Ministry of Higher Education and Scientific Research

University of Kasdi Merbah –Ouargla

Faculty of Modern Information and Communication Technology

Computer science and information technology department

Academic Master Thesis

Field: Computer science and information technology

Branch: Computer science

Specialty: Fundamental computing

Theme

A machine learning technique for emotion detection in social media

Supervised by:

Mrs. W. SAADI

Presented by:

HADJAIDJI Oumaima

MELIK Anouar

2021-2022

Dedication

We thank Almighty God for all the blessings that have been given to us granted.

After writing this note, we give it:

To our dear parents:

We owe you who we are today, thanks to your love, your patience, and you're countless sacrifices. We hope this humble work will give you a little compensation and a little gratitude for what you have done incredibly for us. That Almighty God protects you and grants you health and long life.

To our dear brothers and sisters:

No devotion will be expressed deep enough for how we feel for you, we would just like to say, thank you very much, we love you.

To our dear friends:

As a testimony to the sincere friendship we had and the good times we spent together. We dedicate this work to you; we wish you bright future full good promises.

To our dear teachers:

Who accompanied us throughout our academic journey, thank you a lot.

HADJAIDJI Oumaima.

MELIK Anouar

Thanks

First of all, we thank Almighty God for giving us the strength and courage to carry out this modest work.

We would like to express my warm thanks to my promoter Mrs. W. SAADI for her entire disposition, her judicious advice, and for having directed us all the working period.

We sincerely thank Mr. MEZATI Messaoud for his advice and especially for the precious help he gave us throughout the working period.

We also thank each member of the jury for this honor they do us by agreeing to judge our work.

Before closing our acknowledgments page, we do not forget to thank our parents for their patience and unconditional support without which we could not have completed our work.

We end by thanking our families, our colleagues and all the people who have

Contributed directly or indirectly to their encouragement and moral support at the end of this work.

Abstract

Emotions are known to affect human perception as well as their memory, thinking and imagination. We are living a new era that Microblogging and social networking sites like Twitter and Facebook are challenging sources of information that allow people to share their feelings and thoughts as well as their emotions on a daily basis. Emotion detection (ED) is a special case of sentiment analysis wherein the result is more accurate and it is depicted in more expressions such as sadness, enjoyment, anger, disgust, fear, surprise and happiness. Our work is interested on emotions detection from Arabic text in social media using the NLP technique based on machine learning model.

Keywords: Emotion detection, social media, Machine learning, NLP, LSTM.

Résumé

La perception humaine est influencée par les émotions de mama que la mémoire, la pensé et l'imagination. La nouvelle époque est dominée par les sites de microblogging et de réseaux sociaux comme Twitter et Facebook qui sont des sources d'information stimulantes qui permettent aux gens de partager leurs sentiments et leurs pensées ainsi que leurs émotions au quotidien. La détection des émotions (DE) est un cas particulier de l'analyse des sentiments dans lequel le résultat est plus spécifique et il est représenté dans plus d'expressions telles que la tristesse, la joie, la colère, le dégoût, la peur, la surprise et le bonheur. Notre travail s'intéresse à la détection des émotions à partir de texte arabe dans les médias sociaux en utilisant la technique NLP basée sur le modèle d'apprentissage automatique.

Mots clé : Détection des émotions, Réseaux sociaux, Apprentissage automatique, NLP, LSTM.

الملخص:

من المعروف أن العواطف تؤثر على الإدراك البشري وكذلك على ذاكرتهم وتفكيرهم وخيالهم. نحن نعيش حقبة جديدة حيث تمثل المدونات الصغيرة ومواقع الشبكات الاجتماعية مثل Twitter و Facebook مصادر صعبة للمعلومات التي تسمح للناس بمشاركة مشاعرهم وأفكارهم وعواطفهم بشكل يومي. كشف العاطفة (ED) هو حالة خاصة لتحليل المشاعر حيث تكون النتيجة أكثر دقة ويتم تصويرها بمزيد من التعبيرات مثل الحزن والمتعة والغضب والاشمئزاز والخوف والمفاجأة والسعادة. يهتم عملنا باكتشاف المشاعر من النص العربي في وسائل التواصل الاجتماعي باستخدام تقنية البرمجة اللغوية الطبيعية القائمة على نموذج التعلم الآلي.

الكلمات المفتاحية : اكتشاف المشاعر ، وسائل التواصل الاجتماعي ، التعلم الآلي ، البرمجة اللغوية الطبيعية ، LSTM.

Table Of Contant

List of Tables :

List of Figures :

General Introduction

Social media is a popular means of communication characterized by the ability to reach large audiences in a short time frame. It facilitates interactions between users and create many forms of expression to them to express their feelings and opinions in a variety of ways that can be seen such as facial expressions, gestures, speech, and written text.

Social media platforms like Facebook and Twitter are becoming one of the most popular tools for people to express their thoughts, emotions, reviews and feedback. They are considered as an important information source for many Natural Language Processing (NLP) applications.

The huge data created in social media platforms (social Big Data) can be useful in many different and important aspects in high and accurate professional domains in the manner of researches.

Emotion detection (ED) **[1]** or emotion recognition is a branch of sentiment analysis that deals with the extraction and analysis of emotions. An emotion can be defined as psychological states differently connected with contemplations or as sentiments that result in physical changes reflect one's thoughts and conduct during given state.

Emotion detection from text **[2]** is an attractive research subject, it identifies the emotional tone behind a set of words or a body of text. It is an approach to natural language processing, which involves the use of data analyzing, machine learning (ML) and artificial intelligence (AI) to analyze texts. This field importance is present in various domains and applications.

Some strategies and linguistic pipelines were developed for analyzing English Tweets to automatically extract sentiments and emotions from there text, but Arabic text analysis is still an active research area **[3,4,5]** this is due to the particularity and the complexity of the Arabic language which is classified into two categories: dialectical Arabic and standard Arabic. That's why it is hard to build a system to detect emotions in this language.

The main problematic turn around how to recognize emotion from dialectical Arabic and specially from Algerian dialect, since there is no dedicated Dataset. To overcome this problem, we are implementing a system for detecting emotions from dialectical Arabic Tweets using NLP and Machine learning techniques and specifically LSTM techniques, where, research in this topic is considered few due to the challenges of lack of annotated resources and more complex morphology relative to other languages.

The different concepts related to the context of our work are represented via this dissertation in the following order:

Chapter 1 is concerned with the emotion related concepts precisely emotion models. In the second chapter we will present the most used machine learning techniques related to the concerned field of work. The third chapter illustrate the emotion detection domain with presenting a set of related works. In the last chapter we will present the conception and implementation of our system which is based on a supervised machine learning technique for emotion detection from Twitter specially the Algerian dialectical Tweets. To conclude, we will present our conclusion a perspective.

Chapter 1
Emotion Detection

1. Introduction

Emotions are a very important aspect of our mental life to the quality and meaning of our existence, they are what make life worth living. Most of the great classical philosophers had recognizable theories of emotions which were typically conceived as a subject's phenomenologically salient responses to significant events and as capable of triggering distinctive bodily and behaviours changes. These changes can be detected at faces expression, voices tones, or textual expression on social media. Affective computing is the branch of computing science that deals with the human emotions; one of the most active research fields in this domain is the emotions detection which is generally based on classification techniques.

In this section, we will mention the definition of emotion then the classification of its models and finely the application domain of emotion detection.

2. What is emotion?

Emotions are intense feelings that are directed at someone or something.**[6]** On the other hand, emotion can be used to refer to states that are mild (as in annoyed or content) and to states that are not directed at anything (as in anxiety and depression).

The Lexical definition of emotion is: "*A strong feeling deriving from one's circumstances, mood, or relationships with others.*"**[7]** Emotions are responses to significant internal and external events.

In psychology, emotion can be defined as psychological states differently connected with contemplations. It can also be defined as sentiments that result in physical changes reflect one's thoughts and conduct during that state. **[8]**

The main theories of emotion are grouped into three categories: physiological, neural, and cognitive. Physiological theories suggest that the answers responsible for emotions are inside the human body. Neurological theories suggest that actions within the brain trigger emotional responses. Finally, cognitive theories argue that thoughts and other mental activities play an important role in the formation of emotions.

The simplest and the most representative of common-sense definition of emotions, is that emotions are simply a class of feelings, differentiated by their experienced quality from other sensory experiences like tasting chocolate or proprioception like sensing a pain in one's knee. The idea that emotions are a specific kind of subjective experiences has dominated emotion theory roughly.

3. Emotion vs mood vs feeling

Emotion can be differentiated from a number of similar constructs within the field of affective neuroscience. **[9]** Currently, the terms emotions, feelings**[10]**, and moods**[11]** are used interchangeably, though these terms actually mean different things [14].

- Feeling; not all feelings include emotion, such as the feeling of knowing. In the context of emotion, feelings are best understood as a subjective representation of emotions, private to the individual experiencing them. **[12]**
- Moods are diffuse affective states that generally last for much longer durations than emotions, are also usually less intense than emotions and often appear to lack a contextual stimulus. **[13]**
- Affect is used to describe the underlying affective experience of an emotion or a mood.

Studying emotion is a rich field that includes many disciplines, as well as psychology, computer science, etc. Hence, affective computing is any form of computing that has something to do with emotions. Due to the strong relation with emotions, their correct interpretation (detection) is the cornerstone of this domain. The researchers in this field use various computational models to recognize emotion which can be interpreted in different ways related to the used model.

4. Emotion models

Emotion models are the foundation of the emotion detection process, as these models determine how emotions are represented. Therefore, it is necessary to define the emotion model when performing any emotion-detecting activity.

There are two unique family of models for denoting emotions: the categorical models and the dimensional models.

a. The categorical models

In these models, emotions are labeled into categories or into distinct classes that are basic and universally recognized . Commonly used models in this category are the Robert Plutchik model, the Paul Ekman model, and the OCC Model.

Figure 1 : Six Ekman's basic emotions

i. The Paul Ekman model

In this model, emotions are based on six basic categories (**Figure 1**). They are considered independent and they depend on how the experience perceives the situation. These basic emotions are happiness, sadness, anger, disgust, surprise, and fear **[8].** Each one of them is not a dingle affective state but a group of related states**. [15]**

ii. **The Robert Plutchik model**

In this model, Plutchik proposed eight fundamental emotions. There are varying degrees of intensity for each emotion. The eight emotions (**Figure 2**) are expressed in opposite pairs as: surprise versus anticipation, joy versus sadness, anger versus fear and trust versus disgust.

Figure 2 : Eight Plutchik basic emotions.

iii. The model of Orthony, Clore, and Collins (OCC)

This model classifies Emotions according to their intensity to 22. Where 16 emotions are added to Ekman emotions postulated as primary, such as: envy, comfort, appreciation, self-reproach, shame, reproach, pity, admiration, disappointment, sadness, gratification, emphatic fears, gloating, hope, admiration and dislike.

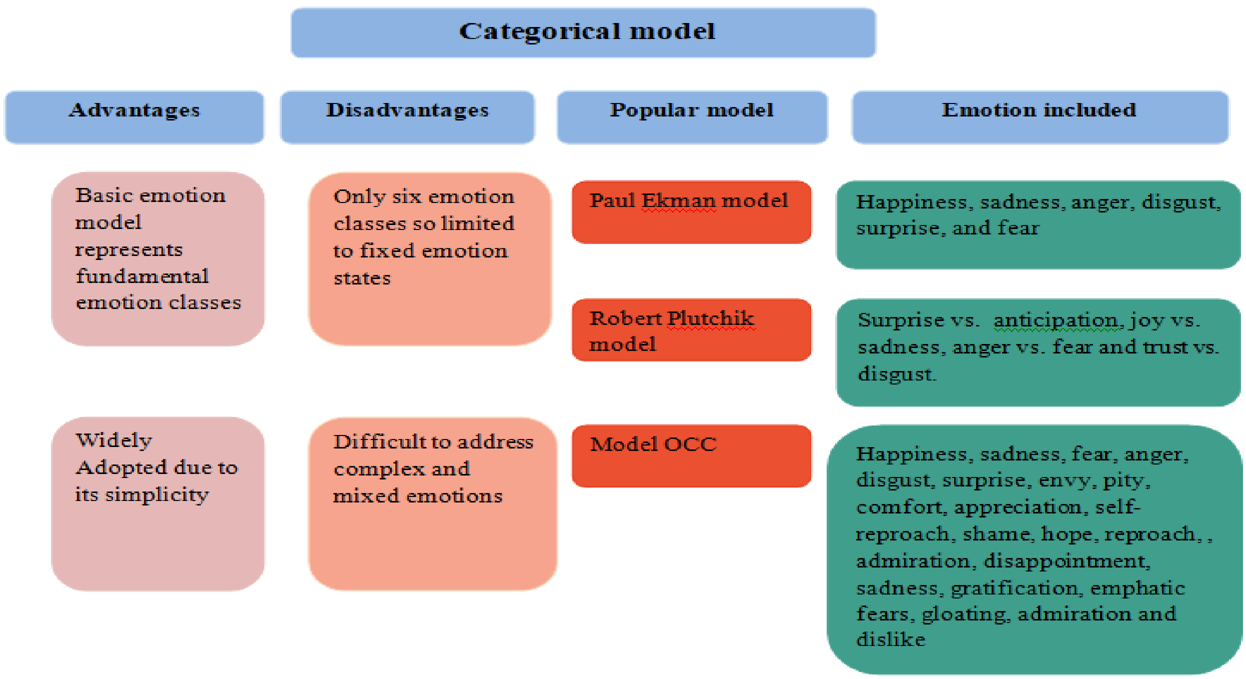


Figure 3 : Categorical model presentation

b. **The dimensional model**

In this model, emotions are considered dependent and there is a relationship between them, as represented in dimensions space to show how they are related to each other.

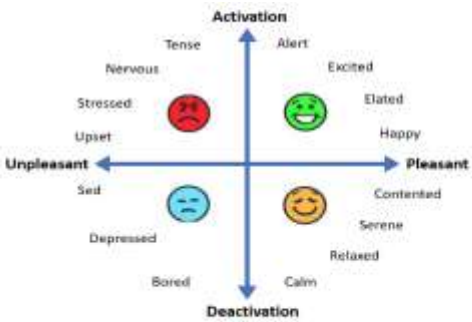


Figure 4 : Circumplex model by Russell

i. **The Russell's model**

The emotions are represented as a circumplex two-dimensional (2D) (**Figure 4**) model in the Arousal-Valence. Where the dimension Arousal (Activation and Deactivation) refers to how excited or apathetic an emotion is, and the dimension Valence (Pleasant and Unpleasant) refer to how positive emotion and negative emotion is.

ii. **Plutchik model**

This model proposed eight fundamental emotions. There are varying degrees of intensity for each emotion (**Figure 5)**. The emotions are expressed in opposite pairs as surprise versus anticipation, joy versus sadness, anger versus fear, and trust versus disgust. As illustrated in Figure 5, the emotions are represented in a wheel which contains a set of concentric circles where the innermost emotions are derived from the eight basic emotions, then the eight core emotions in the middle and finally the groups of basic emotions in the outer parts of the wheel.

