**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: Industrial Computer Science

**Theme**

# Sentiment analysis of Algerian dialect

**Presented by:**

**SAID ABDELAZIZ**

**Before the jury :**

| | | |
|---|---|---|
| **Mrs. Nasra Bouhyaoui** | **Supervisor** | **HST Ouargla** |
| **Dr. MEZATI. Messaoud** | **Examinator** | **University of Ouargla** |
| **Dr. Bekkari Fouad** | **President** | **University of Ouargla** |

**2023/2024**

# Thanks

Firstly, I thank Almighty God for granting me the determination, willpower, and energy to complete this humble work.

I sincerely thank the supervisor of this work, Mrs. BOHIAOUI. N, for her full disposition, wise advice and guidance to me throughout the work.

I express my heartfelt gratitude to Dr. MEZATI. M for his advice, especially for the invaluable assistance he provided me.

I also thank each member of the jury for the honor they give me by agreeing to judge my humble work.

Before closing our acknowledgments page, I would like to thank my family for their patience and unconditional support, without which I would not have been able to complete my work.

I conclude by thanking my colleagues and all individuals who directly or indirectly contributed to their encouragement and moral support at the end of this project. Thank you very much from the bottom of my heart.

## Résumé

Alors que la technologie continue de progresser, l'analyse des sentiments est devenue l'un des domaines de recherche les plus importants en traitement du langage naturel et en apprentissage automatique. L'analyse des sentiments se concentre sur l'étude computationnelle des émotions et des sentiments exprimés dans les textes écrits. Les données sociales sont devenues l'une des sources de données les plus importantes dans ce domaine. Alors que la plupart des recherches actuelles se concentrent sur l'analyse des sentiments des textes en anglais, l'intérêt pour l'analyse des sentiments en arabe, en particulier le dialecte algérien, est limité. Dans ce travail, nous proposons un modèle d'analyse des sentiments pour la classification des dialectes algériens qui comprend deux étapes principales : la première étape est le prétraitement, où les données textuelles brutes sont nettoyées et les emojis sont traduits en texte. La deuxième étape est la classification, où trois algorithmes de classification sont appliqués au texte traité

**Mots-clés:** Analyse des sentiments, Dialecte algérien, Apprentissage automatique, Traitement du langage naturel (NLP), Réseau social en ligne

## Abstract

As technology continues to advance, sentiment analysis has become one of the prominent research areas in natural language processing and machine learning. Sentiment analysis focuses on the computational study of emotions and sentiments expressed in written texts. Social data has become one of the most important sources of data in this field. While most current research focuses on sentiment analysis of English texts, there is limited interest in sentiment analysis of Arabic, particularly Algerian dialect. In this work, we propose a sentiment analysis model for Algerian dialect classification that includes two main steps: the first step is preprocessing, where raw textual data is cleaned and emojis are translated into text. The second step is the classification, where three classification algorithms are applied to the processed text.

**Keywords**: Sentiment analysis, Algerian dialect, Machine learning, Natural Language Processing (NLP), online social network.

# Table of contents

# List of Tables



# List of figures

## General Introduction

The exponential growth of social media can facilitate interactions among users and foster various forms of expression. Social media platforms like Facebook , YouTube and Twitter have become one of the most common tools for individuals to express their thoughts, emotions, reviews, and comments. Additionally, social networking sites and virtual community platforms contribute vast amounts of data. This data can be valuable in numerous different aspects and crucial in high-precision professional fields through research methods.

Machine learning techniques play a vital role in leveraging the vast amount of data generated by social media platforms. These techniques involve the use of algorithms to analyze and interpret this data, enabling companies and researchers to extract valuable insights.

Sentiment analysis (SA) or opinion mining (OM) is an interesting research topic because knowing a person's positive, negative, or neutral feelings from the text they write is a crucial step in gathering information and making decisions. Sentiment analysis identifies the emotional tone behind a set of words or a body of text. It is an approach of natural language processing that involves the use of data analysis, machine learning (ML), and artificial intelligence (AI) to analyze texts for emotions and subjective information. The importance of this field is evident in a variety of domains and applications.

Numerous efforts have been proposed to overcome the challenges encountered in sentiment analysis. However, this task becomes increasingly complex when applied to data from social media platforms, known for their informal and highly noisy nature. Additionally, processing various human languages is not straightforward. While the English language garners significant attention from researchers in sentiment analysis, the same cannot be said for Arabic. There is a limited number of studies focusing on sentiment analysis in Arabic, with most concentrating on Modern Standard Arabic. Only a few delve into Arabic dialects such as Tunisian, Saudi, Algerian, and others.

Although there are many previous studies focusing on text classification in the context of sentiment analysis, there is an extreme deficiency of detailed datasets for sentiment analysis written in Algerian dialect. This deficiency is due to the removal of emoticons (emoji). The aim of our work is to create a model for sentiment classification of Algerian dialect, including the translation of emoticons into text. This model can classify comments into three categories: positive, negative, and neutral. The purpose of this work is the processing and the classification of social media comments written in Algerian dialect using classification algorithms, which are mainly related to sentiment. This work is organized as follows:
Chapter 01: we covered Machine learning from different aspects, beginning with its definition, types, algorithms, application, and deep learning.

Chapter 02: we presented different aspects related to sentiment analysis, definition, levels, application domains, approaches, then we introduced an overview about NLP and Arabic language and its challenges.

Chapter 03: we described our followed method for applying preprocessing techniques, dataset used and classification, and then we defined the used tools in the course of this works in addition to results and discussions.

# Chapter 01: Machine Learning

## 1.1. Introduction

Today, artificial intelligence has become a pivotal field in computer science, benefiting numerous sectors such as the economy, industry, and medicine significantly through the continuous advancement of this science. The progress of artificial intelligence has surpassed expectations. Just a few decades ago, having devices or software approaching human intelligence seemed impossible and unrealistic. However, currently, there are programs that surpass humans in various areas such as video games, decision-making, medical diagnostics, and more.

This remarkable progress in the field of artificial intelligence is closely linked to the rapid development of machine learning, which became the most renowned field in artificial intelligence from the second decade of the 21st century onwards, widely utilized by most researchers. In this context, this chapter will delve into the topic of machine learning and its algorithms.

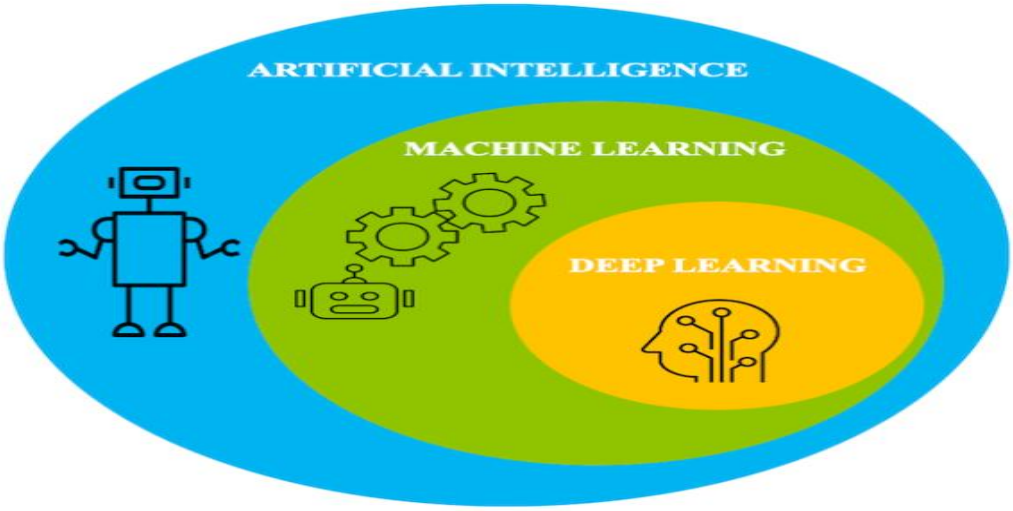



Figure1. 1 Artificial intelligence

## 1.2 Definition of Artificial Intelligence

Artificial Intelligence (AI) is a computer science discipline dedicated to developing intelligent machines with the ability to execute tasks traditionally associated with human intelligence. It encompasses the emulation of human cognitive processes, including perception, reasoning, decision-making, and learning. This is achieved through the application of formal methods such as logical thinking, planning, search, machine learning, and natural language processing. Intelligent machines can be tailored to handle a diverse range of functions, spanning from speech and image recognition to intricate decision-making and the operation of autonomous vehicles. The ultimate objective of artificial intelligence is to design machines capable of autonomously tackling complex problems through reasoning and learning methodologies, minimizing the need for direct human intervention. [1]

## 1.3 Definition of machine learning

Machine learning, as a subset of artificial intelligence (AI), empowers systems to autonomously learn and enhance performance through data-driven processes, eliminating the need for explicit programming. This approach has proven successful in diverse domains, including pattern recognition, computer vision, spacecraft engineering, finance, entertainment, computational biology, as well as biomedical and medical applications.

### 1.3.1 Types of machine learning

There are three types of machine learning: supervised learning, unsupervised learning, and reinforcement learning.

