PEOPLES DEMOCRATIC AND REPUBLIC OF ALGERIA
MINISTRY OF HIGH EDUCATION AND SCIENTIFIC RESEARCH
IBN KHALDOUN UNIVERSITY – TIARET

Master Thesis

Introduced to:

Mathematics and Computer Science faculty
Computer Science department

For obtaining:

MASTER DEGREE

Speciality : Software Engineering.

By:

AOUCI NARIMENE
HAMIDI NESRINE

About:

Opinion Analysis from Arabic Texts:

A Case Study of the Algerian Dialect Comments

Submitted on: 11 / 07 / 2023 in Tiaret in front of the jury:

Mme LAKHDHARI Aicha	MAA	Tiaret University	President
Mr MOKHTARI Ahmed	MAA	Tiaret University	Examiner
Mr BEKKI Khadir	MAA	Tiaret University	Supervisor

Academic year: 2022/2023.

Abstract:

Opinion mining or sentiment analysis has emerged as a dynamic field within natural language processing, focusing on the computational analysis of sentiments and emotions expressed in written human languages. In recent years, sentiment analysis has gained significant traction across various domains including politics, production, services, marketing and others. This interest stems from the understanding that opinions hold substantial impact and contribute to decision-making processes across these domains.

In this study, our main objective is to develop a system that performs sentiment analysis on Arabic texts, specifically focusing on the Algerian dialect. The system aims to analyze the extracted sentiments and classify them into positive, negative, or neutral classes using machine learning techniques and deep learning algorithms.

Keywords: opinion mining, sentiment analysis, Algerian dialect, machine learning, deep learning, neural network.

Résumé:

L'analyse d'opinions ou l'analyse de sentiments a émergé en tant que domaine dynamique dans le traitement du langage naturel, se concentrant sur l'analyse computationnelle des sentiments et des émotions exprimés dans les langues humaines écrites. Ces dernières années, l'analyse de sentiments a gagné en importance dans divers domaines tels que la politique, la production, les services, le marketing, et d'autres. Cet intérêt découle de la compréhension que les opinions ont un impact significatif et contribuent aux processus de prise de décision dans ces domaines.

Dans cette étude, notre objectif principal est de développer un système qui réalise l'analyse de sentiments sur des textes arabes, en se concentrant spécifiquement sur le dialecte algérien. Le système vise à analyser les sentiments extraits et à les classer en catégories positives, négatives ou neutres en utilisant des techniques d'apprentissage automatique et d'apprentissage profond.

Mots clés: fouille d'opinion, analyse des sentiments, dialecte algérien, apprentissage automatique, apprentissage profond, réseaux de neurones.

Thanks:

«Praise to Allah, who has guided us to this; and we would never have been guided if Allah had not guided us»

First, we would like to thank Allah for giving us the strength and the endurance to realize this work.

Second, we would like to express our deepest and sincere gratitude for our supervisor **Bekki Khadir** for the guidance, the thoughtful comments and the unfailing support all along our journey to make this humble work. We're truly grateful for all his efforts in revising and correcting this paper, and for all the advice that he provided.

Our special thanks to the jury members who devoted their precious time and efforts in reading and correcting our work.

We also thank our family for their support.

Dedication:

This modest work is dedicated to:

My dear parents,

For Their support, patience, and encouragement throughout my academic journey

I thank you for everything you have done for me

To my brothers Oussama and Omar,

You are my joy and my support

To my partner Nesrine,

You are my sister, my best friend, and my soulmate. I am incredibly grateful for your support and friendship.

To my family and friends,

Who have always encouraged me.

Narimene.

I dedicate this work to beloved people who have had a great impact on my life. First I want to say a big thank you to my mother for everything she's done for me, My dad has also been my biggest supporter, always encouraging me to succeed in life. I express my gratitude to my family and my siblings, especially Abdou .I will never forget your standing by my side throughout my career. Lastly to my best friends, especially Narimene, who has been my partner in this work, thank you for being patient, understanding and hardworking. I wish you all the best.

Nesrine.

Table of contents:

List of figures:

List of tables:

General Introduction:

In today's digital age, the growing communication and opinion exchange on various platforms, especially social media, are crucial for decision-making in diverse domains.

The development of Web 2.0 has made it possible for us to readily obtain information and analyze it. People now actively express their thoughts and ideas online, discussing a wide range of topics on platforms like discussion groups, blogs, forms, and product review websites. These opinions have become a crucial source of information for businesses to consider during product development and marketing planning.

Within this context, the field of opinion mining or sentiment analysis emerges as a field dedicated to analyzing expressed opinions and extracting valuable insights, particularly in domains where public opinion plays a significant role.

The objective of our work is to analyze Arabic comments, specifically in the Algerian dialect. By employing various approaches, we aim to accurately classify the sentiment express in these comments as positive, negative, or neutral.

By developing a robust system that employs machine learning techniques and deep learning algorithms, our research aims to provide valuable insights into the sentiment patterns prevalent in Arabic comments, thereby enabling a better understanding of public opinion within the Algerian dialect. This contribution can be instrumental for decision-making processes across domains such as politics, production, services, marketing, and beyond, where opinions hold substantial impact.

This thesis is organized into four chapters, which can be summarized as follows:

- **Chapter 1:** This chapter introduces machine learning and provides an overview of commonly used algorithms.
- **Chapter 2:** This chapter introduces the concept of opinion mining, including various approaches and relevant studies.
- **Chapter 3:** This chapter is dedicated to the unique characteristics of the Arabic language, Algerian dialect and Arabizi.
- **Chapter 4:** The main focus of this chapter is the practical implementation and experimentation of our application.

Finally, this thesis concludes with a general conclusion and some perspectives.

Chapter 1:

Machine learning

1. Machine learning

1.1 Introduction:

Machine learning has evolved into a revolutionary force in the current world that is changing fields and our way of life. The development of big data and advancements in processing power have made machine learning an essential tool for obtaining insightful information and directing data-centric decision-making.

Machine learning, a fast developing discipline of artificial intelligence, is concerned with creating algorithms and models that let systems learn from their experiences and get better without having to explicitly design them.

In this chapter, we will delve into the fundamental concepts of machine learning, exploring its definition, various types, wide-ranging application, and some popular machine learning algorithms.

1.2 Machine learning definition:

Tom Mitchell [1] in his book "Machine Learning" defines machine learning as follows:

"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."

In simpler terms, machine learning is the process of developing computer programs that can improve their performance on specific tasks through experience and data. The program learns patterns and relationships from the provided data, allowing it to make accurate predictions or take actions without being explicitly programmed for every scenario.

In general, machine learning [2] is a subfield of artificial intelligence (AI) that enables computers to learn and improve at tasks, without the need for explicit programming. The concept is to allow machines access to data and let them learn on their own; it is based on the machine's comprehension of the structure of data and transforms that data into models that people can comprehend and use.

The ultimate goal is to make the machine or computer capable of providing solutions to complex problems by processing a vast amount of information. This offers the possibility to analyze and uncover correlations that exist between two or more given situations and predict their various implications.

1.3 How does machine learning work?

Machine learning operates by starting with data. A machine learning model will be trained using the data, which has been gathered and prepared to serve as training data. The performance of the application improves as the data volume increases [3].

Once the data is prepared, programmers select a machine learning model and provide the data for training. The computer model then learns from the data, identifying patterns and making predictions. When presented with new data, the trained model utilizes its learned knowledge to generate outputs or predict outcomes.

The quantity and quality of training data impacts the accuracy of these predictions. A system can build more complex models and predict more accurately when it has access to high-quality data sets.

The following diagram (figure 1) demonstrates how the machine learning algorithm functions.

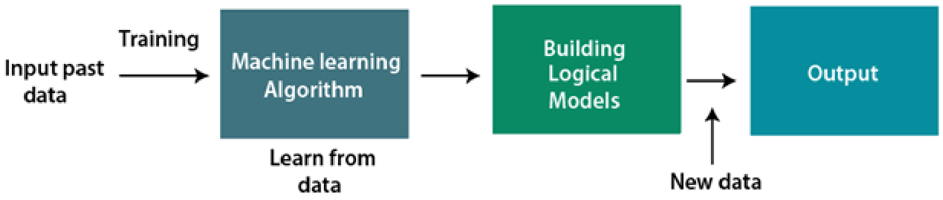


Figure 1: The working of Machine Learning algorithm [4].

1.4 Machine learning applications:

Each day, machine learning advances quite quickly. Without even realizing it, we utilize machine learning in our daily lives through applications like Google Maps, Google Assistant, Alexa, etc. Here are some of the trendiest applications of machine learning [5]:

- Image Recognition, Speech Recognition
- Product recommendations
- Self-driving cars
- Virtual Personal Assistant

- Automatic Language Translation
- Game playing
- Medical diagnosis
- Email management, Robotics
- Natural Language Processing and many more.