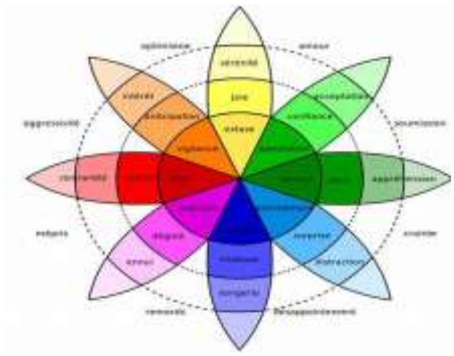


Figure 5 : Plutchik's Wheel of emotions.

iii. **Russell and Mehrabian**

The embodiment of emotions in a 3D (**Figure 6**) model such as pleasure (or positivity), arousal (or response), and power (or dominance). In the two-dimensional representation, emotions are distinguished as arousal (activation and deactivation) and valence (pleasure and aversion). The third dimension of dominance describes the

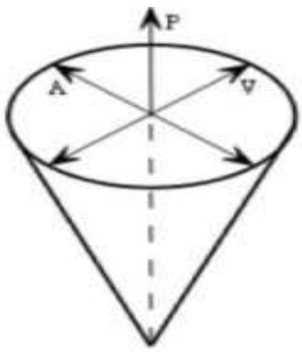


Figure 6 : 3D Emotion Space (Valence, Arousal, and Power)

degree to which the experimenters were in control of their emotions

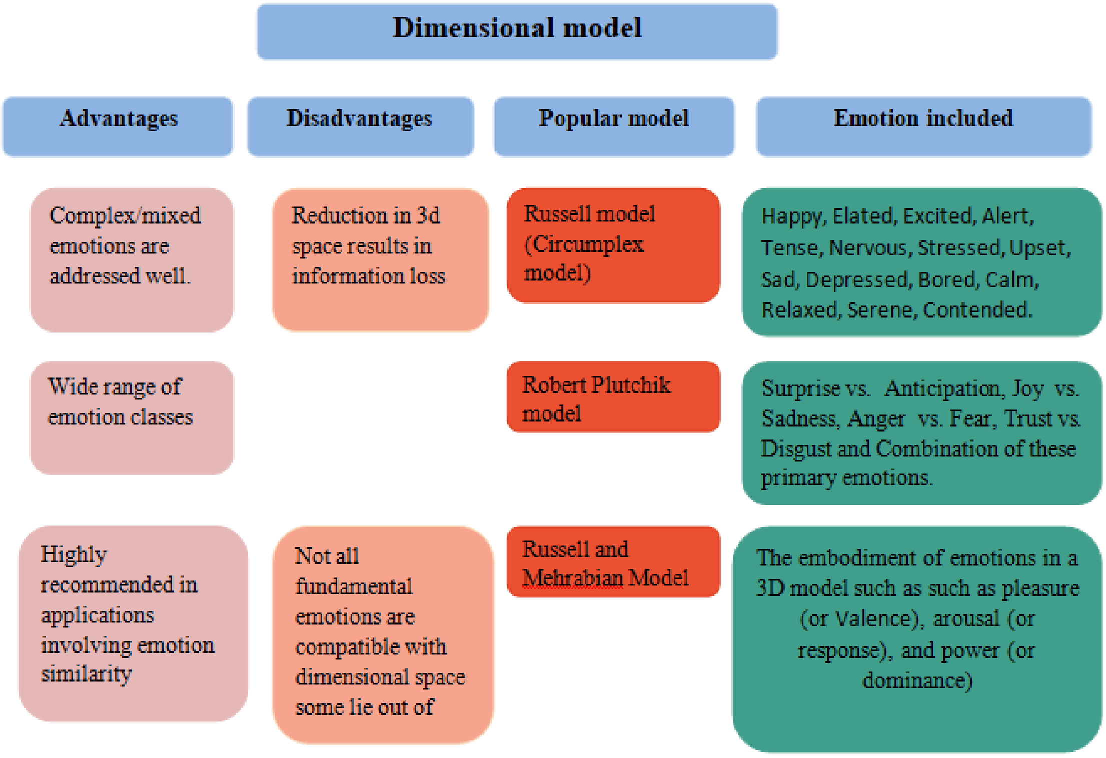


Figure 7 : Dimensional model presentation.

5. Emotion in computer science

Research in computer science aimed at developing devices that recognize human affect display and model emotions which called affective computing the more modern branch of computer science originated with Rosalind Picard's 1995 paper **[]** on affective computing. In this field we can design a computational device proposed to exhibit innate emotional capabilities or simulating emotions.

6. Emotion detection

Emotion detection (ED) called also emotion recognition, is a branch of sentiment analysis that deals with the extraction and analysis of emotions. It has become one of the most important aspects to consider in any project related to Affective Computing.

The difference between Sentiment Analysis and emotion detection is that the first aims to detect positive, neutral, or negative feelings, whereas the second aims to detect and recognize types of feelings, such as happiness, anger, disgust, fear, sadness, and surprise.

Emotion can be conveyed in several forms, such as face and movements, voice, and written language.[1]

6.1. Emotion detection from text

Emotions hold a paramount role in the conversation, as it expresses context to the conversation. Text/word in conversation consists of lexical and contextual meanings. Extracting emotions from text has been an interesting work recent these years.

Text-Based Emotion Detection is one of the fastest growing branches of Natural Language Processing (NLP), it is the process of classifying syntactic or semantic units of a set of data into a given set of emotion classes proposed by a given model.

With the advancement of machine learning techniques and hardware to support the machine learning process, recognizing emotions from a text with machine learning provides promising and significant results.

7. Applications of automated emotional detection

Emotion detection is used for several purposes. We will mention bellow some areas where emotion detection has proved his efficacity:

c. **Security measures**: Emotion recognition is used by schools and other institutions to help prevent violence and improve the overall security of the place.

d. **Human resources assistance**: Emotion recognition is used by companies in applications to assist human resources. The system is useful in determining if a candidate is genuine and genuinely interested in the position by evaluating intonation, facial expressions, and keywords and creating a human recruiting report for final evaluation.

e. **Customer service**: Systems are installed in customer service centers. They have cameras equipped with artificial intelligence, to compare the emotion of the customer before and after entering the center to determine the extent of their satisfaction with the service they received. And if there is a low score, the system can advise employees to improve the quality of service.

f. **Audience Participation:** Emotion recognition is used by companies to determine their business outcomes in terms of audience emotional responses. Apple has also released a new feature in its iPhones where an emoji is designed to mimic a person's facial expressions, called Animoji.

In short, the areas in which emotion recognition can be applied are those that aim to achieve one of the following points:

- Extract information about the emotional state of the user.
- Customer Service System Transformer.
- Detecting cases of intimidation.
- Collect opinion on a subject on a large scale for political and /or marketing purposes.

8. Conclusion

We presented in this chapter the emotion context by mentioning the essential concepts related to emotion. We presented also the emotion detection principal which is generally based on classification techniques, that require a predetermined set of emotion categories. In the next chapter we will present the different concepts of machine learning required to build an emotion detection system.

Chapter 2
Machine learning

1. Introduction

There are a variety of techniques for detecting emotions, hence machine learning techniques which are probably used dozens of times a day in medicine, email filtering, computer vision, self-driving cars, effective web search, and emotion detection in texts; where it is difficult to develop conventional algorithms for the needed tasks.

In this chapter, we will present a brief description of the most used machine learning techniques in the context of emotion detection.

2. Machine Learning

2.1. Machine learning definition

We can define Machine learning (ML) as: "the field of study that gives computers the ability to learn without being explicitly programmed". [16]

Machine learning is a pathway to artificial intelligence. This subset of AI uses statistical learning algorithms to automatically learn insights and recognize patterns from data, applying that learning to make increasingly better decisions, which means machine learning uses learning algorithms to build systems that have the ability to learn and improve from experiences without being explicitly programmed automatically. [17]

Tom M. Mitchell provided a more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." **[18]**

This definition is for machine learning tasks which are concerned and offer a fundamentally operational definition rather than defining the field in cognitive terms.

Modern-day machine learning has two objectives, one is to classify data based on developed models, and the other is to make predictions for future outcomes based on those models. **[19,20]**

Most common people tend to use the terms like artificial intelligence and machine learning as synonymous and they do not know the difference. However, these two terms are actually two different concepts even though machine learning is actually a part of artificial intelligence.

2.2. learning forms

ML algorithms can be broadly classified into four categories: Supervised, Unsupervised, semi-supervised and Reinforcement learning.

Table 1 : The four categories of Machine learning Algorithms

ML algorithms categories	definition	defer
Supervised	Supervised learning [21,22] algorithms (SLA) build a mathematical model of a set of data that contains both the inputs and the desired outputs. In this kind of learning, the machine learning algorithm is supplied with the labelled data to train the algorithm "training data", which means that every data input is tagged with a labelled output also known as a supervisory signal	Regression: linear regression and poison regression Classification: decision trees, logistic regression, naïve bayes, SVM, K-nearest neighbor, GMM
Unsupervised	Unsupervised learning is used to train the ML algorithm, which means supplying the data only having features without any tagging of historical labels and through mimicry, which is an important mode of learning in people, the algorithm figures out the data and according to the data segments, it makes clusters of data with new labels.**[27]**	Clustering: KMC, DBSCAN, Mean Shift, OPTICS, Agglomerative. Association Rule Learning: apriori algorithm, Eclat algorithm. Dimensionally Reduction: PCA, NMF, T-SNE, UMAP
Semi-supervised	Semi-supervised learning is a learning problem that involves a small number of labeled examples and a large number of unlabeled examples. Learning problems of this type are challenging as neither supervised nor unsupervised learning algorithms are able to make effective use of the mixtures of labeled and untellable data. As such, specialized semis-supervised	Generative models Low-density separation Laplacian regularization Heuristic approaches

	learning algorithms are required.[39]	
Reinforcement	Reinforcement learning can be explained as learning by continuously interacting with the environment. It is a type of machine learning algorithm in which an agent learns from an interactive environment in a trial-and-error way by continuously using feedback from its previous actions and experiences.[40]	Q-learning

2.3. Supervised

In the mathematical model, each training example is represented by an array or vector, and the training data is represented by a matrix. **[23]**. SLA learn a function that can be used to predict the output associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data.

The algorithm learns by comparing its actual production with correct outputs to find errors. It then modifies the model accordingly.

Types of supervised-learning algorithms include classification and regression. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. The difference between the two tasks is the fact that the dependent attribute is numerical for regression and categorical for classification.

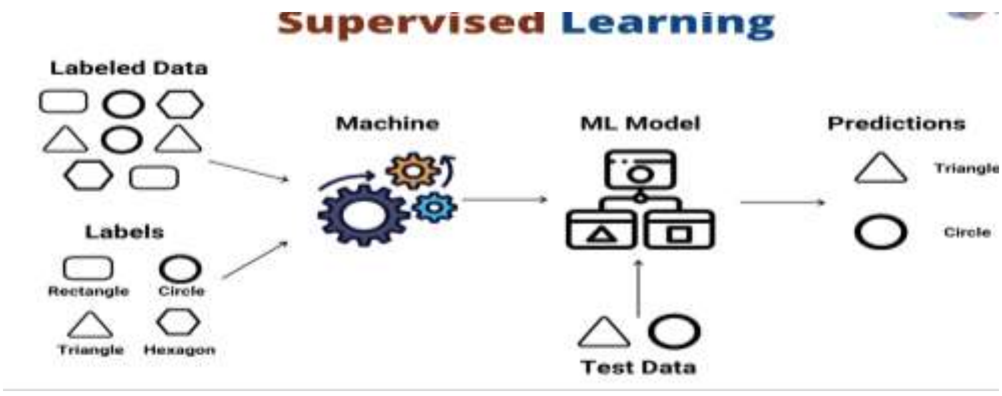


Figure 8 : supervised Learning Example

2.4. Classification

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. There are two types of learners in classification supervised learning:

2.4.1. Lazy learners

Lazy learning is a learning method in which generalization of the training data is, in theory, delayed until a query is made to the system, as opposed to eager learning, where the system tries to generalize the training data before receiving queries. **[24]**

Lazy learners store the training data and wait until testing data appear. that by motiving the data set is continuously updated with new entries. Because of the continuous update, the "training data" would be rendered obsolete in a relatively short time. this is why one cannot really talk of a "training phase". Lazy learners have less training time but more time in predicting. Lazy classifiers are most useful for large, continuously changing datasets with few attributes that are commonly queried (e.g.: k-nearest neighbor, Case-based reasoning).

2.4.2. Eager learns

Eager learning is a learning method in which the system tries to construct a general, input-independent target function during the training of the system, as opposed to lazy learning, where generalization beyond the training data is delayed until a query is made to the system.

[25] Eager learners construct a classification model based on the given training data before receiving data for classification.

The main advantage gained in employing an eager learning method, such as an artificial neural network, is that the target function will be approximated globally during training. Eager learning systems also deal much better with noise in the training data. It is an example of offline learning, in which post-training queries to the system have no effect on the system itself, and thus the same query to the system will always produce the same result. it must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time to train and less time to predict. (e.g.: Decision Tree, Naive Bayes, Artificial Neural Networks).

2.4.3. Classification Algorithms

2.4.3.1.Naïve Bayes

Naive Bayes is a probabilistic classifier inspired by the Bayes theorem under a simple assumption (called naïve) which is that the attributes are conditionally Independent. [26] It implements a naive Bayesian classifier, or naive Bayes classifier, belonging to the family of linear classifiers.

The Bayesian classifier is a linear classifier based on Bayes theory. It assumes that a class has certain characteristics. It then makes it possible to classify a given example according to these characteristics.

The classification is conducted by deriving the maximum posterior which is the maximal P(X|Ci) with the above assumption applying to Bayes theorem. This assumption greatly reduces the computational cost by only counting the class distribution. Even though the assumption is not valid in most cases since the attributes are dependent, surprisingly Naive Bayes has able to perform impressively.

Naive Bayes is a very simple algorithm to implement and good results have obtained in most cases. It can be easily scalable to larger datasets since it takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

2.4.3.2. SVC

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

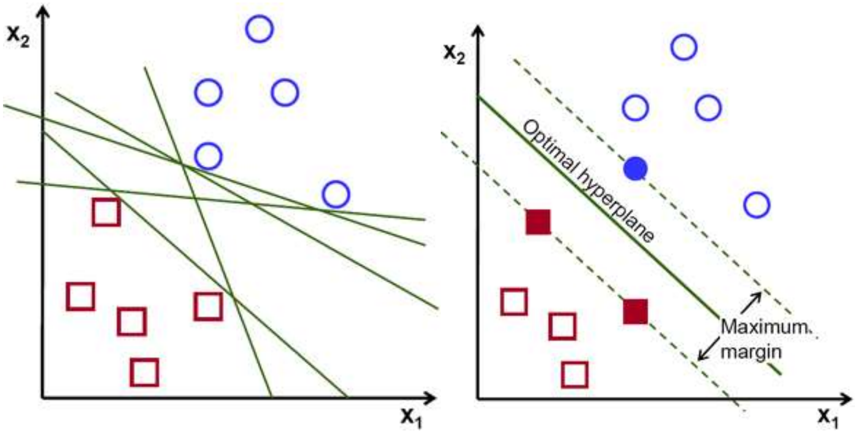


Figure 9 : Possible hyperplanes

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. the objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence, Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes.

2.4.3.3.LR

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example,To predict whether an email is spam (1) or (0). Whether the tumor is malignant (1) or not (0).

Types of Logistic Regression

- Binary Logistic Regression: The categorical response has only two 2 possible outcomes. Example: Spam or Not.
- Multinomial Logistic Regression: Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan).
- Ordinal Logistic Regression: Three or more categories with ordering. Example: Movie rating from 1 to 5.

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

Figure 10 : Bayes theorem

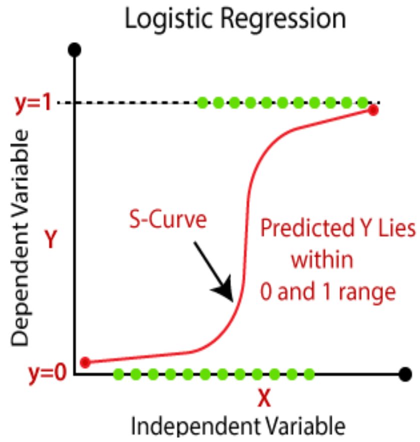


Figure 11 : LR

3. Deep Learning

3.1. Deep Learning Definition

Deep learning is a machine learning technique that is inspired by the way a human brain filters information, it is basically learning from examples. It helps a computer model to filter the input data through layers to predict and classify information. Since deep learning processes information in a similar manner as a human brain does, it is mostly used in applications that people generally do. It is the key technology behind driver-less cars, that enables them to recognize a stop sign and to distinguish between a pedestrian and lamp post. Most of the deep learning methods use neural network architectures, so they are often referred to as deep neural networks. The three fundamental network architectures are in table below:

Table 2 : Deep Learning Approach

Convolutional Neural Networks (CNN)	**Convolutional Neural Network is basically an artificial neural network that is most widely used in the field of Computer Vision for analyzing and classifying images. It is a deep learning algorithm that takes the input image and assigns weights/biases to various aspects or objects in the image, so that it can differentiate one from the other. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers, and normalization layers. The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the**

	Human Brain and was inspired by the organization of the Visual Cortex.[41]
Recurrent Neural Networks	Recurrent Neural Networks (RNN) is a type of neural network architecture that is used in sequence prediction problems and is heavily used in the field of Natural Language Processing. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being dependent on the previous computations. Another way to think about RNNs is that they have a “memory” which captures information about what has been calculated so far. They are networks with loops in them, allowing information to persist.[42]
Recursive Neural Networks	A recursive neural network is a kind of deep neural network created by applying the same set of weights recursively over a structured input, to produce a structured prediction over variable-size input structures, or a scalar prediction on it, by traversing a given structure in topological order.[43]
Convolutional Neural Networks (CNN)	**Convolutional Neural Network is basically an artificial neural network that is most widely used in the field of Computer Vision for analyzing and classifying images. It is a deep learning algorithm that takes the input image and assigns weights/biases to various aspects or objects in the image, so that it can differentiate one from the other. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers, and normalization layers. The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex.**
Recurrent Neural Networks	Recurrent Neural Networks (RNN) is a type of neural network architecture that is used in sequence prediction problems and is heavily used in the field

	of Natural Language Processing. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being dependent on the previous computations. Another way to think about RNNs is that they have a "memory" which captures information about what has been calculated so far. They are networks with loops in them, allowing information to persist.
Recursive Neural Networks	A recursive neural network is a kind of deep neural network created by applying the same set of weights recursively over a structured input, to produce a structured prediction over variable-size input structures, or a scalar prediction on it, by traversing a given structure in topological order.

