ISSN 2723-6471


# An Approach for Emotion Detection in Natural Arabic Audio Files Based on Acoustic and Lexical Features

Ashraf Kaloub[1,*], Eltyeb Abed Elgabar[2]

[1]*Department of Multimedia and Information Technology, Al-Aqsa University, Gaza-Strip, Palestine*

[1]*Department of Information Technology, University of Al-Butana, Sudan*

[2]*Department of Computer Science, University of Al-Neelain, Sudan*

(Received: December 15, 2024; Revised: January 30, 2025; Accepted: March 25, 2025; Available online: July 10, 2025)

**Abstract**

Emotion Detection is a crucial for enhancing human-machine interactions. This paper addresses the challenge of accurately recognizing emotional states from speech, particularly in distinguishing between emotions with similar acoustic characteristics, such as anger, happiness and surprise, which have high pitch and energy. While acoustic features convey significant information about emotional states, they are often inadequate for distinguishing between these emotions. This limitation highlights the need for improved performance in emotion detection systems. The main contribution of this work is the introduction of a multimodal approach that combines both acoustic and lexical features for emotion detection in natural Arabic audio files, focusing on four emotions anger, happiness, sadness and neutral. To the best of our knowledge, this is the first study that employ such a combination in this context, building on our previous work that utilized only acoustic features. Several Machine Learning (ML) classifiers were applied including Sequential Minimal Optimization (SMO), Random Forest (RF), K-Nearest Neighbors (KNN), and Simple Logistic (SL). Two types of experiments were executed: one using only lexical features and another combining various acoustic features sets with lexical features. This approach enhances our previous experiments that used only acoustic features. The experimental results show that SMO classifier achieved the highest performance, with an accuracy 96.11% when using all acoustic features combined with a unigram model, outperforming the other classifiers. These results suggest that combining acoustic and lexical features enhances the performance of emotion detection models, particularly for complex emotions in natural Arabic audio datasets.

*Keywords:* Emotion Detection, Machine Learning, Acoustic Features, Lexical Features, A Multimodal Approach

## 1. Introduction

Emotion Detection is crucial for facilitating natural human-machine interactions. Despite advancements in machine intelligence, accurately understanding human emotions and expressions remains a significant challenge. The primary objective of emotion detection is to identify the emotional state of the human based on their speech. This process typically involves extracting various acoustic features from audio, which are then used as inputs for classification classifiers.

While these acoustic features convey valuable information about emotional states, they are often inadequate for distinguishing between emotions with similar acoustic characteristics, such as anger, happiness and surprise, which are both associated with have high pitch and energy levels. In this case, distinguishing between emotions such as anger, happiness and surprise solely through acoustic features is challenging. Moreover, analyzing only the textual component of speech fails to present a comprehensive view of the emotional content [1]. In addition, Arabic is spoken in different dialects, and every dialect own its characteristics, which can affect how emotions are conveyed [2].

To address these limitations, this paper proposes a multimodal approach that combines acoustic and lexical features to improve the performance of a previously proposed system focused only on acoustic features for recognizing emotions specifically anger, happiness, sadness and neutral from natural Arabic speech [3]. By integrating both acoustic and

***Corresponding author: Ashraf Kaloub (ai.kaloub@alaqsa.edu.ps)**
**DOI: https://doi.org/10.47738/jads.v6i3.617**

ISSN 2723-6471


lexical features we can leverage the strengths of acoustic features as it valuable for capturing prosodic cues like energy, which are important for detection emotions like anger, happiness. On the other hand, lexical features analyze the actual content of the speech, providing insights into the sentiment that conveyed over the choice of words. By combining these two feature sets, we can improve emotion detection of the models.

This paper has five main sections. Section 1 presents the topic and highlights the research problem, followed by a section 2 of related works on emotion detection based on acoustic features, lexical features, or a combination of the two. Section 3 describes the audio dataset used in this work. Section 4 describes the proposed approach. Section 5 presents Experiments and Results. Section 6 comparison with previous works. Finally, section 7 presents the conclusion and future works.

## 2. Related Works

In this section, various related works are studied and investigated. The related works are introduced and analyzed for using both of speech and text in the process of emotion detection. Most of the research for Arabic and other languages relies on acted, elicited and semi-natural audio datasets with limited number of audio recordings. In addition, for each related work, we discuss the limitations identified in the research.

Kaloub and Abed Elgabar [3] introduced a natural Arabic audio dataset constructed using freely accessible YouTube videos on the internet. Each audio file was labelled as angry, happy, sad, or neutral based on the emotional content perceived by human listeners. Aset of acoustic features, including spectral and prosodic features, was extracted for each 1 to 9 second speech segment. Multiple classification classifiers were applied to recognize anger, happiness, sadness, and neutral in the natural Arabic dataset. The highest accuracy performance was 83.82% achieved by both the SMO and SL classifiers, when using a combination 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. The accuracy performance of the classification models still needs improvement. Table 1 shows the performance of the four classifiers in terms of accuracy, using all acoustic feature sets.

**Table 1.** Classification accuracy results for all Acoustic features sets.

| Features Extracted | 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 | 76.00 | 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 +ZCR | 69.23 | 66.68 | 66.11 | 71.20 |
| Mel Spectrogram +Intensity | 74.51 | 73.74 | 69.47 | 75.85 |
| Spectral Contrast +ZCR | 65.00 | 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 acoustic features | 83.82 | 81.71 | 77.34 | 83.80 |

Aljuhani et al. [4] presented a new approach for Arabic Speech Emotion Recognition from Saudi Dialect Corpus. The dataset was created 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 algorithms, the final result of dataset was included of 175 records, contained male and female actors divided 113 chunks for males and 62 for females with total duration 11 minutes. The four 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 emotions in audio dataset. For the classification, spectral features used where MFCC and

ISSN 2723-6471


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. However, the dataset is limited, containing only 175 records and its scope is constrained by concentrating on specific regional dialect (Saudi gulf dialect). In addition, the dataset is semi-natural Audio dataset collected from popular Saudi YouTube channel Telfaz11.

Mohammad and Elhadef [5] presented a method for Arabic Speech Emotion Recognition (SER). They used an audio dataset that contains four emotions (happy, surprised, sad, questioning). Emotion speech audio files were gathered and recorded by humans (5 males and 5 females), each one of them recorded 20 sentences for each type of emotion and the results were 200 collected records. For the recorded audio files, they used the extended WAV files format. After segmenting the signal, the 14 Linear Predictive Coding (LPC) coefficients were extracted from the segmented original signal, along with 11 Perceptual Phase Shift detection (PPSD) coefficient. In this method they applied five Classification algorithms (MLP, KNN, Decision Tree, SVM and Logistic Regression), the experiments done using the same extracted features for all of them. The results achieved by these algorithms were 66.7%, 66.7%, 91%, 75%, 91.7%, respectively. However, this method has multiple limitations, it used only 200 collected records for experiments, the acted dataset not sufficient for multimodal approaches like (speech and text), the emotions are simulated, the obtained accuracy for some of algorithms is an indication of the difficulty of the task and there is no information about the number of instances for each emotion in the dataset.

Klaylat et al. [6] proposed approach to detect emotions from natural audio files, the first a realistic corpus from Arabic TV shows were collected, 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 include Egyptian, Gulf and Lebanese speakers, the videos were different in length, containing both male and female speakers. Listening test was done to label each video, where 18 listeners were asked to listen to each video to perceive one of the three emotion states: happy, angry or surprised. Each video was divvied into smaller segments based on who is speaking the representer or the caller, some pre-processing operations were done to remove Silence, laughs and noisy segments. Every segment was automatically split into 1 sec speech units, the final result of audio dataset was included of 1384 records with 505 happy, 137 surprised and 741 angry segments. Thiry five classification algorithms were used to classify these audio files based on 845 audio features extracted from these audio files. the best result was 95.52% using SMO (Sequential minimal optimization) 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. the limitations of this work that it is used an imbalanced audio dataset, also 1 sec duration for all segments not completely enough to detect some other emotions, limited emotions and dialects. Limited dialectal coverage, specifically focusing on Egyptian, Gulf, and Lebanese dialects. Different classifier performance, while some classifiers performed well, others yielded lower accuracies. Additionally, the dataset consists of 1384 chunks, which might be consider small for developing a robust model.

Ira and Rahman [7] presented a dataset known as Ryerson Audio-Visual Dataset (RAVDESS). This dataset was created using the voice of 24 professional actors having North American accent. The dataset contains voice of both males and females. This dataset has audio, video and audio-visual files. They used only audio files for this work. The audio files of all actors (numbered 01-24) consist of 1440 utterances. This audio dataset includes eight emotions (happy, neutral, calm, angry, surprise, fearful, disgust and sad) for neutral emotion they used 96 segments and 192 segments for each other emotions. They used five audio features for extraction: MFCC, Mel spectrogram, Chroma, Contrast and Tonnetz. For experiments they used six classifiers: MLP, Random Forest, AdaBoost, SVM, Gradient Boosting, and HistGradientBoosting, theses classifiers achieved accuracies of 53%, 59%, 32%, 54%, 56% and 59% respectively. The experiments were performed with different training and testing data ratios. The best accuracy has been achieved from the 90% training data and 10% test data and this result achieved by ensemble method with an accuracy 70%, the ensemble method used combination of Random Forest, Gradient Boosting and Hist Gradient Boosting classifiers. The obtained accuracy for these algorithms is an indication of the difficulty of the task. The acted dataset not sufficient for multimodal approach like (speech and text). The emotions are simulated. In addition, imbalanced dataset, this can lead to bias in the model.

Azmin and Dhar [8] proposed a method for emotion detection based on text, they used a corpus of texts and posts collected from Facebook groups and some public posts of famous bloggers and the comments were collected based on different socio-political issues. The dataset contains 4200 comments and they take on consideration three emotion labels (happy, sad, anger) for each of those comments, in the dataset nearly 3780 of the comments were used for training data and 420 comments as test data. The three emotional states used from the dataset for happy, sad and angry included 1812 comments, 1166 comments and 1222 comments, respectively, a lot of preprocessing operations on data were happened to remove any kind of unnecessary information to be easy for classification. They use multipole features such as n-grams, POS tagger, and TF-IDF to enhance the efficiency. For emotion classification they used Multinomial Naïve Bayes classifier. The results showed an accuracy of 78.6% using Naïve Bayes classifier. However, the method has several limitations, it used only one classification algorithm. Using only (textual dataset) limits the exploration of multimodal approaches, which combine both text and speech for emotion detection.

Wikarsa and Thahir [9] presented a text mining application of emotion classifications of twitters users, the dataset was collected using streaming API and twitter search, with additional filters based on username or keyword, they used 10-fold cross validation to measure the level of accuracy generated by the application. The application can classify the emotions into six categories happiness, sadness, fear, anger, surprise and disgust. The results showed that the accuracy of this applications is 83% using Naïve Bayes Classifier for 105 tweets. The limitations of this method that it used only one classification algorithm. Limited dataset, containing (105 tweets). There is no information about the number of instances for each emotion in the dataset. In addition, using only (textual dataset) limits the exploration of multimodal approaches, which combine both text and speech for emotion detection.