1.5 Machine learning methods:

Machine learning can be classified into three types:

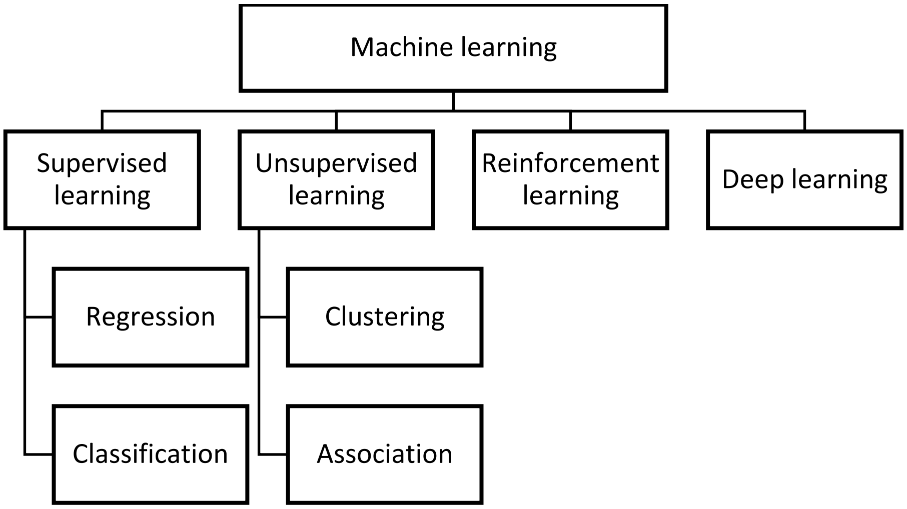


Figure 2: Machine learning methods classification.

1.5.1 Supervised learning:

The objective of supervised learning, a sort of machine learning, is to develop a function that can map input data to appropriate output labels. This is accomplished by giving the model a collection of sample input-output pairings, or labeled training data. The supervised learning algorithms require external assistance in the form of labeled data to learn from. The labeled training data is typically divided into two sets: the training dataset, used to train the model, and the test dataset, used to evaluate the model's performance. The training dataset contains input variables and their corresponding output labels that the model needs to predict or classify. The supervised learning algorithms learn patterns from the training dataset and apply them to the test dataset to make predictions or classifications. The goal is to train the model to generalize well to new, unseen data and accurately predict the output labels [6].

Supervised learning can be categorized into two types of problems: regression and classification [7].

1.5.1.1 **Regression algorithms:** are used to predict continuous output variables, such as predicting house prices based on features like size, location, and number of bedrooms. These algorithms find the relationship between the input variables and the continuous output variable, allowing for accurate predictions. Examples of regression algorithms include linear regression, decision trees, support vector regression, and neural networks.

1.5.1.2 **Classification algorithms:** are used when the output variable is categorical, like classifying emails as spam or ham based on their content. These algorithms learn patterns in the input data to accurately assign input variables to predefined categories. Popular classification algorithms include logistic regression, random forests, naive Bayes, and support vector machines.

1.5.2 Unsupervised learning:

Unsupervised learning is a type of machine learning where algorithms are left to explore and discover the underlying structure in data without being guided by correct answers or a teacher. There are no predefined labels or categories for the data. Instead, the algorithms autonomously learn patterns and features from the data. When new data is introduced, the algorithms use the previously learned features to recognize and classify it [6].

Unsupervised learning algorithms have two main types: clustering and association.

1.5.2.1 **Clustering:** involves grouping objects based on similarities, identifying commonalities and categorizing data accordingly.

1.5.2.2 **Association:** focuses on finding relationships between variables in a dataset, determining sets of items that frequently occur together, and improving marketing strategies.

1.5.3 Semi-supervised learning:

The process of semi-supervised learning combines supervised and unsupervised learning techniques. This strategy involves training a model on a smaller amount of labeled data and a larger collection of unlabeled data. The model's learning from the labeled data and performance are improved by combining it with the additional information from the unlabeled data. This type of learning may be used with regression, prediction, and other methods.

1.5.4 Reinforcement learning:

Reinforcement learning is a type of machine learning where an agent learns by receiving rewards for its actions. Through trial and error, the agent interacts with its environment to achieve the highest total rewards over time. RL has been successfully applied to various tasks like robot control and gaming.

By leveraging historical experiences, reinforcement learning agents learn from their mistakes during the training process and adjust their decision-making accordingly. This iterative learning enables them to make better decisions over time, approaching optimal performance. RL provides a mechanism for agents to learn from their environment, improve their actions, and strive towards making near-perfect decisions [8].

Figure 3 illustrates the interactions between the reinforcement learning components.

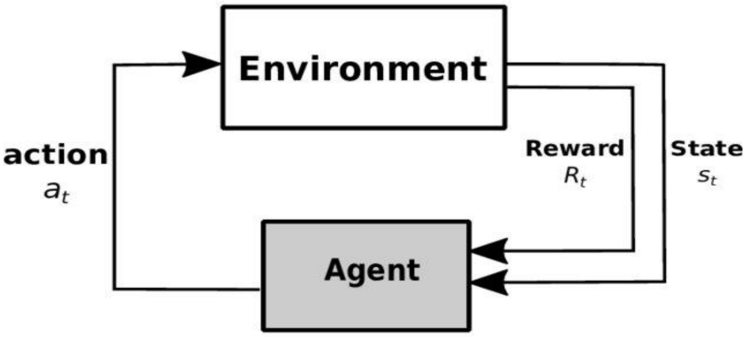


Figure 3: Reinforcement learning [9].

Reinforcement learning has two types of reinforcement: positive and negative. Positive reinforcement involves the addition of stimuli to encourage desired behavior and strengthen it. On the other hand, negative reinforcement focuses on avoiding or removing negative conditions to increase the likelihood of behavior repetition.

1.5.5 Deep learning:

Deep learning is a type of artificial intelligence derived from machine learning where the machine is able of learning by itself. Unlike machine learning, which involves teaching computers to process and learn from data, deep learning involves the computers training by itself to process and learn from data.

Deep learning is based on a network of artificial neurons that draw inspiration from the human brain. This network is composed of tens or even hundreds of layers of neurons, with each layer receiving and interpreting information from the previous layer [10].

Deep learning includes numerous neural network models like: Convolutional Neural Networks (CNNs) [11], Recurrent Neural Networks (RNNs) [12], Long Short-Term Memory Networks (LSTMNs) [13]and Deep Belief Networks (DBNs) [14].

An illustration of a deep neural network's architecture is shown in figure 4 below.

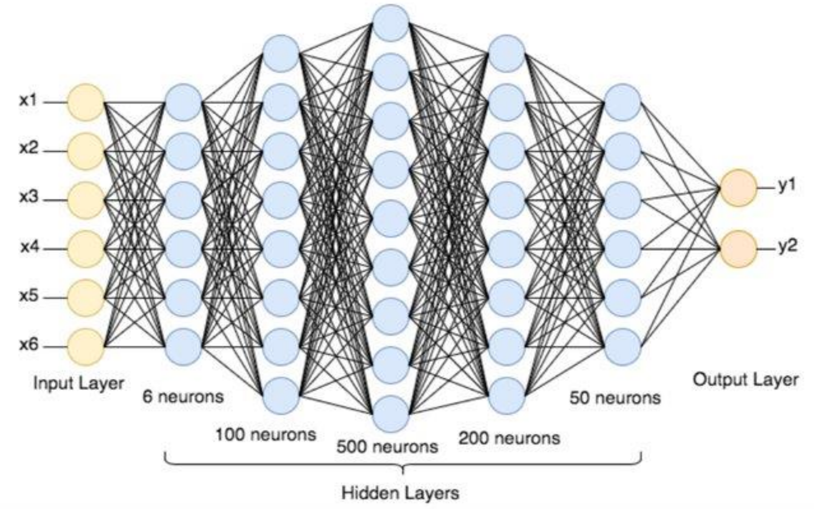


Figure 4: Deep neural network architecture [15].

1.5.5.1 Convolutional Neural Networks (CNNs):

Convolutional Neural Networks are used to analyze visual data like images and videos. They are great at automatically learning and extracting features from the input data. CNNs excel at capturing complex patterns and hierarchical representations within the data. This makes them highly effective for tasks such as image classification, object detection, and images segmentation.

The key complements of a CNN include convolutional layers, pooling layers, and fully connected layers. Here's brief explanation for each:

- **Convolutional layers:** these layers have learnable filters that slide over input data to detect features like edges, textures, or shapes. They produce feature maps
- **Pooling layers:** Used to downsample feature maps, reducing dimensions. Max polling preserves prominent features.
- **Fully Connected layers:** transform high-level features into desired output format.