3.2. LSTM

Long Short-Term Memory networks or simply LSTMs are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many people in following work. They work tremendously well on a large variety of problems, and are now widely used.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

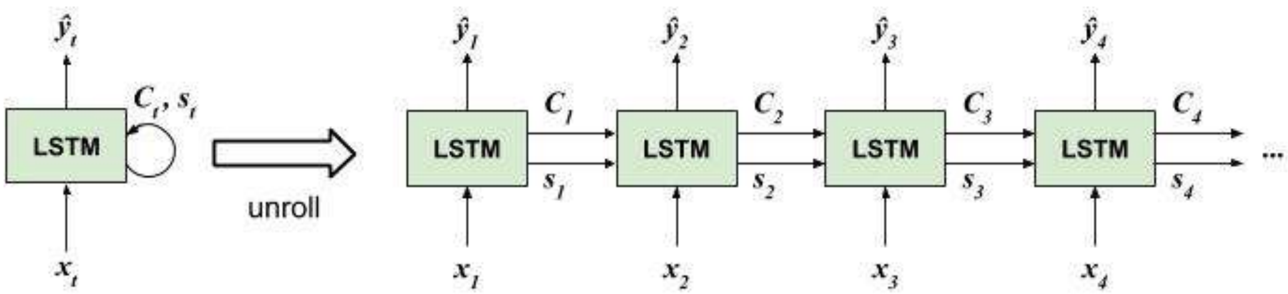


Figure 12 : LSTM

3.3. Bi-LSTM

Bidirectional long-short term memory(bi-lstm) is the process of making any neural network o have the sequence information in both directions backwards (future to past) or forward(past to future). In bidirectional, the input flows in two directions, making a bi-lstm different from the regular LSTM. With the regular LSTM, it can make input flow in one direction, either backwards or forward. However, in bi-directional, it can make the input flow in both directions to preserve the future and the past information.

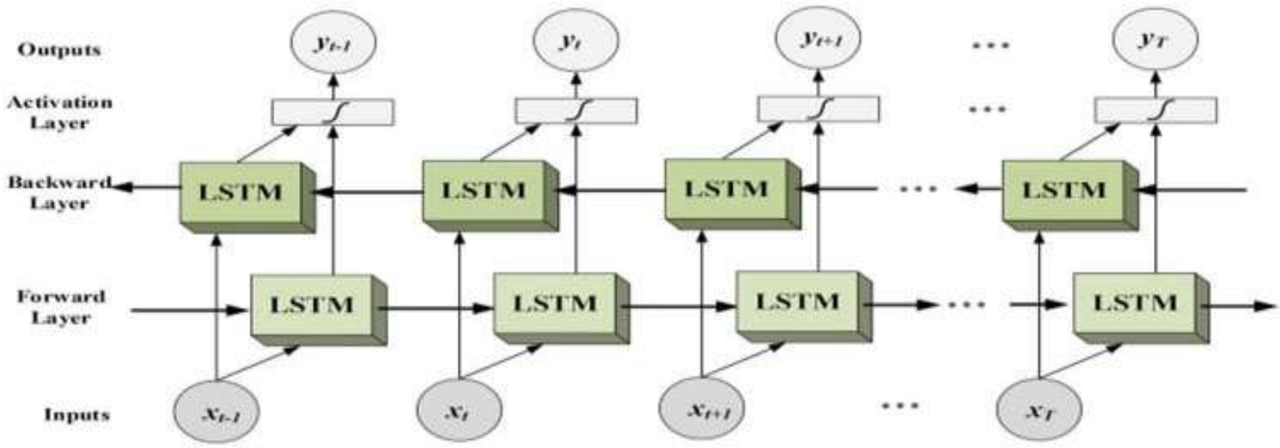


Figure 13 : BI-LSTM

4. Conclusion

Machine learning has become a central part of our lives as consumers, costumers, and hopefully as researchers and practitioners.

In the next chapter, we will present the domain of emotion detection in social Medias.

Chapter 3
Emotion Detection in social media

1. Introduction

Emotion can be conveyed in several forms, such as face and movements, voice, and written language. Emotion detection in text documents is an issue of material – identification based on principles derived from machine learning. In day-to-day life, human emotions play an important role. Emotion can generally be understood as intuition that differs from thought or knowledge.

Many social networking sites generate various textual data containing significant data and perform an ever more significant emotional understanding role [12]. The secure production of cognitive technologies is influenced as a foundation of human-computer emotional communication.

Emotion detection is generally based on classification techniques, which require a predetermined set of emotion categories, in this section, we mention the essential components needed in the context of the emotion detection in social medias like the natural language processing. In the end of the chapter we will present some related works in the domain and we will conclude with a comparative study of most existing datasets related to the emotion detection domain.

2. Why social medias

Social media platforms like Facebook and Twitter have emerged as the most popular tools for people to express their thoughts, emotions, reviews, and feedback about the current surrounding events. They are considered an important information source for many Natural Language Processing applications such as sentiment analysis and emotion detection.

Emotion detection from a text document is fundamentally a content-based classification issue, including notions from NLP and machine learning fields.

3. Natural Language Processing (NLP)

Natural language processing (NLP) is the sub-field of Computer Science especially Artificial Intelligence (AI) that is concerned about enabling computers to understand, interpret and process human language. Natural language processing draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.[5] Technically, the main task of NLP would be to program computers for analyzing and processing huge amount of natural language data.[4]

4. How does NLP work?

NLP employs two distinct processing modes. The data pre-processing state is where AI techniques assist machines in understanding natural language by converting data (speech/text) into automatic interpretation mode. The building of NLP algorithms is the second stage, which entails training algorithms on data so that they can analyze and interpret language as well as perform certain tasks.

5. Data Pre-Processing

The first state in NLP work includes the transformation of data into the machine-interpretable mode. The key to NLP lies with its ability to understand and identify the grammatical structure, and semantic of the language, as well as, to relate words with each other to understand the context of the text/speech. This complex processing is achieved by transforming (cleaning) the dataset and making it more organized. This state includes:

- **Tokenization:** Tokenization involves breaking of text/speech into single clauses or small semantics
- **Part-of-speech-tagging:** This involves the categorization of words into nouns, pronouns, adjectives, adverbs, etc.
- **Stemming and Lemmatization:** This involves standardization of words into their root forms.
- **Stop word removal:** Stop word removal involves filtering of commonly used words which adds little to no information to the text. This mostly includes articles, prepositions, and others.

6. The building of NLP algorithms

Once the data transformation (cleaning) is completed, the next step in Natural Language Processing involves building an NLP algorithm. This state includes training of the NLP algorithm so as to make it efficient in performing specific tasks and interpreting human language.

Two main paradigms used in this state include:

- **A rule-based approach:** This approach heavily relies on the grammatical rules hand-crafted by linguists, knowledge engineers, and grammar experts. This approach was the pioneering approach in the development of NLP algorithms.
- **Machine learning approach:** Machine learning is the latest and advanced approach to build NLP algorithms. In this approach machines learn the rules from training data.

Also, uses statistical and deep learning methods to enable algorithms to learn, and interpret human language.

7. NLP applications

NLP is used in many areas such as social media monitoring, translation tools, smart home devices, survey analytics, etc. The most popular applications of it are chatbots, Voice Assistants, Sentiment & emotion Analysis, Language Translators, speech recognition & synthesis and Email Classification and Filtering.

Figure 14 : Applications of Natural Language Processing

8. Text based emotion detection

Emotions play an essential role in the conversation. They express meanings to the conversation reciprocally with the text of the text. Detecting emotions form text has been an attractive task recent these years. With the advancement of machine learning techniques and hardware to support the machine learning process, detecting emotions from a text with machine learning provides promising and significant results.

A. Process of text-based emotion detection

Detecting emotions from text is one of the most hurdles (spellings, languages, slang), but another source of emotional information to consider. Since the detection of sentiment from texts analyzes the words in the message, the process of analyzing the text takes some more steps than analyzing the face or voice. This processing includes tasks of preprocessing steps Including: tokenization, parsing and part-of-speech tagging, lemmatization, stemming, and many others.

Recently, the process of text-based emotion detection uses artificial intelligence. In any text-based emotion detection system, initially, datasets are created by downloading the data from online social media. After data generation, text pre-processing steps involve making the text suitable for any machine or deep learning algorithm to process it. Text preprocessing involves tokenization, text cleaning, normalization, and creating feature vectors/embedding. Particularly, the text from social media consists of slang words, emojis, hashtags, HTML tags, short text, incomplete words, etc., which requires preprocessing. Next, machine or deep learning is applied to generated feature vectors. These feature vectors are fed into any machine learning algorithm or deep learning neural network where feature vectors with

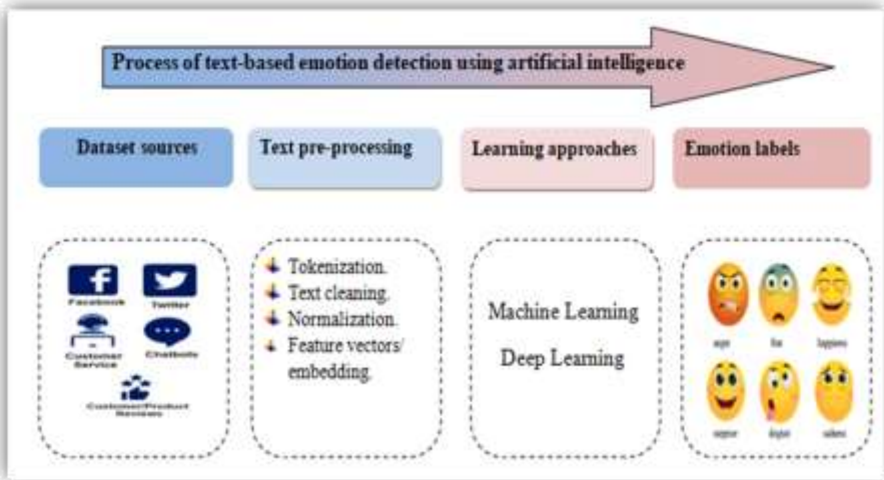


Figure 15 : Text preprocessing along with process flow of text-based emotion detection using artificial intelligence.

emotion labels are trained. Then the trained system is used to classify and predict the labels of text. [20]

B. Text based emotion detection approaches

The general approaches used for detecting emotions from the text are The Machine learning (ML) based approach, the keyword based approach, The hybrid approach , rule-based approach and deep learning based approach.

1. **The ML approach:** Emotion detection from text is based on classification problems involving different models of Machine Learning (ML). Machine learning is categorized into unsupervised learning and supervised learning. This approach generally starts with the text preprocessing step. Then the useful features are extracted from the text, and only features are selected with the most information gained. After that, with the given

feature set and emotion labels, the system is trained. Lastly, the trained system is used to classify the emotion from the unseen text, termed prediction. [6;7]

2. **Keyword based approach:** Keyword discovery technique is the problem of finding duplicates of keywords from a text document and matching them with the labels stored in the data set [7]. To discover the feelings in the text, we go through the following stages:

   Firstly, the list of emotion keywords is determined from standard lexical databases. Then preprocessing is performed on the data set. Then, keywords are discovered between the emotional keywords from the text and the predefined keyword lists. The intensity of emotion keywords is then analyzed. Next, a negation examination is performed to determine the negation signs and the scope of the sign, and finally, the naming of emotion is determined [20]. These emotions are classified into categories according to the chosen emotional model [21].
3. **Hybrid approach:** The hybrid approach is a combination of different approaches. combines with the keyword, the rule-construction, and the ML approaches into a unified model. Thus, drawing from the strength of approaches used while concealing their associated limitations, this approach has a higher probability of transcending the other approaches individually [6, 21].
4. **Rule construction approach:** The rule-based approach is used to manipulate knowledge in order to view the information in an advantageous way. It begins with text preprocessing initially, and that stop word elimination, POS tagging, tokenization, etc. The rules of emotion are then derived using the concepts of statistics, linguistics, and computation. The best rules are selected later. Finally, the rules are applied to emotion datasets to determine the emotion labels. Subsequently, the appropriate rules are chosen. Additionally, the rules are applied to the emotion dataset for determining the emotion labels [7,20].
5. **Deep learnin approach:** Deep learning is a variant of machine learning in artificial intelligence with networks capable of unsupervised learning from unstructured or unlabeled data.

   This approach enables neural networks to learn complicated concepts by building them up from simpler ones. The dataset is subjected to preliminary preprocessing. The embedding layer is next created, in which tokens are represented as numbers. Then, depending on the number of emotion labels, these feature vectors are input into one or more Deep Neural Network layers. Patterns are learned from data and used to predict the labels by using classification.[20]

9. Related works in the text-based emotion detection

Emotion detection from textual sources can be done utilizing notions of Natural Language Processing. NLP techniques improve the performance of learning-based methods by incorporating the semantic and syntactic features of the text. Word embeddings are extensively utilized for several NLP tasks, like machine translation, sentiment analysis, and question answering.

In the next section, we will present some works related to the text-based emotion detection.

- Mentioned Romana Rahman, Tajul Islam, Md. Humayan Ahmed [10] in they work that Shiv N. et.al [1] described emotion detection from textual documents and blogs. They proposed two components: Emotion Detector, Emotion Ontology for finding emotion from the text.
- Romana Rahman, Tajul Islam, Md. Humayan Ahmed [10] in their work to detect emotions from text and symbols, they proposed some methodologies that solve the problem of detecting emotions in the case of sentence and symbol level. Their generated method works on the basis of keyword analysis (KA), keyword negation analysis (KNA), a set of proverbs, emoticons, acronyms, exclamations and so on. To find feelings, they created 25 chapters on feelings. This analysis resulted in a better result for detecting emotion than text and symbol.
- Mentioned Romana Rahman, Tajul Islam, Md. Humayan Ahmed [10] in they work that Abdul Hannan detected emotion from text using NLP (natural language processing).Here he described mainly two kinds of NLP methods .i.e.keywords or pattern matching technique and parsing technique.
- Mentioned Ashritha R Murthy [6] that Srinivas Badugu and Matla Suhasini [55] developed a Rule Based Approach that detects the emotions from the tweets and classify into different emotion categories.
- Mentioned Ashritha R Murthy [6] that Ramalingam et al. developed a hybrid model combined with the keyword and learning based method and obtained high accuracy result for detection of emotions from the text. Angelina et al [81] used twitter dataset with NRC emotion lexicon. And used SVM for multiclass classification and implemented on the software WEKA .

- Mentioned Ashritha R Murthy [6] that Hasan et alin on his work on emotion detection worked on text flow data and using both online and offline messaging. The Support Vector Machine, Naïve Bay, and Decision Tree (DT) were used to detect.
- Shaikh Abdul Salam, Rajkumar Gupta focuses on data obtained from one of the most popular social media - Twitter by analyzing live as well as past feeds and getting emotions from them. The twitter data required in English language is converted into a vector of eight emotions and supervised learning techniques such as K-means, Naive Bayes and SVM is used to determine label identifying one of the basic emotion family. Then give a comparative study of the performance of different classifiers is discussed.
- Bhajibhakare M. M., Ankita Borkar , Simran Naik , Suvarna Solase and Padmasen Kunjir [15] they work on Sentiment Analysis on Youtube & Twitter Data using Machine Learning for classification and analysis we using Naïve Bayes and Support Vector Machine (SVM) techniques are used.
- Kashif khan, Sher Hayat and Muhammad Ejaz khan [16] present recent work in the field of Emotion Detection through text, also present some limitations on text-based emotions to solve somehow the problems which we are facing in emotion detection. They focus on SVM, KNN, HMM, Byes classification and will identify advantages and disadvantages.

