takes the prediction and determines the final result based on the majority vote of these predictions [32].

3) K-Nearest Neighbors (KNN).

KNN consider one of the simplest ML classifiers. It assumes the similarity between the new case and available cases, placing the new case into the category that is most similar to the available categories. K-NN can be used for both regression and classification, although it is more frequently applied in classification problems. K-NN at the training phase only stores the dataset and when it gets new data, then is classifies that data into a category that is most similar to the new data [31].

4) Simple Logistic (SL).

The creation of the SL classifier is influenced by the principles of Logistic Model Trees (LMT). The LMT algorithm combines logistic regression into decision

[4] https://altair.com/altair-rapidminer

tree framework, improving both interpretability and the accuracy of predictions. LMTs are designed to address both binary and multi-class classification problems, offering probabilistic predictions and a high level of interpretability. Logistic Regression a robust statistical approach, used model to the probability of a binary outcome [39].

3.7. Evaluation

For the evaluation process, specific metrics are required to evaluate the ML classifiers performance. A commonly used method to evaluate the performance of the classifiers is by using a confusion matrix. A Confusion matrix is a suitable tool for investigating the classifiers capability to recognize instances of various classes. It contains information about real and predicted classifications [29]. Performance is calculated from the confusion matrix using three evaluation metrics: accuracy, recall and the precision. These metrics defines as follows: [57].

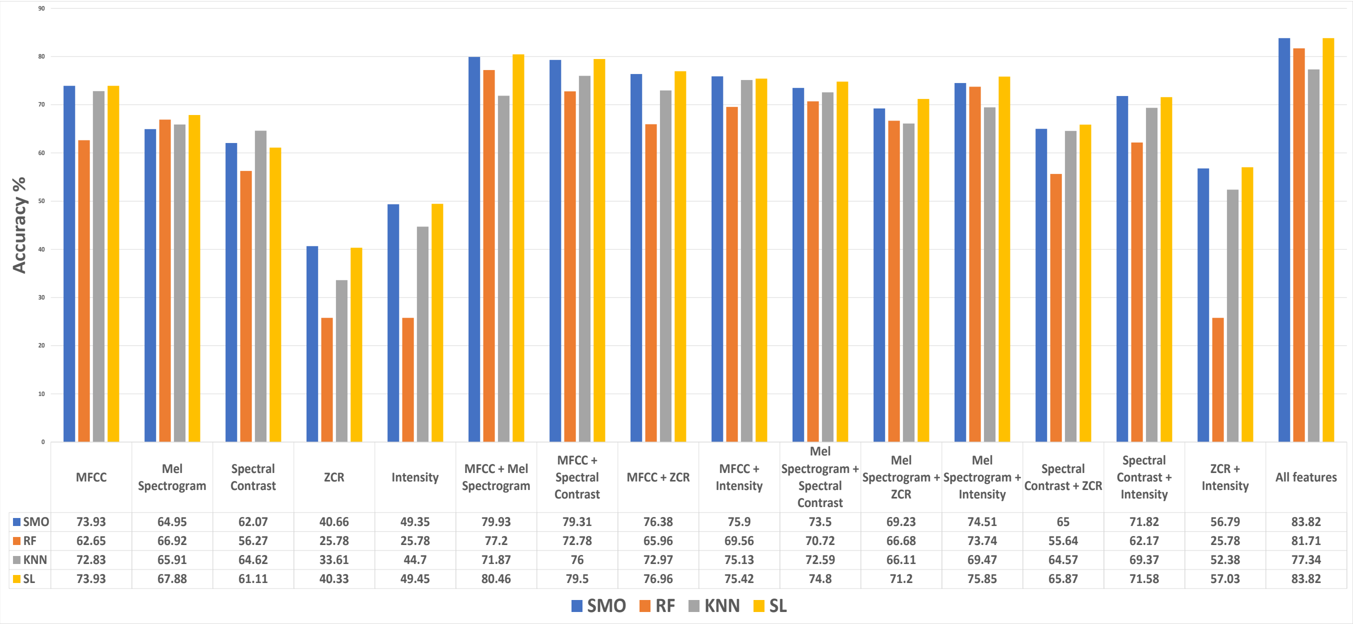
- Accuracy: It refers to the percentage of instances on a given test set is that are correctly classified by the classifier. The associated class label of each test instance is compared with the learned classifiers class prediction for that instance.
- Precision: It is the capability of the model to identify the correctly predicted positives from all the predicted instances labeled as positives.
- Recall: This measures the model capability to identify the correct positive from all the existing positives in the test dataset.

4. Experiments and Results

This section presents the experiments conducted to evaluate and test the acoustic features and the performance of the chosen classifiers to detect emotion in natural Arabic audio files. It presents the experimental results and their evaluation. In addition to that, it discusses the obtained results to justify our proposed approach.

1) Experimental setup.

In this subsection, we explained the experimental process we have used to evaluate our method for the task of emotions detection in audio files. For the audio files classification task experiments, we have used the audio dataset that we collect in order to apply the ML classifiers for the problem of emotions detection in natural Arabic audio files. Our dataset includes a total of 2083 audio files (522 for anger emotion, 518 for happiness emotion, 506 for sadness emotion, 537 for neutral emotion). We implemented all the experiments using 10-fold cross-validation in RapidMiner Platform. To perform the experimentation, we have used numerous classifiers within RapidMiner platform to judge whether the emotional state (anger, happiness, sadness or neutral) of each audio file. These classifiers included: SMO, RF, KNN and SL. These classifiers were selected because they are widely used in SER systems and have shown best results in many audio classification tasks. We have implemented sixteen different experiments in order to evaluate our method. These experiments are grouped according to multiple acoustics features sets. One of the main goals of this work is to evaluate various ML classifiers for emotion detection in natural Arabic audio files as well as the selected acoustics features. To evaluate ML classifiers, we based on calculating accuracy, Precision and Recall, which are commonly used to measure a systems performance in this filed. To compute these metrics, its essential to generate a confusion matrix after the classification process.