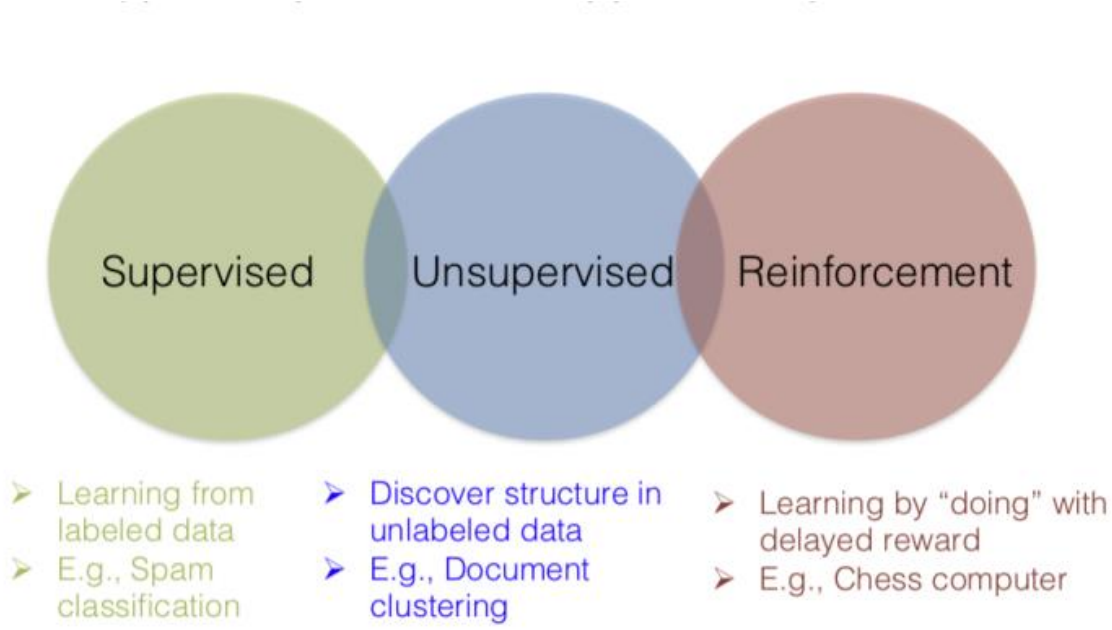



Figure1. 2 types of machine learning [2]

#### 1.3.1.1 Supervised Learning

Supervised learning algorithms construct a mathematical model based on a dataset that includes both inputs and desired outputs [3]. This dataset, known as training data, consists of a set of examples where each example in the training set contains one or more inputs along with the desired output. In the mathematical model, each example is represented by a table or vector, sometimes referred to as a feature vector and the training data is represented by a matrix. Through the training examples, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs [4]. The optimal function provides the algorithm with the ability to accurately determine the output for inputs that were not part of the training data.

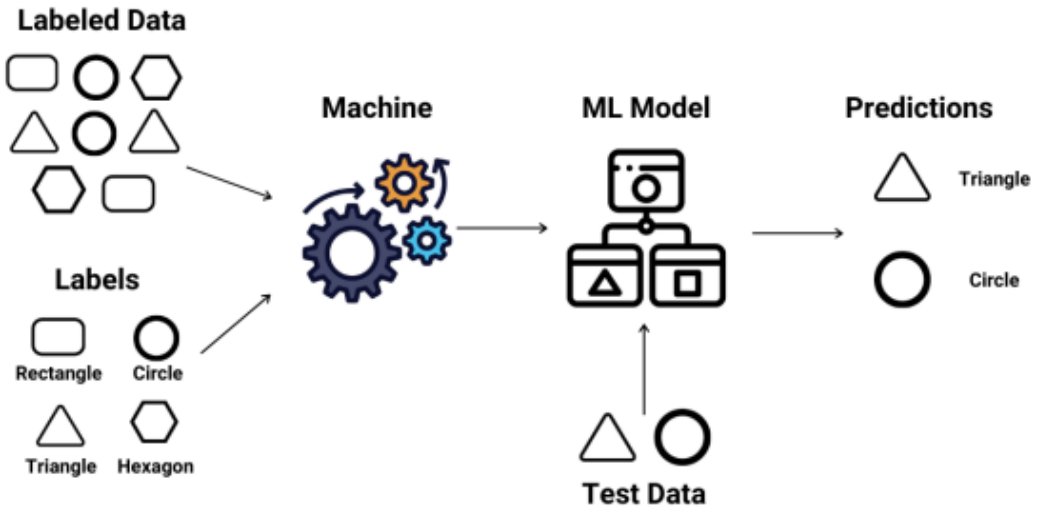



Figure1. 3 supervised learning workflow [5]

In supervised learning, we have two types of algorithms; Regression and classification.

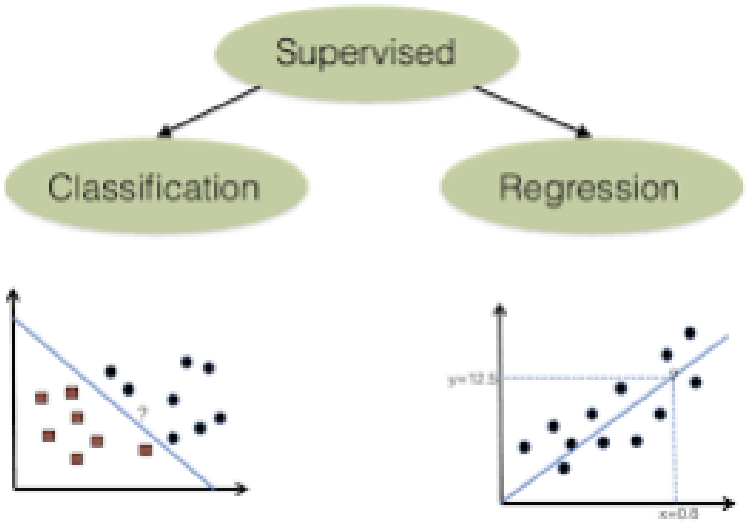



Figure1. 4 The two main subcategories of supervised learning [5]

### Regression

Regression is used to predict a continuous numerical value based on a set of input variables. Several different models can be used, such as polynomial regression, SVR (Support Vector Regression), and regression trees. For example, it can be employed to estimate the price of a house based on factors such as its size, location, and age.

### Classification

Classification is utilized to group similar data points into different categories for labeling. Machine learning is employed to discover rules that explain how to separate various data points. there are multiple ways to uncover rules, all of which focus on using data and responses to discover rules that linearly separate the data.

Linearity in separability is a key concept in machine learning. All it means is, "Can data points be separated by a line?" So, classification methods strive to find the best way to separate data points using a line. The lines drawn between categories are known as decision boundaries. The entire selected region to define a particular category is referred to as the decision surface. The decision surface dictates that if a data point falls within its boundaries, it will be assigned to a specific category. The most well-known classification algorithms are SVM (Support Vector Machine), KNN (K-Nearest Neighbors), and neural networks.

#### 1.3.1.2 Unsupervised Learning

Unsupervised machine learning algorithms are used when the information used to train the model is unclassified or unlabeled. In this case, the model studies its training data with the aim of extracting a function to describe a hidden structure within this data. At any given time, the system does not know the correct output with certainty. Instead, it uses inferences derived from datasets to determine what the output should be.

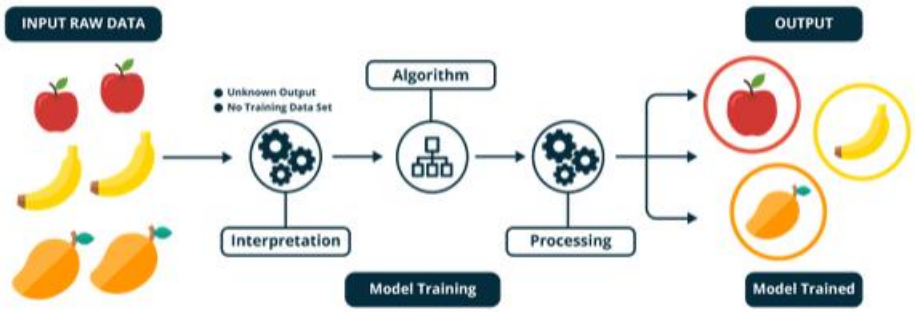



Figure1. 5 unsupervised learning workflow [5]

Common examples of unsupervised learning include clustering, dimensionality reduction, data mining, and association.

**Clustering:** The clustering process involves grouping similar data points into distinct clusters without providing any classification information to the algorithm.

**Dimensionality reduction:** Aims to simplify the data by reducing the number of input variables while preserving essential information.

**Data mining:** involves extracting useful information from large datasets, such as frequency patterns, correlations, and trend models.

**Association:** is a problem where one seeks to solve it by discovering rules that describe significant portions of its data. For instance, in the context of studying the purchasing behavior of a group of customers, it becomes apparent that individuals who purchase a specific product also tend to buy another particular product. [5]

#### 1.3.1.3 Reinforcement learning

Reinforcement learning is a branch of the field of machine learning that focuses on directing the performance of software agents in a specific environment to achieve maximum understanding of a particular concept by accumulating rewards. Due to its diverse applications, this field is explored in various other disciplines such as game theory, control theory, practical research, information theory, simulation-based optimization, multi-agent systems, statistics, and genetic algorithms. Reinforcement learning algorithms do not require knowledge of an exact mathematical model; instead, they are used when creating precise models is impractical. This type of learning is often utilized in areas like self-driving vehicles or learning how to play games against human opponents.

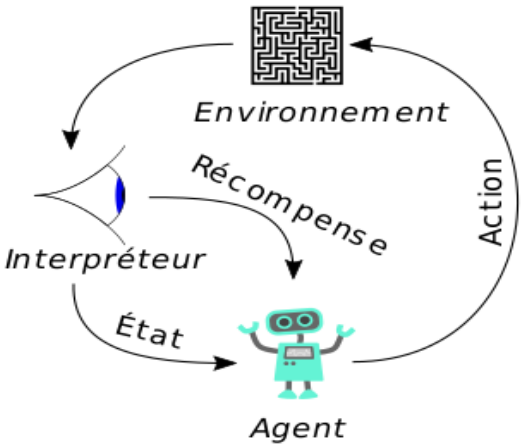



Figure1. 6 Reinforcement learning workflow.[5]

### 1.3.2 The machine learning algorithms

Machine learning algorithms are computer methods that allow computers to learn from data and make decisions or take actions without being explicitly programmed to do so. There are several types of machine learning algorithms, each with its own characteristics and applications. Here are some examples:

**Artificial neural networks:** These algorithms are inspired by the functioning of the human brain and are used for tasks such as image recognition, automatic translation, or prediction of time series.

**Decision trees:** These algorithms allow decisions to be made by following a series of "yes/no" questions to arrive at a conclusion. They are often used for classification tasks.

**Clustering algorithms:** These algorithms allow similar data to be grouped into clusters. They are often used for market segmentation tasks or analysis of unstructured data.

**Regression algorithms:** These algorithms allow a numerical value to be predicted from input data. They are often used for prediction tasks, such as price or sales prediction.

**Classification algorithms:** Those are all common machine learning algorithms used for classification tasks:

- Support Vector Machine (SVM)
- Naïve Bayes
- Random forest
- Decision Tree
- Logistic regression

Each of these algorithms has its strengths and weaknesses, and their performance can vary depending on the nature of the data and the specific problem at hand. It's often a good idea to experiment with multiple algorithms and choose the one that performs best for your particular task.