10. Emotion detection from Arabic text in social medias

Considered the Arabic language as the official language of 22 countries. It is also the mother tongue of over 1.4 billion Muslims worldwide. It's been around for over 2000 years [28]. The Arabic alphabet is made up of 28 letters with no higher or lower case, and writing is done from right to left. Its letters can be written in a variety of shapes depending on where they appear in the word. According to [22, 25], the Arabic language is divided into three groups: Classical Arabic (CA), Modern Standard Arabic (MSA) and Dialectical Arabic (DA).

Arabic language is one of the fastest growing languages on social media twitter specifically. It is hard to build a system to detect sentiments and emotions for this language. It is evident that the Arabic language occupies a distinct position among the languages on the social networking site, despite the cultural, cultural and environmental diversity of users within the Arab world. Most of arb users use dialectical Arabic within social media what amplify the complexity of the text treaatemnt.

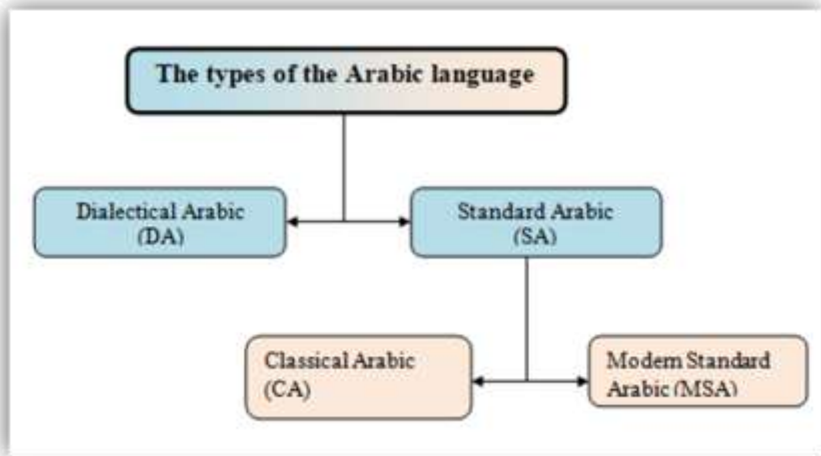


Figure 16 : Types of Arabic language

In the next section, we will present an ensemble of works established in the context of Arabic text emotion detection which is the main goal of our stady.

- Malak Abdullah, Mirsad Hadzikadic and Samira Shaikh [17] create system - SEDAT, to detect sentiments and emotions in Arabic tweets. We use word and document embeddings and a set of semantic features and apply CNN-LSTM and a fully connected neural network architectures.
- Hamdy Mubarak, Kareem Darwish and Walid Magdy [18] they work on detecting abusive language on Arabic social media. We extract a list of obscene words and hashtags using common patterns used in offensive and rude communications. We also classify Twitter users according to whether they use any of these words or not in their tweets.
- Asma Chader, Dihia Lanasri, Leila Hamdad, Mohamed Chemes Eddine Belkheir and Wassim Hennoune [19] they propose a supervised approach for sentiment analysis of Arabizi Algerian dialect using different classifiers such as Naive Bayes and Support Vector Machines. We investigate the impact of several preprocessing techniques, dealing with dialect specific aspects. Experimental evaluation on three manually annotated datasets shows promising performance where the approach yielded the highest classification accuracy using SVM algorithm. Moreover, our results emphasize the positive impact of proposed preprocessing techniques.
- Reima Al-Jarf [21] work on the effect of Facebook on Arabic language attrition, i.e., decrease in language proficiency, as exhibited in the use of Colloquial instead of

Standard Arabic, use foreign words although Arabic equivalents exist, and committing spelling errors.

11. Datasets

The process of text emotion detection requires reliable database named in this case dataset broad enough to fit every need for its application, as well as the selection of a successful classifier which will allow for quick and accurate emotion identification. In the next section we will illustrate a non-exhaustive list of specific data sets related to the domain of emotion detection, at the end of the section a comparative table is presented in the same context.

C. **ISEAR:** The International Survey on Emotion Antecedents and Reactions (ISEAR) database21 constructed by the Swiss National Centre of Competence in Research and lead by Wallbott and Scherer consists of seven emotion labels (joy, sadness, fear, anger, guilt, disgust, and shame) obtained as a result of gathering series of data from cross-cultural questionnaire studies in 37 countries. Three thousand (3000) participants from varying cultural backgrounds were made to fill questionnaires about their experiences and reactions toward events. The final dataset reports a total of 7665 sentences labeled with emotions.

D. SemEval: The Semantic Evaluations (SemEval) is a database consisting of Arabic and English news headlines extracted from news websites such as BBC, CNN, Google News, and other major newspapers. The dataset contains 1250 data in total. The data in this database are rich in emotional content for emotion extraction and it is labeled using the 6 emotional categories (ie, joy, sadness, fear, surprise, anger, and disgust) presented by Ekman.

E. EMOBANK: This dataset consists of over 10 000 sentences annotated dimensionally in accordance to the Valence-Arousal-Dominance (VAD) emotion representation model. These sentences were obtained from news headlines, essays, blogs, newspapers, fiction, letters, and travel guides of writers and readers, thus spanning a wider domain. A subset of the dataset has also been annotated categorically using the Ekman's basic emotion model making it suitable for dual representational designs

F. **SMIC:** dataset contains spontaneous micro-expressions of 16 participants and 164 spontaneous micro-expressions. There are three datasets included in SMIC: SMIC-HS, SMIC-VIS, and SMIC-NIR. Videos were recorded with 100 fps (High Speed — HS), and the last ten subjects were recorded with cameras of 25 fps of both visual (VIS) and near inferred (NIR) light ranges

G. **BAUM:** is a video-based dataset. The dataset contains audio-visual clips of different languages and they are annotated. The clips represent real-world scenarios like different poses, lighting conditions, and subjects of varying ages. An image-based dataset is created with the peak frames from each clip**.**

H. **EMOTIC:** database of images of people taken from the real environments, and annotated with their apparent emotions. They are labeled with 26 categories of emotion. They were labeled using Amazon Mechanical Turk (AMT) platform. The dataset contains 18,313 images and 23,788 annotated people. Some images were collected from Google as well.

I. **AffectNet:** is one of the popular datasets for detecting facial emotions. The dataset contains around one million facial images. They were collected from search engines with a thousand two hundred and fifty keywords from six languages altogether. Around four-hundred and fifty thousand images were manually annotated by 12 experts

J. **FER-2013:** The FER-2013 is a widely used emotion dataset. The images are labeled with seven emotions: neutral, happy, surprise, sad, fear, disgust, and anger. The dataset contains 28,000 of training data, 3,500 of validation data, and 3,500 of test data. The images were collected from google. The images were collected in a way that they vary in the pose, age, and occlusion.

K. **Comparative table of the deferent datasets**

Table 3 : Comparative table of the deferent datasets

dataset	**field**	**Emotion model**		**model**	**approach**
		Categorical	**dimensional**		
ISEAR	**Text**	**X**		**Ekman**	**ML**
EMOBANK	**text**		**X**	Russell and Mehrabian Ekman	
SemEval	**text**	**X**		**Ekman**	**ML (LSTM)**
SMIC	**video**	**X**		Ekman	**ML**

EMOTIC	**image**	**X**	**X**	Ekman Russell and Mehrabian	**ML (CNN)**
AffectNet	**image**	**X**	**X**	Ekman Russell	**ML (DL)**
FER-2013	**image**	**X**		Ekman	**ML**

12. Conclusion

In this chapter, we presented the essential concepts of the field of emotion detection; in conclusion we noticed that the emotion detection domain is a growing up domain specially in the number of used techniques but we noticed also the pure number of datasets related to the Arabic language and precisely the dialectical ones. Unfortunately, there is no available data set destinated to the Algerian dialect. In the next chapter, we will present our system. Witch is concerned with the emotion detection from Arabic text.

Chapter 4 Conception and implementation

1. Introduction

Our project's aim is to explore the world of Natural Language Processing (NLP) by building what is known as an Emotion detection Model. An Emotion detection model is a model that analyses a given set of text and predicts emotion recognition: anger, joy, surprise, or sad.

In this chapter we will present the different components of the architecture of our system.

2. Architecture of the system EDIDA

2.1. The global architecture of the system EDIDA

The system architecture is divided into several parts: data set, preprocessing, Learning Model, and results.

1. Data-set: ensemble of tweets in dialectical Algerian language and dialectical Egyptian language.
2. Preprocessing: in this step we clean the database form punctuation, stop words, Emoji, duplicates removal… etc.
3. Learning Models: using classification algorithms Logistic regression, Naïve Bayes (NB), Long-short term memory (LSTM) and SVC.
4. Result: one of the four emotions and its accuracy.

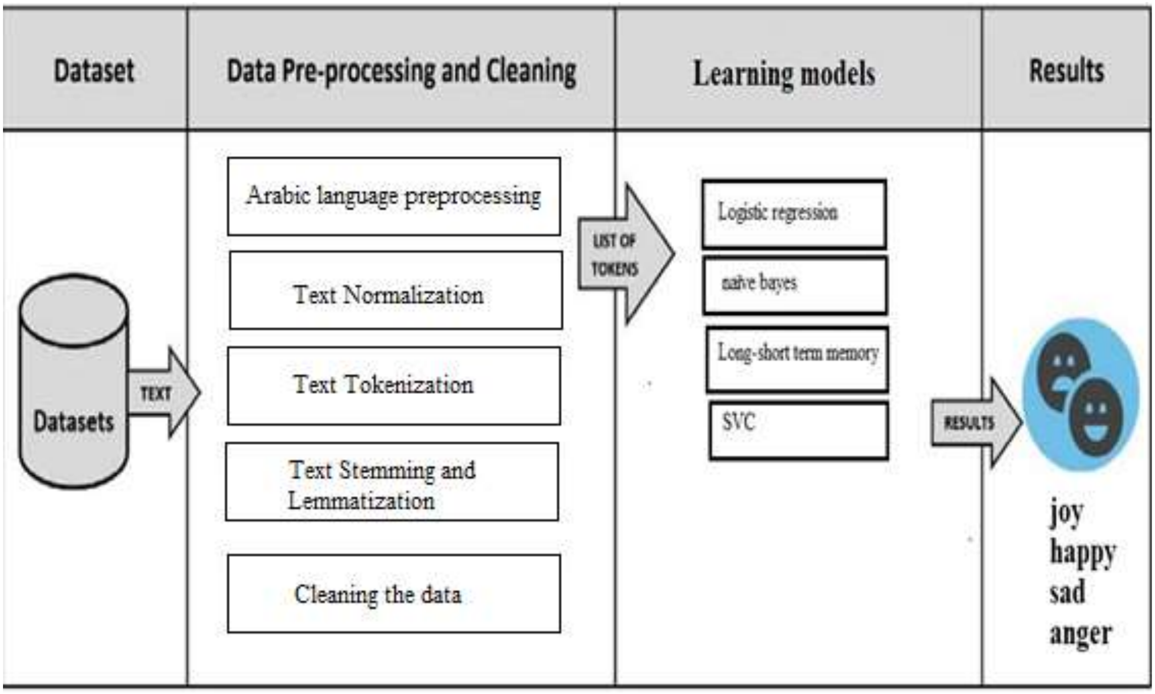


Figure 17 : The Global Architecture Of EDIDA

2.2. Detailed architecture of EDIDA:

2.2.1. Importing and Discovering the Data-set

We used the Egypt dialectical language data-set and Algerian dialectical language containing data collected from Twitter. An impressive feature of the Egyptian dataset is that it is perfectly balanced (i.e., the number of examples in each class is equal) 10065 rows but there is 8 emotions in it; contrary to The Algeria dataset 3180 rows with only 4 emotions ,it's not as efficient for the model as the Egypt dataset.

2.2.2 Preprocessing: Cleaning and Processing the Data

2.2.2.1. Arabic language preprocessing First step in the Arabic language processing is deleting the Latin words and the emoji from our sentences. In cases where the sentence is composed with only Latin words plus emoji, it will all be deleted.

2.2.2.2. Text Normalization It is a process that converts a list of words to a uniform sequence by transforming the words to a standard format. It includes removing numbers, removing punctuations, remove diacritics, reduce or decompose characters to normal forms and many other tasks. In addition, the normalization process might also compromise an NLP task such as removing stop words.

2.2.2.3. Text Tokenization There is a series of cleaning and data processing, In order to feed our text data to a classification model, we first need to tokenize it: "Tokenization is the process of splitting up a single string of text into a list of individual words, or tokens".

2.2.2.4. Text Stemming and Lemmatization

Word stemming in Arabic is the process of removing all of a word's prefixes and suffixes, or the conversion of plural to singular, or the derivation of a verb from the gerund form to produce the stem or root. It goes after finding the origin (root) of words in the natural Arabic language by getting rid of any additions in words, because Arabic words may have more complicated forms than any other language with such additions. This model used nltk's WordNetLemmatizer to accomplish this task. This lemmatizer however takes as input two arguments: a list of tokens to be lemmatized as well as their corresponding part of speech. The most common parts of speech in English are nouns and verbs. In order to extract each token's part of speech, we will utilize nltk's post_tag function, that takes an input a list of tokens, and returns a list of tuples, where each tuple is composed of a token and its corresponding position tag. Various position tags can be outputted from the pos_tag function, however the most notable ones are:

- NNP: Noun, proper, singular
- NN: Noun, common, singular or mass.
- VBG: Verb, gerund or present participle.
- VBN: Verb, past participle.

2.2.2.5. Cleaning the data

a custom function is defined in order to fine-tune the cleaning of the input text. This function is highly dependent on each use case. third, Only include misspellings or abbreviations of commonly used words. Including many minimally present cases would negatively impact the performance. Then, a function will be our all-in-one noise removal function. Eliminate the token if it is a link, if it is a mention and if its length is less than 3, if it is a punctuation or if it is a stop word, Previewing the remove_noise() output. As the Naive Bayesian classifier accepts inputs in a dict-like structure, we have to define a function that transforms our data into the required input structure .A final point, Removing noise from all the data and Transforming the data to fit the input structure of the Naive Bayesian classifier

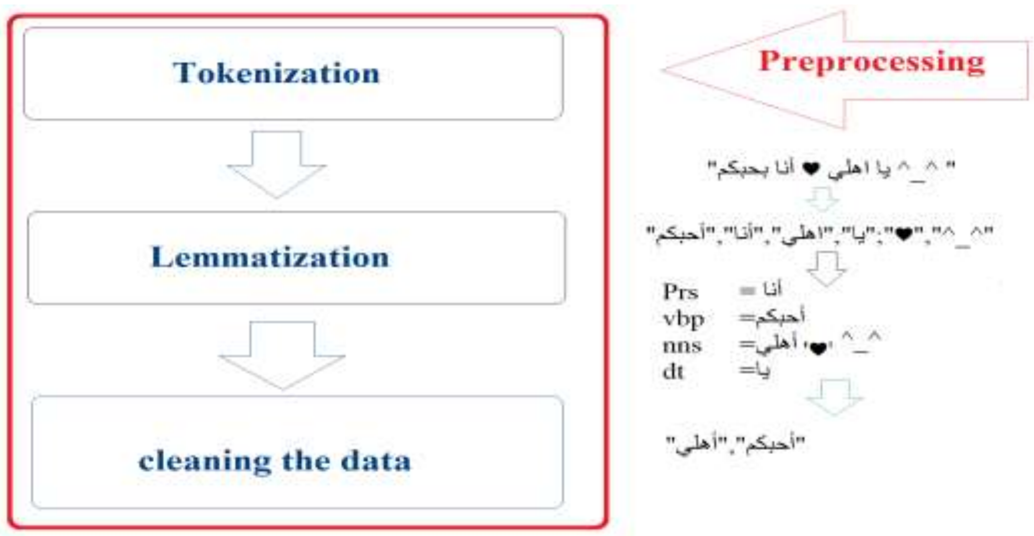


Figure 18 : Detailed architecture of EDIDA preprocessing steps LSTM Model

2.2.3. Splitting the Data

First, we need to split our data into two sets: Training and Testing sets. Train Data is data used in order to build and train our classification model. Test Data is data, that our classifier model has never seen before, used in order to assert the accuracy and test our classification model. As our data is currently ordered by label, we have to shuffle it before splitting it.