Aljwari [10] proposed a method for emotion detection based on texts, they utilized the dataset of Arabic tweets presented in SemEval-2018, which is a publicly available benchmark dataset. They used only (934) tweets from this dataset and by taking on consideration four emotions labels (fear, anger, sadness, and joy), multiple preprocessing operations on data were conducted to remove any kind of unnecessary information to be easy for classification such as: removing stop words, repeating chars, English characters, mentions, punctuation marks, and Arabic diacritics. The experiments were conducted using five machine learning classifications algorithms: Decision Tree (DT), KNN, Naive Bayes (NB), Multinomial Naive Bayes (NB), and Support Vector Machine (SVM) to classify emotions. The dataset was split into two parts: the training and the testing part, 747 samples for training and 182 samples for testing. The results showed that Decision Tree and K- Nearest Neighbour classifiers have achieved the best accuracy of 74%, While the NB and Multinomial NB classifiers achieved accuracy of 69%, and the SVM achieved accuracy of 63%. However, the study has several limitations, including the obtained accuracy for these algorithms is an indication of the difficulty of the task. The use only textual dataset limits the exploration of multimodal approaches, which combine both text and speech for emotion detection. Limited textual dataset, it contains only 934 tweets, and there is no information about the number of instances for each emotion in the dataset.

Kurniawati et al. [11] presented a new approach to detect emotions in Indonesian spoken language based on acoustic and lexical features, they used a corpus of Indonesian video recordings from television talk shows, where video recordings of the talk shows stripped down to audio only. They choose recording of the corpus from three Indonesian talk shows to cover a broader range of emotional content in the collected data. The final result of the dataset was 1854 segments or utterances, 1576 utterances for the training data and 278 utterances for the test data, where the training data has distribution as follows: 355 utterances labelled as happiness, 303 as sadness, 350 as anger, 329 as fear, 11 as disgust, and 228 as surprise. They used three classification algorithms SVM, Random Forest, and Multinomial Naïve Bayes. The final results showed that SVM outperforms the RF and MNB algorithms. It achieved an average F- measure of 71.3% for 6 emotion classes by combining both acoustic and lexical features. However, the method has some limitations, the obtained accuracy for these algorithms is an indication of the difficulty of the task. In addition, it is imbalanced dataset, where only 11 segments are labelled as disgust.

The reviewed studies indicate a growing interest in emotion detection in audio, particularly in Arabic and other languages. Most of these studies rely on acted, elicited or semi-natural audio datasets. Furthermore, most of these studies focus solely on acoustic features. This study aims to address these gaps by focusing on these specific challenges, by constructed the first natural Arabic audio dataset based on combined acoustic and lexical features.

 ISSN 2723-6471


## 3. The Audio Dataset

The natural Arabic audio dataset used in this paper was constructed by 1103 videos ranging from 1 to 50 minutes that were downloaded from various freely YouTube channels. We used the "Opposite Direction" program on Al-Jazeera YouTube channel as a source to collect the audio files to represent the anger emotion, and used winning in sports and achieving success in tawjihi exams topics as a source to represent happiness emotion. For sadness emotion we used topics related to loss, the sources of these files are meetings with individuals who lost some of their families in wars, accidents or natural deaths. For neutral emotions, our sources included podcasts, news and documentary programs. All these YouTube channels provided natural audio files, and we have explained this more extensively in our previous work in [3]. Since the audio dataset was designed to capture emotional content from different and large natural sources, the distribution of dialects into the dataset has not specifically documented. This diversity of emotional content, alongside the different dialects represented, improve dataset robustness in detection emotions across different Arabic dialects. Furthermore, the collected dataset consists of audio files spoken in Modern Standard Arabic (MSA), different DA, or a mix of both. providing a diverse linguistic range for every video, 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 data 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 audio files were extracted independently from the videos as the focus on speech data. All audio files were segmented into smaller chunks based on the emotional content of every single audio file. Noise, including background music, was removed from each chunk to maintain the appropriate quality for every audio file. These audio files were then labelled by three human listeners (2 females and 1 male) as happy, angry, sad and neutral on the emotions perceived, and hence a corpus composed of 2083 audio files classified by emotion (522 for anger, 518 for happiness, 506 for sadness, and 537 for neutral), this distribution ensures a balanced representation for every emotional state, which is important for training robust emotion detection models. Before feature extraction, all audio files were normalized to ensure consistent loudness levels for all audio files, this a very important step for accurate features extraction, especially in different recording 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. Next, 326 acoustic features, categorized into two types (spectral and prosodic) were extracted for each speech unit. These acoustic features include seven types of raw speech features: Mel Frequency Cepstral Coefficient (MFCC), Mel Spectrogram, Spectral Contrast, Chroma, Zero Crossin Rate (ZCR), Pitch and Intensity. Following extraction, features were normalized 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. Finally, feature selection techniques were applied to select the best appropriate features that have information to obtain better performance of the classifiers, feature's selection was separated into two parts: Attribute evaluator and search method and each part have several techniques from which to select. For audio normalization, features extraction, and features normalization, we used librosa library, which consider a Python package for music and audio analysis [12].

Figure 1 Illustrates screenshot of using Weka during features selection. At the end, we obtained 26 acoustic features, excluding Pitch and Chroma, which are related to MFCC, Mel Spectrogram, Spectral Contrast, ZCR and Intensity, as shown in table 2. The reduction from 326 acoustic features to 26 was executed using WrapperSubsetEval with the BestFirst search method, ensuring that only the most informative features were chosen for optimal classifier performance.

 ISSN 2723-6471


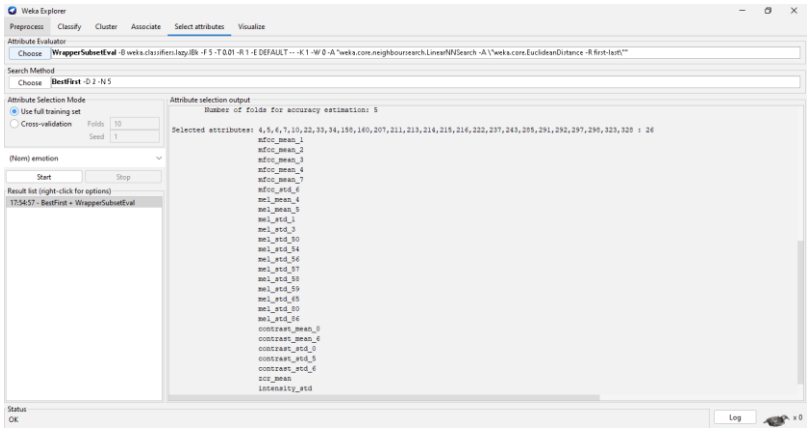


**Figure 1.** Screenshot of the WEKA during feature selection.

**Table 2.** Distribution of 26 acoustic features after features selection process

| Category | Raw Feature | Derived Features | Category | Raw Feature | Derived Features |
|---|---|---|---|---|---|
| Spectral | MFCC | Mfcc_mean_1 | Spectral | Mel Spectrogram | Mel_std_56 |
| | | Mfcc_mean_2 | | | Mel_std_57 |
| | | Mfcc_mean_3 | | | Mel_std_58 |
| | | Mfcc_mean_4 | | | Mel_std_59 |
| | | Mfcc_mean_7 | | | Mel_std_65 |
| | | Mfcc_std_6 | | | Mel_std_80 |
| | | Mfcc_mean_1 | | Spectral Contrast | Contrast_mean_0 |
| | Mel Spectrogram | Mel_mean_4 | | | Contrast_mean_6 |
| | | Mel_mean_5 | | | Contrast_std_0 |
| | | Mel_std_1 | | | Contrast_std_5 |
| | | Mel_std_3 | | | Contrast_std_6 |
| | | Mel_std_50 | Prosodic | ZCR | Zcr_mean |
| | | Mel_std_54 | | Intensity | Intensity_std |

Moreover, we utilized the features selection methods to address the computational complexity of our multimodal approach, emphasizing the important of features selection techniques which reduce the dimensionality and redundancy. By executing the feature selection methods, we decreased the computational effort, making the model more efficient in processing data. In addition, reducing the number of features allow our model to handle larger datasets, ensuring our model works well and remains accurate across various datasets.

## 4. The Proposed Approach

This section presents our proposed a multimodal approach for detecting emotions in natural Arabic audio files from freely YouTube channels. The approach combines both acoustic and lexical features, extended our previous work [3], which focused on experiments and discussions related solely to acoustic features.

In this work, we expand on that work by applying experiments with lexical features and then combining them with the previously analyzed acoustic features to enhance emotion detection. Specifically, after concentrating on acoustic features experiments in our previous work, we now investigate lexical features and their combination with acoustic features for a more robust and accurate emotion detection method. The overall process of our proposed approach is depicted in figure 2. To conduct experiments with lexical features, each of the following steps should be performed first.

 ISSN 2723-6471


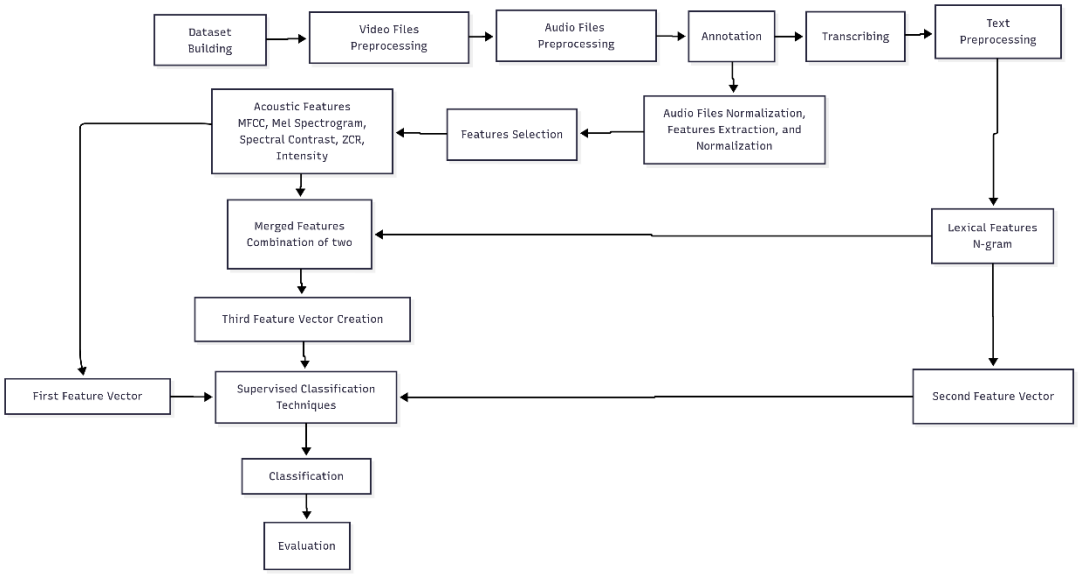


**Figure 2.** The overall process of our proposed approach

## 4.1. Transcribing

In this step, all annotated audio files were converted into the corresponding text. Transcribing was performed on all speech utterances obtained from the annotation process. This was done manually to ensure consistency and maintain the quality of the transcriptions for all speech utterances. In the end we obtained 2083 transcriptions representing four emotions distributed as follow: 522 transcriptions for anger, 518 transcriptions for happiness, 506 transcriptions for sadness and 537 transcriptions for neutral. Samples of transcriptions obtained from our experimental Dataset can be seen in table 3.