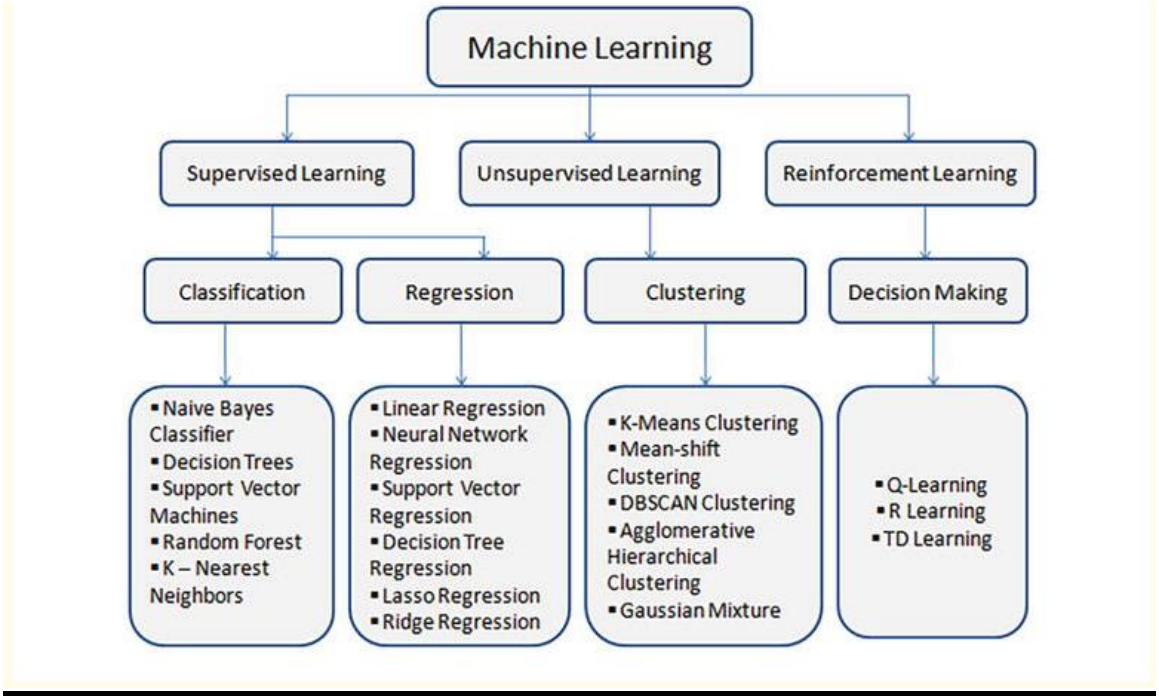



Figure1. 7 Machine Learning Algorithms [6]

### 1.3.3 Application of machine learning

Machine learning is a field of study and an approach to problem-solving, distinguished by the diversity of applications it can be applied to. Here are some key uses of machine learning techniques and methods:

**Natural Language Processing (NLP):** NLP involves understanding interactions between computers and natural languages. Machine learning can be applied to areas such as speech recognition, natural language understanding, and natural language generation.

**Insurance Claim Analysis:**Machine learning is used in the insurance industry to provide predictions about future claims and determine insurance premium costs, as well as to detect fraud.

**Bioinformatics and Medical Diagnosis:** Methods of machine learning are developed to efficiently store biological data and intelligently extract insights, aiding in the analysis of medical data and disease diagnosis.

**Image Processing and Pattern Recognition:** Machine learning can classify images and identify objects within them, facilitating pattern recognition without the need for specific programming for each object.

**Search Engines:** Machine learning is utilized in web search engines to improve search results and understand user queries, resulting in more accurate and effective search results. [7]

These examples represent only a fraction of the multitude of machine learning applications. The potential applications are limitless, and advancements in technology continually lead to the development of new applications in various domains.

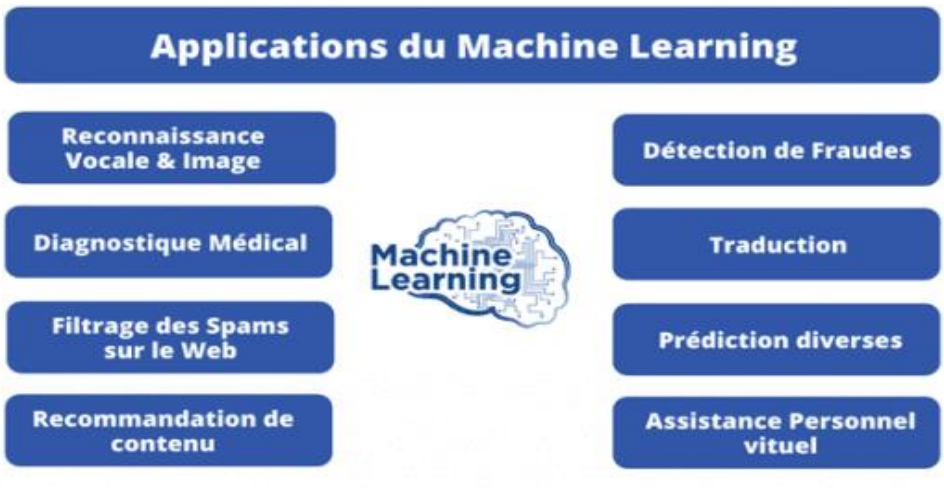



Figure1. 8 Applications of machine learning [8]

## 1.4 Deep learning

Deep learning, also often referred to as neural networks, refers to a set of algorithms and approaches inspired by the workings of the human brain. Deep learning architectures offer numerous benefits for text classification, as they can achieve excellent performance with low-level engineering and computation. The use of deep learning algorithms typically requires a large amount of training data compared to traditional machine learning algorithms. However, unlike traditional algorithms such as SVM and NB, deep learning classifiers do not require a learning threshold for the training data. Therefore, the more data used, the more effectively the algorithm is trained.

The most popular deep learning algorithms are: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), Deep Belief Networks (DBN), and many others. [9]

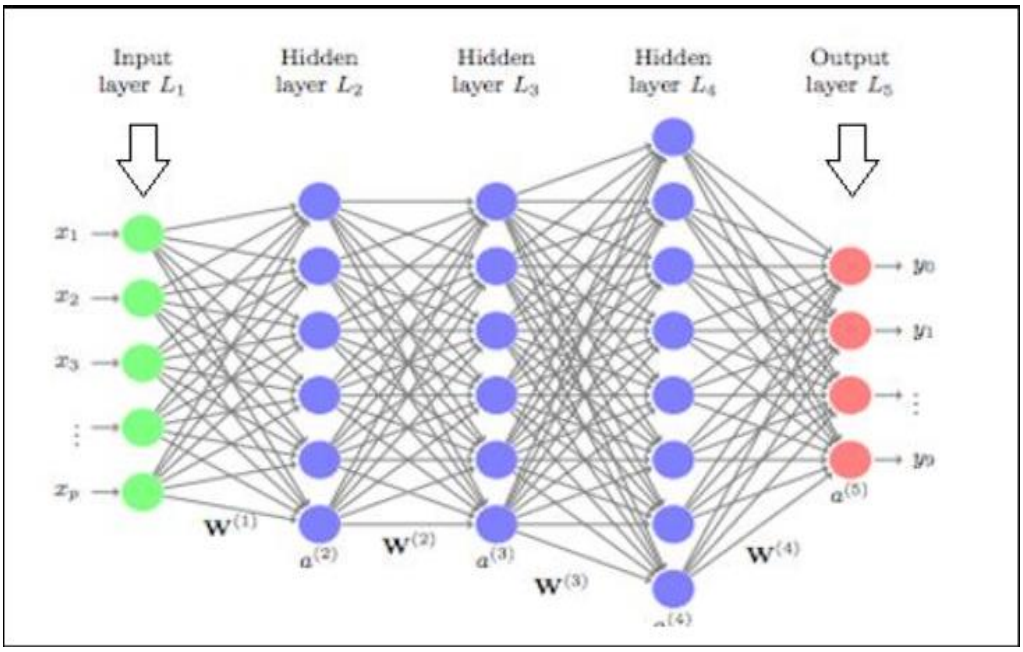



Figure1. 9 Deep neural network architecture [10]

## 1.5 Conclusion

In this chapter, we addressed the concepts of artificial intelligence and machine learning, providing a general overview of the fundamentals additionally, we reviewed various types of learning and algorithms associated with them, elucidating the crucial role of machine learning play in advancing the field of artificial intelligence and their positive impact on various technological domains. The following chapter provides an overview of Sentiment analysis, Natural Language Processing and Arabic dialects.

# Chapter 02: Sentiment Analysis

## 2.1 Introduction

Sentiment analysis, also known as opinion mining, is a branch of computational techniques that extracts and summarizes large amounts of data opinions that are challenging to process manually. This analysis begins in the realm of social media by examining current research trends and delving into the social and psychological factors contributing to social network interactions. [11] The complexity of sentiment analysis in these networks is highlighted due to their vast and constrained nature, short messages, noise, context dependence, and dynamics. The chapter also demonstrates how sentiment analysis tools can be applied in specific contexts and domains to achieve the best results in understanding the data.

Sentiment analysis relies on methodologies, theories, and techniques from various scientific and computational fields, including psychology, sociology, natural language processing, machine learning, big data, and statistics. In this context, the chapter provides a definition of sentiment analysis and explores its diverse application areas, with a focus on natural language processing and the challenges in understanding different languages. Additionally, the chapter discusses the approaches and techniques used in sentiment analysis and categorize previous works based on the employed techniques.

## 2.2 Sentiment analysis

Sentiment Analysis or Opinion Mining [12] is a technique aimed at determining the emotional tone behind a series of words. This involves building automatic tools capable of extracting subjective information from texts in natural language, with the goal of creating organized and exploitable knowledge that can be utilized by a decision support system or decision maker. Sentiment analysis is used to better understand the perceptions, opinions, emotions, and feelings expressed in online reports.

Firstly, it is essential to focus on the concept of sentiments. Simply put, sentiments express the feelings resulting from a comment or review and answer questions such as: Does someone support or oppose a specific topic? And do they like or dislike something?

Therefore, sentiment analysis is a process that includes examining, preprocessing, transforming, and classifying a dataset of natural language texts to extract the subjective information and determine its sentiment.

When a sentence in a text is objective, as in "It is raining," no additional basic task is required. However, when a sentence is subjective, as in "I love the rain!" the sentiment is usually described as a binary opposition, but it is often more complex. Some comments and reviews neither offer a positive nor negative opinion, known as a neutral opinion. Therefore, classifying whether the sentence expresses objective information or subjective views and opinions is the task of subjectivity classification, and determining the polarity is the task that identifies sentences expressing positive,

negative, or neutral polarities. Sentiment analysis encompasses various other tasks such as opinion retrieval, opinion summarization, opinion holder identification, topic/sentiment dynamics tracking, opinion spam detection, and many more.