Random(140) randomizes our data with seed = 140. This guarantees the same shuffling for every execution of our code; altering this value or even omitting it to have different outputs for each code execution. we decided to split our data as 95% train data and 5% test data.

2.2.4. Training the Model

Outputting the model accuracy on the train and test data .Outputting the words that provide the most information about the emotion of a tweet. We can notice the high volume of negative to positive (0:1) informative features. This is it means that negative tweets have a much more concentrated and limited vocabulary when compared to positive tweets.

2.2.5. Testing the Model

Nevertheless, this model has various shortcomings. As the model only evaluates sentences at an independent word level, it performs very poorly when it comes to negations and other multi-word constructs. Finally, it is worth mentioning one more weakness of such a model: it does not generalize well. The model would perform greatly on data similar to the data it trained on. For example, if the model learned that I love football and I love cooking curry a positive emotion, it would be pretty easy for it to classify I love machine learning as a sentence carrying a positive sentiment. However, if the model were to classify I adore embeddings, it would most probably miss-classify it. As it never encountered any of these words before, therefore it is unable to properly classify them, and would simply output a random choice.

3. Deep Learning Model - LSTM

3.1. Data Pre-processing

In order to feed our text data to our LSTM model, we'll have to go through several extra preprocessing steps. Most neural networks expect numbers as inputs. Thus, we'll have to convert our text data to numerical data using Glove for Arabic language.

3.2. Word embeddings

Word embeddings are basically a way for us to convert words to representational vectors. What I mean by this is that, instead of mapping each word to an index, we want to map each word to a vector of real numbers, representing this word.

The goal here is to be able to generate similarly or close representational vectors for words that have a similar meanings.

3.3. Global Vectors for word Representation (Glove)

GloVe embeddings come in various flavors. They basically differ depending on the type of data they were trained on, the length of the vocabulary, the size of the representational vectors and so on.

note that the first steps in this process is similar to process previously preformed for the naive bayes model, Lr and SVC in Data transformation and building model and training model

4. Algorithme of preprocessing

```
def preprocess(textdata):
    processedText = [] # create list tweet processed
    wordLemm = WordNetLemmatizer() # word  Lemmatizer for lemmatize words of
tweets
    for item in textdata:
        a = remove_noise(item)
        word = wordLemm.lemmatize(' '.join(a))
        processedText.append(word)
    return processedText
```

Figure 19 : function proccess one tweet

```
def remove_noise(tweet_tokens):
    stemmer = ISRIStemmer()
   # tknzr = TweetTokenizer()
    tk = TweetTokenizer(reduce_len=True)
    mylist = str(tweet_tokens).split()
    res=[]
    for index in mylist:
      newList = [i for i in index if ((not i.isdigit()) )] # Remove digit
      newList = ''.join([i.lower() for i in newList if not i.startswith(('@', '#'))]) # Remove mentions and hashtags
      newList = re.sub(r"http\S+", "",newList) # Remove links
      newList = re.sub(r"-+", "",newList) # Remove links
      newList = stemmer.norm(newList, num=1) # remove diacritics
      newList = re.sub(r'[A-Za-z]','', newList)
      newList = re.sub(r'[^\w\s]','', newList)  # Remove punctuation
      newList = re.sub(r"""ّ    | # Tashdid
                          َ    | # Fatha
                          ً    | # Tanwin Fath
                          ُ    | # Damma
                          ٌ    | # Tanwin Damm
                          ِ    | # Kasra
                          ٍ    | # Tanwin Kasr
                          ْ    | # Sukun
                          ـ    # Tatwil/Kashida
                        """,'', newList)  # Remove punctuation
      newList = re.sub(r"[,]|[.]|[%]|[']|[#]|[é]|[â]|[à]|[ô]|[Ø]|[ê]|[è]|[í]|[ó]|[…]|[~]",'', newList)
```

```
    for e in t:
        if(len(re.sub("\s+", "", e))==1):
            continue
        r+=e+" "
    newList = ''.join(i for i, _ in itertools.groupby(r)) # Remove consecutive duplicate
    if(re.sub("\s+", '', newList)=='' or len(re.sub("\s+", '', newList))==1):
      continue
    a=tk.tokenize( newList)
   # print ("aaaaaazzz", a)
    res.append(a[0])
  PosTokens = [nltk.pos_tag(e) for e in [res]]
  chunks = nltk.ne_chunk_sents(PosTokens)
  aa=res
  return aa
```

Figure 20 : remove noises in tweets

5. **Architecture Models:** We adopted in These codes:

- **NB :** the architectur of naive baise model

```
from sklearn.naive_bayes import BernoulliNB # import library
BNBmodel = BernoulliNB(alpha = 2) #create model
BNBmodel.fit(X_train, y_train) #fitting model
```

Figure 21 : the architectur of naive baise model

- **LR** : the architectur of Logestic regression model

```
from sklearn.linear_model import LogisticRegression # import library

LRmodel = LogisticRegression(C = 2, max_iter = 10, n_jobs=-1) #create model

LRmodel.fit(X_train, y_train) #fitting model
```

Figure 22 : the architectur of Logestic regression model

- **SVC :** the architectur of sepport vector classification model

```
from sklearn.svm import SVC # import library
SVCmodel = SVC(probability=True) #create model
SVCmodel.fit(X_train, y_train) #fitting model
```

Figure 23 : architectur of sepport vector classification model

- **LSTM :** the architectur of Long short term memory model

```
def pretrained_embedding_layer(word_to_vec_map, word_to_index, max_len):
    vocab_len = len(word_to_index) + 1
    emb_dim = word_to_vec_map["unk"].shape[0]  # 50
    emb_matrix = np.zeros((vocab_len, emb_dim))
    for word, idx in word_to_index.items():
        emb_matrix[idx, :] = word_to_vec_map[word]
    embedding_layer = Embedding(vocab_len, emb_dim, trainable=False,
input_shape=(max_len,))
    embedding_layer.build((None,))
    embedding_layer.set_weights([emb_matrix])
    return embedding_layer
```

Figure 24 : embedding layer

6. Implementation.

6.1. Development environment:

Spyder : is a free and open source scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It features a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package. **[29]**

6.2. Hardware architecture: We developed our application on a PC with the following characteristics:

- o Processor: Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz 2.50GHz.
- o RAM: 8GB.
- o Operating system: Windows 10.

6.3. Language used:

6.3.1. Python :is an interpreted programming language; it runs on an interpreter system, where the code can be executed as soon as it is written, which offers rapidity in prototyping. Python has a simple, easy to learn syntax similar to the English language, this syntax allows developers to write programs with fewer lines than some other programming languages, and this syntax also emphasizes readability and therefore reduces the cost of program maintenance. **[30]**

6.3.2. Python libraries used:

- **Pandas** : is a library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical arrays and time series. Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool. **[31]**
- **NLTK** (**Natural Language Toolkit**): is an open-source platform and a suite of Python modules (libraries and programmes) for natural language processing. It is a leading platform for building Python programs to work with human language data. NLTK provides easy-to-use interfaces to over 50 corpora and lexical resources, along with a suite of text processing libraries for classification, tokenization, stemming, and tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. **[32]**
- **Scikit-learn** : is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities. **[33]**
- **Numpy**: is the fundamental package needed for scientific computing with Python. Besides its obvious scientific uses, NumPy can also be used as an efficient multidimensional generic data container. Arbitrary data types can be defined. This allows NumPy to seamlessly and quickly integrate with a wide variety of databases. **[34]**
- **TensorFlow** : is an end-to-end open source machine learning platform. It offers a comprehensive and flexible ecosystem of tools, libraries, and community resources

that allow researchers to advance in the field of machine learning, and developers to easily create and deploy applications that use this technology. **[35]**

- **PyQt4**: : PyQt **[36]** is a set of Python v2 and v3 bindings for the Qt application framework from The Qt Company and works on all platforms supported by Qt, including Windows, macOS, Linux, iOS and Android.

6.3.3 Google Colab : Google Colab or Colaboratory **[37]** is and is a free Jupyter notebook environment that runs entirely in the online cloud service. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members - just the way you edit documents in Google Docs. Colab supports many popular machine learning libraries which can be easily loaded into your notebook.

6.4 CSV extension file :

The acronym CSV means command-separared values**[38]** and designates a computer type computer file, the values of which are separated by commas:

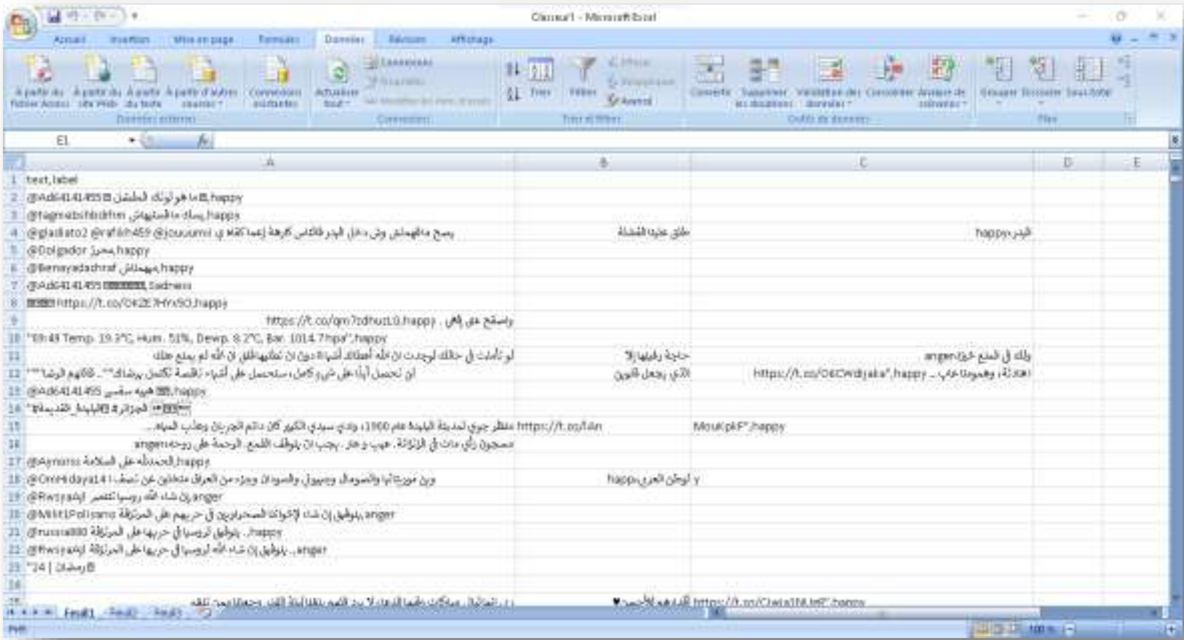


Figure 25 : Dataset of Algerian tweets

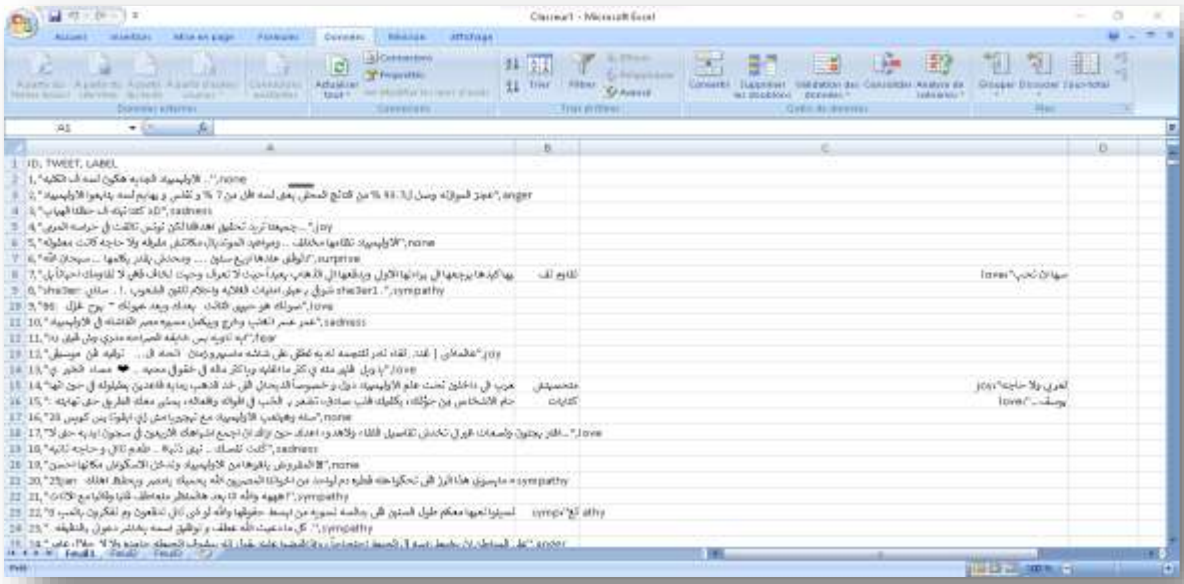


Figure 26 : Dataset of Egyptian tweets

6.5 Presentation interface of our system EDIDA

Our desktop application is with a simple user interface witch contined text zone to put in tweets for detect emotion from it, next ther is a menu to choose between Algeria dataset and Egypt dataset, and button userd for obtain the detection .In order to present the results in the area designated for it , As shown in the following figure :

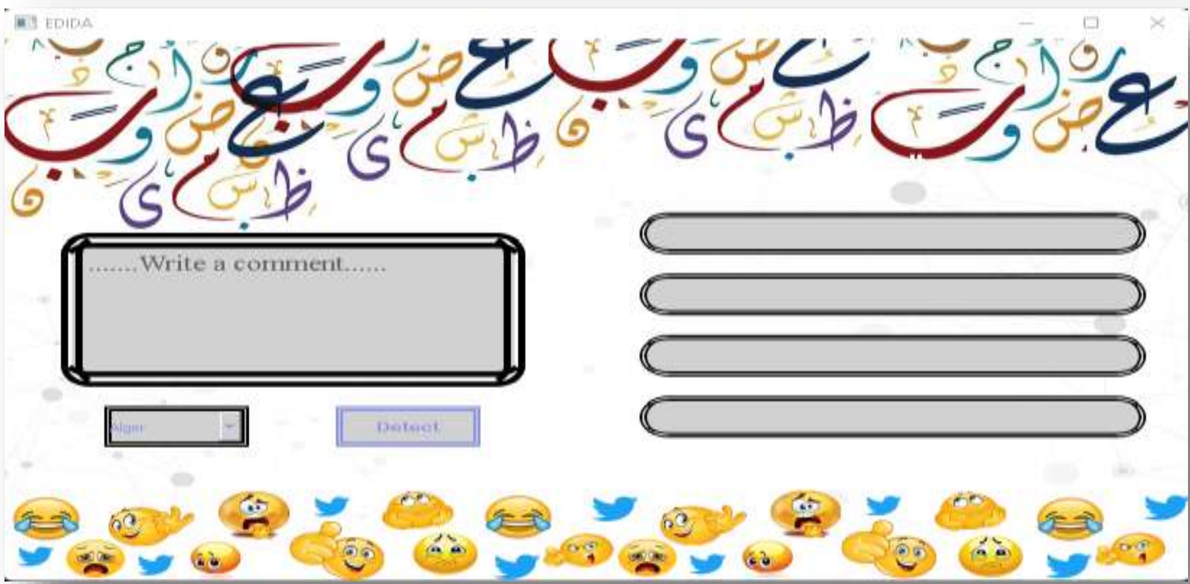


Figure 27 : User interface

After writing the tweet to be detected and classified, we define the dialect (as for Algerian or Egyptian), we click on Detect.