	MFCC	Mel Spectrogram	Spectral Contrast	ZCR	Intensity	MFCC + Mel Spectrogram	MFCC + Spectral Contrast	MFCC + ZCR	MFCC + Intensity	Mel Spectrogram + Spectral Contrast	Mel Spectrogram + ZCR	Mel Spectrogram + Intensity	Spectral Contrast + ZCR	Spectral Contrast + Intensity	ZCR + Intensity	All features
■ SMO	73.93	64.95	62.07	40.66	49.35	79.93	79.31	76.38	75.9	73.5	69.23	74.51	65	71.82	56.79	83.82
■ RF	62.65	66.92	56.27	25.78	25.78	77.2	72.78	65.96	69.56	70.72	66.68	73.74	55.64	62.17	25.78	81.71
■ KNN	72.83	65.91	64.62	33.61	44.7	71.87	76	72.97	75.13	72.59	66.11	69.47	64.57	69.37	52.38	77.34
■ SL	73.93	67.88	61.11	40.33	49.45	80.46	79.5	76.96	75.42	74.8	71.2	75.85	65.87	71.58	57.03	83.82

Figure 3. Classification performance of the four classifiers on various acoustic feature sets in terms of accuracy.

2) Experimental results and discussion.

This subsection presents and discusses the results of the multiple experiments that have been performed. The purpose for these experiments was to evaluate the best acoustic features sets and ML classifiers that work well for emotion detection in natural Arabic audio files, we implemented these experiments using various ML classifiers this including: SMO, RF, KNN and SL. Sixteen experiments were conducted. These features set contain26 distinct features. Tables 4, 5, 6, and 7 show the confusion matrices of the four ML classifiers SMO, RF, KNN and SL for the combinations of acoustic features set, respectively. In these tables, (A) corresponds to Anger, (H) corresponds to Happiness, (S) corresponds to Sadness, and (N) corresponds to Neutral. In addition, the tables display the classification results of the acoustic features experiments for these classifiers. The values in the tables represent the accuracy, precision and recall for every classifier. where accuracy represent overall correct classification across all classes, while precision and recall represent the performance for each class. Values in bold indicate the best results for the classifiers for the acoustic features set experiments.

Figure 3 illustrates a graphical summary of the classification performance of the four classifiers (SMO, RF, KNN, and SL) a cross sixteen different experiments in terms of accuracy percentage. Each bar represents the accuracy achieved by every classifier.

As we can see from Figure 3, which represents the accuracy of acoustic feature sets in a bar graph of Tables 4, 5, 6 and 7, the highest overall accuracy achieved was 83.82% using the SMO and SL classifiers, based on combinations of all acoustic features (MFCC, Mel spectrogram, Spectral contrast, ZCR and intensity). The RF and KNN classifiers yielded competitive results with accuracies of 81.71% and 77.34%, respectively. From the total of 2083 instances, both SMO and SL classifiers correctly classified 1746 instances, while 337 were incorrectly classified. For the RF classifier 1702 instances were correctly classified and 381 were incorrectly classified. The KNN classified 1611 of the instances correctly and 472 were incorrectly classified. SMO and SL are achieved the highest accuracy rates, substantially outperformed both the RF and KNN classifiers. In addition to the overall accuracy metrics, Tables 4, 5, 6, and 7 show measures of recall and precision for each emotion class to evaluate the performance of each classifier using the combination of all acoustic features sets. For the SMO classifier, the highest recall and precision were achieved for anger emotion with precision at 93.54% and recall at 94.25%, while the lowest was achieved for sadness emotion with precision at 75.5% and recall at 72.53%. For SL classifier, the highest recall and precision were achieved for the anger emotion with precision at 92.96% and recall at 96.17%, while the lowest was achieved for sadness emotion with precision at 74.74% and recall at 71.94%. The RF classifier, while also showing strong performance in detecting anger 89.76% precision and 94.06% recall, showed the lowest performance for the sadness emotion with precision at 74.54% and recall at 63.64%. For the KNN classifier, the highest recall and precision were achieved for anger emotion with precision at 77.35% and recall at 91.57%, while the lowest performance was for sadness emotion with precision at 73.63% and recall at 66.21%. This consistency over different classifiers confirms the robustness of anger detection in our natural Arabic audio dataset. However, While the classifiers were effective correctly identifying a subset of the sadness instances (high precision), they missed a considerable number of true sadness cases (low recall), suggesting difficultly in correctly identifying sadness, potentially because its acoustic features are more complex. Comparing our results with existing research, such as the study by [38], we find that challenges in distinguishing specific emotions, like sadness in our study and surprise in theirs, are common, also emotions like sadness have more subtle acoustic features compared to more distinct emotions like anger, and these subtle acoustic features can overlap with those of other emotions, such as neutral, causing misclassification. To address this problem, we suggest incorporating a multimodal approach by adding lexical features alongside acoustic features. Lexical features can provide additional emotional cues that may not be fully captured by acoustic features alone. Both studies highlight the difficulties attaining balance between precision and recall across different emotions. Figure 4 illustrate the performance metrics for four classifiers (SMO, RF, KNN, SL) across four emotions (anger, happiness, sadness, neutral) using all feature combinations. Each bar represents the Precision and Recall achieved by every classifier. As we can see from Figure 4, all four classifiers demonstrated strong performance in detecting anger, happiness and neutral emotions, with high precision and recall values indicating that these emotions were accurately identified. The performance for the sadness was significantly lower for both precision and recall across all classifiers.

Table 4. Confusion matrix and accuracy details of SMO classifier.