**Table 3.** Samples of transcriptions of utterances

| Emotion Type | English | Arabic |
|---|---|---|
| Happy | A thousand million congratulations to the great Al-Ahly fans. | ألف مليون مبروك لجماهير الأهلي العظيمة. |
| Happy | An indescribable joy, definitely happy to have my family and relatives with me in this celebration. | فرحة لا توصف, أكيد سعيدة بوجود أهلي وأقاربي معاي في هالفرحة. |
| Angry | They all escaped from your hell and your crimes. | كلهم هاربين من جحيمكم من إجرامكم. |
| Angry | They are the ones who conspired against the Palestinian cause, they are the ones who killed it. | هم من تآمروا على القضية الفلسطينية, هم من قتلوا القضية الفلسطينية. |
| Sad | May god have mercy on him, and hopefully God accepts him as a martyr in heaven at the highest ranks. | الله يرحمه, إن شاء الله ربنا بتقبله شهيد في الجنة وفي أعلى الدرجات. |
| Sad | I want to tell him that despite the pain of his departure, he is a martyr. | بدي أقله رغم الألم إلا بوجعني على فراقه إلا إنه هو شهيد. |
| Neutral | Geography, lifestyle, and environment all determine what we eat. | الجغرافيا ونمط الحياة والبيئة كل ذلك يُحدد ما نأكل. |
| Neutral | As I mentioned, the economic delegation is discussing further investment between the two countries. | الوفد الإقتصادي كما ذكرت يتناول مزيد من الإستثمار بين البلدين. |

## 4.2. Text Preprocessing

The transcriptions in the dataset that obtained from the process of transcription contained a lot of useless and duplicate data. Consequently, applying emotion detection directly to these textual data may lead to poor results. That is why preprocessing techniques are important in order to improve the value of the data. Some preprocessing in the Arabic text dataset were applied. It includes tokenizing strings to words, applying stop words removal, applying the suitable term stemming as shown in figure 3. We used the open-source machine learning tool Rapid Miner for text preprocessing [13].

ISSN 2723-6471


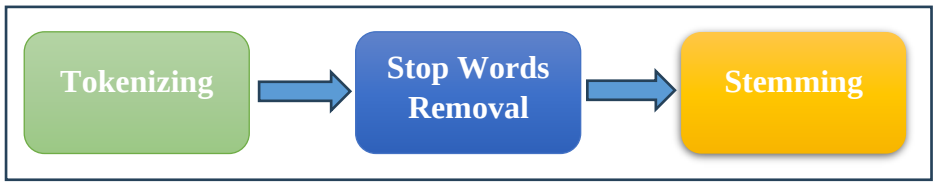


**Figure 3.** The text preprocessing steps [13]

### 4.2.1. Tokenization

Tokenization is the process of dividing a document, paragraph or just one sentence into chunks of words called tokens. For example, consider the sentence "this place is so beautiful"; after tokenization, it will split into tokens like 'this,' 'place,' 'is,' 'so,' 'beautiful.' Normalizing the text is a crucial step for achieving uniformity in data by converting the text into standard form and correcting the spelling of words. In this work, the tokenization process is responsible for defining word boundaries such as white spaces from transcriptions [14]. The open-source machine learning tool RapidMiner has been used for text tokenization as shown in figure 4 [13].

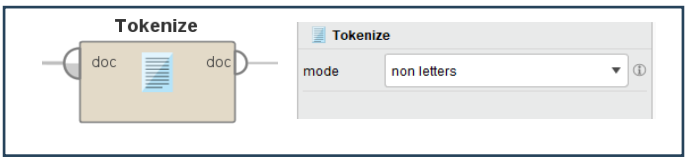


**Figure 4.** Text tokenization using RapidMiner [13]

### 4.2.2. Stop Words Removal

Unnecessary words that do not contribute in emotion detection should be removed [14]. In the case of Arabic, the list of stop words comprises articles, prepositions, conjunctions, pronouns, days of week, and months of the year [15]. So, these need to be removed to reduce redundant computations. In addition, it is wise to discard frequently occurring words as they have little information content. An effective Arabic Stop words list is found in Arabic stemmer package with some modifications and additional words are added into this list [16]. The open source machine learning tool RapidMiner has been used for stopwords removal as shown in figure 5 [13].

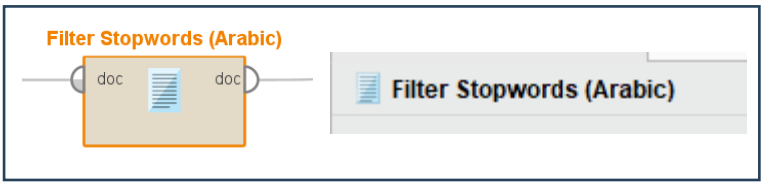


**Figure 5**. Stopwords removal using RapidMiner [13]

### 4.2.3. Stemming

In the Arabic language, two primary methods of morphological analysis are used: root-based stemming and light stemming. Root-based stemming removes all affixes and return each Arabic word to its root form. While, light stemming removes only common affixes (prefixes and suffixes) without changing the origin root of a word. The light stemmer is considered simple and fast since it does not need any grammatical analysis to identify the root, and instead reduce every word to its shortest possible form while preserving its meaning. The rood-based stemmer, however, reduces dictionary size as it matches many words to the same root. In this study, we use the light stemming approach because many words which share the same root have completely different meanings, consequently we maintain the correct meaning of the word. Moreover, the light stemming has the least preprocessing time [17], [15]. Additionally, the experiments report that a light stemmer outperforms the root-based stemmer in information retrieval as the latter affects the words meanings [16]. The open-source machine learning tool RapidMiner has been used for light stemming as shown in figure 6.

 ISSN 2723-6471


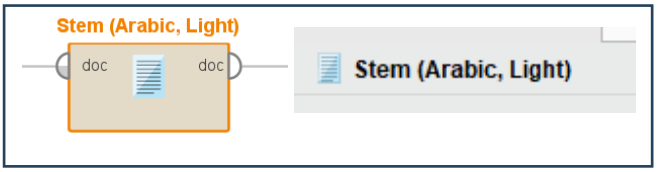


**Figure 6.** Light stemming using RapidMiner [13]

## 4.3. Features Extraction

This section explains the features extraction methods that we have applied during this work for emotion detection, which consider a classification problem that can be solved by using the ML techniques. ML classifiers understands text in terms of numerical data, and the process of converting the text or words into numerical vectors is called word vectorization. This feature extraction technique involves breaking a document into sentences, and then further into words, after that the feature matrix is built [14]. One of the primary objectives of this paper is to extract and understand which features are most relevant when combining audio and text information, as well as identifying the possible overlap between features.

Two types of features are extracted in this multimodal approach: acoustic and lexical features. Acoustic features were discussed in detail in our previous paper [3]. For lexical feature extraction, one of the most commonly methods used is the N-gram model, often combined with TF-IDF weighting technique. The N-gram model is an effective approach to resolve the order of words in sentence vector representation. It refers to a sequence of words within a sentence, where N indicates the size (number of words) of a sequence. The commonly used sizes of N-grams include unigram (N = 1) bigram (N = 2) and trigram (N = 3). In the case of the unigram model, each distinct word in the dataset is considered as a feature. To represent the candidate features, TF-IDF weighting is applied to the unigram model to enhance a given text representation as feature vector. The TF-IDF model, a statistical measure used to evaluate how important a word is within a document relative to a collection of documents. The TF-IDF model has been chosen in this work because it is one of the most effective weighting methods that is used in text mining, as it emphasizes relevant terms among all sentences [14], [19]. As a result of this process, a set of 3919 distinct feature was obtained. The open-source machine learning tool RapidMiner has been used for this purpose.

## 4.4. Classification

For the text files classification tasks, we have to select a suitable supervised classifier that can achieve higher classification results. We have selected four ML classifiers within RapidMiner Platform to judge whether the text files represent one of the following emotions: anger, happiness, sadness or neutral (See table 4). These classifiers include: 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 emotions detection tasks in textual data. The training process includes feeding these classifiers with lexical features vectors generated from the text files obtained from the process of transcription of audio files to classify every text file into one of the emotional states: anger, happiness, sadness, or neutral.

**Table 4.** classifiers and their rational

| Classifier | Rationale |
|---|---|
| SMO | Efficient for handling high dimensional data, suitable for training SVM quickly and effectively. Leveraging analytical solutions by dividing large problems into smaller problems [21]. |
| RF | Improve predictive accuracy in complex problems by combining several decision trees outputs, effective for handling diverse and noisy data as found in natural emotion datasets [10]. |
| KNN | It is suitable for the tasks where the similarity of new instances to known instances is a reliable method for classification [11]. |
| SL | Provides clear probabilistic outputs, making it valuable for understanding feature contributions and outcomes in multimodal emotion detection [16]. |

### 4.4.1. SMO

The SMO algorithm is a heuristic approach used to optimize two variables at a time in the process of variable selection process, ensuring that at least one variable violates the Karush-Kuhn-Tucker (KKT) conditions. The SMO algorithm

mainly solves the dual problem of convex quadratic programming of nonlinear SVM. The dual formulation of the SVM optimization problem solved by SMO is given as follows:

maxα∑i=1nαi−12∑i=1n∑j=1nyiyjK(xi,xj)αiαj,

$$\max_{\alpha} \Sigma_{\{i=1\}}^{n}\alpha_i - \left(\frac{1}{2}\right)\Sigma_{\{i=1\}}^{n}\Sigma_{\{j=1\}}^{n}y_i y_j K(x_i, x_j)\alpha_i\alpha_j \quad (1)$$

*with the following constraints*:

$$0 \leq \alpha_i \leq C, for\ all\ i = 1, 2, \ldots, n,$$

$$\Sigma_{\{i=1\}}^{n} y_i \alpha_i = 0$$

C is an SVM hyperparameter and K (xi, xj) is the kernel function, both supplied by the user; and the variables αi are Lagrange multipliers [20].