We see in the results, both the Algorthms used in detection, classifying the predominant feelings, and its percentage .As shown in the following figure :

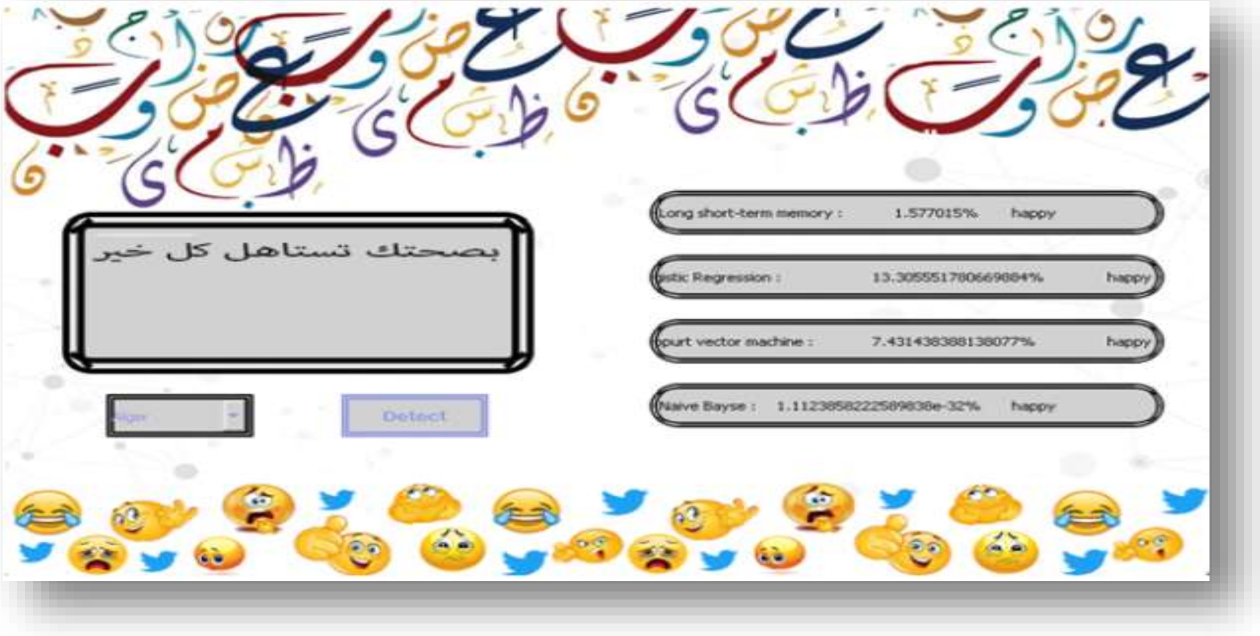


Figure 28 : Detect using dialect Alger

Figure 29 : Detect using dialect Egypt

These facades were created in order to facilitate the use of models.

7. Result and discussion

This section is intended to compare the performance of the proposed model described in previous section with the state-of-art Arabic emotion detection methods. Additionally, we compare our results of a deep learning model (LSTM) with three baselines prevalent in traditional machine learning, namely the support vector classifier (support vector machine), Naïve Bayes classifier and the Logistical regression classifier.

A. In Egyptian dataset:

Table 4 presents the comparison between our proposed approach EDIDA models on the Egyptian dataset. It is clear that our Logistical regression model improved the performance of emotion classification, and achieves an Accuracy of 57.73%, Naive bayes accuracy is 15.2%, SVC achieved 62.5% which outperforms the EDIDA models, and for our deep learning model LSTM its accuracy was 24.4%.

EDIDA Models	Accuracy
LR	0.5773809523809523
SVC	0.625
NB	0.15277777777777778
LSTM	0.2444

Table 4 : EDIDA models on Egyptian Dataset

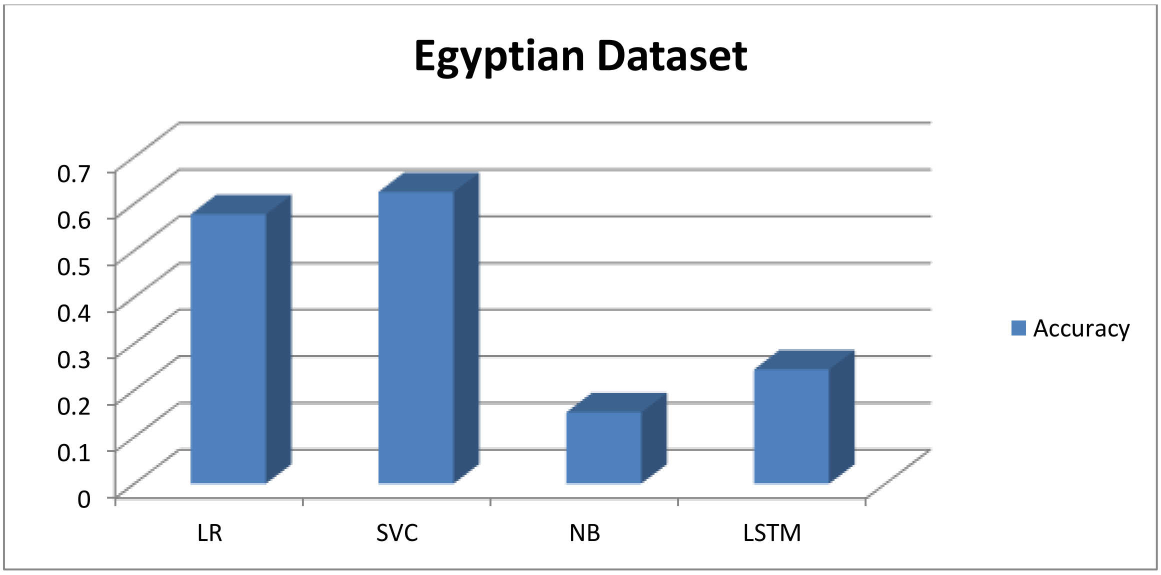


Figure 30 : EDIDA models on Egyptian Dataset

Table 4 and Figure 30 shows the detailed performance of our machine learning model on each of the four emotions in Egyptian data. We compare our results of the LSTM model with the three baselines prevalent in traditional machine learning, namely the Logistic Regression (LR) classifier, Naive Bayes (NB) and the Support Vector Machine (SVC) classifier. Our results in Table 4 consistently reveal superior performance through the use of our SVC model over the other models.

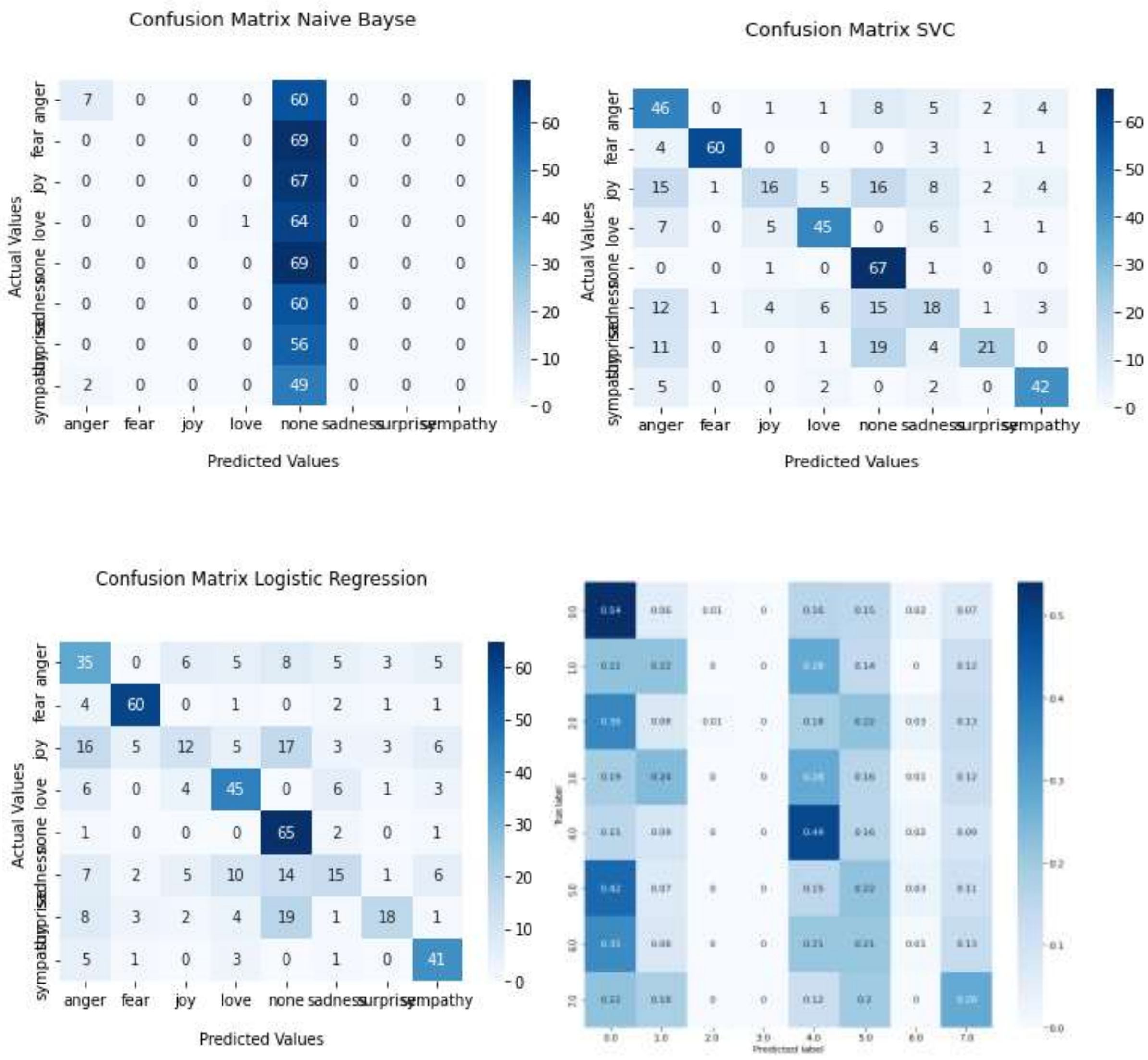
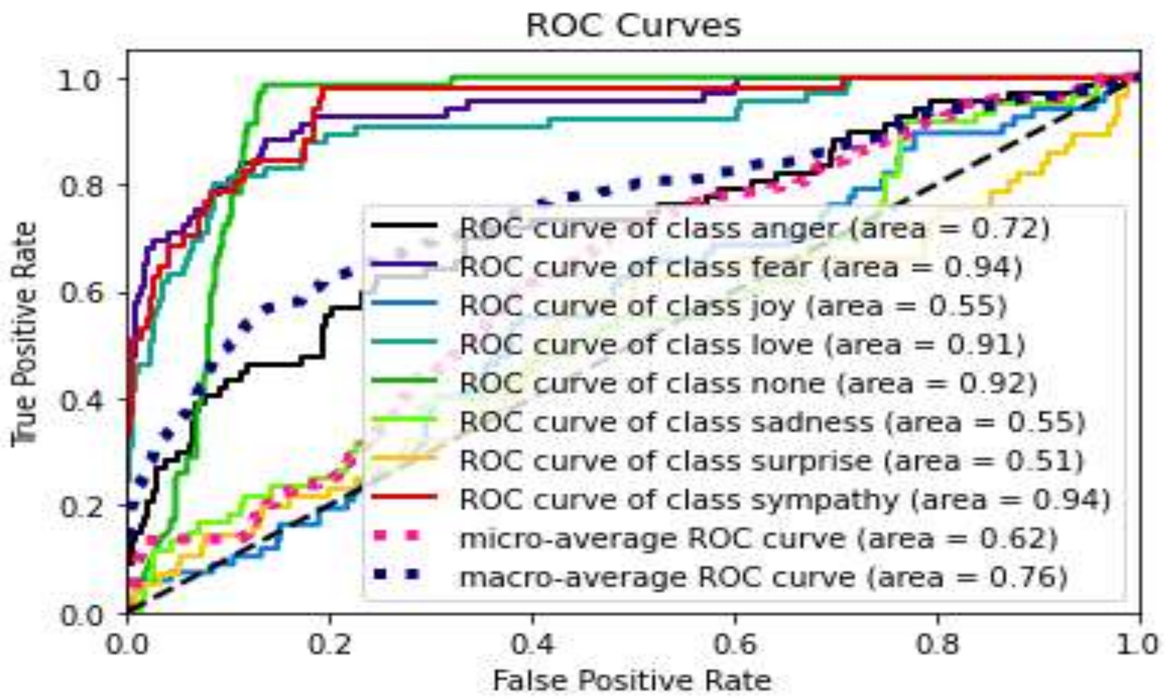


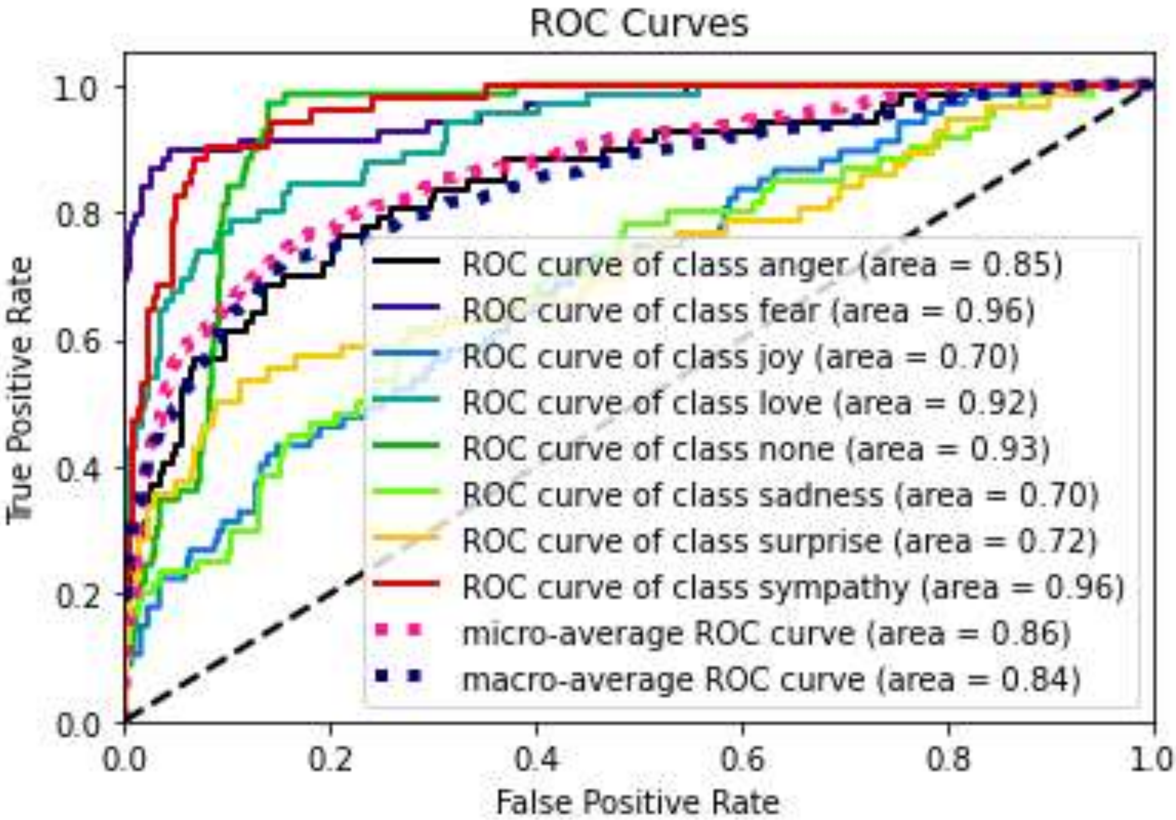
Figure 31 : confusion metrics of EDIDA (LR, SVC, NB and LSTM)