## 2.3 Sentiment analysis levels

To apply sentiment analysis, we need to define the text that will be analyzed in the case of a study. Sentiment analysis can occur at three general levels[13] :

### 2.3.1 Document level sentiment analysis

Analyzes review texts to determine whether they have a positive or negative sentiment. It supports any sentiment-bearing text and identifies the overall opinion of the document. Document-level sentiment analysis assumes that each document expresses opinions about a single entity.

### 2.3.2 Sentence level sentiment analysis

Determines whether each sentence has expressed an opinion. This level distinguishes between objective sentences expressing factual information and subjective sentences expressing opinions.

In this type of sentiment analysis, it first identifies whether the sentence has expressed an opinion or not, and then assesses the polarity of that opinion.

### 2.3.3 Aspect-based sentiment analysis

Refers to categorizing opinions by aspect and identifies the sentiment related to each. First, a system identifies the attitude targets mentioned in a given sentence, a process known as aspect extraction. Once these aspects are identified, the system determines the attitude associated with each target in a process known as aspect-level sentiment analysis.

Rule-based strategies that leverage predefined text classifiers are a common technique for aspect extraction. Various approaches have been developed to understand the relationship between attitude targets and their context.

Sentiment analysis is of great importance in monitoring social media because it provides an overview of the public's opinions on specific topics. The ability to extract insights from social media data is a widely adopted practice by companies worldwide. Therefore, the use of sentiment analysis extends and proves to be highly effective.

## 2.4 Sentiment analysis applications

Sentiment analysis, also known as opinion mining, is a natural language processing technique used to determine the sentiment or subjective information expressed in a piece of text. This analysis can be valuable in various applications across different industries. Here are some common applications of sentiment analysis:

- **Social Media Monitoring:** Sentiment analysis is widely used by businesses to monitor social media platforms like Twitter, Facebook, and Instagram to gauge public opinion about their products, services, or brand in real-time. It helps companies understand customer feedback, complaints, and overall sentiment towards their brand.
- **Brand Monitoring and Reputation Management:** Brands monitor online mentions and conversations about their brand to manage their reputation effectively. Sentiment analysis helps in identifying positive and negative sentiment towards the brand, allowing companies to address issues promptly, respond to customer concerns, and maintain a positive brand image.
- **Financial Trading and Stock Market Analysis:** Sentiment analysis is employed in financial trading and stock market analysis to gauge market sentiment and predict stock price movements. By analyzing news articles, social media feeds, and financial reports, traders and investors can assess market sentiment and make investment decisions accordingly.
- **Political Analysis and Public Opinion Tracking:** Sentiment analysis is used in political campaigns and public opinion tracking to analyze sentiment towards political candidates, parties, and policies. It helps political organizations and policymakers understand public sentiment, identify key issues, and tailor their messaging and strategies accordingly.
- **Product and Service Reviews:** Sentiment analysis is applied to analyze product and service reviews on e-commerce platforms, review websites, and forums. It helps businesses understand customer opinions, identify product strengths and weaknesses, and improve overall product quality and customer satisfaction.
- **Healthcare and Patient Feedback Analysis:** Sentiment analysis is used in healthcare to analyze patient feedback, reviews, and surveys. It helps healthcare providers understand patient satisfaction levels, identify areas for improvement in service delivery, and enhance patient experience.

## 2.5 Sentiment analysis approaches

In recent years, the field of sentiment analysis has been well studied by researchers, with many different methods and techniques developed and tested through various tasks and at different levels. However, there is still much work to be done. Sentiment analysis is entirely different from simple text classification due to the numerous challenges in this field. Three types of techniques have been used to classify opinions: the machine learning-based approach, natural language processing (NLP) technique, lexical resources-based approach, and the hybrid approach.

### 2.5.1 The lexicon-based approach

Relies on uncovering a lexicon of opinion words for the analysis of textual content. Its objective is to ascertain the polarity of a text through the utilization of two distinct word sets: positive words signifying desired expressions and negative words representing undesired expressions. This methodology requires the utilization of a sentiment lexicon, which can be generated either manually or through a semi-automatic process.

In this model, the text undergoes evaluation by tallying the occurrences of positive and negative words. The cumulative count yields an overall assessment of the sentiment expressed in the text. The input text is tokenized using the Tokenizer of the NLP system, and each token is compared with the lexicon present in the dictionary or corpus. If a positive indication is identified, the result is incorporated into the total score for the input text, thereby increasing the overall score. Conversely, upon identifying a negative indication, the score is decremented, or the word is labeled as negative. The text is considered potentially neutral if the counts are equal.

This technique is guided by two distinct approaches: the dictionary-based method and the corpus-based method [9]. Both methodologies can be implemented using statistical or semantic techniques. The dictionary-based approach entails identifying opinion root or seed words and subsequently searching for their synonyms and antonyms in a dictionary. Conversely, the corpus-based approach commences with a roster of seed opinion words and delves into a substantial corpus to pinpoint opinion words with context-specific orientations.

### 2.5.2 Machine learning approach

Analyzing sentiments and categorizing text into positive, negative, or neutral categories necessitates the application of practical techniques. Hence, the adoption of machine learning techniques with their fully automated implementation and capability to handle extensive datasets becomes essential. Machine learning techniques prove to be highly effective for sentiment analysis, given the robust and accurate steps involved in training a dataset by learning documents and subsequently testing to validate performance. Various machine learning algorithms are employed for text classification.

Machine learning-based sentiment classification methods can be broadly categorized into supervised and unsupervised approaches. In sentiment analysis, supervised learning is predominantly utilized and typically involves four key steps: data collection, pre-processing, training data, classification, and result generation.

The process initiates with the compilation of training data in the form of a tagged corpus. Subsequently, the classifier undergoes training on this dataset, generating a series of feature vectors. Consequently, a model is created based on the training dataset, which is then applied to new text for classification purposes. The results are generated based on the selected representation type, and performance tuning and precision execution are conducted before the algorithm's release.

### 2.5.3 The Hybrid approach

Integrates the strengths of both previous methods, leveraging the accuracy of machine learning and the speed of the lexical approach. This strategy considers all the linguistic processing aspects of the lexical approach before initiating the learning process, similar to statistical approaches. By combining the high accuracy of machine learning with the stability of the lexicon-based approach, the hybrid approach achieves a balanced performance.

In lexicon-based approaches, tools and techniques rely on dictionaries and lexicons as primary sources for sentiment classification. These lexicons contain predefined semantic orientations that are subsequently compared with the input dataset for classification. On the other hand, machine learning-based approaches utilize learning algorithms to construct a training dataset. Subsequently, the inputs are compared and classified based on the trained dataset, determining whether they exhibit positive, negative, or any other sentiment.

## 2.6 Sentiment analysis challenges

The challenges related to sentiment analysis are diverse, presenting obstacles in understanding the precise meaning of sentiments and discovering the appropriate context for them. Some common challenges in sentiment analysis include subjectivity and tone, context and polarity, irony and sarcasm, comparisons, emoticons, defining neutrality, and many more.

As sentiment analysis relies on the application of natural language processing methods and algorithms, as well as text analysis techniques to identify and extract subjective information from text, the primary challenges are associated with the field of natural language processing. In the next section, we will delve into natural language processing and elucidate the implementation of its key algorithms.

## 2.7 Natural Language Processing (NLP)

NLP is currently utilized across a diverse range of applications, playing a pivotal role in various common tasks such as Language Translation, Word Processing, and Personal Assistant functionalities. Additionally, Sentiment Analysis stands out as a crucial field within NLP, focusing on extracting meaningful patterns from text data.

Understanding and manipulating language is a complex process, and NLP employs different techniques to address the challenges inherent in natural language. The primary methods employed in natural language processing are syntactic analysis and semantic analysis.

**Syntactic analysis:** involves identifying grammatical rules within a sentence to derive meaning by applying these rules to a group of words. This process illustrates how natural language aligns with grammatical structures.

**Semantic analysis**: involves analyzing the grammar of a sentence, encompassing tasks such as word segmentation, which divides text into units, and morphological segmentation, which groups words. Applying computer algorithms to comprehend the meaning and structure of words and sentences is crucial in semantic analysis, though it remains a challenging aspect of NLP.

Implementation of these techniques often relies on programming languages such as Java and Python. Python, recognized for its simplicity and robust capabilities in linguistic data processing, is frequently utilized. The Natural Language Toolkit (NLTK) in Python serves as a platform defining infrastructure for building NLP programs. It offers basic classes to represent human language data and includes text processing libraries for implementing various NLP algorithms. [14]

Despite the advancements, NLP remains a challenging problem in computer science. Human language's inherent complexity, including abstract language use, sarcasm detection, and the impact of stress on a sentence's meaning, poses difficulties. Additionally, the diversity of human languages, with their ambiguity and precise characteristics, further complicates NLP tasks for machines, making mastery a formidable challenge.

## 2.8 Arabic Natural Language Processing

Arabic serves as the official language in 22 countries and is spoken by a population exceeding 400 million individuals. It holds the distinction of being recognized as the fourth most utilized language on the Internet. Arabic can be categorized into three primary varieties.1)Classical Arabic(CA), a variant of the Arabic language employed in literary texts and the Quran. 2) Modern Standard Arabic (MSA), utilized for both written communication and formal discussions. 3) Arabic Dialects (AD) employed in everyday communication, informal exchanges, and similar contexts.[15]

Figure2. 1 Map of the Arab world [15]

### 2.8.1 Arabic dialects

Arabic dialects, also known as colloquial Arabic, are, in contrast, the informal native linguistic forms. They are typically limited in use to informal daily communication and are often not taught in schools or standardized, despite the rich culture of colloquial language, including folk tales, songs, movies, and TV shows. Dialects are primarily spoken and not extensively documented in writing [16]. Here are some common Arabic dialects:

- Egyptian Arabic: The language of Egypt and Sudan.

- Levantine Arabic: Encompassing Lebanon, Syria, Jordan, Palestine.

- Gulf Dialectal Arabic: Including languages spoken in Kuwait, the United Arab Emirates, Bahrain, Qatar, Saudi Arabia (with various sub-dialects), and Oman.

- North African Arabic (or Maghrebi Dialectal Arabic): Covering Morocco, Algeria, Tunisia, Mauritania, and Libya.