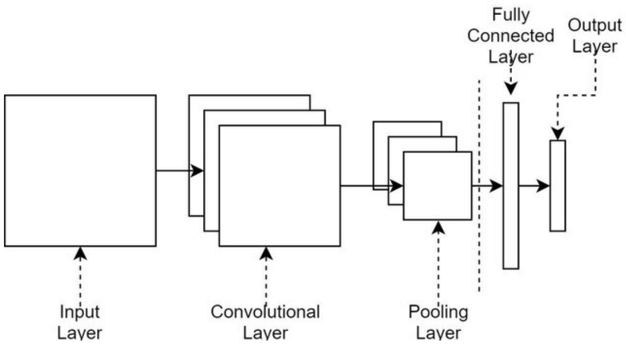


Figure 5: Convolutional Neural Network Architecture [16].

1.5.5.2 Long Short-Term Memory Network (LSTMN):

Long Short-Term Memory Network is type of recurrent neural network (RNN) designed for modeling sequential data. It handle dependencies and retain information over time using memory cells. LSTM capture long-term relationship by remembering or forgetting at each step. It is used in speech recognition, machine translation, sentiment analysis, and more.

LTSMs consist of several essential components that enables their unique functionality for sequential data analysis.

- **Memory cell:** stores and updates information over time, retaining important data and discarding irrelevant information.
- **Input gate:** controls the flow of information into the memory cell, determining which parts of the input should be stored.
- **Forget data:** decides which information to remove from the memory cell, selectively erasing irrelevant information that is no longer needed.
- **Output gate:** determines the output form the memory cell.

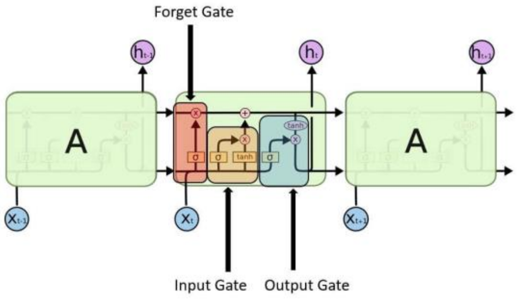


Figure 6: Long Short-Term Memory Network Architecture [17].

1.6 Machine learning algorithms:

There are several algorithms in the different types of machine learning. Some algorithms include:

1.6.1 Support Vector machine (SVM):

SVM is a powerful non-probabilistic classifier that can effectively separate data points linearly or nonlinearly, making it versatile for various types of variables. It is known for its solid theoretical foundation and high accuracy in classification tasks, outperforming many other algorithms in different applications. SVM is particularly suitable for text classification. The main objective of an SVM classifier is to identify the optimal hyperplane that can best separate classes. The quality of separation is determined by maximizing the margin between the hyperplane and the closest training points from each class. A larger margin helps minimize the generalization error of the classifier, leading to improved performance [8].

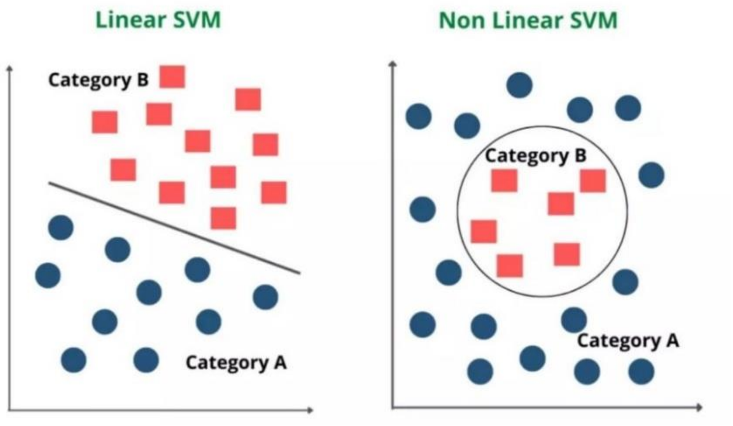


Figure 7: Types of SVM separation [18].

1.6.2 Naïve Bayes (NB):

The Naïve Bayes classifier is a supervised machine learning algorithm. It is a classification technique that relies on Bayes' Theorem and assumes independence among predictors. In simpler terms, it assumes that the presence of one feature in a class is not related to the presence of any other feature. Naive Bayes is commonly used in the text classification industry. It is primarily employed for tasks such as clustering and classification, where the main focus is on determining the conditional probability of an event occurring based on the available features [6].

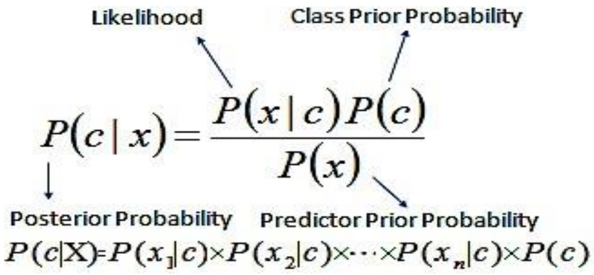

Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$

Figure 8: Naive Bayes [6].

1.6.3 Decision Tree (DT):

Decision tree is a type of supervised learning algorithm used in machine learning and decision-making processes. It is a tree-shaped diagram that represents a series of decisions and their possible consequences. It follows a hierarchical structure that includes a root node, branches, internal nodes, and leaf nodes. Decision trees can be used for both classification and regression problems. In sentiment analysis, decision trees can be used to classify text data based on their sentiment polarity by recursively splitting the data based on the most informative features [6].

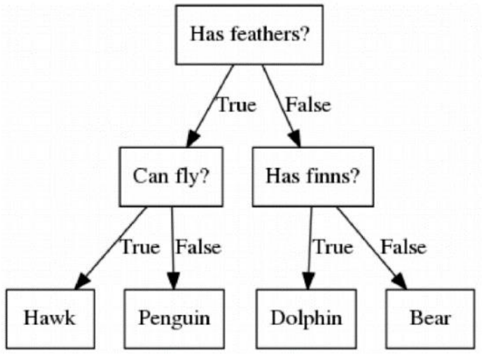


Figure 9: Decision Tree example [19].

1.6.4 Random Forest (RF):

Random Forest is a machine learning algorithm that was first proposed by Leo Breiman and Adele Cutler in 2001 [20]. It is an extension of the decision tree algorithm and is widely used for classification and regression tasks. The algorithm works by creating multiple decision trees and combining their predictions to produce a final output.

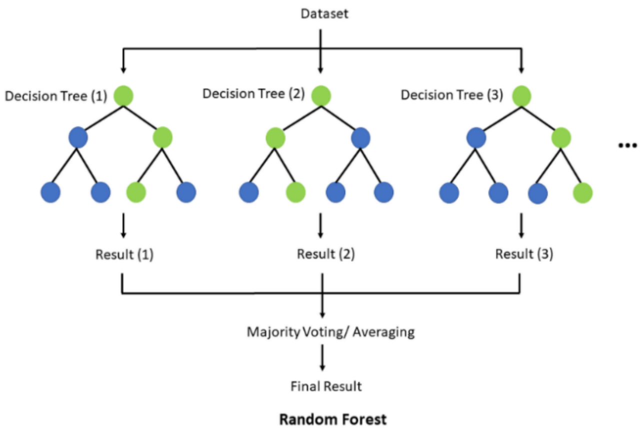


Figure 10: Random Forest classification.

1.6.5 Artificial Neural Network (ANN):

Neural Network learning algorithm, or an artificial neural network is a system of computational learning that utilizes a network of functions to comprehend input data and process it into the desired output is referred to as a neural network [21]. It comprises of units or neurons that are inspired by the biological neurons present in the human nervous system. Alexander Bain (1873) and William James (1890) independently laid the initial theoretical foundation for modern neural networks [22], recognizing that both cognitive and physical activity resulted from the interactions between neurons in the brain. A neural model-based artificial system's potential was discovered as a result of this realization.

The fundamental unit of a neural network is the neuron. Every neuron has an activation function that bases its output on input and weights, as well as a set of weights that regulate how strongly it links to other neurons. Layers made up of the neurons in a neural network each have a distinct function for processing incoming input. A hidden layer or layers that perform intermediate processing on the input data receive the output of the input layer and pass it to them. The output layer receives the output of the last hidden layer, which creates the network's final output. The equation y = f (w1. x1 + b), where y is the output, w1 is the weight of the neuron, x1 is the input it received, and b is the bias, describes how a neuron produces its signal.

Neural networks, a popular tool for model recognition, classification and prediction, can be developed using supervised, unsupervised and reinforcing learning approaches. The neural network algorithm acquires knowledge through processing labeled data during training, improving its accuracy as it learns to identify key characteristics of input data.

Neutral networks have achieved success in various domains, such as image and speech recognition, natural language processing and predictive analytics.

The architecture of a layered neural network is commonly depicted in Figure 11 with the connections between neurons indicated by lines or arrows, and each layer comprising a certain number of neurons.