### 4.4.2. RF

The Random Forest 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. Instead of relying on a single decision tree, Random Forest enhances predictive accuracy by averaging predictions from individual trees. Each tree depends on a random vector sampled independently and with the same distribution for all trees in the forest. As the number of trees increases, the generalization error converges to a limit, and the overall performance of the model is influenced by both the strength of individual trees and the correlation between them. The formula for Random Forest in class prediction is:

$$\hat{y} = arg\ max_{c}\left(\sum_{t=1}^{T} I(h_{t(x)} = c)\right) \quad (2)$$

where each tree (ht) in the forest makes a prediction for an input x, and the function I(ht(x)=c) checks if tree t predicted class c. If yes, it counts as 1, otherwise 0, the formula adds up these counts across all T trees, the arg max part means: choose the class c that got the highest count (i.e., the most votes from trees) [21].

### 4.4.3. KNN

KNN is a non-parametric supervised learning algorithm used for both classification and regression. It operates on the concept of feature similarity to classify new data. In this, the new data will be assigned a class based on how closely it matches the data in the training set [22].

It allocates the feature variable to the designated class based on a distance measure such as the Euclidean distance. The Euclidean distance between two vectors, $p = (p1, p2 \ldots \ldots \ldots pm)$ and $q = (q1, q2 \ldots \ldots \ldots qm)$, is calculated as:

$$d(p, q) = \sqrt{\sum_{j=1}^{m}(p_j - q_j)^2} \quad (3)$$

### 4.4.4. Simple Logistic (SL)

The creation of the Simple Logistic classifier is influenced by the principles of Logistic Model Trees (LMT). The LMT algorithm combines logistic regression into decision 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 [23].

## 4.5. Evaluation

For the evaluation process, specific metrics are required to evaluate the ML classifiers performance. A commonly used method used 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 [24]. Performance is calculated from the confusion matrix using three evaluation metrics: accuracy, recall and the precision [18].

ISSN 2723-6471


## 5. Experiments and Results

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

### 5.1. Experimental setup

In this subsection, we explained the experimental process we have used to evaluate our method for the task of emotions detection. For our multimodal approach (using both acoustic and lexical features) in classification task experiments, we have used the audio dataset and the transcriptions of this audio files, 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 with their transcriptions (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 the field of emotion detection and have demonstrated best results in many experiments involving acoustic and lexical features. We carried out three groups of experiments in order to evaluate the effectiveness of our approach for detection emotions in the collected natural Arabic audio dataset. These experiments were grouped based on feature sets: acoustic features, lexical features and a combination of both acoustic and lexical 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.

### 5.2. 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 feature sets and ML classifiers that work well for emotion detection in natural Arabic audio files.

#### 5.2.1. Experiments with the Lexical Features (Unigram)

This subsection presents and discussed the results of the various experiments performed to evaluate the best ML classifiers that work well with lexical features that derived from transcriptions of the audio dataset. we utilized a unigram model, where each word in the transcriptions represents an individual feature. The selection of unigram model was mainly based on its effectiveness in capturing the most relevant lexical cues without the computational complexity and potential overfitting associated with higher order in Ngrams. We combined Ngram model with TF-IDF weighting because this method effectively highlights significant words, making it easy to identify important text patterns. Other techniques were excluded to maintain computational efficiency. In addition, the preliminary experiments showed accepted results using Ngram model combines with TF-IDF, which then used with the combinations with various acoustic features experiments done in the previous work.

The experiments were implemented using multiple ML classifiers including: SMO, RF, KNN and SL. The primary purpose of these experiments is to evaluate the effect of combining lexical features from transcriptions with the feature model. In addition, this analysis helps to investigate the impact of using these features alone for the first time in emotion detection on a natural Arabic audio dataset.

Table 5 shows the confusion matrices of the four ML classifiers SMO, RF, KNN and SL based on the lexical features extracted from transcriptions. In this table, "A" represents Anger, "H" represents Happiness, "S" represents Sadness, and "N" represents Neutral. In addition, the table display the classification results of the Lexical features experiments for these classifiers. As shown in the table, out of a total of 2083 instances, SMO classifier correctly classified 1873 instances, while 210 were incorrectly classified. For the RF classifier 1782 instances were correctly classified and 301 were incorrectly classified. The KNN classifier classified 1753 of the instances correctly and 330 were incorrectly classified. The SL classifier classified 1815 of the instances correctly and 268 were incorrectly classified. Moreover, the table display the classification results of the Lexical features experiments for these classifiers. The values in the

 ISSN 2723-6471


table represent the accuracy, precision and recall for every classifier. Values in bold indicate the best results for the classifiers for the lexical features experiments.

**Table 5.** Confusion matrix and accuracy details for the four classifiers of Unigram experiments

| Features Extracted | Class | Classified as | | | | Precision (%) | Recall (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|
| | | A | H | S | N | | | |
| | | | | ***SMO Classifier*** | | | | |
| Unigram | A | 486 | 16 | 29 | 82 | 79.28 | 93.10 | **89.92** |
| | H | 1 | 487 | 4 | 4 | 98.19 | 94.02 | |
| | S | 7 | 8 | 459 | 10 | 94.83 | 90.71 | |
| | N | 28 | 7 | 14 | 441 | 90.00 | 82.12 | |
| | | | | ***RF Classifier*** | | | | |
| Unigram | A | 489 | 12 | 41 | 135 | 72.23 | 93.68 | 85.55 |
| | H | 6 | 497 | 27 | 10 | 92.04 | 95.95 | |
| | S | 10 | 7 | 429 | 25 | 91.08 | 84.78 | |
| | N | 17 | 2 | 9 | 367 | 92.91 | 68.34 | |
| | | | | ***KNN Classifier*** | | | | |
| Unigram | A | 478 | 3 | 22 | 90 | 80.61 | 91.57 | 84.16 |
| | H | 11 | 494 | 48 | 57 | 80.98 | 95.37 | |
| | S | 14 | 18 | 425 | 34 | 86.56 | 83.99 | |
| | N | 19 | 3 | 11 | 356 | 91.52 | 66.29 | |
| | | | | ***SL Classifier*** | | | | |
| Unigram | A | 410 | 5 | 19 | 38 | 86.86 | 78.54 | 87.13 |
| | H | 0 | 485 | 4 | 1 | 98.98 | 93.63 | |
| | S | 13 | 8 | 433 | 11 | 93.12 | 85.57 | |
| | N | 99 | 20 | 50 | 487 | 74.24 | 90.69 | |

Figure 7 illustrate a graphical summary of the classification performance of the four classifiers (SMO, RF, KNN, and SL) for lexical features, showing the accuracy percentage for every classifier. Each bar represents the accuracy achieved by every classifier. As we can see from figure 7, which represent the accuracy of lexical features in a bar graph of table 5, the highest overall accuracy achieved was 89.92%, using the SMO classifier. The RF, KNN, and SL classifiers showed comparable classification performance, with a slight difference: RF achieved 85.55%, KNN 84.16%, and SL 87.13%. which is albeit lower comparted to the classification performance of the SMO classifier.

This indicate that ML approach is suitable for the detection of emotion. In addition, it suggests that the lexical features are informative for the emotion detection task. Moreover, these lexical features have outperformed the acoustic features that were previously discussed, this refer how our model leverages Unigram and TF-IDF to capture specific key words that are highly indicative of emotions such as happiness, anger or sadness. These lexical features often include terms associated with emotional expressions, especially our audio files collecting from different domains and every domain has its own subjects and discussions.

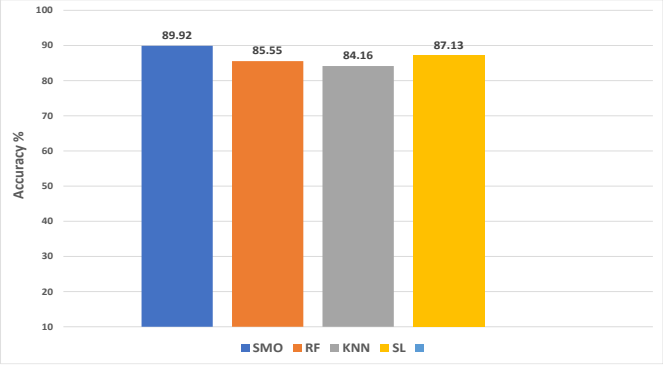


**Figure 7.** Classification performance of the four classifiers based on lexical features (unigram) in terms of accuracy.

 ISSN 2723-6471


In addition, figure 8 illustrate the performance metrics for four classifiers (SMO, RF, KNN, SL) across four emotions (anger, happiness, sadness, neutral) using lexical features. Each bar represents the Precision and Recall achieved by every classifier. As we can see from figure 8, all four classifiers demonstrated strong performance in detecting happiness and sadness emotions, with high precision and recall. This mean the classifiers are able to identify the true instances of these emotions very accurately without misclassifying them. This shows that the classifiers are effectively detect the specific features related to happiness and sadness. The performance for anger was also good, but neutral emotion showed significantly lower for recall across all classifiers, particularly in the KNN and RF classifiers, with the exception of SL classifier, this means these models struggle to detect the instances of neutral emotions accurately, and this refer to the potential complexity associated with neutral lexical features.

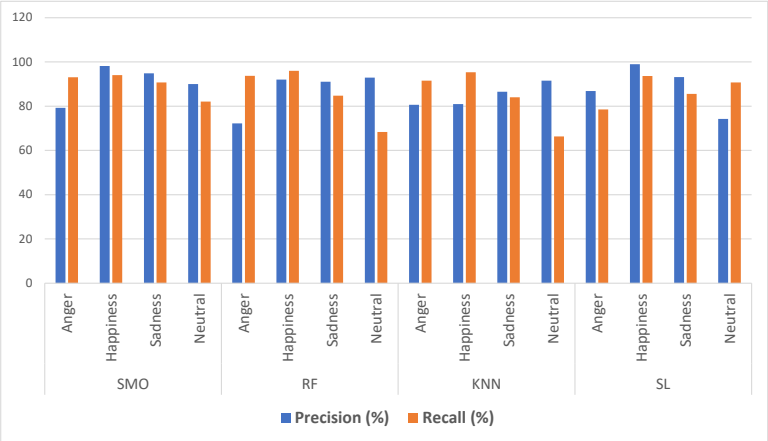


**Figure 8.** Precision and recall of four Classifiers (SMO, RF, KNN, SL) across four Emotions (anger, happiness, sadness, neutral) Using Lexical Features (unigram).

Moreover, we noted that the results of certain classifiers presented high precision and recall and the corresponding accuracy values were less expected, this returns to misclassifications among similar emotions states, such as sadness or neutral, which share common features. These misclassifications have a significant effect on overall accuracy. Where accuracy measures all correct predictions both true positives and true negatives across all classes. For example, when sadness is incorrectly identifies as neutral, it effects not just the precision and recall of these specific classes but the overall accuracy. In the following experiments, we will combine the lexical features results with those from the acoustic features results to help us in judging the different features sets.