- Iraqi Dialectal Arabic: A hybrid of Levantine and Gulf dialects.

- Yemenite Dialectal Arabic: Often considered a distinct dialect.

On social media, the Arabic language may take different forms, being written in Arabic script or Romanized script (known as Arabizi).

### 2.8.2 Algerian dialect

Algerian Arabic, or the Algerian dialect, represents a group of North African Arabic dialects (Maghrebi) intertwined with various languages spoken in Algeria [17]. This dialect is significantly influenced by the Amazigh language, in addition to the influences of French, Classical Arabic, Modern Standard Arabic.[16]

**Challenges of the Arabic Language and Dialects**

Several challenges arise when dealing with the Arabic language, whether it is Modern Standard Arabic or Arabic dialects. Some of these challenges include:

- A single dialect, such as the Algerian dialect, may contain several sub-dialects.
- There is a significant difference between Modern Standard Arabic and Arabic dialects.
- The repetition of a letter several times to intensify the meaning or feeling.
- The presence or absence of diacritics, which can completely change the meaning of words.
- The use of negation words to deny actions in the past or present, completely altering the meaning.

## 2.9 Related works

There are numerous studies related to sentiment analysis in the Arabic language, with a focus on case studies of Arabic dialects. The Arabic language is characterized by a significant diversity of dialects, where various Arabic languages are used in most daily conversations, in addition to Modern Standard Arabic, which serves as the official language. With the emergence of social media and various electronic networks, Arab users can express their opinions using diverse Arabic dialects. Therefore, researchers have explored the need to understand the characteristics of the written forms of these diverse dialects by generating this content.[18]

Table 1: Related works

| REFERENCES | Dataset | Data cleaning | Normalization | Stop Words | Stemming |
|---|---|---|---|---|---|
| [19] | 100,000 comments from Facebook in Algerian dialect | - | -Similarity Regrouping | Arabic stop words | Phonetic Regrouping for dialect stemming |

| | | | | | |
|---|---|---|---|---|---|
| [20] | 7800 comments in Algerian dialect | -Deleted Diacritics | - Removing elongation<br>-Replace أ إ آ with ا<br>-Substitution of URLs by the <url> tag | - | Arabic stemmer |
| [21] | 22550 tweets in Jordanian dialect | -Deleted URLS<br>-Deleted Punctuation<br>-Deleted Diacritics | Normalization of Hamzaa & Alef & ya | Used Arabic list Stop words | Removing suffixes, prefixes, |
| [22] | 34576 tweets in Moroccan dialect | - | Strength of words by calculate the repeated letters | Moroccan Dialect 200 Stop words | - |
| [23] | 5,500 tweets in Saudi dialect | -Deleted Punctuation<br>-Deleted number | -Remove diacritics<br>-Replace أ إ آ with ا<br>-Replace ة with ه<br>-Replace ى with ي | Remove definite article (ال) Remove inseparable conjunction | -Remove suffixes and prefixes |

**a) Algerian Dialect**

- The work of Abdelli, A., Guerrouf, F., Tibermacine,O. in[19],Presents a supervised method for analyzing sentiments in the Arabic Algerian dialect. Two supervised methods are applied to a substantial annotated dataset.Their approach involves three phases:

  1. **Data Collection:** Gathering a substantial dataset from diverse Arabic Algerian sources and annotating the collected data.

  2. **Data Preprocessing:** Preparing the collected data through preprocessing.

  3. **Data Training:** Training and testing the two models.

  They gathered more than 100,000 comments from popular Algerian Facebook pages, categorizing over 10,000 comments as positive or negative. Additional datasets were utilized and combined into a comprehensive dataset. For Word2Vec, a large text corpus of 1.4 gigabytes was compiled.

After training the two models, the Support Vector Machines model exhibited superior performance compared to the Long Short-Term Memory model.

- The research [20] demonstrated the effectiveness of RNN and CNN in classifying comments related to the 2019 Hirak (a popular protest in Algeria during 2019) expressed in the Algerian dialect and retrieved from social networks. The experiment was conducted on a dataset of 7800 comments. The "comments preprocessing" phase comprises 11 stages:

1. Text segmentation into words.

2. Keeping only Arabic and French characters, deleting numbers, and removing unknown characters and extra spaces.

3. Substitution of URLs (≪http≫ or ≪https≫) by the <url> tag.

4. Substitution of user mentions and email addresses by the <user> tag.

5. Substitution of emoticons by the <emoticon> tag.

6. Hashtags (≪#≫) composed of concatenated words are substituted by their separated word version.

7. Lowercasing text.

8. Keeping stopwords and punctuation marks to avoid destructing possible obfuscated words.

9. Additionally, we carry out specific Arabic tasks such as unifying letters that are written differently. Example: {ء، أ، إ، آ} are substituted by {ا}.

10. Deleting all Arabic diacritics such as (fatha, damma, kasra, shadda, etc.).

11. Removing elongation as in (e.g., مـــــاذا becomes ماذا).

**b) Jordanian Dialect**

The study by Atoum and Nouman (2019) in [21] introduces a model designed for sentiment analysis of tweets in the Arabic Jordanian dialect. Their proposed model encompasses four distinct phases: Collecting Tweets, Tweets Extraction, Cleaning and Tweets Annotations, and Tweets Preprocessing.

During the "Collecting Tweets" phase, the objective is to amass a corpus of Jordanian dialect tweets. The "Tweets Extraction" phase involves extracting crucial content from each tweet. In the "Cleaning and Tweets Annotations" phase, the focus is on eliminating links and specific symbols from the collected tweets.

The "Tweets Preprocessing" phase comprises 5 stages:

1. **Cleaning Stage:**Tweets containing special symbols and various characters, such as emoticons, are assigned new classifications.

2. **Normalization:** Extra spaces are removed, and any unnormalized letter is replaced by its standardized form.

3. **Tokenization**

4. **Removing Stop Words:** Using Arabic list Stop words

5. **Stemming:** This involves removing any attached suffixes, prefixes, and/or infixes from words in the tweets.

For the classification task, the study employs Support Vector Machines and Naïve Bayes machine learning algorithms. To assess and compare the performance of these classifiers, the researchers conduct several experiments. The results indicate that the Support Vector Machines classifier consistently outperforms the Naïve Bayes classifier in terms of performance.

**c) Moroccan dialect**

Moroccan dialect using a machine learning approach: by Abdeljalil, Mohcine, Hafdalla, and Fatima-Zahra [22]. This research begins by gathering comments and annotations through crowdsourcing and a group of volunteers to determine the context of the comments, whether they are positive or negative. This task is followed by a text processing phase to extract abbreviated Arabic words to their roots. These words are utilized to construct input variables, automatically retrieved from the composite formed by preprocessed comments.

To classify Facebook comments, three supervised classification algorithms were employed (implemented using R software): Naïve Bayes (NB), Random Forests (FA), and Support Vector Machines (SVM).

**d) Saudi dialect**

Researchers Abdullah Mohsen, QubaylAlqubayl, and Abdulaziz A created a sentiment lexicon for analyzing sentiments in Saudi dialect tweets [23], which they named "SauDiSenti". This lexicon consists of 4431 words and phrases manually extracted from Modern Standard Arabic and Saudi dialects, based on previously categorized tweets collected from social media platforms in Saudi Arabia. Additionally, a test dataset containing 1500 tweets was evenly distributed into three categories: positive, negative, and neutral.

To construct the SauDiSenti lexicon, the researchers utilized a pre-classified dataset known as the "Saudi Dialect Twitter Corpus," comprising 5400 tweets in both Saudi dialect and Modern Standard Arabic. With the assistance of multiple annotators, all negative words and phrases provided by each annotator were added, duplicates were removed, and a score of -1 was assigned. The same procedure was applied to positive words and phrases, with a score of +1 assigned.

The performance of the SauDiSenti lexicon was evaluated based on four threshold values. A tweet is classified as positive if its score is greater than or equal to 0, strictly greater than 0, greater than or equal to 1, or strictly greater than 1. Otherwise, the tweet is classified as negative.

## 2.10 Conclusion

Sentiment analysis is a process where the positive, negative, or neutral context of a specific text is determined. This is often applied on social media platforms, particularly on user posts, comments, tweets, or even messages. In this chapter, we provided a definition of sentiment analysis. In addition to explaining and clarifying the wide application fields of sentiment analysis and its impact on our daily lives, we also defined the various methods of sentiment analysis. While sentiment analysis is a process applied to human-written texts, Natural Language Processing (NLP) was a significant focus in our chapter, where we explained how the NLP system operates along with

various techniques. Moreover, we discussed some of the broad challenges faced by NLP, especially in Arabic languages and their different dialects. Finally, we categorized some relevant works according to the methodology employed. The next chapter will review the Conception and implementation of our work and the associated results.

# Chapter 03: Conception and implementation

## 3.1 Introduction

Analyzing sentiment on social media data is considered important across a wide range of fields, yet it's not an easy task. In this chapter, we will delve into the details of our work, which revolves around sentiment analysis in the Algerian dialect. We will explain the preprocessing stages of the data, classification, and the application of some of its algorithms.