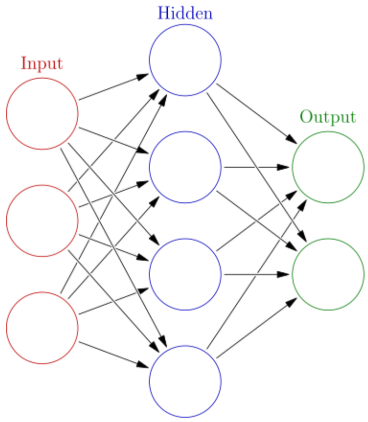


Figure 11: Artificial neural network layers.

1.7 Evaluation measures:

1.7.1 Accuracy:

Accuracy measures the proportion of correct predictions made by model.

- Accuracy = (TP + TN)/ (TP + TN + FP + FN)

True Positive (TP): The model correctly predicted the positive class.

True Negative (TN): The model correctly predicted the negative class.

False Positive (FP): The model incorrectly predicted the positive class when the actual class was negative.

False Negative (FN): The model incorrectly predicted the negative class when the actual class was positive.

1.7.2 Precision:

Precision measures the proportion of true positive predictions of all positive predictions made by a model. It focuses on minimizing false positives.

- Precision= TP / (TP + FP)

1.7.3 Recall:

Recall measures the proportion of true positive predictions out of all actual positive instances in the data.

- Recall = TP / (TP + FN)

1.7.4 F1 Score:

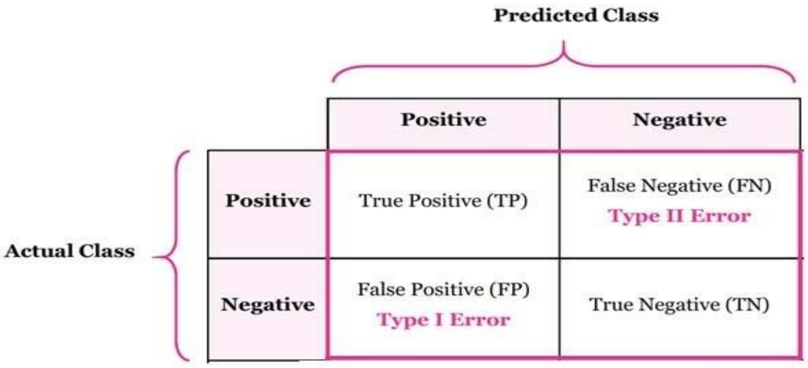
The F1 score is the harmonic mean of precision and recall.

- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

1.7.5 Confusion Matrix:

A confusion matrix is a table or matrix that summarizes the performance of a classification model by displaying the counts of different types of predictions made by the model compared to the actual labels in the dataset. It helps assess how well a model is performing in terms of correctly and incorrectly classifying instances.

In a binary classification problem, the confusion matrix typically has two rows and two columns, representing the predicted and actual classes. The four cells of the matrix represent different outcomes:



		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)

Figure 12 : confusion matrix.

1.8 Conclusion:

In this chapter we have introduced machine learning and listed its applications and types. Alongside, we had explained some of the most popular and useful machine learning algorithms.

In the next chapter, we will provide an overview of opinion mining.

Chapter 2:

Opinion mining

2. Opinion mining

2.1 Introduction:

Opinion mining, also known as sentiment analysis, utilizes computational linguistics and natural language processing techniques to automatically detect and extract sentiments or opinions from textual data. This analysis categorizes sentiments into positive, negative, neutral, or other relevant categories. This field has gained significant interest lately due to its ability to provide various tools for analyzing public opinion across a range of different subjects.

A range of techniques and methodologies, such as machine learning, lexicon-based, and hybrid approaches, are used in opinion mining. These techniques may be used to categorize text into positive, negative, or neutral thought and to pinpoint the emotions it conveys.

The goal of this chapter is to provide an overview of opinion mining and the various techniques and used in this filed. We will discuss the levels of opinion mining. We will also review the applications of opinion mining and highlight some of the recent research in this area.

2.2 Opinion:

An opinion refers to the judgement, appraisal, position, viewpoint, or thoughts held by an individual or a group of people towards something or someone. It is predominantly based on their feelings and beliefs [23].

2.2.1 Types of opinion:

In general, opinions can be evaluated in two ways: through direct opinions and comparisons.

- **Direct opinions:** provide a positive or negative evaluation of the object without comparing it to other objects (directly), such as saying "The picture quality of this camera is great."
- **Comparisons:** involve comparing the object to similar objects, such as saying "Car x is cheaper than y" [24].

2.3 Opinion mining definition:

Opinion mining or sentiment analysis, is a field of study that examines and analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards various entities. These entities can include products, services, individuals, issues, events, topics, and their attributes. It encompasses various tasks and is often referred to by different names such as sentiment analysis, review mining, opinion extraction, sentiment mining, and subjectivity analysis [25].

The goal of opinion mining is to extract meaningful insights from textual data and understand the overall sentiment express by individuals towards specific entities or topics.

The internet has made the web a vital source of information, where people express their opinions and sentiments. Analyzing this data is crucial for monitoring public opinion and assisting decision-making. Sentiment analysis has become a fast-growing area of research, with an increase in studies focused on the analysis of opinions and sentiments.

Google Trends data further illustrates the rising popularity of sentiment analysis, as depicted in Figure 13.

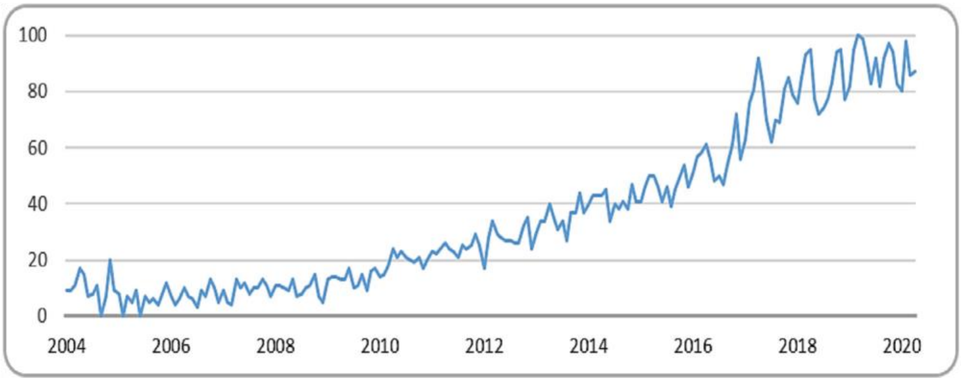


Figure 13: Interest in ‘‘Sentiment Analysis’’ since 2004 according to Google Trends [8].

2.4 Classification in Opinion mining:

There are several methods available for sentiment or opinion analysis classification. One common approach is to classify the polarity of the content, which involves analyzing people's opinions expressed in the text. Polarity classification typically involves three categories: positive, negative, and neutral [26].

Or we have the option of performing fine-grained sentiment analysis, which involves using a more extensive range of categories than the traditional three classes. These categories may include:

- Very positive
- Positive
- Neutral
- Negative
- Very negative

This approach is frequently utilized in opinion polls or surveys, where a rating system is employed, with "Very Positive" being represented by five stars and "Very Negative" by one star [27]**.**

2.5 Levels of Opinion mining:

In opinion mining, the objective is to classify the sentiment of a given text as either positive, negative, or neutral. This classification can be carried out at various levels such as document, sentence, or feature/aspect level.

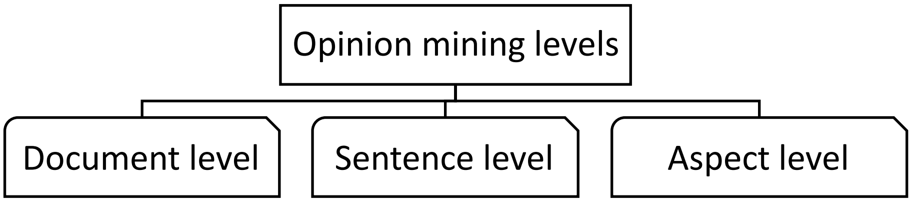


Figure 14: Opinion mining levels.

2.5.1 Document level:

Document level sentiment analysis focuses on evaluating the overall sentiment expressed in a document, rather than analyzing individual sentences or phrases. By considering the document as a unified entity, this approach determines whether the overall sentiment is positive, negative, or neutral. It finds applications in various domains such as product or service reviews, social media monitoring, and market research. Document level sentiment analysis provides valuable insights into the general sentiment conveyed in the document, aiding in decision-making and understanding public opinion.

2.5.2 Sentence level:

This level of analysis focuses on identifying the sentiment or opinion of individual sentences within a document. Each sentence is considered as a separate entity and analyzed individually. The results obtained from each sentence are then summarized to provide an overall sentiment for the document [28].