### 5.2.2. Experiments with a Multimodal Approach Combining Acoustic and Lexical Features (Unigram)

In these experiments, we try to compare between the various feature combinations, focusing on various acoustic features sets and lexical features. The acoustic features used include MFCC, Mel Spectrogram, Spectral Contrast, Zero-Crossing Rate (ZCR), and Intensity, which were previously tested as individual features and in combinations (e.g., MFCC + Mel Spectrogram or Spectral Contrast+ Intensity). We noted from our previous experiments that acoustic features such as MFCC, Mel Spectrogram and Spectral Contrast not only enhance the model performance when used individually, but also adding additional benefits when combined with other acoustic features, this returns to the nature of these features. For instance, the high noise resistance of MFCC confirms their reliability in different recording conditions. The Mel Spectrogram feature is useful of identifying patterns in the speech that correspond to emotional expressions, capturing nuances, which are often subtle. In addition, Spectral Contrast has the ability to analyze the strength and variance between peaks and valleys, which can highlight emotional intensity in speech.

In order to perform our multimodal experiments, we form sixteen feature sets from different feature combinations, using both acoustic features and unigram model. The primary purpose of these experiments is to find the effect of adding unigram model to every acoustic features individually or combinations with other acoustic features. Additionally, this analysis tries to find the most effective feature combination. Finally, it helps to investigate the impact of using these features for the first time on natural Arabic audio dataset for emotion detection. We performed these experiments using various ML classifiers including: SMO, RF, KNN and SL.

 ISSN 2723-6471 

Table 6, table 7, table 8 and table 9 show the confusion matrices of the four ML classifiers SMO, RF, KNN and SL for the combinations of various acoustic features set and unigram model. The values in the tables represent the accuracy, precision and recall for every classifier. Values in bold indicate the best results for the classifiers for the combinations of various acoustic features set and unigram model experiments. In table 6, out of a total of 2083 instances, SMO classifier correctly classified 2002 instances, while 81 instances were incorrectly classified using a combination of all acoustic features + unigram.

**Table 6.** Confusion matrix and accuracy details of SMO classifier

| Feature Extracted | Class | Classified as | | | | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|---|---|
| | | A | H | S | N | | | |
| MFCC + Unigram | A | 510 | 13 | 11 | 25 | 94.38 | 91.23 | 97.70 |
| | H | 4 | 503 | 8 | 4 | | 96.92 | 97.10 |
| | S | 1 | 1 | 466 | 21 | | 95.30 | 92.09 |
| | N | 7 | 1 | 21 | 487 | | 94.38 | 90.69 |
| Mel Spectrogram+ Unigram | A | 510 | 18 | 22 | 49 | 93.23 | 85.14 | 97.70 |
| | H | 3 | 494 | 5 | 0 | | 98.41 | 95.37 |
| | S | 5 | 2 | 461 | 11 | | 96.24 | 91.11 |
| | N | 4 | 4 | 18 | 477 | | 94.83 | 88.83 |
| Spectral Contrast+ Unigram | A | 513 | 16 | 29 | 26 | 93.57 | 87.84 | 98.28 |
| | H | 0 | 492 | 4 | 5 | | 98.20 | 94.98 |
| | S | 4 | 6 | 450 | 12 | | 95.34 | 88.93 |
| | N | 5 | 4 | 23 | 494 | | 93.92 | 91.99 |
| ZCR+ Unigram | A | 490 | 13 | 23 | 74 | 90.64 | 81.67 | 93.87 |
| | H | 1 | 486 | 6 | 1 | | 98.38 | 93.82 |
| | S | 6 | 6 | 458 | 8 | | 95.82 | 90.51 |
| | N | 25 | 13 | 19 | 454 | | 88.85 | 84.54 |
| Intensity+ Unigram | A | 488 | 5 | 12 | 38 | 92.85 | 89.87 | 93.49 |
| | H | 1 | 491 | 6 | 4 | | 97.81 | 94.79 |
| | S | 10 | 14 | 476 | 16 | | 92.25 | 94.07 |
| | N | 23 | 8 | 12 | 479 | | 91.76 | 89.20 |
| MFCC + Mel spectrogram+ Unigram | A | 512 | 10 | 6 | 15 | 95.06 | 94.29 | 98.08 |
| | H | 4 | 502 | 8 | 5 | | 96.72 | 96.91 |
| | S | 1 | 3 | 467 | 18 | | 95.50 | 92.29 |
| | N | 5 | 3 | 25 | 499 | | 93.80 | 92.92 |
| MFCC + Spectral Contrast+ Unigram | A | 514 | 10 | 8 | 13 | 95.39 | 94.31 | 98.47 |
| | H | 2 | 504 | 10 | 5 | | 96.74 | 97.30 |
| | S | 3 | 2 | 467 | 17 | | 95.50 | 92.29 |
| | N | 3 | 2 | 21 | 502 | | 95.08 | 93.48 |
| MFCC + ZCR+ Unigram | A | 514 | 8 | 7 | 21 | 94.96 | 93.45 | 98.47 |
| | H | 2 | 505 | 8 | 5 | | 97.12 | 97.49 |
| | S | 1 | 3 | 471 | 23 | | 94.58 | 93.08 |
| | N | 5 | 2 | 20 | 488 | | 94.76 | 90.88 |
| MFCC + Intensity + Unigram | A | 508 | 4 | 5 | 8 | 95.39 | 96.76 | 97.32 |
| | H | 4 | 507 | 9 | 6 | | 96.39 | 97.88 |
| | S | 4 | 4 | 472 | 23 | | 93.84 | 93.28 |
| | N | 6 | 3 | 20 | 500 | | 94.52 | 93.11 |
| Mel spectrogram + Spectral Contrast + Unigram | A | 514 | 12 | 18 | 24 | 94.86 | 90.49 | 98.47 |
| | H | 2 | 501 | 7 | 2 | | 97.85 | 96.72 |
| | S | 3 | 3 | 460 | 10 | | 96.64 | 90.91 |
| | N | 3 | 2 | 21 | 501 | | 95.07 | 93.30 |
| Mel spectrogram + ZCR+ Unigram | A | 514 | 14 | 20 | 37 | 94.05 | 87.86 | 98.47 |
| | H | 2 | 497 | 5 | 1 | | 98.42 | 95.95 |
| | S | 4 | 4 | 459 | 10 | | 96.23 | 90.71 |
| | N | 2 | 3 | 22 | 489 | | 94.77 | 91.06 |
| Mel spectrogram + Intensity+ Unigram | A | 514 | 10 | 8 | 17 | 95.10 | 93.62 | 98.47 |
| | H | 3 | 498 | 8 | 4 | | 97.08 | 96.14 |
| | S | 3 | 4 | 471 | 18 | | 94.96 | 93.08 |
| | N | 2 | 6 | 19 | 498 | | 94.86 | 92.74 |
| Spectral Contrast + ZCR+ Unigram | A | 515 | 14 | 24 | 24 | 93.76 | 89.25 | 98.66 |
| | H | 0 | 494 | 6 | 8 | | 97.24 | 95.37 |
| | S | 2 | 6 | 451 | 12 | | 95.75 | 89.13 |
| | N | 5 | 4 | 25 | 493 | | 93.55 | 91.81 |
| Spectral Contrast + Intensity+ Unigram | A | 510 | 4 | 9 | 14 | 95.15 | 94.97 | 97.70 |
| | H | 2 | 502 | 6 | 11 | | 96.35 | 96.91 |
| | S | 8 | 9 | 473 | 15 | | 93.66 | 93.48 |

 ISSN 2723-6471


| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | N | 2 | 3 | 18 | 497 | | 95.58 | 92.55 |
| ZCR + Intensity + Unigram | A | 500 | 4 | 10 | 34 | 93.33 | 91.24 | 95.79 |
| | H | 1 | 489 | 6 | 3 | | 98.00 | 94.40 |
| | S | 6 | 13 | 472 | 17 | | 92.91 | 93.28 |
| | N | 15 | 12 | 18 | 483 | | 91.48 | 89.94 |
| All acoustic features + Unigram | A | 516 | 3 | 6 | 4 | **96.11** | 97.54 | 98.85 |
| | H | 2 | 510 | 9 | 5 | | 96.96 | 98.46 |
| | S | 2 | 3 | 472 | 24 | | 94.21 | 93.28 |
| | N | 2 | 2 | 19 | 504 | | 95.64 | 93.85 |

In table 7, out of a total of 2083 instances, RF classifier correctly classified 1908 instances, while 175 instances were incorrectly classified using combinations of spectral contrast + intensity + unigram.