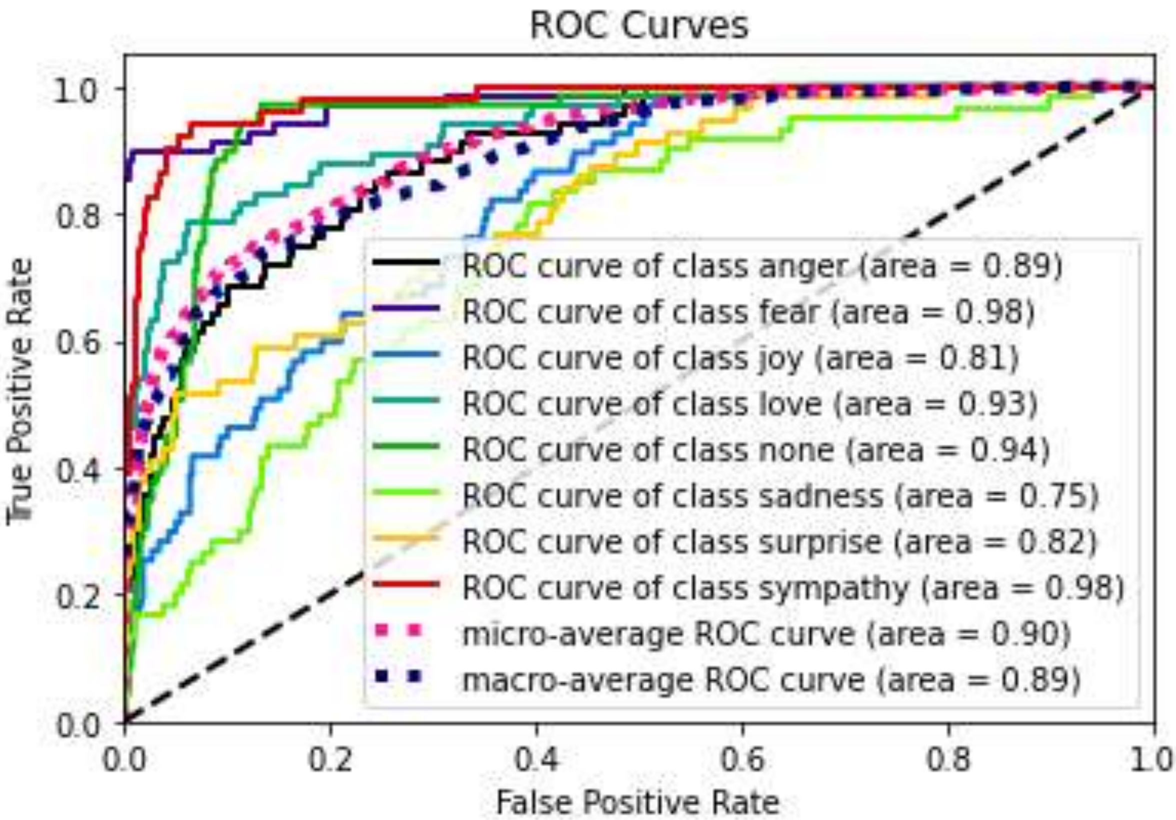
Table 5 : EDIDA MODELS

		precision	recall	f1-score	accuracy
DATASET EGYPTE	LR				
	Sadness	0.43	0.25	0.32	57.73%
	anger	0.43	0.52	0.47	
	joy	0.41	0.18	0.25	
	surprise	0.67	0.32	0.43	
	svc				
	Sadness	0.38	0.30	0.34	62.5%
	anger	0.46	0.69	0.55	
	joy	0.59	0.24	0.34	
	surprise	0.75	0.38	0.50	
	NB				
	Sadness	0.00	0.00	0.00	15.27 %
	anger	0.78	0.10	0.18	
	joy	0.00	0.00	0.00	
	surprise	0.56	0.00	0.00	

Additionally, In Figure 32 we calculate the ROC score for each baseline traditional machine learning method and Deep Learning model on Egyptian dataset used in the experiment. As shown, ROC scores of LSTM and NB generally have the lowest ROC score in the different datasets compared with the other machine learning model LR in general and SVC based model in particular.




ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class anger (area = 0.85)
ROC curve of class fear (area = 0.96)
ROC curve of class joy (area = 0.70)
ROC curve of class love (area = 0.92)
ROC curve of class none (area = 0.93)
ROC curve of class sadness (area = 0.70)
ROC curve of class surprise (area = 0.72)
ROC curve of class sympathy (area = 0.96)
micro-average ROC curve (area = 0.86)
macro-average ROC curve (area = 0.84)


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class anger (area = 0.89)
ROC curve of class fear (area = 0.98)
ROC curve of class joy (area = 0.81)
ROC curve of class love (area = 0.93)
ROC curve of class none (area = 0.94)
ROC curve of class sadness (area = 0.75)
ROC curve of class surprise (area = 0.82)
ROC curve of class sympathy (area = 0.98)
micro-average ROC curve (area = 0.90)
macro-average ROC curve (area = 0.89)

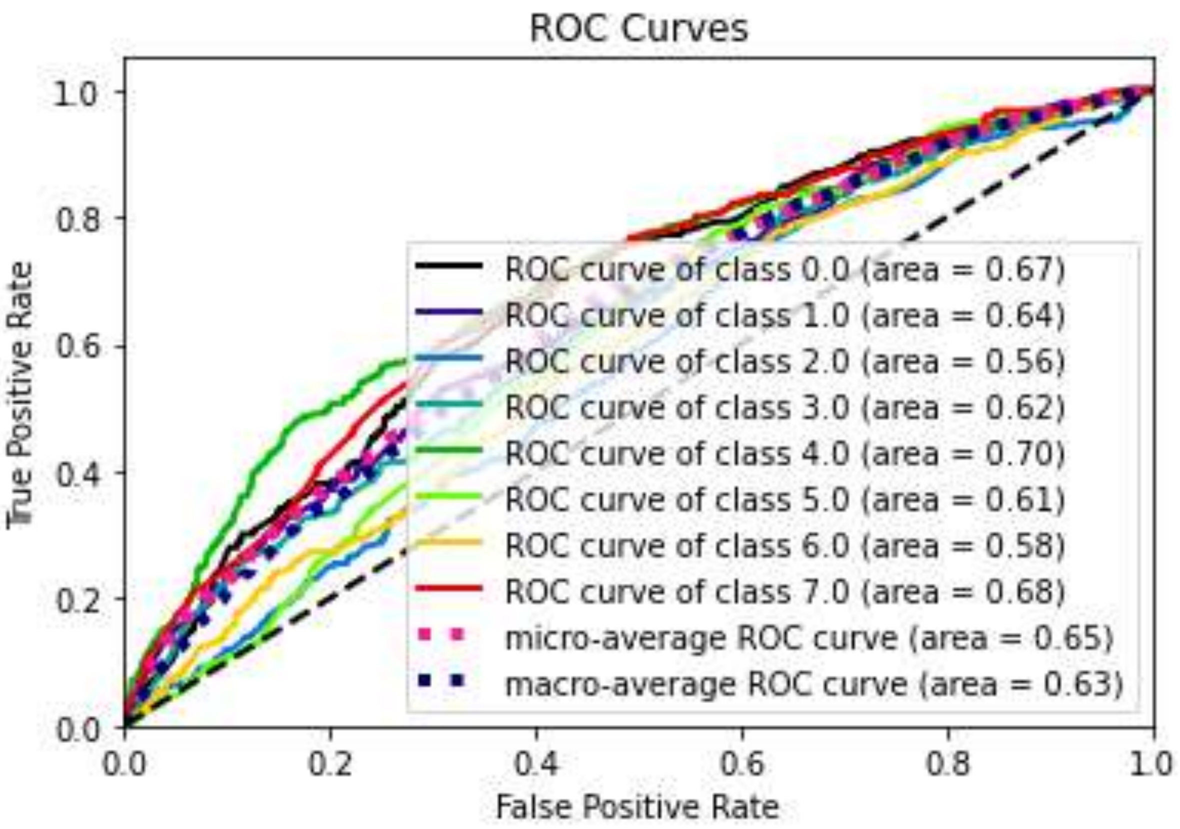


Figure 32 : NB, LR, SVC and LSTM ROC for Egyptian dataset

The performance for emotion classification varies from one approach to another while maintaining the same data (in this case we have the Egyptian data set), we note that the lstm-based approach ranks third compared to the rest of the approaches, while the second place goes back to LR, and the first place is for the classification based on SVC .

So the LSTM-based approach is not a good option in this case because it does not give good results, unlike SVC, which has provided good results

B. .In Algerian Dataset:

we compare our results of the SVC model with two baselines prevalent in traditional machine learning: the Naïve Bayes classifier and the Logistical regression classifier and deep learning model (LSTM).

Table 6 presents the comparison between our proposed approach EDIDA models on Algerian dataset. It is clear that our Logistical regression model improved the performance of emotion classification, and achieves an Accuracy of 87.42%, Naive bayes accuracy is 56.66%, SVC achieved 85.53% which outperforms the EDIDA models, and for our deep learning model LSTM(BI-LSTM) its accuracy was 61.48%.

Table 6 : Accuracy of Algothms

Algorthms	Accuracy
Naive bayse	0.5660377358490566
support vector machine	0.8553459119496856
Logistic Regression	0.8742138364779874
Long short-term memory	0.6148

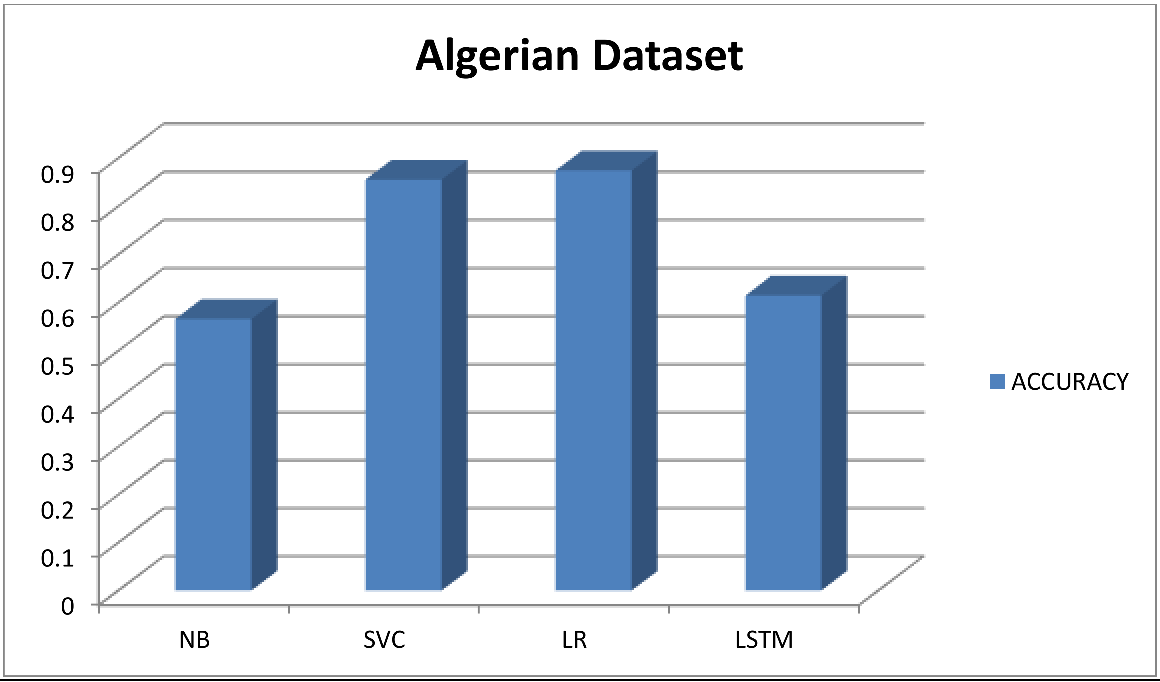


Figure 33 : Accuracy of Algothms

Table 6 and Figure 33 shows the detailed performance of our machine learning model on each of the four emotions in Algerian data. We compare our results of the Bi-LSTM model with the three baselines prevalent in traditional machine learning, namely the Logistic Regression (LR) classifier, Naive Bayes (NB) and the Support Vector Machine (SVC) classifier. Our results in Table 6 consistently reveal superior performance through the use of our SVC model over the other models.

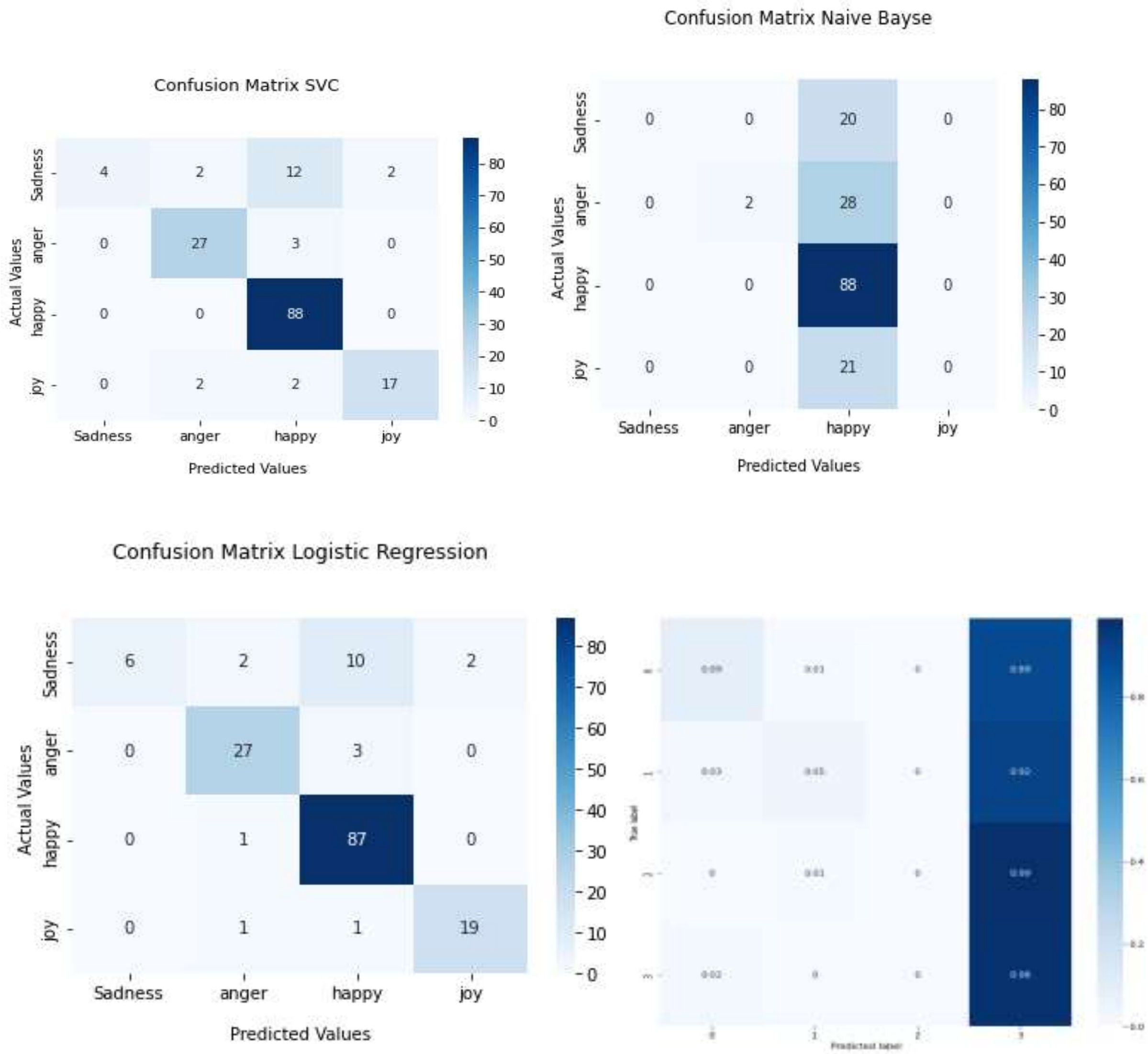
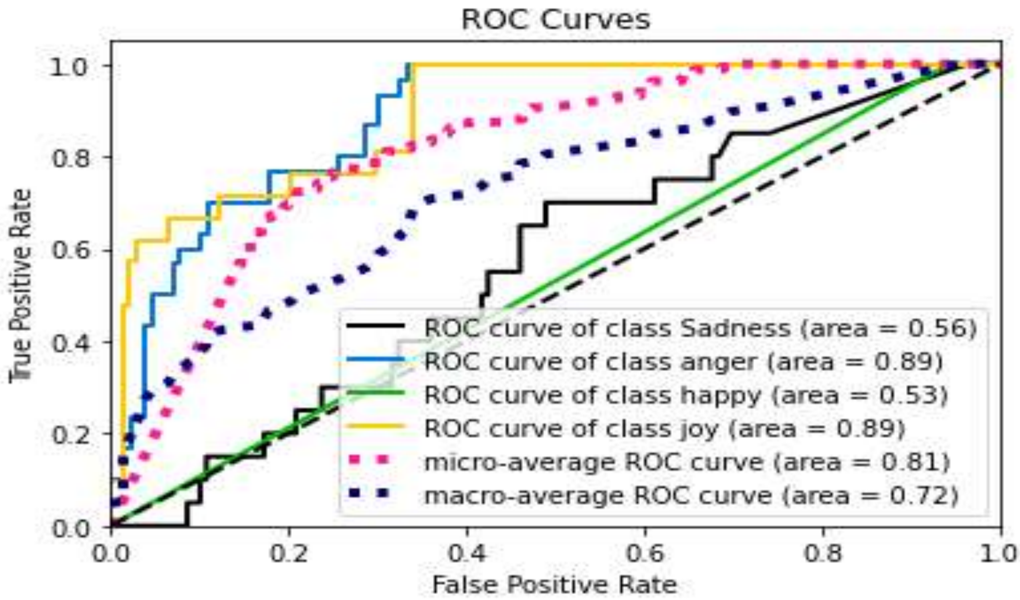