2.5.3 Aspect level:

While document and sentence level analyses do not provide a precise understanding of what people like or dislike, aspect level analysis offers a more detailed approach by focusing on specific aspects and determining the polarity relative to them. Additionally, the task of opinion mining at the feature level involves extracting features of the commented object, determining the overall sentiment towards them, and grouping synonyms to produce a summary report. For example, the following sentence: "The camera is great, but the battery life is poor." In this case, aspect level opinion mining would involve identifying and analyzing the sentiment towards specific features of the mobile phone, such as the camera and battery life. The analysis may reveal a positive sentiment towards the camera feature, but a negative sentiment towards the battery life feature [24].

The table 1 below provides a comprehensive breakdown of Arabic studies across different processing levels. According to the table, most of the sampled works are focused exclusively on conducting sentiment analysis at the document level.

S/N	Processing Level	Studies
1	Sentence-level	[2], [3], [23], [52], [69], [70]
2	Document-level	[1], [4], [5], [6], [7], [9], [10], [11], [13], [19], [20], [21], [22], [27], [28], [29], [32], [33], [39], [41], [42], [46], [48], [49], [51], [65],
3	Document-level + Sentence-level	[24]

Table 1: Arabic studies levels [29].

2.6 Opinion mining approaches:

Opinion mining can be approached in several ways, such as the widely used Lexicon-Based Approach, Machine Learning Approach, and Hybrid Approach. However, despite these methods being popular, researchers are constantly exploring ways to improve the precision of opinion mining.

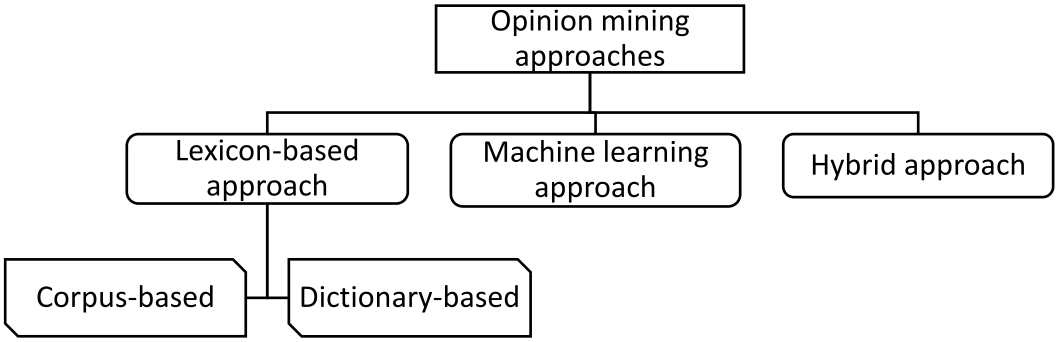


Figure 15: Opinion mining approaches.

2.6.1 Lexicon-based approach:

The Lexicon-Based approach, also known as the knowledge-based approach, is a prominent method used in sentiment analysis. It relies on a lexical resource called an opinion lexicon, which consists of a predefined list of words associated with their semantic orientation, such as positive or negative, using scores. These scores can be simple polarity values like +1 for positive, -1 for negative, or 0 for neutral, or they can represent sentiment strength or intensity. To determine the overall sentiment of a document, the semantic orientation values of the words within it are calculated. The document is first tokenized into individual words or micro phrases, and then sentiment values from the opinion lexicon are assigned to each element. Various formulas or algorithms, such as sum or average, can be applied to derive the overall sentiment of the given document [8].

The dictionary-based method involves identifying opinion words in the text and finding their synonyms and antonyms from a pre-existing dictionary [30].

On the other hand, the corpus-based method involves starting with a list of opinion words and analyzing a large corpus of text to find other related opinion words in a specific context [31].

2.6.2 Machine learning approach:

Machine learning is an approach in which algorithms are trained using manually labeled data to classify new data. In the context of sentiment analysis, this approach involves training a machine learning model to predict the sentiment polarity of new, unlabeled data based on the patterns learned during training. The quality and coverage of the labeled training data can have a significant impact on the performance of the machine learning model. This approach can achieve higher accuracy than the Lexicon-Based Approach, but it requires a large and diverse database to be effective [32].

The Machine Learning Approach is highly effective for sentiment analysis due to its powerful and accurate training steps that learn from labeled data and validate the model's performance.

2.6.3 Hybrid approach:

The hybrid approach in sentiment analysis combines the strengths of both the lexicon-based and machine learning approaches. It begins by using the lexicon-based approach to classify text at a high level, and then employs machine learning algorithms to further improve the classification using specific patterns and features within the text. By leveraging the advantages of both approaches, the hybrid approach can achieve greater accuracy in sentiment analysis. However, implementing this approach successfully requires a combination of domain expertise and technical skills to ensure effective implementation.

The figure 16 illustrates the distribution of documents across different opinion mining approaches. The three main approaches represented are lexicon-based, machine learning-based, and hybrid.

The figure shows that the majority of the documents are analyzed using the machine learning-based approach, followed by the hybrid approach, and finally the lexicon-based approach. This suggests that researchers and practitioners in the field of opinion mining are favoring machine learning-based approaches due to their accuracy and ability to handle complex data.

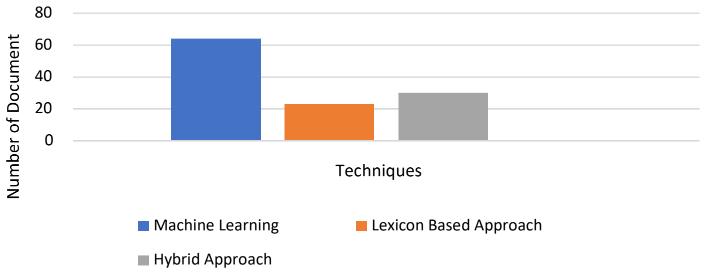


Figure 16: Opinion mining approaches by document [33].

2.7 Methods used in Opinion mining:

Numerous methods and techniques have been proposed in existing research to address the problem of Arabic opinion mining (AOM).The findings indicate that SVM and NB are the most frequently utilized methods in the articles analyzed. According to their Systematic Literature Review [34] Abdullatif, Abdulqader, and Yousef Ali report that SVM has been employed in the sentiment classification of several previous studies, with 74 out of 108 papers adopting this approach. Meanwhile, NB was used in 71 papers. Notably, the application of the SVM classifier in earlier studies has been found to be superior or comparable to other classifiers, such as NB. The Figure 17 highlights the most commonly used methods in AOM.

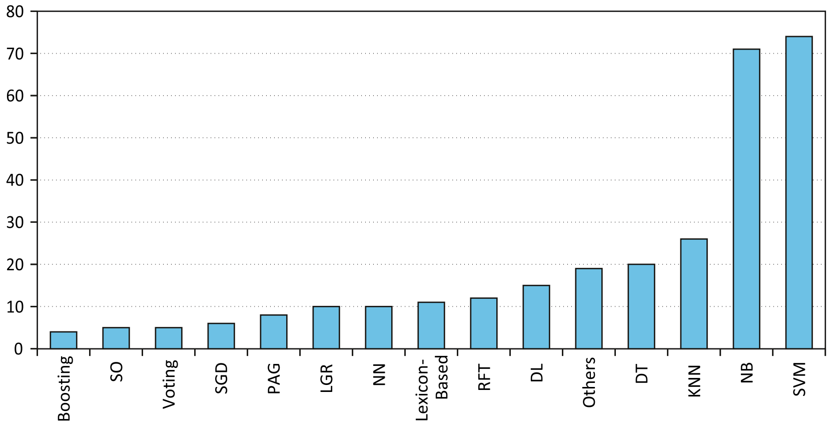


Figure 17: The most methods used in Arabic opinion classification [34].

2.8 Application of Opinion mining:

The importance of opinion mining is evident in several domains, and numerous applications have emerged in this context. Some applications are briefly mentioned below.

- **Marketing**: Opinion mining may help marketers better grasp the tastes and viewpoints of their target market. Businesses may utilize opinion mining to pinpoint client demands and modify their marketing strategies to better connect with their target market.
- **Social media:** On social media sites like Twitter, Facebook, and Instagram, opinion mining may be used to track user sentiment on a variety of subjects and issues. Businesses may learn more about client satisfaction and spot possible problems with the use of this study.
- **Voice of the customer analysis:** Opinion mining may be used to examine client feedback posted on blogs, forums, and e-commerce websites. Businesses may use this study to enhance their offerings and pinpoint problems that need to be fixed.
- **Politics:** Public opinion and sentiment regarding political topics and politicians may be analyzed using opinion mining in politics. By understanding the public's thoughts and concerns, politicians and political parties may better target their messaging and campaigns.

- **Healthcare:** Opinion mining may be utilized in the industry to examine customer evaluations and comments as well as to keep track of the caliber of healthcare services. Sentiment analysis is a tool that hospitals and healthcare professionals may use to evaluate patient satisfaction and pinpoint areas for development.
- **Finance:** Sentiment analysis has a valuable application in the financial industry, allowing investors to track their preferred companies and monitor sentiment data in real-time. This technique enables investors to easily obtain business news and consolidate the information for better financial decision-making. Consequently, sentiment analysis can assist investors in making informed decisions that lead to better returns on investment [35].