SMO classifier								
Feature extracted	**Class**	**Classified as**				**Accuracy (%)**	**Precision (%)**	**Recall (%)**
		A	H	S	N			
MFCC	A	464	68	26	22	73.93	80	88.89
	H	44	404	80	17		74.13	77.99
	S	7	41	290	116		63.88	57.31
	N	7	5	110	382		75.79	71.14
Mel spectrogram	A	427	174	96	81	64.95	54.88	81.80
	H	55	250	18	1		77.16	48.26
	S	33	75	292	71		62	57.71
	N	7	19	100	384		75.29	71.51
Spectral contrast	A	379	131	95	28	62.07	59.87	72.61
	H	67	276	64	42		61.47	53.28
	S	68	87	237	66		51.75	46.84
	N	8	24	110	401		73.85	74.67
ZCR	A	376	219	221	83	40.66	41.82	72.03
	H	99	140	115	117		29.72	27.03
	S	9	29	17	23		21.79	3.36
	N	38	130	153	314		49.45	58.47
Intensity	A	380	18	37	63	49.35	76.31	72.80
	H	11	348	256	170		44.33	67.18
	S	3	15	17	21		30.36	3.36
	N	128	137	196	283		38.04	52.70
MFCC+Mel spectrogram	A	482	55	14	6	79.93	86.54	92.34
	H	28	402	50	14		81.38	77.61
	S	7	53	345	81		70.99	68.18
	N	5	8	97	436		79.85	81.19
MFCC+Spectral contrast	A	470	44	28	8	79.31	85.45	90.04
	H	31	421	55	18		80.19	81.27
	S	17	44	330	80		70.06	65.22
	N	4	9	93	431		80.26	80.26
MFCC+ZCR	A	486	46	24	9	76.38	86.02	93.10
	H	29	419	83	17		76.46	80.89
	S	2	46	293	118		63.83	57.91
	N	5	7	106	393		76.91	73.18
MFCC+Intensity	A	467	26	19	5	75.90	90.33	89.46
	H	34	430	89	20		75.04	83.01
	S	15	53	290	118		60.92	57.31
	N	6	9	108	394		76.21	73.37
Mel spectrogram+Spectral contrast	A	443	122	43	30	73.50	69.44	84.87
	H	48	330	34	18		76.74	63.71
	S	25	50	331	62		70.73	65.42
	N	6	16	98	427		78.06	79.52
Mel spectrogram+ZCR	A	452	154	73	30	69.23	63.75	86.59
	H	42	269	15	7		80.78	51.93
	S	21	81	308	87		61.97	60.87
	N	7	14	110	413		75.92	76.91
Mel spectrogram+Intensity	A	444	26	24	10	74.51	88.10	85.06
	H	32	370	47	29		77.41	71.43
	S	38	96	331	91		59.53	65.42
	N	8	26	104	407		74.68	75.79
Spectral contrast+ZCR	A	421	105	85	18	65	66.93	80.65
	H	51	291	68	52		62.99	56.18
	S	44	96	241	66		53.91	47.63
	N	6	26	112	401		73.58	74.67
Spectral contrast+Intensity	A	444	36	30	8	71.82	85.71	85.06
	H	44	388	114	55		64.56	74.90
	S	29	76	261	71		59.73	51.58
	N	5	18	101	403		76.47	75.05
ZCR+Intensity	A	427	21	38	55	56.79	78.93	81.80
	H	22	324	225	80		49.77	62.55
	S	47	75	99	69		34.14	19.57
	N	26	98	144	333		55.41	62.01
All features	A	492	24	7	3	83.82	93.54	94.25
	H	22	443	47	19		83.43	85.52
	S	6	42	367	71		75.51	72.53
	N	2	9	85	444		82.22	82.68

Table 5. Confusion matrix and accuracy details of RF classifier.

RF classifier								
Feature extracted	**Class**	**Classified as**				**Accuracy (%)**	**Precision (%)**	**Recall (%)**
		A	H	S	N			
MFCC	A	376	84	78	37	62.65	65.39	72.03
	H	42	397	169	42		61.08	76.64
	S	3	16	106	32		67.52	20.95
	N	101	21	153	426		60.77	79.33
Mel spectrogram	A	415	133	95	44	66.92	60.41	79.50
	H	68	300	40	9		71.94	57.92
	S	18	58	244	49		66.12	48.22
	N	21	27	127	435		71.31	81.01
Spectral contrast	A	397	151	158	50	56.27	52.51	76.05
	H	112	326	108	82		51.91	62.93
	S	6	7	46	2		75.41	9.03
	N	7	34	194	403		63.17	75.05
ZCR	A	0	0	0	0	25.78	0	0
	H	0	0	0	0		0	0
	S	0	0	0	0		0	0
	N	522	518	506	537		25	100
Intensity	A	0	0	0	0	25.78	0	0
	H	0	0	0	0		0	0
	S	0	0	0	0		0	0
	N	522	518	506	537		25	100
MFCC+Mel spectrogram	A	469	63	24	14	77.20	82.28	89.85
	H	31	378	63	14		77.78	72.97
	S	8	68	305	53		70.78	60.28
	N	14	9	114	456		76.90	84.92
MFCC+Spectral contrast	A	465	69	49	23	72.78	76.73	89.08
	H	44	389	89	27		70.86	75.10
	S	5	41	232	57		69.25	45.85
	N	8	19	136	430		72.51	80.07
MFCC+ZCR	A	421	79	73	41	65.96	68.57	80.65
	H	37	389	138	34		65.05	75.10
	S	5	29	149	47		64.78	29.45
	N	59	21	146	415		64.74	77.28
MFCC+Intensity	A	442	32	31	31	69.56	82.46	84.67
	H	68	442	178	46		60.22	85.33
	S	4	20	145	40		69.38	28.66
	N	8	24	152	420		69.54	78.21
Mel spectrogram+Spectral contrast	A	430	110	67	37	70.72	66.77	82.38
	H	65	327	36	9		74.83	63.13
	S	11	57	275	50		69.97	54.35
	N	16	24	128	441		72.41	82.12
Mel spectrogram+ZCR	A	424	127	100	37	66.68	61.63	81.23
	H	69	296	43	12		70.48	57.14
	S	12	69	234	53		63.59	46.25
	N	17	26	129	435		71.66	81.01
Mel spectrogram+Intensity	A	467	37	42	13	73.74	83.54	89.46
	H	41	390	91	43		69.03	75.29
	S	6	62	248	50		67.76	49.01
	N	8	29	125	431		72.68	80.26
Spectral Contrast+ZCR	A	431	182	173	47	55.64	51.74	82.57
	H	81	298	122	86		50.77	57.53
	S	3	7	28	2		70.00	5.53
	N	7	31	183	402		64.53	74.86
Spectral Contrast+Intensity	A	427	38	45	28	62.17	79.37	81.80
	H	88	446	252	98		50.45	86.10
	S	0	1	11	0		91.67	2.17
	N	7	33	198	411		63.33	76.54
ZCR+Intensity	A	0	0	0	0	25.78	0	0
	H	0	0	0	0		0	0
	S	0	0	0	0		0	0
	N	522	518	506	537		25	100
All features	A	491	30	21	5	81.71	89.76	94.06
	H	25	428	61	17		80.60	82.63
	S	4	52	322	54		74.54	63.64
	N	2	8	102	461		80.45	85.85