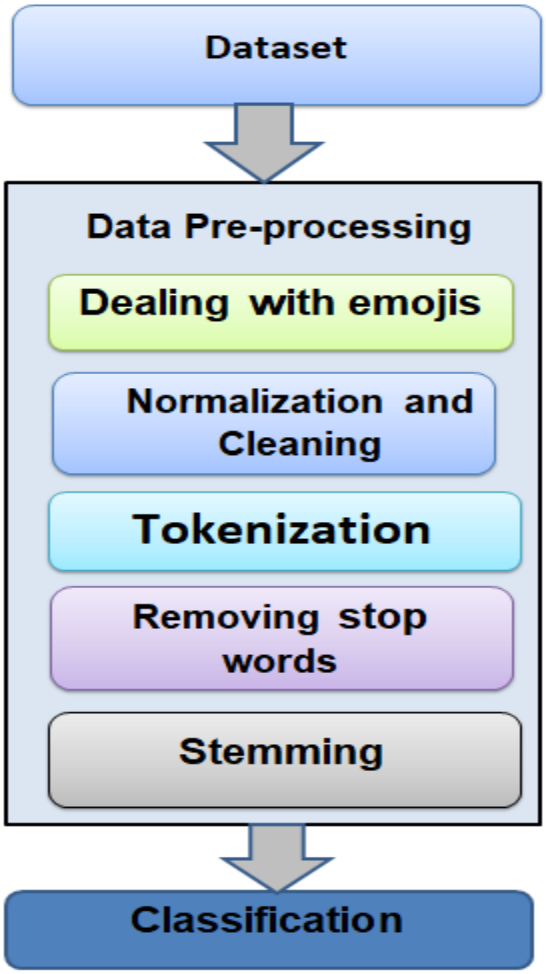



Figure3. 1 Global proposed architecture

## 3.2 Dataset

To save time, we utilized the dataset used in the work by Abdeldjalil Bouzenzana [24]. They built their own dataset in the Algerian dialect. The dataset contains approximately 27,800 manually labeled comments about the "Algerian car market," extracted from YouTube videos discussing the Algerian car market. The comments have been classified as positive, negative, or neutral, as shown in the figure, making it an ideal resource for sentiment analysis and market research.

```
1 df['sentiment'].value_counts()
```

```
negative    12122
neutral      8203
positive     7431
Name: sentiment, dtype: int64
```

Figure3. 2 Dataset statistics

## 3.3 Data Preprocessing

Preprocessing is a set of operations performed on raw data, aimed at facilitating analysis or training machine learning models. These operations are designed to clean the data, improve its quality, and prepare it in an appropriate format for use in various applications. Preprocessing involves a variety of steps, including:

- Dealing with emojis
- Normalization and cleaning
- Tokenization
- Removing stop words
- Stemming

### 3.3.1 Dealing with emojis

The emoji, small images or symbols, are used on the internet for adding additional emotions or meanings to text. Emojis first appeared in Japan in the late 1990s and since then have become an integral part of digital communication worldwide [25].

Most previous works related to sentiment analysis in the Algerian dialect removed emojis. In this work, we chose not to remove them but to translate them into text for two main reasons:

- Some comments consist only of emojis, and removing them would mean deleting the entire comment.
- Translating the emojis makes the comments clearer, which facilitates the classification process.

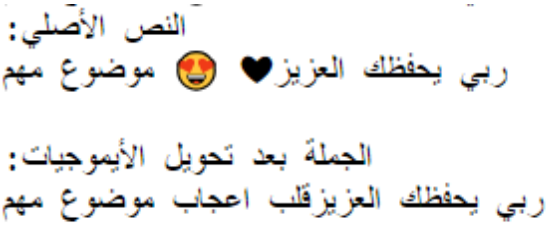
In this process, emojis are first converted into text in English, then translated into Arabic. After that, we remove repeated words like word (وجـــه) and replace long sentences with a single word that conveys the full meaning of the sentence.

To ensure this process is executed systematically, a dictionary was created with emojis as keys and their textual translations as values, as shown in the following figure.


Figure3. 3 Emojis translation file

Then, any emoji found in the sentence is replaced with its corresponding translation.

Figure3. 4 translate emojis into text




### 3.3.2 Normalization and Cleaning

Normalization is the process of converting textual data into a standard format to facilitate its processing. In this step, unnecessary words are reduced for vocabulary that doesn't provide any information about the polarity of the texts, by removing hashtags, usernames discovered by "@" and URLs starting with "https".

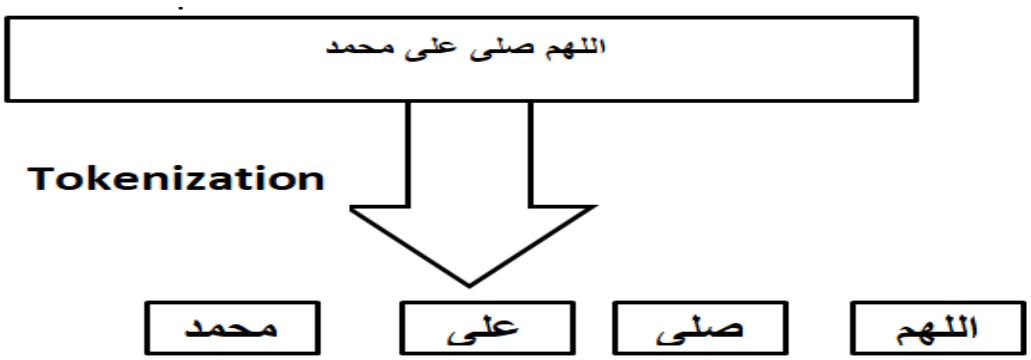
- replacing “hmazatedalif” (‘أ’,‘آ’ ;‘إ’) with “alif bare (‘ا’) ”.
- replacing “alif maqsura ”(‘ىيييييي’) with “yaa (‘ي’)“.
- replacing repeated letters with a single letter, for example: (‘اااااا’)with(‘ا’).

- Deleting all Arabic diacritics ’Tashkeel’ like( fatha, damma, kasra, tashdid ...etc.).
- Deleting elongationlike (‘الــعربــيــة’into ‘العربية’).

### 3.3.3 Tokenization

Tokenization (Tokenization of words)is the process of breaking down a single string of text into a list of individual words or distinct tokens. Tokenization is done as shown in Figure:

Figure3. 5 Tokenization example




### 3.3.4 Removing stop words

Stop words are the words and letters that do not change the meaning of the sentence when removed, such as prepositions, conjunctions, demonstratives, and temporal and spatial adverbs. In this step, stop words are removed. First, all single characters are deleted as they are considered empty words. Second, Arabic stop words are removed, and to identify these words, we use a natural language processing algorithm called stopwords from the NLTK library, which contains Arabic stop words such as prepositions and conjunctions, etc. As for the Algerian dialect, there are few lists containing its specific stop words. I have created a list of stop words for the Algerian dialect, and here are some of them:

['وهو', 'يا', 'يمالى', 'قل', 'كثر', 'ألي', 'ليك', 'ليكم', 'نتيا'].

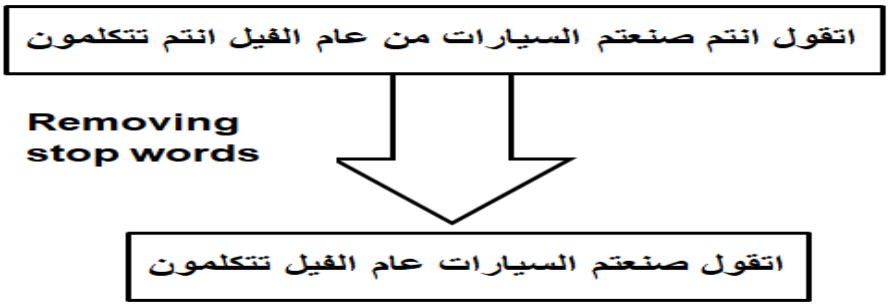
We remove stop words for several reasons, including:

- They do not contribute to classification.

- Increasing algorithm efficiency: Some machine learning algorithms and text analysis techniques are more effective when used with concise and focused texts.

- By removing them, we can reduce the amount of data that needs to be processed, which improves system performance and speeds up processing operations.

Figure3. 6 Removing stop words




### 3.3.5 Stemming

Stemming in Arabic involves removing all prefixes and suffixes from the word, converting plural forms to singular, or deriving verbs from their root forms to produce the stem or root. After identifying the root (stem) of words in natural Arabic, this is done by eliminating any affixes in the words, as Arabic words can have more complex forms with these affixes than any other language. In this step, each word in the list is stemmed, after tokenization, to its root or base form. We used the ISRIStemmer for this process.

ISRIStemmer is an algorithm for stemming Arabic words to their roots, developed as part of the NLTK (Natural Language Toolkit). This algorithm is specifically designed for processing the Arabic language and is considered one of the popular tools for analyzing and processing Arabic texts.

The simple example following explain the general idea of stemming: the words in " اخدم ربي يعونك" will be stemmed to "عون", "رب", "خدم".

### 3.3.6 Algorithm of Cleaning and preprocessing data

```
import re
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import ISRIStemmer
from pyarabic.araby import strip_tashkeel
import pandas as pd
def load_emoji_translations(file_path):
    emoji_translations = {}
    with open(file_path, 'r', encoding='utf-8') as file:
        for line in file:
            parts = line.strip().split(',')
            if len(parts) == 2:
                emoji, translation = parts
                emoji_translations[emoji] = translation
    return emoji_translations
emoji_translations = load_emoji_translations('emojis.txt')
stop_words = set(stopwords.words('arabic'))
stop_words.update([' راكم', 'واش', 'شحال', 'نشالله', 'معندوش', 'وشنو', 'ماهوش', 'لازم', 'عادي', 'بالزاف'
'خاصها', 'قاع', 'دائما', 'بلاش', 'كيفها', 'شكون', 'شويا', 'زيد', 'نهار', 'كيفاه', 'يعني', 'كيما', 'برك', 'هوما', 'حتى', 'مرة',
'هاك', 'هاذي', 'بزاف', 'معاكم', 'ياخي', 'بصح', 'شوف', 'والا', 'والو', 'هكذا', 'هكا', 'داير', 'واحد', 'قدام', 'لازم''])
stemmer = ISRIStemmer()
def translate_emojis_to_text(sentence):
    translated_sentence = sentence
    for emoji, translation in emoji_translations.items():
        translated_sentence = translated_sentence.replace(emoji, translation)
    return translated_sentence
```

```
def clean_text(text):
        text = strip_tashkeel(text)
        text = re.sub(r'ـ+', ' ', text)
        text = re.sub(r'ر' ,'+ر', text)
        text = re.sub(r'ا','+اا',text)
        text = re.sub(r'و','+ووو',text)
        text = re.sub(r'ههه','+ههه',text)
        text = re.sub(r'ة','+ةة',text)
        text = re.sub(r'ي','+ييي',text)
        text = re.sub(r'ا','أ',text)
        text = re.sub(r'ا','آ',text)
        text = re.sub(r'ا','إ',text)
        text = re.sub(r'ه','ة',text)
        text = re.sub(r'ي','ى',text)
        text = " ".join(text.split())
        return text
def clean_and_normalize_text(text):
        text = re.sub(r'#\S+|@\S+|https?:\S+', '', text)
        text = re.sub(r'[!""\$%&\'()*+,\-.\’/:;=?\\\[\\\\]^_`{|}~]', '', text)
        text = clean_text(text)
        text = re.sub(r'\w*\d\w*', '', text)
        text = re.sub(r'\d+', '', text)
        text = re.sub(r'[A-Za-z]', '', text)
        return text
def tokenize_text(text):
    tokens = word_tokenize(text)
     return tokens
```

```
def remove_stopwords(tokens):
        filtered_tokens = [word for word in tokens if word not in stop_words]
        return filtered_tokens


def stem_text(tokens):
        stemmed_tokens = [stemmer.stem(word) for word in tokens]
        return stemmed_tokens
df = pd.read_csv('dataset1.csv')
results = []
for index, row in df.iterrows():
      if pd.notnull(row['text']):
          sentence = str(row['text'])
          translated_sentence = translate_emojis_to_text(sentence)
          cleaned_and_normalized_sentence = clean_and_normalize_text(translated_sentence)
          tokens = tokenize_text(cleaned_and_normalized_sentence)


       tokens_without_stopwords = remove_stopwords(tokens)
      stemmed_tokens = stem_text(tokens_without_stopwords)
   results.append({
      'text': ' '.join(stemmed_tokens),
      'sentiment': row['sentiment']
   })
results_df = pd.DataFrame(results)
results_df.to_csv('res2.csv', index=False)
```