**Table 7.** Confusion matrix and accuracy details of RF classifier

| Feature Extracted | Class | Classified as | | | | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|---|---|
| | | A | H | S | N | | | |
| MFCC+ Unigram | A | 501 | 9 | 16 | 26 | 91.17 | 90.76 | 95.98 |
| | H | 6 | 495 | 14 | 7 | | 94.83 | 95.56 |
| | S | 3 | 8 | 445 | 46 | | 88.65 | 87.94 |
| | N | 12 | 6 | 31 | 458 | | 90.34 | 85.29 |
| Mel spectrogram+ Unigram | A | 488 | 34 | 29 | 28 | 87.37 | 84.28 | 93.49 |
| | H | 18 | 458 | 17 | 2 | | 92.53 | 88.42 |
| | S | 11 | 18 | 409 | 42 | | 85.21 | 80.83 |
| | N | 5 | 8 | 51 | 465 | | 87.90 | 86.59 |
| Spectral Contrast+ Unigram | A | 497 | 16 | 37 | 31 | 90.45 | 85.54 | 95.21 |
| | H | 9 | 496 | 26 | 7 | | 92.19 | 95.75 |
| | S | 10 | 3 | 416 | 24 | | 91.83 | 82.21 |
| | N | 6 | 3 | 27 | 475 | | 92.95 | 88.45 |
| ZCR+ Unigram | A | 486 | 11 | 43 | 72 | 88.00 | 79.41 | 93.10 |
| | H | 4 | 494 | 24 | 13 | | 92.34 | 95.37 |
| | S | 11 | 8 | 420 | 19 | | 91.70 | 83 |
| | N | 21 | 5 | 19 | 433 | | 90.59 | 80.63 |
| Intensity+ Unigram | A | 505 | 8 | 27 | 84 | 88.14 | 80.93 | 96.74 |
| | H | 3 | 490 | 33 | 10 | | 91.42 | 94.59 |
| | S | 10 | 15 | 429 | 31 | | 88.45 | 84.78 |
| | N | 4 | 5 | 17 | 412 | | 94.06 | 76.72 |
| MFCC + Mel spectrogram+ Unigram | A | 492 | 16 | 16 | 19 | 88.24 | 90.61 | 94.25 |
| | H | 16 | 477 | 20 | 7 | | 91.73 | 92.08 |
| | S | 8 | 18 | 406 | 48 | | 84.58 | 80.24 |
| | N | 6 | 7 | 64 | 463 | | 85.74 | 86.22 |
| MFCC + Spectral Contrast+ Unigram | A | 505 | 13 | 14 | 11 | 91.02 | 93 | 96.74 |
| | H | 7 | 495 | 22 | 8 | | 93.05 | 95.56 |
| | S | 4 | 5 | 417 | 39 | | 89.68 | 82.41 |
| | N | 6 | 5 | 53 | 479 | | 88.21 | 89.20 |
| MFCC + ZCR+ Unigram | A | 495 | 5 | 10 | 25 | 90.97 | 92.52 | 94.83 |
| | H | 4 | 499 | 27 | 7 | | 92.92 | 96.33 |
| | S | 4 | 10 | 429 | 33 | | 90.13 | 84.78 |
| | N | 19 | 4 | 40 | 472 | | 88.22 | 87.90 |
| MFCC + Intensity+ Unigram | A | 504 | 9 | 10 | 24 | 91.36 | 92.14 | 96.55 |
| | H | 7 | 496 | 18 | 8 | | 93.76 | 95.75 |
| | S | 7 | 8 | 443 | 45 | | 88.07 | 87.55 |
| | N | 4 | 5 | 35 | 460 | | 91.27 | 85.66 |
| Mel spectrogram + Spectral Contrast+ Unigram | A | 483 | 30 | 22 | 20 | 88.14 | 87.03 | 92.53 |
| | H | 20 | 464 | 20 | 2 | | 91.70 | 89.58 |
| | S | 11 | 18 | 413 | 39 | | 85.86 | 81.62 |
| | N | 8 | 6 | 51 | 476 | | 87.99 | 88.64 |
| Mel spectrogram + ZCR+ Unigram | A | 482 | 27 | 29 | 23 | 87.18 | 85.92 | 92.34 |
| | H | 20 | 467 | 14 | 3 | | 92.66 | 90.15 |
| | S | 14 | 20 | 401 | 45 | | 83.54 | 79.25 |
| | N | 6 | 4 | 62 | 466 | | 86.62 | 86.78 |
| Mel spectrogram + Intensity+ Unigram | A | 500 | 14 | 19 | 16 | 88.86 | 91.07 | 95.79 |
| | H | 11 | 478 | 23 | 12 | | 91.22 | 92.28 |
| | S | 6 | 17 | 406 | 42 | | 86.20 | 80.24 |
| | N | 5 | 9 | 58 | 467 | | 86.64 | 86.96 |
| Spectral Contrast + ZCR+ Unigram | A | 506 | 16 | 30 | 20 | 90.73 | 88.46 | 96.93 |
| | H | 5 | 494 | 28 | 15 | | 91.14 | 95.37 |

 ISSN 2723-6471


| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | S | 5 | 5 | 409 | 21 | | 92.95 | 80.83 |
| | N | 6 | 3 | 39 | 481 | | 90.93 | 89.57 |
| Spectral Contrast + Intensity+ Unigram | A | 510 | 7 | 26 | 28 | **91.60** | 89.32 | 97.70 |
| | H | 9 | 500 | 29 | 16 | | 90.25 | 96.53 |
| | S | 2 | 8 | 424 | 19 | | 93.60 | 83.79 |
| | N | 1 | 3 | 27 | 474 | | 93.86 | 88.27 |
| ZCR + Intensity + Unigram | A | 500 | 7 | 20 | 70 | 88.86 | 83.75 | 95.79 |
| | H | 2 | 494 | 36 | 14 | | 90.48 | 95.37 |
| | S | 9 | 13 | 426 | 22 | | 90.64 | 84.19 |
| | N | 11 | 4 | 24 | 431 | | 91.70 | 80.26 |
| All acoustic features + Unigram | A | 509 | 13 | 16 | 4 | 90.16 | 93.91 | 97.51 |
| | H | 8 | 479 | 21 | 8 | | 92.83 | 92.47 |
| | S | 3 | 23 | 406 | 41 | | 85.84 | 80.24 |
| | N | 2 | 3 | 63 | 484 | | 87.68 | 90.13 |

In table 8, out of a total of 2083 instances, KNN classifier correctly classified 1888 instances, while 195 instances were incorrectly classified using combinations of MFCC + Spectral Contrast+ unigram.

**Table 8.** Confusion matrix and accuracy details of KNN classifier

| Feature Extracted | Class | Classified as | | | | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|---|---|
| | | A | H | S | N | | | |
| MFCC+ Unigram | A | 489 | 5 | 12 | 42 | 89.82 | 89.23 | 93.68 |
| | H | 14 | 499 | 37 | 25 | | 86.78 | 96.33 |
| | S | 9 | 13 | 439 | 26 | | 90.14 | 86.76 |
| | N | 10 | 1 | 18 | 444 | | 93.87 | 82.68 |
| Mel spectrogram+ Unigram | A | 488 | 9 | 11 | 39 | 86.32 | 89.21 | 93.49 |
| | H | 22 | 469 | 28 | 5 | | 89.50 | 90.54 |
| | S | 8 | 33 | 394 | 46 | | 81.91 | 77.87 |
| | N | 4 | 7 | 73 | 447 | | 84.18 | 83.24 |
| Spectral Contrast+ Unigram | A | 484 | 9 | 13 | 21 | 89.01 | 91.84 | 92.72 |
| | H | 17 | 499 | 45 | 27 | | 84.86 | 96.33 |
| | S | 11 | 8 | 411 | 29 | | 89.54 | 81.23 |
| | N | 10 | 2 | 37 | 460 | | 90.37 | 85.66 |
| ZCR+ Unigram | A | 472 | 2 | 15 | 82 | 84.40 | 82.66 | 90.42 |
| | H | 14 | 494 | 48 | 60 | | 80.19 | 95.37 |
| | S | 18 | 18 | 433 | 36 | | 85.74 | 85.57 |
| | N | 18 | 4 | 10 | 359 | | 91.82 | 66.85 |
| Intensity+ Unigram | A | 478 | 2 | 10 | 80 | 84.93 | 83.86 | 91.57 |
| | H | 8 | 497 | 50 | 60 | | 80.81 | 95.95 |
| | S | 15 | 17 | 431 | 34 | | 86.72 | 85.18 |
| | N | 21 | 2 | 15 | 363 | | 90.52 | 67.60 |
| MFCC + Mel spectrogram+ Unigram | A | 497 | 10 | 7 | 21 | 87.90 | 92.90 | 95.21 |
| | H | 19 | 473 | 27 | 8 | | 89.75 | 91.31 |
| | S | 4 | 31 | 401 | 48 | | 82.85 | 79.25 |
| | N | 2 | 4 | 71 | 460 | | 85.66 | 85.66 |
| MFCC + Spectral Contrast+ Unigram | A | 499 | 10 | 12 | 14 | **90.64** | 93.27 | 95.59 |
| | H | 12 | 498 | 41 | 17 | | 87.68 | 96.14 |
| | S | 8 | 8 | 414 | 29 | | 90.20 | 81.82 |
| | N | 3 | 2 | 39 | 477 | | 91.55 | 88.83 |
| MFCC + ZCR+ Unigram | A | 489 | 5 | 12 | 42 | 89.82 | 89.23 | 93.68 |
| | H | 14 | 499 | 37 | 25 | | 86.78 | 96.33 |
| | S | 9 | 13 | 439 | 26 | | 90.14 | 86.76 |
| | N | 10 | 1 | 18 | 444 | | 93.87 | 82.68 |
| MFCC + Intensity+ Unigram | A | 493 | 5 | 10 | 39 | 90.11 | 90.13 | 94.44 |
| | H | 12 | 499 | 38 | 27 | | 86.63 | 96.33 |
| | S | 9 | 13 | 441 | 27 | | 90.00 | 87.15 |
| | N | 8 | 1 | 17 | 444 | | 94.47 | 82.68 |
| Mel spectrogram + Spectral Contrast+ Unigram | A | 482 | 18 | 11 | 7 | 87.13 | 93.05 | 92.34 |
| | H | 29 | 463 | 28 | 8 | | 87.69 | 89.38 |
| | S | 7 | 30 | 394 | 46 | | 82.60 | 77.87 |
| | N | 4 | 7 | 73 | 476 | | 85 | 88.64 |
| Mel spectrogram + ZCR+ Unigram | A | 488 | 9 | 10 | 39 | 86.37 | 89.38 | 93.49 |
| | H | 22 | 469 | 28 | 5 | | 89.50 | 90.54 |
| | S | 8 | 33 | 395 | 46 | | 81.95 | 78.06 |
| | N | 4 | 7 | 73 | 447 | | 84.18 | 83.24 |
| Mel spectrogram + Intensity+ Lexical | A | 492 | 7 | 8 | 33 | 86.75 | 91.11 | 94.25 |
| | H | 21 | 470 | 28 | 7 | | 89.35 | 90.73 |

 ISSN 2723-6471


| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | S | 6 | 34 | 396 | 48 | | 81.82 | 78.26 |
| | N | 3 | 7 | 74 | 449 | | 84.24 | 83.61 |
| Spectral Contrast + ZCR+ Unigram | A | 484 | 9 | 13 | 21 | 89.01 | 91.84 | 92.72 |
| | H | 17 | 499 | 45 | 27 | | 84.86 | 96.33 |
| | S | 11 | 8 | 411 | 29 | | 89.54 | 81.23 |
| | N | 10 | 2 | 37 | 460 | | 90.37 | 85.66 |
| Spectral Contrast + Intensity+ Unigram | A | 484 | 8 | 8 | 19 | 89.34 | 93.26 | 92.72 |
| | H | 19 | 500 | 47 | 27 | | 84.32 | 96.53 |
| | S | 11 | 8 | 414 | 28 | | 89.80 | 81.82 |
| | N | 8 | 2 | 37 | 463 | | 90.78 | 86.22 |
| ZCR + Intensity + Unigram | A | 483 | 2 | 11 | 78 | 85.31 | 84.15 | 92.53 |
| | H | 8 | 496 | 51 | 61 | | 80.52 | 95.75 |
| | S | 13 | 17 | 434 | 34 | | 87.15 | 85.77 |
| | N | 18 | 3 | 10 | 364 | | 92.15 | 67.78 |
| All acoustic features + Unigram | A | 490 | 17 | 5 | 6 | 87.76 | 94.59 | 93.87 |
| | H | 26 | 461 | 29 | 9 | | 87.81 | 89 |
| | S | 4 | 33 | 401 | 46 | | 82.85 | 79.25 |
| | N | 2 | 7 | 71 | 476 | | 85.61 | 88.64 |

In table 9, out of a total of 2083 instances, SL classifier correctly classified 1966 instances, while 117 instances were incorrectly classified using combinations of all acoustic features + unigram.