2.9 Related works:

There are various perspectives from which the current literature on sentiment analysis can be categorized, such as the techniques employed, the language of the text being analyzed, the level of detail of text analysis, the source of data and other factors. This section discusses the related research on sentiment analysis based on the previous approaches identified, including the lexicon approach, machine learning approach, and hybrid approach.

2.9.1 A Proposed Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic: In their work [36], Mhamed Mataoui ,Omar and Mediha focused on developing a sentiment analysis approach for the Algerian dialect (ALGD) that is commonly used in social networks. Their approach was based on lexicons, and to create their model, they developed three lexicons: a lexicon of keywords, a lexicon of negation words, and a lexicon of intensification words. They also utilized two other resources, a list of emoticons with assigned polarities and a dictionary of common expressions in ALGD. The keywords lexicon contains 3,093 words with 713 positive and 2,380 negative polarities. To evaluate their approach's effectiveness in identifying sentiment in ALGD text data, they collected and annotated their own dataset containing 7,698 Facebook comments to evaluate the effectiveness of their approach.

2.9.2 Lexicon-based approach for sentiment analysis of Arabic tweets: by Mahmoud Al, Safaa and Izzat [37].They constructed a sentiment lexicon consisting of approximately 120,000 Arabic terms, where each sentiment term value ranged from 0% to 100%. Positive, neutral, and negative words had sentiment values in the ranges of (60%-100%), (40%-60%), and (0%-

40%), respectively. The process of building the sentiment lexicon was divided into three steps: collecting Arabic stems, translating them into English, and determining the sentiment value of each word using online English sentiment lexicons. The collection of tweets was carried out through one of the two application programming interfaces (APIs) that Twitter offers, namely the REST API or the Streaming API.

After preprocessing and stemming the tweets, they built a sentiment analysis tool based on predicate calculus that mapped sentences into a sentiment vector. This vector combined the individual word sentiment values to compute the sentence's overall sentiment orientation. To achieve this, they opted to formulate words and sentences using a variant of predicate logic and employed predicate calculus to determine the tweet's overall sentiment orientation.

To perform their experiments, they manually labeled dataset of tweets. They ensured that the testing dataset had an equal number of tweets in each of the three sentiment classes (positive, negative, and neutral) to maintain perfect balance. Additionally, to prevent any potential biases, they selected tweets that were of similar length in terms of both word and character count.

2.9.3 Sentiment lexicon for sentiment analysis of Saudi dialect tweets: by Abdul Mohsen, Qubayl and Abdulaziz [38],a created SauDiSenti as a sentiment lexicon to analyze the sentiment of Saudi dialect tweets. They manually extracted 4431 words and phrases from modern standard Arabic (MSA) and Saudi dialects from a pre-labeled dataset of tweets obtained from trending hashtags in Saudi Arabia.

The SauDiSenti lexicon was built using the Saudi dialect twitter corpus (SDTC), which contained 5,400 tweets that included Saudi dialect and MSA. To score the words and phrases, two annotators provided negative and positive entries, which were combined and duplicates were removed. A score of -1 was given for negative entries and +1 for positive entries.

They also created a testing dataset consisting of 1500 tweets evenly distributed over three sentiment classes: positive, negative, and neutral.

Based on four threshold values, they evaluated the performance of the SauDiSenti lexicon. A tweet was considered positive if its score was higher than 0, equal to 1, or fully higher than 0, or if it was higher than 1. A tweet was labeled as negative, on the other hand, if its score didn't fit the requirements for positive categorization.

2.9.4 Sentiment Analysis of Arabic Tweets in e-Learning: by Hamed, Renxi, Khalid and Dayou [39], they conducted a study that utilized two machine learning algorithms, namely Support

Vector Machine (SVM) and Naive Bayes (NB), to develop a framework for sentiment analysis of Twitter "tweets" in the field of education. The aim of the study is to develop a framework to analyze Twitter “tweets” as having negative, positive or neutral sentiments in education or, in other words, to illustrate the relationship between the sentiments conveyed in Arabic tweets and the students’ learning experiences at universities.

They collected tweets by an application was developed in C# and used Twitter's official Developers API to download them. The tweets were stored in a database and manually labeled as negative (-1), positive (1), or neutral (0) after filtration.

They employed Rapidminer for data pre-processing by applying Tokenization, Stop-word removal, Light stemming, and Filtering tokens based on length. The classification models were built using SVM and NB algorithms to categorize the tweets as negative, positive, or neutral. Finally, precision and recall methods were applied to assess the classification outcomes.

2.9.5 Sentiment Analysis of Facebook comments published in standard Arabic or Moroccan dialect using a machine learning approach:

2.9.5 Sentiment Analysis of Facebook comments published in standard Arabic or Moroccan dialect using a machine learning approach: by Elouardighi, Mohcine, Hafdalla and Fatima-Zahra [40], they conducted a study on the Moroccan dialect to analyze sentiments from Facebook comments. Their work included identifying the properties of the Moroccan dialect and the challenges it poses for sentiment analysis. They also proposed techniques for preprocessing Facebook comments written in the Moroccan dialect to improve sentiment analysis.

To collect data for their study, they targeted Moroccan newspapers that published online comments about the Moroccan legislative elections that took place on October 7, 2016. Using Facebook Graph API, they collected 10,254 comments over a period of 70 days. They then preprocessed the comments by cleaning and normalizing the text, removing unnecessary words and symbols, and dividing the text into tokens.

In the end, 6581 comments were annotated as negative and 3673 as positive. The input variables are automatically extracted from the corpus formed from the preprocessed comments using n-gram extraction and TF/TF-IDF weighting schemes.

They utilized three supervised classification algorithms (implemented on R software), namely Naïve Bayes (NB), Random Forests (FA), and Support Vector Machines (SVM) to classify

Facebook comments**.**

2.9.6 **Sentiment analysis and opinion extraction for e-commerce sites:** application to the Arabic language: by Mohamed Ali ,Houssem ,Rami and Mounir [41]. They propose theimplementation of a sentiment analysis tool that aims to detect the polarity of opinions extracted from websites specializing in e-commerce or product reviews in the Arabic language.

They collected their corpus manually from various web resources such as reviewzat1, jawal1232, jumia3, etc. The corpus consists of a set of text documents, where each document represents a product with its type, name, and reviews (comments) about it. They selected five types of products to form the corpus: Camera, Laptop, Mobile phone, Tablet, and Television. The corpus contains 250 documents, 2812 sentences, and 15466 words.

Their prototype consists of collecting reviews from the internet through e-commerce websites, followed by data pre-processing, stemming, and finally detecting the polarity of opinions using the SVM classifier, either positive or negative.

They tested several types of classification algorithms, such as Support Vector Machines (SVM), Naïve Bayes (NB), and K-nearest neighbors (KNN). These algorithms were applied on different combinations of preprocessed data. The testing phase for evaluating the performance of classifiers on the corpus was done using the open-source tool Weka, which is a popular suite of machine learning software. It is written in Java and was developed at the University in New Zealand.

2.9.7 **An hybrid scheme for Arabic Tweets Sentiment Analysis:** by Haifa and Aqil [42], which combines semantic orientation and machine learning techniques. Through this approach, the lexical-based classifier will label the training data, a time-consuming task often prepared manually. The output of the lexical classifier will be used as training data for the SVM machine learning classifier.

The data was preprocessed by performing various steps such as cleaning, normalization, stop-word removal, elimination of speech effect, and stemming. Then, they used SentiWordNet to extract some sentiment words after translating them into Arabic. This was followed by adding their own list of essential sentiment words, resulting in a sentiment lexicon consisting of 1500 words. The overall polarity of the tweets was determined by the cumulative score of the positive degree of all the sentiment words in each tweet. To extract features, they used n-gram models.

A Support Vector Machine (SVM) classifier was utilized to anticipate the polarity category of unclassified tweets that were not classified by the lexical-based classifier. The SVM classifier builds a model from the tweets that were already categorized by the lexical classifier.

The table (table 2) presented below classifies prior research studies according to their approach, language studied, dataset size, and social network type that was used as the data source. Additionally, the table provides classification algorithm employed in each study and the accuracy if it was stated.

Type approach	**Article**	**Social network type**	**Dataset size**	**Language**	**The classification algorithm applied**	**Accuracy**
Lexicon-based approach	[36]	Facebook	7698	Arabic(MSA and ALGD)	**	79.13 %
	[37]	Twitter	900	Arabic (MSA)	**	86.89%
	[38]	Twitter	4431	Arabic (MSA and Saudi dialects)	**	**
Machine learning approach	[39]	Twitter	1121	Arabic (MSA)	SVM NB	(SVM) 73.15%

	[40]	Facebook	10254	Arabic (MSA and Moroccan dialects)	SVM NB FA	(SVM) 78%
	[41]	Websites	2812	Arabic	SVM NB KNNS	(SVM) 91.2%
Hybrid approach	[42]	Twitter	1103	Arabic (MSA and Saudi dialects)	SVM	84.01%

Table 2: Classification of the previous related works.