## 3.4 Text Vectorization

The scikit-learn library provides a transformer called Tf-idf-Vectorizer in the feature_extraction. text module to vectorize documents using TF-IDF values. The Tf-idf-Vectorizer works in the background by using the CountVectorizer estimator, which is used to produce bag-of-words encoding by counting the occurrences of tokens. It then uses the Tf-idf-Transformer to adjust these counts by the inverse document frequency.[26]

Figure3. 7 TfIdf vectorization

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer()

X_train_tfidf = tfidf.fit_transform(X_train)

X_test_tfidf = tfidf.transform(X_test)

```

## 3.5 Classification

Classification is a fundamental task in machine learning where the goal is to predict the category or class of a given input data point. This involves training a model on a labeled dataset, where each data point is associated with a specific classification. The model is then used to predict the class for new and unseen data points. In our work, we have relied on the following algorithms:

### 3.5.1 Naïve Bayes

Naïve Bayes is one of the recognized machine learning algorithms primarily used in classification tasks. It is a simple probabilistic classifier based on the application of Bayes' theorem, facilitating the rapid construction of machine learning models for quick predictions. The term "naive" in this context refers to the simplified assumption made by the algorithm that each feature is independent of the others.

As for Bayes' theorem, it describes the probability of an event occurring based on prior knowledge of the conditions associated with it. The formula for this theorem is as follows:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Where: P(A|B) represents the conditional probability of event A given that event B has occurred. It is also referred to as the posterior probability. P(A) and P(B) represent the probabilities of events A and B independently. [27]

### 3.5.2 Decision Tree

This is a type of supervised learning algorithm that is mainly used for classification problems. In this algorithm, we split the population into two or more homogeneous sets. This is done based on the most significant/independent attributes to create groups that are as distinct as possible [28]

### 3.5.3 Random Forest

In this algorithm, a collection of decision trees (referred to as a "forest") is used. To classify a new object based on its attributes, each tree provides a classification, and the tree "votes" for that class. The forest then chooses the classification that receives the most votes (across all trees in the forest) [29].

## 3.6 Working environment

### 3.6.1 Hardware environment

We developed our application on a PC with the following characteristics:

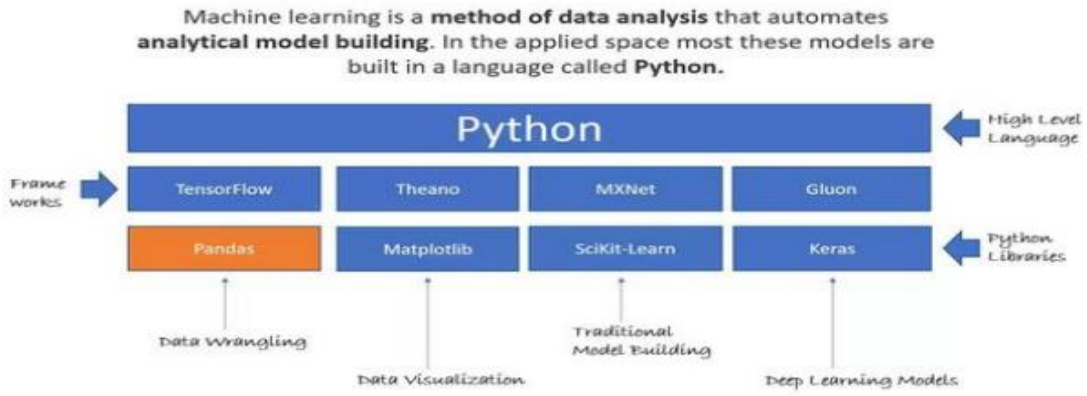
- ✓ Processor: Intel(R) Core(TM) i3-2328M CPU @ 2.20GHz 2.20 GHz
- ✓ RAM: 4GB.
- ✓ Hard drive: 256 GB
- ✓ Operating system: Windows 10.

### 3.6.2 Software environment and Libraries

- **Python language**

**Python** is a multi-paradigm programming language and the dominant programming language in data science with numerous implementations, making it even more interesting. Regarding the field of machine learning, Python stands out particularly by offering a plethora of high-quality libraries, covering all available types of learning, which combine ease of use and learning with the power of the libraries they possess.

Figure3. 8 Aperçu des Framework et libraires de python[30]




- **Anaconda**

Anaconda is a Python distribution designed for data science and machine learning tasks. It's a free, open-source software package that includes numerous libraries. One of the key benefits of Anaconda is its role as a centralized hub for essential libraries required for data processing, predictive analysis, and scientific computations.

- **Jupyter Notebook**

Jupyter Notebook is a programming environment that supports multiple programming languages, including Python. Jupyter Notebook allows us to create documents containing code snippets, equations, visualizations, and text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and more.

- **Libraries used**

**NLTK (Natural Language Toolkit):** is an open-source platform that includes a collection of Python modules (libraries and programs) dedicated to natural language processing. This platform is a leader in building Python programs that work with human language data. NLTK offers easy-to-use interfaces to over 50 corpora and lexical resources, as well as a variety of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. It also includes wrappers for robust NLP libraries and an active discussion forum.[31]

**RE (Regular Expressions):** library in Python is a built-in library used for Regular Expressions. It is used to search for specific text patterns within strings and perform operations such as searching, replacing, and verifying matches.

The `re` library provides powerful functions for regular expressions, allowing for the search of complex patterns and the identification of specific parts of the text. This feature is useful for text analysis, cleaning, and checking for various purposes such as data formatting, email validation, and many other uses.

**Pyarabic:** A specific Arabic language library for Python, provides basic functions to manipulate Arabic letters and text, like detecting Arabic letters, Arabic letters groups and characteristics, remove diacritics etc.[32]

**sklearn.linear** is a library in the Python language. It is one of the most useful libraries for machine learning. The sklearn library contains many effective tools for machine learning and statistical modeling, including classification, regression, clustering, and dimensionality reduction.

## 3.7 Source codes examples

In this section, we will present some examples of source codes.

The instructions for loading the dataset

Figure3. 9 Loading dataset

```
import pandas as pd
df = pd.read_csv('dataset1.csv')
df['text'].dropna(inplace=True)
```

Instructions for displaying the dataset can be seen

Figure3. 10 Dataset

| | sentiment | text |
|---|---|---|
| 0 | negative | فرنسيه تاعك طربق 😂 😂فات شهر والو |
| 1 | negative | واش انتم ولاد فرنسا |
| 2 | negative | ...قنوات الجزاير بدون ترجمه نتوما خير سبوره والله |
| 3 | negative | جاز شهر ماكانش الاسعار كيما قلتو علاه تكذبو |
| 4 | negative | اللغه الفرنسيه تجري عروقهم لاحول قوه الا بالله |
| ... | ... | ... |
| 27769 | positive | اذوما هوما لفيديو ليجبودا نحنا دزبريه |
| 27770 | positive | رب يحفظك خويا علي المعلومات القيمه |
| 27771 | positive | كون تعرف شحال نبغي لاقودا |
| 27772 | positive | مليحه وموتورها ميضيعش |
| 27773 | positive | حابه نشري |

27774 rows × 2 columns

The instructions for calling the Classifiers NB, Decision Tree (DT) and Random Forest (RF)

Figure3. 11 The training code for the NB classifier

```
 #Naive Bayes Model
from sklearn.naive_bayes import MultinomialNB

NB = MultinomialNB()
NB.fit(X_train_tfidf, y_train)
predictions_NB = NB.predict(X_test_tfidf)

```

Figure3. 12 The training code for the DT classifier

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
DT = DecisionTreeClassifier()
DT.fit(X_train_tfidf, y_train)
predictions_DT = DT.predict(X_test_tfidf)
```

Figure3. 13 The training code for the RF classifier

```
from sklearn.ensemble import RandomForestClassifier

#  Random Forest
RF = RandomForestClassifier()
RF.fit(X_train_tfidf, y_train)
predictions_RF = RF.predict(X_test_tfidf)
```

## 3.8 Experiment and evaluation

Sentiment Analysis (SA) can be considered a sentiment classification challenge from the perspective of machine learning, where the classification is multinomial in our case. Four metrics were used to present the experimental results: accuracy, precision, recall, and F1-Score. which are calculated as follows [33]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\mathrm{F1-Score} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

$$\mathrm{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

TP: True Positive TN: True Negative FP: False Positive FN: False Negative

We classified the dataset using Decision Tree DT, NB, and RF algorithms before and after preprocessing. We noticed an improvement in the results after preprocessing.

## 3.9 Discussion of Results Before and After Preprocessing

### 3.9.1 Presentation of Results

Below are the experiment results before and after data preprocessing:

Table 2: Results before and after Preprocessing

| Algorithms | Categories | Before Preprocessing | | | After Preprocessing | | |
|---|---|---|---|---|---|---|---|
| | | Precision (%) | Recall (%) | f1-score (%) | Precision (%) | Recall (%) | f1-score (%) |
| Random Forest | negative | 75 | 78 | 77 | 75 | 83 | 79 |
| | neutral | 65 | 50 | 56 | 62 | 55 | 58 |
| | positive | 67 | 80 | 73 | 81 | 74 | 77 |
| Accuracy (%) | | 70 | | | 75 | | |
| Decision Tree | negative | 73 | 66 | 69 | 73 | 70 | 71 |
| | neutral | 54 | 50 | 52 | 51 | 57 | 54 |
| | positive | 60 | 74 | 66 | 77 | 69 | 71 |
| Accuracy (%) | | 64 | | | 68 | | |
| Naive Bayes | negative | 68 | 93 | 79 | 68 | 96 | 80 |
| | neutral | 74 | 44 | 55 | 76 | 38 | 51 |
| | positive | 86 | 72 | 78 | 86 | 72 | 79 |
| Accuracy (%) | | 73 | | | 73 | | |

### 3.9.2 Analysis of Results

- **Precision**

Before preprocessing: All algorithms showed varying performance in terms of precision. For example, the precision of the neutral class in the Random Forest algorithm was low (65%), while the precision of the positive class in the Naive Bayes algorithm was high (86%).