Table 6. Confusion matrix and accuracy details of KNN classifier.

KNN classifier								
Feature extracted	**Class**	**Classified as**				**Accuracy (%)**	**Precision (%)**	**Recall (%)**
		A	H	S	N			
MFCC	A	488	90	26	17	72.83	78.58	93.49
	H	21	361	76	17		76.00	69.69
	S	8	61	271	106		60.76	53.56
	N	5	6	133	397		73.38	73.93
Mel spectrogram	A	389	123	62	38	65.91	63.56	74.52
	H	72	276	31	22		68.83	53.28
	S	42	93	301	70		59.49	59.49
	N	19	26	112	407		72.16	75.79
Spectral contrast	A	384	132	53	16	64.62	65.64	73.56
	H	95	306	101	42		56.25	59.07
	S	29	53	237	60		62.53	46.84
	N	14	27	115	419		72.87	78.03
ZCR	A	215	153	129	61	33.61	38.53	41.19
	H	137	129	131	111		25.39	24.90
	S	113	100	104	113		24.19	20.55
	N	57	136	142	252		42.93	46.93
Intensity	A	370	27	49	91	44.70	68.90	70.88
	H	14	216	156	108		43.72	41.70
	S	44	159	144	137		29.75	28.46
	N	94	116	157	201		35.39	37.43
MFCC+Mel spectrogram	A	461	116	52	15	71.87	71.58	88.31
	H	46	294	38	14		75.00	56.76
	S	11	99	311	77		62.45	61.46
	N	4	9	105	431		78.51	80.26
MFCC+Spectral contrast	A	470	81	32	7	76	79.66	90.04
	H	40	380	89	23		71.43	73.36
	S	9	43	282	56		72.31	55.73
	N	3	14	103	451		78.98	83.99
MFCC+ZCR	A	490	89	25	17	72.97	78.90	93.87
	H	20	362	73	17		76.69	69.88
	S	8	61	273	108		60.67	53.95
	N	4	6	135	395		73.15	73.56
MFCC+Intensity	A	495	39	16	9	75.13	88.55	94.83
	H	17	407	84	17		77.52	78.57
	S	6	63	263	111		59.37	51.98
	N	4	9	143	400		71.94	74.49
Mel spectrogram+Spectral contrast	A	445	136	39	15	72.59	70.08	85.25
	H	57	306	48	13		72.17	59.07
	S	15	59	320	68		69.26	63.24
	N	5	17	99	441		78.47	82.12
Mel spectrogram + ZCR	A	392	123	63	33	66.11	64.16	75.10
	H	71	276	31	21		69.17	53.28
	S	42	95	299	73		58.74	59.09
	N	17	24	113	410		72.70	76.35
Mel spectrogram+Intensity	A	412	70	34	30	69.47	75.46	78.93
	H	52	308	37	22		73.51	59.46
	S	40	106	318	76		58.89	62.85
	N	18	34	117	409		70.76	76.16
Spectral contrast+ZCR	A	385	132	54	16	64.57	65.59	73.75
	H	95	305	100	42		56.27	58.88
	S	28	54	236	60		62.43	46.64
	N	14	27	116	419		72.74	78.03
Spectral contrast+Intensity	A	424	72	19	14	69.37	80.15	81.23
	H	67	342	112	42		60.75	66.02
	S	17	76	259	61		62.71	51.19
	N	14	28	116	420		72.66	78.21
ZCR+Intensity	A	432	23	47	45	52.38	78.98	82.76
	H	14	232	180	95		44.53	44.79
	S	38	151	145	115		32.29	28.66
	N	38	112	134	282		49.82	52.51
All features	A	478	107	26	7	77.34	77.35	91.57
	H	38	343	47	14		77.60	66.22
	S	4	55	335	61		73.63	66.21
	N	2	13	98	455		80.11	84.73

Table 7. Confusion matrix and accuracy details of SL classifier.