2.10 Conclusion:

In this chapter, we explored the world of opinion mining and its various components. We began by discussing the different types of opinions and how they can be classified in opinion mining. We then delved into the levels of opinion mining. We defined the approaches to opinion mining. We also discussed the algorithms used in the machine learning approach and the most commonly used methods in opinion mining. Finally, we provided examples of applications of opinion mining.

In the following chapter, we will study some aspects of the specificity of the Arabic language, including dialects and Arabizi.

Chapter 3:

The Arabic language

3. The Arabic language

3.1 Introduction:

Arabic, a complex and diverse language spoken by millions in the Middle East, North Africa and more than, presents unique challenges for natural language processing, particularly in opinion mining.

This chapter aims to explore these challenges and shed light on Arabic opinion mining. We will begin by providing an overview of the language, delving into its intricate morphology, syntax, and semantic features. We will then focus on the Algerian dialect and its impact on Arabic opinion mining. The dialect encompasses colloquialisms and regional variations that pose challenges in classification and analysis.

Next, we will discuss Arabizi, which is commonly used in social media and informal communication, and its impact on Arabic opinion mining.

3.2 The Arabic language:

Arabic language or "al Arabiya" as it’s called in Arabic. Arabic is a Semitic language and the fifth most widely spoken language in the world with over 400 millions speakers. It’s an official language in 25 countries including Algeria, Egypt, Jordan, Iraq, Morocco, and Tunisia [43]. It is also one of the six official languages of the United Nations [44] and as the language of the Quran.

Today's Arabic language can be categorized in three categories. [45]:

- **Classical Arabic:** also known as Quranic Arabic, was used to reveal the Quran and is an enhanced version of medieval Arabic. It served as the basis for Modern Standard Arabic but has different vocabulary and context. Its unique symbols help emphasize important phrases, making it ideal for understanding the Quran
- **Modern Standard Arabic (MSA):** is the prevalent Arabic used globally across various media, including literature, movies, and news. MSA bridges regional dialectal differences and contains a broader range of day-to-day vocabulary than the Quran. It is not truly "modern" but evolved from classical Arabic and is the same everywhere, making it a universal language. MSA was used to write scholarly texts, including religious, scientific, and mathematical works, during the Islamic golden age. Learning MSA can unlock the entire Arab world, but many still speak their local dialects and use MSA selectively.

- **Colloquial Arabic:** also known as dialects, varies from region to region, making it difficult for speakers of different dialects to understand each other. MSA serves as a universal Arabic that resolves the issue. Colloquial Arabic has simpler grammar and distinct expressions that differ across dialects. Though colloquial Arabic is mainly spoken, it's occasionally used in creative writing to represent dialogue. With numerous countries and regional differences within them, the number of Arabic dialects varies, and some may be similar while others entirely distinct.

3.3 The richness of the Arabic language:

The Arabic language is written using 28 letters. It is a very rich language, with 100 different terms to identify love, 500 for the lion, 1000 for the camel and 80 for the sword [46]. It has the largest number of words with 12.3 million words, while English, French, and Russian languages have around 600,000 words, 150,000 words, and 130,000 words in its vocabulary respectively [47].

3.4 Algerian Arabic:

Also known as Algerian dialect (ALGD), is a challenging Arabic dialect to comprehend due to its unique features. Unlike written Arabic, ALGD is heavily influenced by Berber substrates and contains many loanwords from French, Turkish, and Spanish. The original Arabic words are modified phonologically, and case endings are dropped in written language. Consequently, ALGD is not widely used in formal education, television, or newspapers, where standard Arabic or French are more commonly used. Nevertheless, ALGD is frequently heard in Algerian households, on the streets, and in songs. It is a part of the Maghreb Arabic dialect continuum and has a similar vocabulary throughout Algeria. While the eastern dialects sound more like Tunisian Arabic, the western dialects resemble Moroccan Arabic [48].

The Algerian dialect has unique features. One of its main characteristics is the use of code-switching, where words from multiple languages are combined, particularly French and Arabic such as “top خويا” [49]. Another feature involves the use of Arabic expressions written in foreign expressions (predominantly French) written in Arabic letters such as “Sahit khoya rabi ykhalik”. Algerians commonly use specific forms of words in short messages, often using Arabic numerals to represent Arabic letters for instance “Mli7 rabi ywef9ek”.

3.5 Arabizi:

Arabizi, also known as Arabglish or Arabglizi [50], has become increasingly popular among the younger generation with the advent of technology and social media. It is a unique language that blends English and Arabic words together. In Arabizi, the Arabic text is written using a transliterated form using Latin characters and numbers that represent specific Arabic alphabets, effectively enabling the writing of Arabic using English script. This linguistic fusion has gained immense popularity due to the widespread availability of English keyboards and the preference for Arabic dialect in communication. This language fusion has made it easier for people to communicate with one another among friends and family.It is frequently used in social media platforms, SMS messaging, and chat applications.

Some key features of Arabizi include [51]:

- **Use of vowels:** Unlike traditional Arabic script, Arabizi uses Latin script symbols for vowels to represent both short and long vowels.
- **Use of consonants:** Arabizi uses a single Latin letter to represent multiple Arabic phonemes, and some pairs of letters can map to a single Arabic letter or pairs of letters for instance "ذ d and ض D". Digits are also used to represent certain Arabic letters or sounds 3, 5, 7 and 9 are used to represent the sounds of the letters ع, خ, ح and ق respectively.
- **Abbreviations:** Arabizi may also use abbreviations for common phrases and expressions, such as "mrc" for "merci" meaning "thank you".

3.6 Challenges of Arabic opinion mining:

Arabic opinion analysis faces several challenges that are unique to the Arabic language. The most typical challenges are listed below:

- **Dialectal variation:** Arabic is a language with significant dialectal variation, which makes it difficult to develop a single opinion lexicon that can accurately analyze opinion in all dialects. Each dialect has its own vocabulary, grammar, and syntax, which makes it challenging to develop a sentiment analysis model that can accurately analyze sentiment across dialects.
- **Morphological complexity:** Arabic is a morphologically rich language, with a given root having several forms depending on context. Due of its richness, it is difficult to reliably detect sentiment-bearing words and phrases, which causes errors in sentiment analysis such (أحبُ, يحب, يحبون, أحبو, تحب).

- **Ambiguity:** Arabic has many homonyms and synonyms words that can have multiple meanings depending on the context. This makes it challenging to accurately identify the correct sentiment polarity of a word or phrase. For instance, the phrase "جميل" may signify "good" or beautiful depending on the context. For instance, the sentiment is positive in a statement like "هذا المنظر جميل", which meaning "The view is beautiful", since the context suggests that the view is aesthetically pleasant. However, in another sentence like " هذا الطعام جميل" which means "this food is good", the sentiment is positive as the context could suggest that the food is tasty or simply that it is of good quality.
- **Limited resources:** Compared to other languages, there are fewer resources available for Arabic sentiment analysis, including sentiment lexicons, annotated datasets, and pre-trained models. This limits the accuracy and effectiveness of sentiment analysis algorithms for Arabic.
- Code-switching: Many Arabic speakers use English, French, or Spanish in addition to Arabic in their daily communications. This makes it difficult to appropriately assess sentiment in texts with many languages. For example the sentence "C'est bon يعطيك الصحة", which means "It's good thank you". The user in this case mixed French with Arabic.
- **Negation and sarcasm:** When it comes to negation and sarcasm in Arabic, specific words are used to express negation, such as "لن" ,"ما", and "لا", which indicate the meaning of "not". It is important to accurately detect and handle negation in Arabic sentiment analysis because it can completely change the meaning of a sentence and lead to an opposite polarity result [52].

These challenges make Arabic opinion analysis more complex than opinion analysis in other languages. Addressing these challenges requires the development of specialized techniques and tools that can accurately analyze opinion in Arabic texts.

3.7 Conclusion:

We also discussed the richness of the Arabic language, with its complex grammar and vast vocabulary, which poses both opportunities and challenges in opinion mining.

Our focus has been on two specific variations of Arabic: the Algerian dialect and Arabizi, each posing distinct challenges in the realm of opinion mining. The Algerian dialect is known for its colloquialisms and regional variations, making classification and analysis a complex task. On the other hand, Arabizi, a hybrid language blending Arabic and Latin script, poses difficulties for conventional text analysis methods.

The next chapter focuses on providing a comprehensive description of the techniques, algorithms, and tools used in our work, specifically in the design of our AS tool.