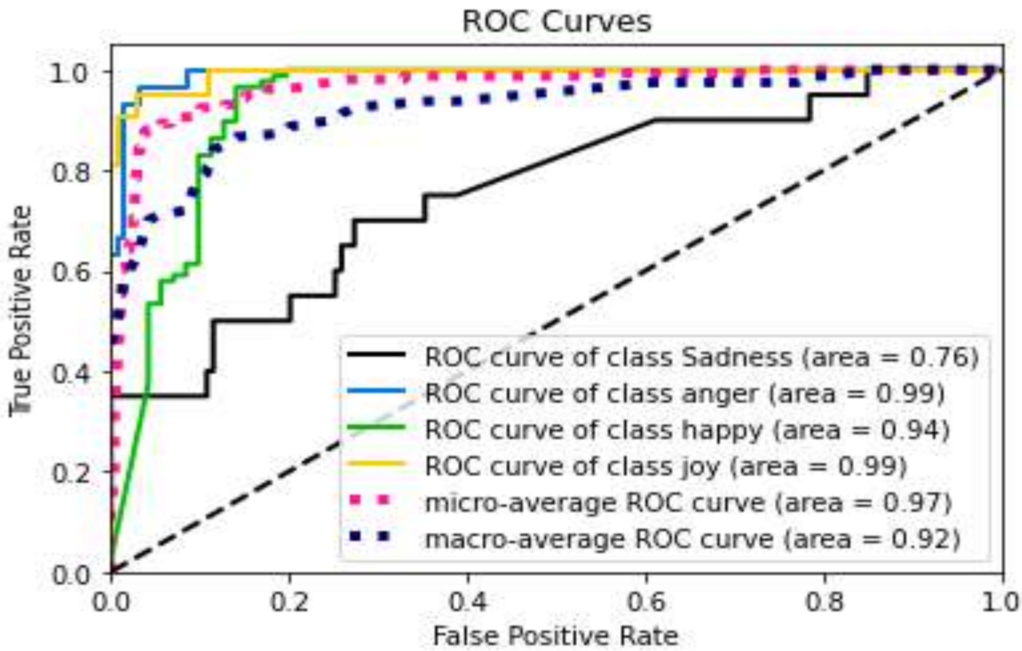
Figure 34 :confusion metrics of EDIDA (LR, SVC, NB and LSTM)

		precision	recall	f1-score	accuracy
DATASET ALGER	LR				
	Sadness	1.00	0.30	0.46	87.42%
	anger	0.87	0.90	0.89	
	joy	0.90	0.90	0.90	
	happy	0.86	0.99	0.92	
	svc				
	Sadness	1.00	0.20	0.33	85.53%
	anger	0.87	0.90	0.89	
	joy	0.89	1.00	0.91	
	happy	0.84	0.81	0.85	
	NB				
	Sadness	0.00	0.00	0.00	56.60%
	anger	1.00	0.07	0.12	
	joy	0.00	1.00	0.00	
	happy	0.56	0.00	0. 72	

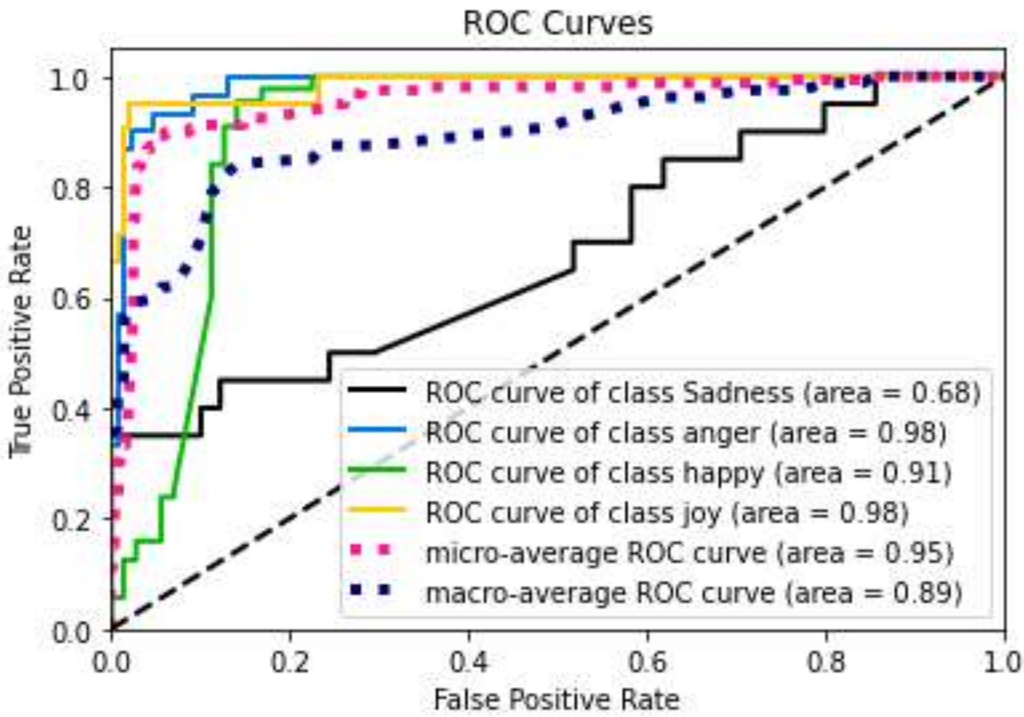
Additionally, In Figure 35 we calculate the ROC score for each baseline traditional machine learning method and Deep Learning model on Egyptian dataset used in the experiment. As shown, ROC scores of LSTM and NB generally have the lowest ROC score in the different datasets compared with the other machine learning model LR in general and SVC based model in particular.



SVC :



LR :



LSTM :

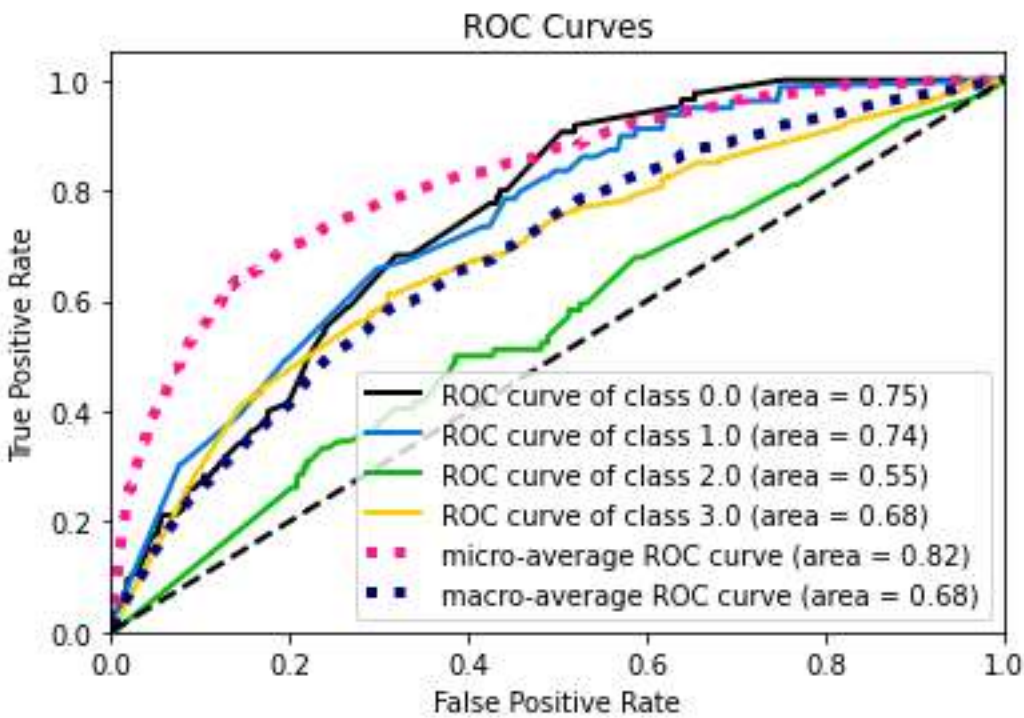


Figure 35 :NB, LR, SVC and LSTM ROC for algerien dataset..

The performance for emotion classification varies from one approach to another while maintaining the same data (in this case we have the Algerian data set), we note that the lstm-based approach ranks third compared to the rest of the approaches, while the second place goes back to SVC, and the first place is for the classification based on LR.

So the LSTM-based approach is not a good option in this case because it does not give good results, unlike LR, which has provided good results.

8. Conclusion

LSTM in Egyptian Dataset is Relatively weaker compared to LSTM in Algerian Dataset, but both of them are lees compared to some other approaches, and this is related to LSTM performance which is related to the reliability of Data.

General Conclusion

In this thesis, we have presented our system EDIDA that uses deep learning architectures LSTM for detecting the intensity of emotions in Arabic tweets. The performance of the system Doesn't surpasses the performance of the baseline's model but indicated that our approach is promising.

In this system, we uses word and document embedding models with feature vectors extracted from Algerian dialectical Arabic and Egyptian dialectical Arabic in multiple model LSTM, LR, NB and SVC which proved a high proficiency in detecting emotions.

where the main challenges in Arabic text-based approach for emotion detection are Arabic language traits, and Lack of data and tools. While currently most popular used methods in Arabic text-based emotion detection are classifiers, Long short term memory, Naive Bayes, Logestic regression and SVC. These methods proved to be the most effective Arabic textbased approaches for emotion detection.

There is no emotional lexicons that covers a wide variety of dialectical Arabic. Thus, much research can be carried out in the future that involve creating and annotating emotional lexicons for different dialects of Arabic language. Also, there is a need to annotate emotional tweets that can help training the model and improving the performance of prediction. we discovered some issues and challenges. Regardless of the fact that we recognize that the NLP field is rapidly evolving and that both Arabic NLP researchers and practitioners recognize the importance of incorporating Arabic into language technologies, the Emotion detection still needs more effort to drive the field of social media classification tasks more in that direction.

References

- **[1]**: Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch ,Wioleta Szwoch, “Emotion Recognition and Its Applications “, Advances in Intelligent Systems and Computing · July 2014.

- **[2]**: Ashritha R Murthy1, Anil Kumar K M2,” A Review of Different Approaches for Detecting Emotion from Text “in IOP Conf. Series: Materials Science and Engineering ,2021.

- **[3]**: Hamdy Mubarak, Kareem Darwish,” Abusive Language Detection on Arabic Social Media”.

- **[4]**: M.Abdullah, M.Hadzikadic and S.Shaikh,“SEDAT: Sentiment and Emotion Detection in Arabic Text using CNN-LSTM Deep Learning”in 17th IEEE International Conference on Machine Learning and Applications,2018.

- **[5]**: Fatma Mallek, Billal Belainine, Fatiha Sadat,”Arabic Social Media Analysis and Translation”, 3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017.

- **[6]**: Hume, D. Emotions and Moods. Organizational Behavior, 258–97.

- **[7]**: Emotion. . Available on: https://www.lexico.com/definition/emotion . (accessed 22/05/2022).

- **[8]**: Schacter, D.L., Gilbert, D.T., Wegner, D.M., & Hood, B.M. “Psychology” (European ed.). Basingstoke, Palgrave Macmillan , 2011.

- **[9]**: Tajul Islam, Romana Rahman Ema and Md. Humayan Ahmed,” Detecting Emotion from Text and Emoticon”.

- **[10]:** Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues. Available on: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723854/ (accessed 22/05/2022).

- **[11]**: Feeling. . Available on: https://www.merriam-webster.com/dictionary/feeling. (accessed 22/05/2022).

- **[12]**: Mood. . Available on https://www.merriam-webster.com/dictionary/mood . (accessed 22/05/2022).
- **[13]**: the experience of emotion. . Available on: https://opentextbc.ca/introductiontopsychology/chapter/10-1-the-experience-of-emotion/ (accessed 22/05/2022).
- **[14]**: ookson LJ. Differences between feelings, emotions and desires in terms of interactive quality. ASSRJ. 2015
- **[15]**:P.Ekman." An Argument for Basic Emotions" (1992)
- **[16]**: Arthur L Samuel. "Some studies in machine learning using the game of checkers". In: IBM Journal of research and development 3.3 (1959), pp. 210–229
- **[17]**: J. Mech. Cont.& Math. Sci., Special Issue, No.- 7, February (2020) pp 86-94
- **[18]**: Mitchell, T. (1997). Machine Learning. McGraw Hill. p. 2.
- **[19]**: Harnad, Stevan (2008), "The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence", in **[20]**: Epstein, Robert; Peters, Grace (eds.), The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer, Kluwer, pp. 23–66, 2012.
- **[21]**: "Introduction to AI Part 1". Edzion. 2020-12-08. Retrieved 2020-12-09.
- **[22]**: Russell, Stuart J.; Norvig, Peter (2010). Artificial Intelligence: A Modern Approach (Third ed.). Prentice Hall.
- **[23]:** Stuart J. Russell, Peter Norvig (2010) Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall
- **[24]:** Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012). Foundations of Machine Learning. The MIT Press.
- **[25]:** Hendrickx, Iris; Van den Bosch, Antal (October 2005). "Hybrid algorithms with Instance-Based Classification". Machine Learning: ECML2005. Springer. pp. 158–169.

- **[26]:** McCallum, Andrew. "Graphical Models, Lecture2: Bayesian Network Representation" (PDF). Retrieved 22 October 2019.

- **[27]:** Hinton, Geoffrey; Sejnowski, Terrence (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press.

- **[28]:** Uhmann, J.; Kuhnel, H. (1992). "Unsupervised and supervised data clustering with competitive neural networks". [Proceedings 1992] IJCNN International Joint Conference on Neural Networks. Vol. 4. IEEE. pp. 796–801

- **[29]:** Spyder. Available on: Home — Spyder IDE (spyder-ide.org) (accessed 22/05/2022).

- **[30]:** Python. About [online]. Available on: What is Python? Executive Summary | Python.org (accessed 21/05/2022).

- **[31]:** Pandas. Available on: pandas - Python Data Analysis Library (pydata.org) (accessed 22/05/2022).

- **[32]:** NLTK-Natural Language Toolkit. NLTK 3.5 Documentation [online]. Available on: NLTK :: Natural Language Toolkit (accessed 22/05/2022).

- **[33]:** Scikit-learn. Getting started [online]. Available on: scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation (accessed 22/05/2022).

- **[34]:** NumPy . Available on: NumPy (accessed 22/05/2022).

- **[35]:** Tensorflow . Available on: TensorFlow (accessed 22/05/2022).

- **[36]:** PyQt intro, Available on: Riverbank Computing | Introduction (accessed 22/05/2022).

- **[37]:** Colaboratory. Bienvenue dans Colaboratory [online]. Available on: https://colab.research.google.com/notebooks/intro.ipynb (accessed 22/05/2022).

- **[38]:** File CSV. Available on: https://debitoor.fr/termes-comptables/fichier-csv(accessed 22/05/2022).

📖 **[39]**: Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (2006). Semi-supervised learning. Cambridge, Mass.: MIT Press.

📖 **[40]**: Kaelbling, Leslie P.; Littman, Michael L.; Moore, Andrew W. (1996). "Reinforcement Learning: A Survey". Journal of Artificial Intelligence Research. 4: 237–285.

📖 **[41]** : Valueva, M.V.; Nagornov, N.N.; Lyakhov, P.A.; Valuev, G.V.; Chervyakov, N.I. (2020). "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation". Mathematics and Computers in Simulation. Elsevier BV. 177: 232–243.

📖 **[42]:** Dupond, Samuel (2019). "A thorough review on the current advance of neural network structures". Annual Reviews in Control. 14: 200–230.

📖 **[43]**: Goller, C.; Küchler, A. (1996). "Learning task-dependent distributed representations by backpropagation through structure". Proceedings of International Conference on Neural Networks (ICNN'96). Vol. 1. pp. 347–352.