**Table 9.** Confusion matrix and accuracy details of SL classifier

| Feature Extracted | Class | Classified as | | | | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|---|---|
| | | A | H | S | N | | | |
| MFCC+ Unigram | A | 503 | 16 | 11 | 25 | 91.98 | 90.63 | 96.36 |
| | H | 2 | 492 | 10 | 2 | | 97.23 | 94.98 |
| | S | 6 | 6 | 428 | 17 | | 93.65 | 84.58 |
| | N | 11 | 4 | 57 | 493 | | 87.26 | 91.81 |
| Mel spectrogram+ Unigram | A | 496 | 25 | 27 | 31 | 90.97 | 85.66 | 95.02 |
| | H | 4 | 480 | 7 | 0 | | 97.76 | 92.66 |
| | S | 5 | 8 | 430 | 17 | | 93.48 | 84.98 |
| | N | 17 | 5 | 42 | 489 | | 88.43 | 91.06 |
| Spectral Contrast+ Unigram | A | 505 | 23 | 54 | 19 | 91.50 | 84.03 | 96.74 |
| | H | 1 | 485 | 4 | 2 | | 98.58 | 93.63 |
| | S | 5 | 5 | 407 | 7 | | 95.99 | 80.43 |
| | N | 11 | 5 | 41 | 509 | | 89.93 | 94.79 |
| ZCR+ Unigram | A | 473 | 13 | 37 | 61 | 88.09 | 80.99 | 90.61 |
| | H | 0 | 476 | 2 | 1 | | 99.37 | 91.89 |
| | S | 7 | 10 | 420 | 9 | | 94.17 | 83 |
| | N | 42 | 19 | 47 | 466 | | 81.18 | 86.78 |
| Intensity+ Unigram | A | 455 | 2 | 16 | 38 | 89.01 | 89.04 | 87.16 |
| | H | 0 | 485 | 4 | 1 | | 98.98 | 93.63 |
| | S | 4 | 6 | 424 | 8 | | 95.93 | 83.79 |
| | N | 63 | 25 | 62 | 490 | | 76.56 | 91.25 |
| MFCC + Mel spectrogram+ Unigram | A | 509 | 15 | 6 | 12 | 92.80 | 93.91 | 97.51 |
| | H | 6 | 486 | 10 | 2 | | 96.43 | 93.82 |
| | S | 1 | 9 | 435 | 20 | | 93.55 | 85.97 |
| | N | 6 | 8 | 55 | 503 | | 87.94 | 93.67 |
| MFCC + Spectral Contrast+ Unigram | A | 505 | 15 | 17 | 11 | 93.47 | 92.15 | 96.74 |
| | H | 3 | 490 | 7 | 2 | | 97.61 | 94.59 |
| | S | 5 | 9 | 442 | 14 | | 94.04 | 87.35 |
| | N | 9 | 4 | 40 | 510 | | 90.59 | 94.97 |
| MFCC + ZCR+ Unigram | A | 513 | 9 | 10 | 9 | 92.94 | 94.82 | 98.28 |
| | H | 0 | 494 | 12 | 3 | | 97.05 | 95.37 |
| | S | 0 | 6 | 424 | 20 | | 94.22 | 83.79 |
| | N | 9 | 9 | 60 | 505 | | 86.62 | 94.04 |
| MFCC + Intensity+ Unigram | A | 510 | 11 | 15 | 8 | 92.75 | 93.75 | 97.70 |
| | H | 2 | 494 | 12 | 3 | | 96.67 | 95.37 |
| | S | 2 | 7 | 420 | 18 | | 93.96 | 83 |
| | N | 8 | 6 | 59 | 508 | | 87.44 | 94.60 |
| Mel spectrogram + Spectral Contrast+ Unigram | A | 506 | 24 | 21 | 17 | 92.61 | 89.08 | 96.93 |
| | H | 4 | 484 | 6 | 0 | | 97.98 | 93.44 |
| | S | 5 | 7 | 432 | 13 | | 94.53 | 85.38 |
| | N | 7 | 3 | 47 | 507 | | 89.89 | 94.41 |
| Mel spectrogram + ZCR+ Unigram | A | 508 | 26 | 26 | 21 | 91.31 | 87.44 | 97.32 |
| | H | 4 | 477 | 9 | 1 | | 97.15 | 92.08 |
| | S | 1 | 8 | 419 | 17 | | 94.16 | 82.81 |

 ISSN 2723-6471


| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | N | 9 | 7 | 52 | 498 | | 87.99 | 92.74 |
| Mel spectrogram + Intensity+ Unigram | A | 505 | 14 | 18 | 5 | 92.17 | 93.17 | 96.74 |
| | H | 2 | 481 | 9 | 1 | | 97.57 | 92.86 |
| | S | 6 | 12 | 422 | 19 | | 91.94 | 83.40 |
| | N | 9 | 11 | 57 | 512 | | 86.93 | 95.34 |
| Spectral Contrast + ZCR+ Unigram | A | 510 | 17 | 38 | 13 | 92.17 | 88.24 | 97.70 |
| | H | 0 | 479 | 6 | 1 | | 98.56 | 92.47 |
| | S | 2 | 6 | 414 | 6 | | 96.73 | 81.82 |
| | N | 10 | 16 | 48 | 517 | | 87.48 | 96.28 |
| Spectral Contrast + Intensity+ Unigram | A | 507 | 7 | 21 | 11 | 93.57 | 92.86 | 97.13 |
| | H | 2 | 489 | 5 | 3 | | 98 | 94.40 |
| | S | 5 | 13 | 438 | 8 | | 94.40 | 86.56 |
| | N | 8 | 9 | 42 | 515 | | 89.72 | 95.90 |
| ZCR + Intensity+ Unigram | A | 483 | 4 | 20 | 41 | 89.44 | 88.14 | 92.53 |
| | H | 0 | 479 | 3 | 1 | | 99.17 | 92.47 |
| | S | 0 | 7 | 415 | 9 | | 96.29 | 82.02 |
| | N | 39 | 28 | 68 | 486 | | 78.26 | 90.50 |
| All acoustic features + Unigram | A | 518 | 5 | 6 | 2 | **94.38** | 97.55 | 99.23 |
| | H | 1 | 493 | 8 | 2 | | 97.82 | 95.17 |
| | S | 1 | 9 | 441 | 19 | | 93.83 | 87.15 |
| | N | 2 | 11 | 51 | 514 | | 88.93 | 95.72 |

Figure 9 illustrate a graphical summary of the classification performance of the four classifiers (SMO, RF, KNN, and SL) across sixteen various experiments for every classifier in terms of accuracy percentage. Each bar represents the accuracy achieved by every classier. As we can see from figure 9, which represents the accuracy of combinations of various acoustic features set and unigram model in a bar graph of table 6, table 7, table 8, and table 9. The highest overall accuracy achieved was 96.11% using SMO classifier, based on combinations of all acoustic features + unigram. The SL classifier yielded 94.38%, based on combinations of all acoustic features + unigram. The RF classifier yielded 91.60%, based on combinations of Spectral Contrast +Intensity + unigram. The KNN classifier yielded 90.64%, based on combinations of MFCC+ Spectral Contrast+ unigram.

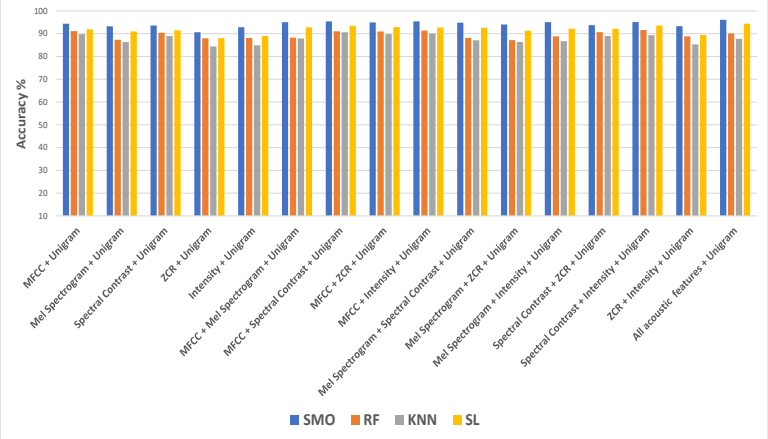


**Figure 9.** Accuracy of four classifiers using various acoustic features combinations and lexical features (unigram).

In addition, figure 10 illustrate the performance metrics for four classifiers (SMO, RF, KNN, SL) across four emotions (anger, happiness, sadness, neutral) using various combinations of acoustic features combined with the unigram model. Each bar represents the Precision and Recall achieved by every classifier. The discussion focused on the classifiers and feature combinations that achieved the highest accuracy. This consistency over different classifiers confirms the robustness detection both of anger, happiness and neutral emotions 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 and this return to the subtle characteristics of sadness emotion, also sadness often shares common features with neutral emotion, such as lower energy. This overlap can cause classifiers to misidentify sadness as neutral.

Furthermore, we noticed that SMO and SL classifiers achieved the highest accuracy 96.11% and 94.38%, respectively, when using all acoustic features + unigram, this indicate that these two classifiers benefit from a comprehensive feature set to effectively handle the complexity of combining various acoustic features with lexical features to identify

 

emotions. The RF and KNN classifiers achieved the highest accuracy 91.60% and 90.64%, respectively, with Spectral Contrast +Intensity + unigram for RF classifier, and MFCC+ Spectral Contrast+ unigram for KNN classifier, this indicate that RF and KNN achieved their best results with specific feature combinations. Additionally, we noticed that combination of spectral contrast and lexical features in The RF and KNN classifiers, yielding the best results because these features complement each other in capturing distinct aspects of emotional expression, where spectral contrast effectively analyses the strength of peaks and valleys in the spectrum and the variance between them, while lexical features provide Semantic information from transcriptions of the audio files.

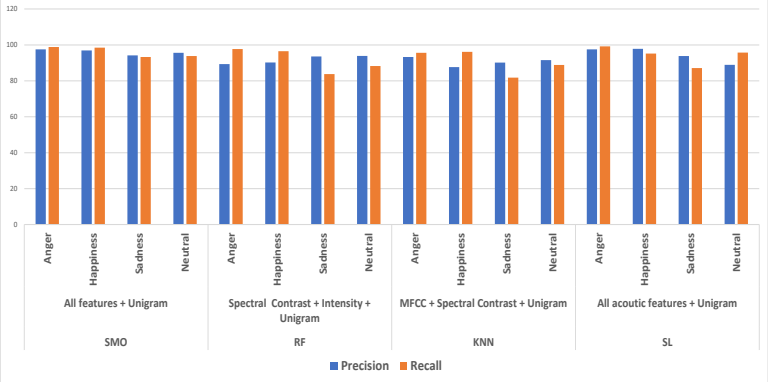


**Figure 10.** Precision and recall of Four Classifiers (SMO, RF, KNN, SL) Across Four Emotions Using Acoustic Features Combined with lexical features (unigram).