SL classifier								
Feature Extracted	**Class**	**Classified as**				**Accuracy (%)**	**Precision (%)**	**Recall (%)**
		A	H	S	N			
MFCC	A	455	58	29	23	73.93	80.53	87.16
	H	46	403	71	17		75.05	77.80
	S	9	49	296	111		63.66	58.50
	N	12	8	110	386		74.81	71.88
Mel spectrogram	A	416	135	75	44	67.88	62.09	79.69
	H	56	294	34	4		75.77	56.76
	S	31	69	299	84		61.90	59.09
	N	19	20	98	405		74.72	75.42
Spectral contrast	A	374	127	101	22	61.11	59.94	71.65
	H	72	272	68	49		59.00	52.51
	S	67	93	229	68		50.11	45.26
	N	9	26	108	398		73.57	74.12
ZCR	A	378	223	223	89	40.33	41.40	72.41
	H	89	114	93	95		29.16	22.01
	S	13	45	24	29		21.62	4.74
	N	42	136	166	324		48.50	60.34
Intensity	A	400	22	42	77	49.45	73.94%	76.63
	H	14	359	274	185		43.15%	69.31
	S	2	13	7	11		21.21%	1.38
	N	106	124	183	264		39.00%	49.16
MFCC+Mel spectrogram	A	484	49	12	6	80.46	87.84	92.72
	H	30	407	52	15		80.75	78.57
	S	3	55	343	74		72.21	67.79
	N	5	7	99	442		79.93	82.31
MFCC + Spectral contrast	A	469	43	29	11	79.50	84.96	89.85
	H	33	426	48	15		81.61	82.24
	S	16	41	331	81		70.58	65.42
	N	4	8	98	430		79.63	80.07
MFCC + ZCR	A	484	36	23	8	76.96	87.84	92.72
	H	29	425	75	19		77.55	82.05
	S	5	50	297	113		63.87	58.70
	N	4	7	111	397		76.49	73.93
MFCC+Intensity	A	468	30	19	5	75.42	89.66	89.66
	H	32	423	86	24		74.87	81.66
	S	15	55	292	120		60.58	57.71
	N	7	10	109	388		75.49	72.25
Mel spectrogram+Spectral contrast	A	439	115	35	15	74.80	72.68	84.10
	H	48	342	35	18		77.20	66.02
	S	26	50	340	67		70.39	67.19
	N	9	11	96	437		79.02	81.38
Mel spectrogram+ZCR	A	439	132	56	16	71.20	68.27	84.10
	H	56	304	32	9		75.81	58.69
	S	15	69	315	87		64.81	62.25
	N	12	13	103	425		76.85	79.14
Mel spectrogram+Intensity	A	460	33	31	5	75.85	86.96	88.12
	H	29	379	50	23		78.79	73.17
	S	26	78	322	90		62.40	63.64
	N	7	28	103	419		75.22	78.03
Spectral Contrast+ZCR	A	437	84	92	16	65.87	69.48	83.72
	H	46	311	78	57		63.21	60.04
	S	31	96	231	71		53.85	45.65
	N	8	27	105	393		73.73	73.18
Spectral Contrast+Intensity	A	449	42	30	8	71.58	84.88	86.02
	H	31	376	112	57		65.28	72.59
	S	37	81	260	66		58.56	51.38
	N	5	19	104	406		76.03	75.61
ZCR+Intensity	A	449	22	45	55	57.03	78.63	86.02
	H	23	329	238	85		48.74	63.51
	S	26	71	86	73		33.59	17
	N	24	96	137	324		55.77	60.34
All features	A	502	27	8	3	83.82	92.96	96.17
	H	14	435	47	16		84.96	83.98
	S	4	46	364	73		74.74	71.94
	N	2	10	87	445		81.80	82.87

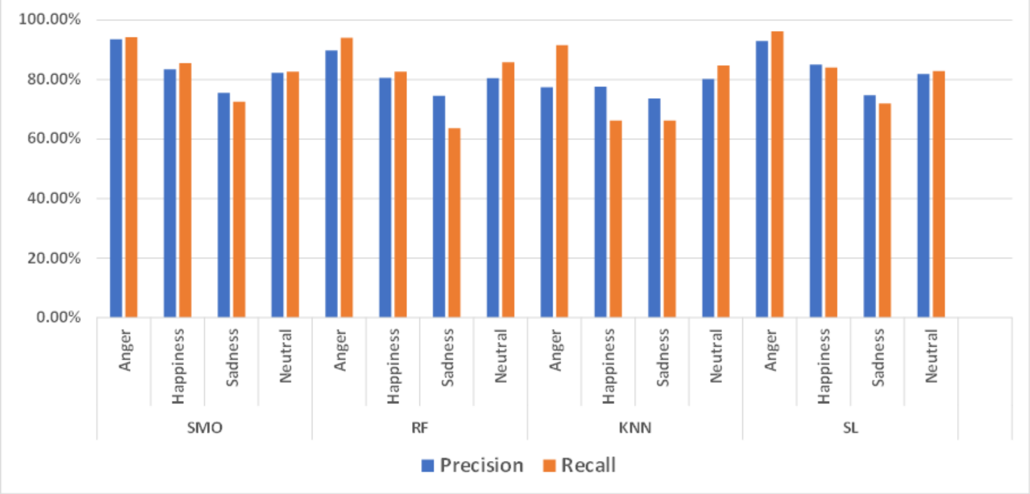


Figure 4. performance metrics for four classifiers (SMO, RF, KNN, SL) across four emotions (anger, happiness, sadness, neutral) using all feature combinations: Precision and recall.

To further discover the difference in classifiers performance, statistical testing was conducted using paired t-tests [67]. The t-test was conducted for accuracy results for all features set as shown in Table 8. The results of these t-tests are summarized in Table 9, where the t-statistic and p-value compare the performance of the classifiers across all feature sets. These t-test results indicate that, although there are some performance differences between classifiers, none of comparisons yielded statistically significant results at the common alpha level of 0.05. For example, the comparison between SMO and RF showed a t-statistic of 1.5883 and a p-value of 0.1227, indicating that while SMO has a higher mean accuracy, the difference was not statistically significant.

Table 8. Classification accuracy results for all features set.