After preprocessing: The precision of the positive class in Random Forest improved (81%), while the precision in Naive Bayes remained stable. In contrast, no significant improvement was observed in the precision of the Decision Tree.

- **Recall**

Before preprocessing: The Naive Bayes algorithm excelled in recall for the negative class (93%), but performed poorly in the neutral class (44%).

After preprocessing: The recall for the negative class in Naive Bayes increased to (96%), and Random Forest also showed improvement in the negative class. However, the Decision Tree showed only a slight improvement.

- **F1-score**

Before preprocessing: The F1-score in Naive Bayes outperformed the other algorithms, particularly for the positive class (78%).

After preprocessing: The F1-score in Random Forest for the positive class increased to (77%) and showed slight improvement for the neutral class, while Naive Bayes maintained a good level for the negative class.

- **Accuracy**

Before preprocessing: Naive Bayes achieved the highest accuracy at (73%), followed by Random Forest at (70%) and Decision Tree at (64%).

After preprocessing: The overall accuracy of Random Forest improved to (75%), Naive Bayes remained stable at (73%), and the accuracy of the Decision Tree slightly improved to (68%).

## 3.10 Conclusion

In this chapter, we examined sentiments on a dataset containing 27,756 comments in Algerian dialect classified as follows: 7,431 positive comments, 12,122 negative comments, and 8,203 neutral comments. We utilized three different classifiers: Decision Tree (DT), Naive Bayes (NB), and Random Forest (RF). We compared the results of the three classifiers before and after data preprocessing. We found that the results varied for each classifier.

## General Conclusion

The use of social media has become a crucial daily activity in today's culture, where it is used for social interaction, accessing news and information, making decisions, and expressing opinions. The excessive use of social media generates a massive amount of data, which contains valuable information that can be extracted and employed in various tasks. Text data is a significant type of this information, appearing in natural language through which humans express and communicate with each other. The goal of understanding natural language by machines has driven researchers to apply automated techniques to language. Natural Language Processing (NLP) is a branch of computer science, more specifically, a branch of artificial intelligence (AI). NLP provides machines with the ability to interpret text and natural language in a way similar to how humans do, using machine learning techniques. Text classification is one of the tasks in NLP.

The aim of this study is to detect polarity in social network posts in two ways: positive posts and negative posts. The goal of the work is to create a Python application that uses two data sources (the first in Dataset.CSV format and the second in emojis.txt). The first contains texts annotated with positive, negative, and neutral values for the purpose of classifying these texts, while the second contains emojis and their translations for use in preprocessing.

We began by defining some of the terms used in this work. After that, we focused on related works. We performed preprocessing with the addition of translating emojis into text. Then, we conducted sentiment analysis on a dataset containing 27,756 comments in the Algerian dialect. Decision Tree (DT), Naive Bayes, and Random Forest were among the classifiers used. Finally, we presented the experimental results of our study.

For future work, preprocessing techniques, such as emoji translation, will contribute to providing an approximate standard for understanding and interpreting the underlying meanings of these symbols, making their classification more accurate and efficient. By using machine learning algorithms such as neural networks or Naive Bayes algorithms.

## References

[1] Cigref, « Livre Blanc CIGREF (, 28 septembre 2016) « Gouvernance de l'Intelligence Artificielle dans les entreprises » », Cigref. Available on:https://www.cigref.fr/livre-blanc-cigref-gouvernance-de-l-intelligence-artificielle-dans-les-entreprises(accessed 03/02/2024).

[2] E. Viennet, « Apprentissage et Fouille de Données Visuelles », machine learning.

[3] Russell, Stuart J.; Norvig, Peter (2010). Artificial Intelligence: A Modern Approach (Third ed.). Prentice Hall.ISBN 9780136042594.

[4] Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012). Foundations of Machine Learning.The MIT Press. ISBN 9780262018258.

[5] « L'apprentissage supervisé : définition et exemples - Blent.ai ». Available on:https://blent.ai/apprentissage-supervise-definition/(accessed 07/02/2024).

[6] « Biochar and Application of Machine Learning: A Review | IntechOpen ». https://www.intechopen.com/chapters/84407.(accessed 07/02/2024).

[7] https://deepai.org/machine-learning-glossary-and-terms/random-forest(accessed 11/02/2024).

[8] « Qu'est ce que le machine learning ? Définition, Examples ».Available on:https://superdatacamp.co/le-machine-learning/.(accessed 11/02/2024).

[9] JASON Brownlee.( 2019). Long Short-Term Memory Networks with Python [online]. Edition 1v.5.Machine Learning Mastery, 246p. Available on: https://books.google.dz/books?id=m7SoDwAAQBAJ&dq=Jason+Brownlee,+Long+Short-Term+Memory+Networks+With+Python,+2019.&hl=fr&source=gbs_navlinks_s(accessed 11/02/2024).

[10] ANDEREAS Holzinger, BERND Malle, PETER Kieseberg et al. (2017). Machine Learning and Knowledge Extraction in Digital Pathology Needs an Integrative Approach. Towards Integrative Machine Learning and Knowledge Extraction [online], DOI: 10.1007/978-3-319-69775-8_2 (pp.13-50). Available on: https://www.researchgate.net/publication/320687279_Machine_Learning_and_Knowledge_Extraction_in_Digital_Pathology_Needs_an_Integrative_Approach (accessed 19/09/2024).

[11] FEDERICO Alberto Pozzi, ELISABETTA Fersini and ENZA Messina et al. (2016) Sentiment Analysis in Social Networks. First Edition. Elsevier Science. 284p

[12] BING Liu.( 2012). Sentiment analysis and opinion mining [online]. Morgan & Claypool Publishers, 167 pages. Available on:https://books.google.dz/books/about/Sentiment_Analysis_and_Opinion_Mining.html?id=Gt8g72e6MuEC&redir_esc=y .(accessed 18/02/2024).

[13] Sentiment analysis levels Available on: https://symbl.ai/developers/blog/sentiment-analysis/ (accessed 18/02/2024).

[14]STEVEN Bird, EWAN Klein and EDWARD Lope.( 2009 ). Natural Language Processing with Python: Analyzing Text with the Natural [online]. First Edition.Gravenstein highway north, Sebastopol: O'Reilly Media, 504 pages. Available on: https://www.nltk.org/book/ (accessed 21/02/2024).

[15] Journal of King Saud University - Computer and Information Sciences June 2021, Pages 497-507.Available on: https://www.sciencedirect.com/science/article/pii/S1319157818310553 (accessed 02/03/2024).

[16] Habash, N. Y. (2010). Introduction to Arabic Natural Language Processing.

[17] Adouane, W. and Dobnik, S. (2017). Identification of languages in algerianarabic multilingual documents. In Proceedings of the Third Arabic Natural Language Processing Workshop, pages 1–8.

[18] Mataoui, M., Zelmati, O., and Boumechache, M. (2016). A proposed lexicon-based sentiment analysis approach for the vernacular algerianarabic. Research on computing science, 110(1):55–70.

[19] Bettiche, M., Mouffok, M. Z., & Zakaria, C. (2018, June). Opinion Mining in Social Networks for Algerian Dialect. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 629-641).

[20] C. Mazari and A. Djeffal (2021), “Deep learning-based sentiment analysis of algerian dialect during hirak 2019,” in 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH). ,( pp. 233–236). IEEE

[21] Najadat, H., Al-Abdi, A., & Sayaheen, Y. (2018, April). Modelbased sentiment analysis of customer satisfaction for the Jordanian telecommunication companies. In 2018 9th International Conference on Information and Communication Systems (ICICS) (pp. 233-237). IEEE.

[22] ZARRA, Taoufiq, CHIHEB, Raddouane, MOUMEN, Rajae, et al.( 2017). Topic and sentiment model applied to the colloquial Arabic: a case study of Maghrebi Arabic. In : Proceedings of the 2017 international conference on smart digital environment. p. 174-181.

[23] LARKEY, Leah S. et CONNELL, Margaret E. Arabic information retrieval at UMass in TREC-10. In : TREC. 2001 LARKEY, Leah S., BALLESTEROS, Lisa, et CONNELL, Margaret E. Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. p. 275-282.

[24] comparative study between different techniques used for sentiment analysis of Algerian Arabic dialect [online]. Available on: sentiment-analysis-arabic-algerian/sentiment dataset/Final.csv at main · djalil21/sentiment-analysis-arabic-algerian · GitHub (accessed 09/03/2024).

[25] General knowledge about the history of emoji and its prevalence in digital communications. [online]. Available on: https://www.wired.com/story/guide-emoji/ (accessed 02/06/2024)

[26] Algerian Dialect text clustering based on Emotion detection [online]. Available on: https://dspace.univ-ouargla.dz/jspui/handle/123456789/30899 (accessed 06/06/2024).

[27]Opinion mining in social networks: Algerian dialect as case study [online]. Available on: http://dspace.univtebessa.dz:8080/xmlui/handle/123456789/788#:~:text=Opinion%20mining%20in%20social%20networks (accessed 06/05/2024).

[28] Decision Tress | Towards data science [online]. Available on: https://towardsdatascience.com/decision trees-in-machine-learning-641b9c4e8052 , (accessed 06/05/2024).

[29] Random Forest – Data Analytics, [online]. Available on: https://dataanalyticspost.com/Lexique/random-forest/ , (accessed 06/05/2024)..

[30] Gandhi, R.to wards data science.(2018).Support Vector 79 Machine | Introduction to Machine Learning Algorithms.[online]. Available on:Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith Gandhi | Towards Data Science(accessed 10/05/2024).

[31] NLTK-Natural Language Toolkit. NLTK 3.5 Documentation [online]. Available on: NLTK ::Natural Language Toolkit — NLTK 3.2.5 documentation (accessed 20/05/2024).

[32] Pyarabic [online]. Available on:PyArabic — PyArabic: Python Library for Arabic 0.6.12 documentation (accessed 20/05/2024).

[33] https://www.omnicalculator.com/statistics (accessed 22/09/2024).