In all cases, combining the lexical information with various acoustic features sets has improved the performance. This improvement might be coming from the valuable information that has been added by lexical features to the classifiers. This also suggests that these combined features are coherent. The highest obtained results indicate that the ML approach is suitable for emotion detection in natural Arabic audio dataset. In addition, this suggest that combining lexical features with various sets of acoustic features are powerful and informative in emotion identification task.

## 6. Comparison with Previous Works

By comparing our results with those of the previous studies, we discovered that our multimodal approach using both acoustic and lexical features outperformed other methods by achieving an accuracy results 96.11% using SMO classifier, when combining all acoustic features with lexical features. The comparison is made with a study based only on natural Arabic audio files, while the remaining studies related to acted, semi-natural or textual datasets. Table 10 shows the result comparison for our multimodal approach and the previous study.

**Table 10.** Comparison results between the prior study and the proposed study

| Study | Approach | Features Extracted | Emotions | Classification classifiers | Bags (limitations) | Results |
|---|---|---|---|---|---|---|
| Klaylat et al. [6] | Audio only | 845 acoustic features extracted | Anger, happiness, Surprise (3) | 35 classification algorithms used | Imbalanced audio dataset can bias classifiers towards more frequently occurring emotions. Limited emotion range, only three emotions used. Using 1 second segments not completely enough to detect some other emotions. Limited dialectal coverage, specifically focusing on Egyptian, Gulf, and Lebanese dialects. 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. | 95.52 % |
| Our study | Multimodal (Audio + Text) | 26 acoustic features and 3919 lexical features | Anger, happiness, sadness, neutral (4) | SMO, RF, KNN, SL | -- | 96.11% |

For these limitations of the previous study in the table, such as Limited emotions range, imbalanced dataset, using only 1 second segments of audio files, limited dialectal coverage, our approach addresses these issues by using more diverse audio files and emotional categories. In addition, the segments of audio files range from 1-9 seconds this allow to capture a broader range of emotional states in audio files. Moreover, our audio dataset is balanced, ensuring that every emotion represented in a suitable way.

## 7. Conclusion and Future Work

In this paper, we have introduced a multimodal approach that combines various acoustic and lexical features for emotion detection in natural Arabic audio files. To the best of our knowledge, this is the first study of its type in Arabic that used natural audio files and based on acoustic and lexical features. Our method consists of three modules: Acoustic Features Module, which was completed in our previous work. Lexical Features Module includes transcribing, text preprocessing, features extraction, supervised learning classification, and performance evaluation. Finally Combined Features Module includes combining both acoustic and lexical features, classification using supervised learning and performance evaluation based on the combined feature set. The dataset was used for the experiments implemented in this work was collected 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). The dataset consists of audio files spoken in Modern Standard Arabic (MSA), different DA, or a mix of both. We have conducted multiple experiments to investigate the best 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. Various acoustic features sets and lexical features were used in the supervised machine learning approach. For evaluation purposes, three common effective measures were used Accuracy, precision and recall.

The experiments gave promising results. The best results for combination of all acoustic features sets + lexical features achieved using the SMO and SL classifiers with the overall accuracies equal to 96.11% and 94.38%, respectively, which these results are quite high especially regarding natural Arabic audio files. The highest obtained results indicate that the machine learning approach is suitable for the detection of emotion in Arabic language. In addition, this suggests that various acoustic features set combinations with lexical features powerful and informative in the emotion detection task. This also indicate that the dataset of natural Arabic audio files well annotated.

Furthermore, we found that SMO classifier outperformed other classifiers RF, KNN and SL in three types of experiments: acoustic features experiments, Ngram experiments, and the combinations of acoustic features and Ngram features, this return to that SMO classifier robust to overfitting by keeping a balance between model complexity and performance. In addition, SMO known for its scalability, when dealing with diverse data types, such as those combining acoustic and lexical features.

Additionally, this research addresses the challenges associated with the Arabic language dialectal diversity, which significantly impacts emotion detection accuracy. The distinct nuances of regional dialects need robust feature extraction and selection methods capable of capturing the varied prosodic characteristics essential to each dialect. Our multimodal approach that combines acoustic and lexical features enhances model adaptability to these dialectal variations, ensuring more reliable emotion detection across different Arabic speaking regions. In general, the integration of dialect and prosody considerations significantly improves our model effectiveness, establishing it as a foundational effort in the field of Arabic speech emotion detection. This not only enhances the models to be more practical but also leads to further advancements in emotion detection systems for Arabic language, which is considered complex.

For the future work, we intend to do data augmentation techniques to improve the diversity and size of our audio dataset. These techniques will include addition background noise, speed variations as examples. Also, we plan to increase the size of our dataset by collecting more natural audio files, where our dataset 2083 audio files, is already considered large compared to natural Arabic audio datasets. Moreover, we plan to compare this work with others datasets this include acted, semi-natural and elicited audio datasets such as RAVDESS audio dataset, which was mentioned in our related works section. The comparison allows us to evaluate the generalizability of our model across

different data sources. Lastly, we look forward to experiment with other methods to enhance on the recognition accuracy. These methods include using deep learning models such as CNNS.

## 8. Declarations

### 8.1. Author Contributions

Conceptualization: A.K., E.A.E.; Methodology: E.A.E.; Software: A.K.; Validation: A.K. and E.A.E.; Formal Analysis: A.K. and E.A.E.; Investigation: A.K.; Resources: E.A.E.; Data Curation: E.A.E.; Writing Original Draft Preparation: A.K. and E.A.E.; Writing Review and Editing: E.A.E. and A.K.; Visualization: A.K. All authors have read and agreed to the published version of the manuscript.

### 8.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 8.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### 8.4. Institutional Review Board Statement

Not applicable.

### 8.5. Informed Consent Statement

Not applicable.

### 8.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] J. Bhaskar, K. Sruthi, and P. Nedungadi, "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining," *Procedia Computer Science*, vol. 46, no. Dec, pp. 635-643, 2015.

[2] A. Almahdawi and W. J. Teahan, "A New Arabic Dataset for Emotion Recognition," in *Intelligent Computing, Advances in Intelligent Systems and Computing*, vol. 2019, no. Jul, pp. 200-216, July 2019.

[3] A. Kaloub and E. Abed Elgabar, "Speech-Based Techniques for Emotion Detection in Natural Arabic Audio Files," *The International Arab Journal of Information Technology*, vol. 22, no. 1, pp. 139-157, 2025.

[4] R. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic Speech Emotion Recognition from Saudi Dialect Corpus," *IEEE Access*, vol. 9, no. Sept, pp. 127081-127085, 2021.

[5] O. Mohammad and M. Elhadef, "Arabic Speech Emotion Recognition Method Based on LPC and PPSD," in *Proceedings of the 2nd International Conference on Computing, Automation and Knowledge Management*, vol. 2021, no. Jan, pp. 31-36, 2021.

[6] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Emotion Recognition in Arabic Speech," in *Proceedings of the Sensors Networks Smart and Emerging Technologies*, vol. 2017, no. Sept, pp. 1-4, 2017.

[7] N. Ira and M. Rahman, "An Efficient Speech Emotion Recognition Using Ensemble Method of Supervised Classifiers," in *Emerging Technology in Computing, Communication and Electronics (ETCCE)*, vol. 2020, no. Dec, pp. 1-5, 2020.

[8] S. Azmin and K. Dhar, "Emotion detection from Bangla text corpus using Naïve Bayes classifier," in *Proceedings of the 4th International Conference on Electrical Information and Communication Technology*, vol. 2019, no. Dec, pp. 1–5, 2019.

[9] L. Wikarsa and S. Thahir, "A text mining application of emotion classifications of Twitter users using Naïve Bayes method," in *1st International Conference on Wireless and Telematics (ICWT)*, vol. 2015, no. Nov, pp. 1–6, 2015.

[10] F. Aljwari, "Emotion detection in Arabic text using machine learning methods," *International Journal of Information System and Computer Science (IJISCS)*, vol. 6, no. 3, pp. 175-185, 2022.

 ISSN 2723-6471


[11] P. Kurniawati, D. Lestari, and M. Khodra, “Speech emotion recognition from Indonesian spoken language using acoustic and lexical features,” in *Proceedings of the Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*, vol. 2017, no. Nov, pp. 189-195, 2017.

[12] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in Python,” in *Proceedings of the 14th Python in Science Conference*, vol. 2015, no. Jul, pp. 18-25, 2015.

[13] S. Angra and S. Ahuja, "Analysis of Students Data using Rapid Miner," *Journal on Todays Ideas-Tomorrows Technologies*, vol.4, no. 2, pp. 109-117, 2016.

[14] P. Nandwani and R. Verma, “A review on sentiment analysis and emotion detection from text,” *Social Network Analysis and Mining*, vol. 11, no. 81, pp. 1-19, 2021.

[15] L. Larkey, L. Ballesteros, and M. Connell, “Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis,” in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, vol. 2002, no. Aug, pp. 275-282, 2002.

[16] M. Abdullah, I. Makki, M. Almasawa, and M. Alsolmi, “Emotions extraction from Arabic tweets,” *International Journal of Computers and Applications*, vol. 42, no. 7, pp. 1-15, 2018.

[17] N. Abdulla, N. Ahmed, M. Shehab, M. Al-Kabi, M. Al-Ayyoub, and S. Al-Rifai, “Towards improving the lexicon-based approach for Arabic sentiment analysis,” *International Journal of Information Technology and Web Engineering*, vol. 9, no. 3, pp. 55-71, 2014.

[18] B. Salian, O. Narvade, R. Tambewagh, and S. Bharne, “Speech emotion recognition using time distributed CNN and LSTM,” in *Proceedings of the International Conference on Automation, Computing and Communication*, vol. 2021, no. Jan, pp. 1-6, 2021.

[19] L. Jing, H. Huang, and H. Shi, “Improved feature selection approach TFIDF in text mining,” in *Proceedings of the International Conference on Machine Learning and Cybernetics*, vol. 2002, no. Nov, pp. 944-946, 2002.

[20] M. Hao, Y. Tianhao, and Y. Fei, "The SVM based on SMO optimization for Speech Emotion Recognition," *2019 Chinese Control Conference (CCC), Guangzhou, China*, vol. 2019, no. Jul, pp.7884-7888, 2019.

[21] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, Oct. 2001.

[22] S. Subudhiray, H. K. Palo, and N. Das, "K-nearest neighbor based facial emotion recognition using effective features," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 1, pp. 57-65, Mar. 2023.

[23] N. Landwehr, M. Hall, and E. Frank, “Logistic model tree,” *Machine Learning*, vol. 59, no. 1, pp. 161-205, 2005.

[24] J. Han and M. Kamber, Data mining: Concepts and Techniques, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2006.