Feature combination	SMO accuracy (%)	RF accuracy (%)	KNN accuracy (%)	SL accuracy (%)
MFCC	73.93	62.65	72.83	73.93
Mel spectrogram	64.95	66.92	65.91	67.88
Spectral contrast	62.07	56.27	64.62	61.11
ZCR	40.66	25.78	33.61	40.33
Intensity	49.35	25.78	44.70	49.45
MFCC+Mel spectrogram	79.93	77.20	71.87	80.46
MFCC+Spectral contrast	79.31	72.78	76	79.50
MFCC+ZCR	76.38	65.96	72.97	76.96
MFCC+Intensity	75.90	69.56	75.13	75.42
Mel spectrogram+Spectral contrast	73.50	70.72	72.59	74.80
Mel Spectrogram+ZC	69.23	66.68	66.11	71.20
Mel spectrogram+Intensity	74.51	73.74	69.47	75.85
Spectral contrast+ZCR	65	55.64	64.57	65.87
Spectral contrast+Intensity	71.82	62.17	69.37	71.58
ZCR+Intensity	56.79	25.78	52.38	57.03
All features	83.82	81.71	77.34	83.80

Table 9. T-test comparisons for classifiers.

Comparison	t-statistic	p-value
SMO vs RF	1.5883	0.1227
SMO vs KNN	0.7086	0.4840
SMO vs SL	- 0.1209	0.9046
RF vs KNN	-1.0278	0.3123
RF vs SL	-1.6743	0.1045
KNN vs SL	-0.8227	0.4172

In the same way, the SMO vs SL comparison produced a very low t-statistic of -0.1209 with a p-value of 0.9046, suggesting that their performances were almost identical. The experimental results demonstrate the effectiveness of combining multiple acoustic features, particularly MFCC, Mel Spectrogram, and Spectral Contrast, for emotion detection in natural Arabic audio files. The classifiers, especially SMO and SL, consistently outperformed RF and KNN, achieving the highest accuracies and best overall emotion detection performance. However, no statistically significant differences were found between the classifiers, as confirmed by the t-test comparisons. This suggests that the choice of feature sets has more significant impact on performance than the choice of the classifier. The t-test was conducted using the Python programming language.

5. Conclusions and Future work

In this paper, we have introduced a method for emotion detection in natural Arabic audio files. Our method consists of multiple stages: including audio dataset collection, pre-processing of audio files, annotation, audio normalization, features extraction and normalization, features selection, supervised learning classification, and evaluation. The dataset was used for the experiments implemented in this research was gathered from several Arabic YouTube channels on the internet. This dataset contains 2083 audio files (522 for anger emotion, 518 for happiness emotion, 506 for sadness emotion, 537 for neutral emotion). Sixteen experiments have been conducted to check the best acoustic feature sets and ML classifiers that work well for emotion detection in natural Arabic audio files. Four ML classifiers, including: SMO, RF, KNN and SL were applied. For evaluation purposes, three common effective measures were used Accuracy, precision and recall. The experiments yielded competitive results for emotion detection in natural Arabic audio files. The best results for all acoustic features set attained using the SMO and SL classifiers with accuracy 83.82%. Our contributions in this work include the following: building and evaluating a natural Arabic audio dataset with multiple emotional states from a large number of speakers and dialects. Another contribution is the size of the natural Arabic audio dataset and the duration of audio files which range from 1-9 seconds for every audio file. In addition, this work contributed not only to the domain of natural Arabic audio files, but also to existing natural dataset. For the future work, the task of building natural Arabic audio datasets with multiple emotional states is still a significant challenge. We intend to increase the audio dataset by adding more natural Arabic audio files. In addition, comparison this work with others datasets this include acted, semi-natural and elicited audio datasets. Lastly, we plan to improve model performance by incorporating multimodal emotional cues in our research. Specifically, we will add lexical features to combine with the acoustic features that are already used in this research. This combination is expected to capture a wider range of emotional expressions, leading to enhance accuracy in emotion detection in natural Arabic audio dataset.

References

[1] Abdel-Hamid L., "Egyptian Arabic Speech Emotion Recognition Using Prosodic, Spectral and Wavelet Features," *Speech Communication*, vol. 122, pp. 19-30, 2020. https://doi.org/10.1016/j.specom.2020.04.005

[2] Ahel Al Himmeh, Ahel Al Himmeh Channel, https://www.youtube.com/@HemmehJU/videos, Last Visited, 2024.

[3] Akhbar Al Nar, Akhbar Al Nar News Channel, https://www.youtube.com/@3lnar.newstv514/videos, Last Visited, 2024.

[4] AL Arabiya Arabic, Al Moqabala Al Arabiya, https://www.youtube.com/watch?v=N2zRSs4f8cA&list=PLOFBlNCrlrW5FRUDcFQ_K7CzRgWJu3QQo&index=81, Last Visited, 2024.

[5] Al Araby, Al Araby Channel, https://www.youtube.com/c/AlArabyAr, Last Visited, 2024.

[6] Al Hadath, Al Hadath Channel, https://www.youtube.com/c/AlHadath, Last Visited, 2024.

[7] Al Jazeera Arabic, Mozaein, https://www.youtube.com/watch?v=8ap6vdC0hrc&list=PLJyrzEL-wvYKmmASkEQKjs7_3RGab2Bp3, Last Visited, 2024.

[8] Al Jazeera Arabic, Podcast Al Jazeera, https://www.youtube.com/watch?v=Dz8VL_8hVXk&list=PLJyrzEL-wvYLgvwn54g28ML-Y4fxEeO2O&index=10, Last Visited, 2024.

[9] Al Kofiya Channel, Al Kofiya Channel, https://www.youtube.com/c/alkofiyatv, Last Visited, 2024.

[10] Al Qalah News, Al Qalah News Channel, https://www.youtube.com/@Alqalah_news/videos, Last Visited, 2024.

[11] Al Quds Today, Al Quds Today Channel, https://www.youtube.com/@-alqudstoday2300/videos, Last Visited, 2024.

[12] Al Quds Today, Al Quds Today Channel, https://www.youtube.com/channel/UCpZa_lVdcx0uRcATZi4CCGQ/videos, Last Visited, 2024.

[13] Al Wakeel News, Al Wakeel News Channel, https://www.youtube.com/@ALWAKEEL_NEWS, Last Visited, 2024.

[14] Al Watan Syria, Al Watan Syria Newspaper Channel, https://www.youtube.com/@Alwatan_Sy, Last Visited, 2024.

[15] Alamri H. and Alshanbari H., "Emotion Recognition in Arabic Speech from Saudi Dialect Corpus Using Machine Learning and Deep Learning Algorithms," *International Journal of Computer Science and Network Security*, vol. 23, no. 8, pp. 1-10, 2023. https://doi.org/10.21203/rs.3.rs-3019159/v1

[16] Alghad TV, Alghad TV Channel, https://www.youtube.com/c/alghadtv, Last Visited, 2024.

[17] Al-Jazeera Arabic, The Opposite Direction, https://www.youtube.com/watch?v=W6KTfe2W4n8&list=PLJyrzEL-wvYJS4SlCjPBUvdcQVBpaTy6B, Last Visited, 2024.

[18] Al-Jazeera Mubasher, Al Jazeera Mubasher Channel, https://www.youtube.com/c/ajmubasher, Last Visited, 2024.

[19] Aljuhani R., Alshutayri A., and Alahdal S., "Arabic Speech Emotion Recognition from Saudi Dialect Corpus," *IEEE Access*, vol. 9, pp. 127081-127085, 2021. https://doi.org/10.1109/ACCESS.2021.3110992

[20] Al-Salam TV, Salam TV Channel, https://www.youtube.com/@salamtv1, Last

Visited, 2024.

[21] Asas Platform, Asas Platform for Tawjihi in Jordan, https://www.youtube.com/@ripple2024_live_/videos, Last Visited, 2024.

[22] Azmin S. and Dhar K., "Emotion Detection from Bangla Text Corpus Using Naïve Bayes Classifier," *in Proceedings of the 4th International Conference on Electrical Information and Communication Technology*, Khulna, pp. 1-5, 2019. https://doi.org/10.1109/EICT48899.2019.9068797

[23] Baali M. and Ghenim N., "Emotion Analysis of Arabic Tweets Using Deep Learning Approach," *Journal of Big Data,* vol. 6, no. 89, pp. 1-12, 2019. https://doi.org/10.1186/s40537-019-0252-x

[24] Dijlah TV, Dijlah TV Channel, https://www.youtube.com/@DijlahTv, Last Visited, 2024.

[25] Dooz Nablus, Dooz Nablus Channel, https://www.youtube.com/@DoozNablus/videos, Last Visited, 2024.

[26] Dorry M., Emotion Identification from Spontaneous Communication, Master Thesis, Addis Ababa University, College of Natural Sciences, 2016. http://thesisbank.jhia.ac.ke/id/eprint/6109

[27] Fajer TV, Fajer TV Channel, https://www.youtube.com/@fajertv/videos, Last Visited, 2024.

[28] Gunes H., Schuller B., Pantic M., and Cowie R., "Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey," *in Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, pp. 827-834, 2011. https://doi.org/10.1109/FG.2011.5771357

[29] Han J. and Kamber M., *Data Mining: Concepts and Techniques*, San Francisco: Morgan Kaufmann, 2006.

[30] Horkous H. and Guerti M., "Recognition of Emotions in the Algerian Dialect Speech," *International Journal of Computing and Digital Systems*, vol. 10, no. 1, pp. 245-254, 2021. http://dx.doi.org/10.12785/ijcds/100125

[31] JavaTPoint, K-Nearest Neighbour Algorithm, https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning, Last Visited, 2024.

[32] JavaTPoint, Random Forest Algorithm, https://www.javatpoint.com/machine-learning-random-forest-algorithm, Last Visited, 2024.

[33] Jemdia Agency, Jemdia Agency Channel, https://www.youtube.com/channel/UChAtATk7PT0tW3YQkCnXJKQ, Last Visited, 2024.

[34] Jiang D., Lu L., Zhang H., Tao J., Cai L., "Music Type Classification by Spectral Contrast Feature," *in Proceedings of the IEEE International Conference on Multimedia and Expo*, Lausanne, pp. 113-116, 2002. https://doi.org/10.1109/ICME.2002.1035731

[35] Khabar News Agency, Khabar News Agency Channel, https://www.youtube.com/@khbrpress1/videos, Last Visited, 2024.

[36] Khalil A., Al-Khatib W., El-Alfy M., and Cheded L., "Anger Detection in Arabic Speech Dialogs," *in Proceedings of the International Conference on Computing Sciences and Engineering*, Kuwait, pp 1-6, 2018. http://doi.org/10.1109/ICCSE1.2018.8374203

[37] Klaylat S., Osman Z., Hamandi L., and Zantout R., "Emotion Recognition in Arabic Speech," *in Proceedings of the Sensors Networks Smart and Emerging Technologies*, Beirut, pp. 1-4, 2017. https://doi.org/10.1109/SENSET.2017.8125028

[38] Klaylat S., Osman Z., Hamandi L., and Zantout R., "Emotion Recognition in Arabic Speech," *Analog Integrated Circuits and Signal Processing*, vol. 96, pp. 337-351, 2018. https://doi.org/10.1007/s10470-018-1142-4

[39] Landwehr N., Hall M., and Frank E., "Logistic Model Tree," *Machine Learning*, vol. 59, no. 1, pp. 161-205, 2005. https://doi.org/10.1007/s10994-005-0466-3

[40] Lee C., "Toward Detecting Emotions in Spoken Dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293-303, 2005. https://doi.org/10.1109/TSA.2004.838534

[41] Ma'an Network, Ma'an Network Channel, https://www.youtube.com/c/MaanNetwork, Last Visited, 2024.

[42] McFee B., Raffel C., Liang D., Ellis D., McVicar M., Battenberg E., and Nieto O., "Librosa: Audio and Music Signal Analysis in Python," *in Proceedings of the 14th Python in Science Conference*, Texas, pp. 18-24, 2015. https://doi.org/10.25080/Majora-7b98e3ed-003

[43] Meddeb M., Karray H., and Alimi A., "Building and Analyzing Emotion Corpus of the Arabic Speech," *in Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition*, Nancy, pp. 134-139, 2017. https://doi.org/10.1109/ASAR.2017.8067775

[44] Meftah A. and Zakariah M., "Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus," *International Journal of Simulation: Systems, Science and Technology*. vol. 21, no. 2, pp. 1-6, 2020. https://doi.org/10.5013/IJSSST.a.21.02.21

[45] Milton A., Roy S., and Selvi S., "SVM Scheme for Speech Emotion Recognition Using MFCC Feature," *International Journal of Computer Applications*, vol. 69, no. 9, pp 34-39, 2013. https://doi.org/10.5120/11872-7667

[46] Mohammad O. and Elhadef M., "Arabic Speech

Emotion Recognition Method Based on LPC and PPSD," *in Proceedings of the 2nd International Conference on Computing, Automation and Knowledge Management*, Dubai, pp. 31-36, 2021. http://doi.org/10.1109/ICCAKM50778.2021.9357769

[47] Msdr News, Msdr News Channel, https://www.youtube.com/@MsdrNews, Last Visited, 2024.

[48] Mutasem Al Shesh, Mutasem Al Shesh Channel, https://www.youtube.com/@MutasemAlshesh937089, Last Visited, 2024.

[49] Nandwani P. and Verma R., "A Review on Sentiment Analysis and Emotion Detection from Text,' *Social Network Analysis and Mining*, vol. 11, no. 81, pp. 1-19, 2021. https://doi.org/10.1007/s13278-021-00776-6

[50] Osama Al Kahlout, Video by Osama Al Kahlout, https://www.youtube.com/watch?v=aOMh4cAdM7Y&t=217s, Last Visited, 2024.

[51] Palestine TV, Palestine TV Channel, https://www.youtube.com/@palestinetvchannel/videos, Last Visited, 2024.

[52] Palo H. and Mohanty M., "Classification of Emotions of Angry and Disgust," *Smart CR Review*, vol. 5, no. 3, pp. 151-158, 2015. https://doi.org/10.6029/smartce.2015.03.003

[53] Panacea Hu, Panacea Hu Channel, https://www.youtube.com/@panaceahu1198/videos, Last Visited, 2024.

[54] Quds News Network, Quds News Network Channel, Last Visited, 2024. https://www.youtube.com/c/QudsNPS

[55] Roya TV, Roya TV Channel, https://www.youtube.com/@royatv, Last Visited, 2024.

[56] Rum Online News, Rum Online News Channel, https://www.youtube.com/@rumonlinenews, Last Visited, 2024.

[57] Salian B., Narvade O., Tambewagh R., and Bharne S., "Speech Emotion Recognition Using Time Distributed CNN and LSTM," *in Proceedings of the International Conference on Automation, Computing and Communication*, Nerul, pp. 1-6, 2021. https://doi.org/10.1051/itmconf/20214003006

[58] Saraya News Agency, Saraya News Agency Channel, https://www.youtube.com/@sarayanewstv/videos, Last Visited, 2024.

[59] Snd News Agency, Snd News Agency Channel, https://www.youtube.com/channel/UCS1fWGLmwc0Fo4KSpKXo4FA, Last Visited, 2024.

[60] Swain M., Routray A., and Kabisatpathy P., "Databases, Features and Classifiers for Speech Emotion Recognition: A Review," *International Journal of Speech Technology*, vol. 21, pp. 93-120, 2018. https://doi.org/10.1007/s10772-018-9491-z

[61] Tajalsir M., Andez S., and Mohammed, F., "ASERS-CNN: Arabic Speech Emotion Recognition System Based on CNN Model," *Signal and Image Processing: An International Journal*, vol. 13, no. 1, pp. 45-53, 2022.

[62] Venkataramanan K. and Rajamohan H., "Emotion Recognition from Speech," *arXiv Preprint*, vol. arxiv:1912.10458, 2019. https://doi.org/10.48550/arXiv.1912.10458

[63] Wadhwa M., Pandey P., and Gupta A., Speech Emotion Recognition (SER) through Machine Learning, Analytics Insight, 2021. https://www.analyticsinsight.net/search?q=Speech%20Emotion%20Recognition%20(SER)%20through%20Machine%20Learning, Last Visited, 2024.

[64] Wafa Agency, Wafa Agency Channel, https://www.youtube.com/c/WafaAgency/videos , Last Visited, 2024.

[65] Wattan News Agency, Wattan News Agency Channel, https://www.youtube.com/c/WattanNews/videos, Last Visited, 2024.

[66] Wikipedia, Sequential Minimal Optimization', https://en.wikipedia.org/wiki/Sequential_minimal_optimization, Last Visited, 2024.

[67] Wikipedia, Student's t-test, https://en.wikipedia.org/wiki/Student%27s_t-test, Last Visited, 2024.

[68] World Population Review, Arabic Population, https://worldpopulationreview.com/country-rankings/arab-countries, Last Visited, 2024.

[69] Zinab R. and Majid M., "Emotion Recognition Based on EEG Signals in Response to Bilingual Music Tracks," *The International Arab Journal of Information Technology*, vol. 18, no. 3, pp. 286-296, 2021. https://doi.org/10.34028/iajit/18/3/4

Ashraf Kaloub received his M.Sc. Degree in Information Technology from the Islamic University of Gaza in 2013. He is currently a Lecturer in the Department of Multimedia and Information Technology at Al-Aqsa University. His research interests include Machine Learning, Multimedia Computing, Data Mining, Natural Language Processing, Speech Processing, 3D Modeling and Animation.

Eltyeb Abed Elgabar received his Bachelor Degree in Computer Science from University of Al-Neelain, master in Computer Science from Al-Neelain University, and PhD from University of Al-Neelain, Sudan. Currently he is an Associated Professor in Faculty of Computer Science and Information Technology, University of Al-Neelain Sudan. His research interest includes Machine Learning, Text Analysis, Database Systems, and Sentiment Analysis.