Chapter 4:

Our approach

4. Our approach

4.1 Introduction:

In this chapter, we will introduce our proposed method and provide more detailed information about our work. We will present the classifiers and dataset used for classification, explain the preprocessing tasks, and finally present the obtained results. Additionally, we will demonstrate our method selection by conducting tests and discussing the outcomes.

4.2Global Architecture:

The architecture of our project is depicted in the following figure (figure 18).

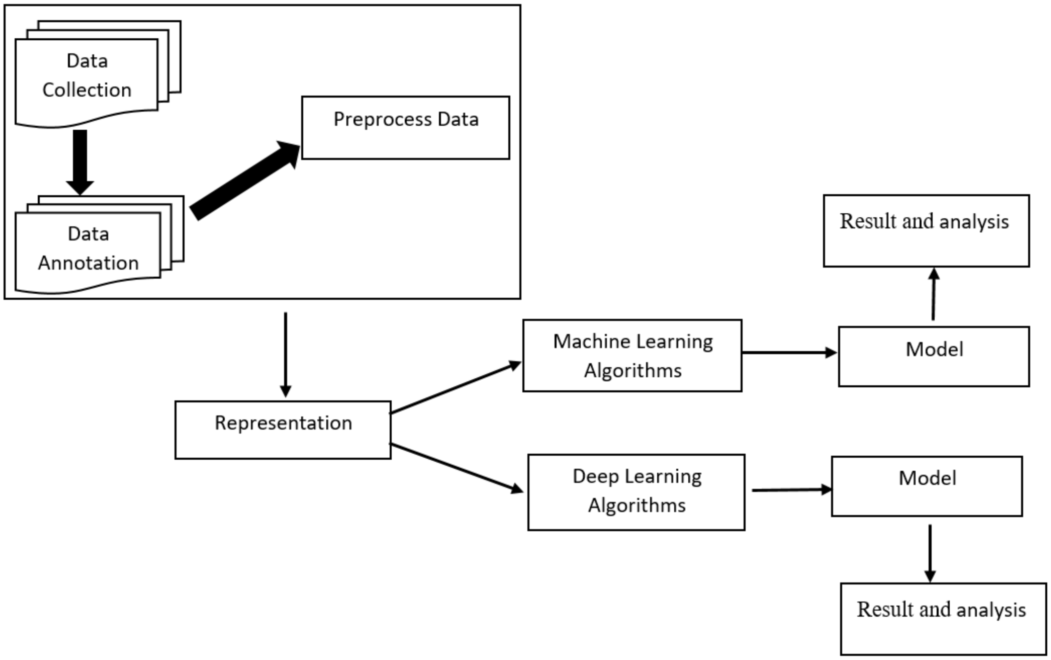


Figure 18: Global proposed architecture.

4.3 Data Collection:

In order to compile our dataset, we employed two distinct methods. The first method involved utilizing a variety of existing datasets specifically tailored for Algerian dialect. We incorporated the Algerian dialect review for sentiment analysis [58] and the Algerian corpus [59]. Additionally, we integrated dataset Mataoui, Madar, BrandtDz [60],and the Algerian Opinion Mining dataset [61], which further enriched our dataset with diverse perspectives and opinions.

For the second method, we employed the web scraper tool Instant Data Scraper to collect comments from various Facebook pages. We targeted popular pages such as Ooredoo, Condor, Mobilis, 1001 night Tech, TOP Commentaire, Stream, and others. This approach enabled us to capture real-time user-generated content, providing valuable insights for our research.

By combining these two methods, we were able to compile a robust dataset that represented Algerian dialect and captured user sentiments and opinions from social media platforms.

4.4 Data annotation:

We manually annotated the dataset collected through Instant Data Scraper. We reviewed and labeled each data point extracted from the Facebook comments. By going through the comments one by one, we categorized and labeled them based on their content and context.

4.5 Dataset collection and annotation results:

Our study involved working with a dataset that included three sentiment classes: positive, negative, and neutral. The dataset contains 24329 comments; it comprised 9894 positive comments, 7883 negative comments, and 6552 neutral comments. This illustration (table 3, figure 19) represents a distribution of the classes.

Total	Positive	Negative	Neutral
24329	9894	7883	6552

Table 3: Distribution of the classes.

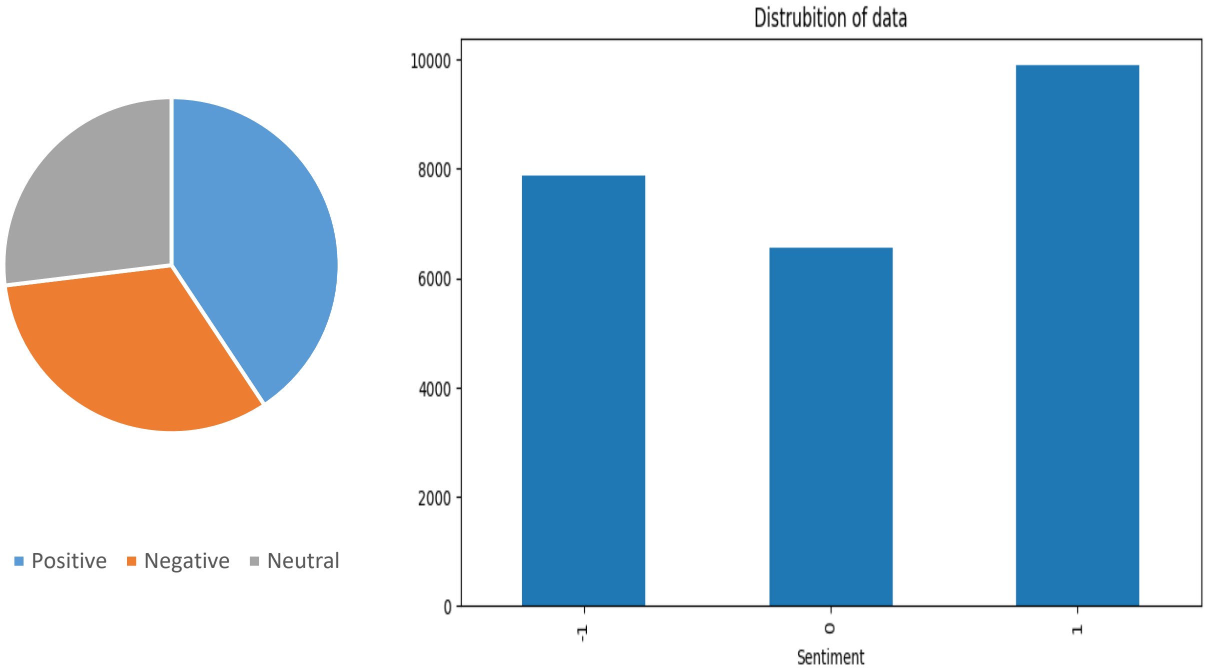


Figure 19: Distribution of the classes.

4.6 Data preprocessing for training:

Pre-processing is an essential step in natural language processing, especially when analyzing Arabic language data as it addresses the unique challenges posed by Arabic's characteristics allowing for the transformation of raw text into a format that machine learning models can more easily comprehend and analyze accurately.

Figure 20 illustrates the step-by-step process of data preprocessing.

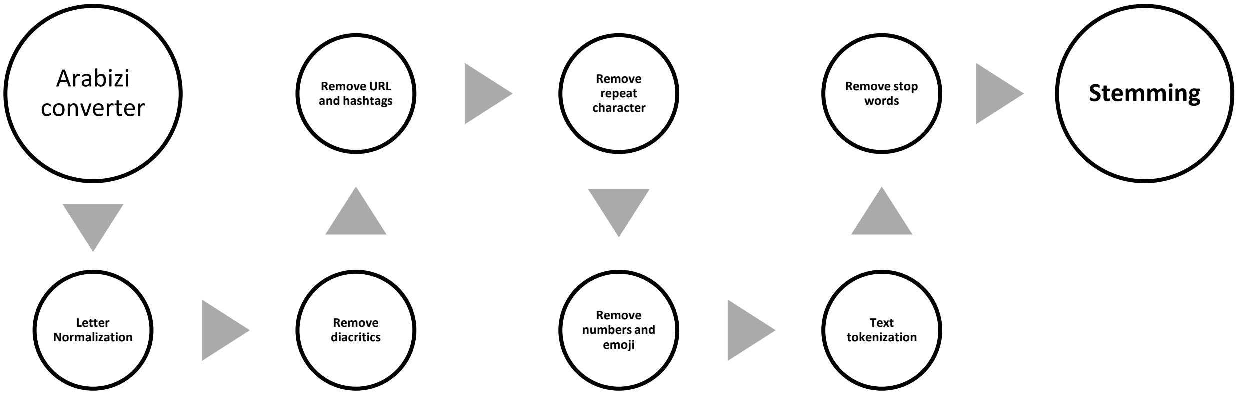


Figure 20: Preprocessing processes.

- **Arbizi Converter:** In our pre-processing, the initial step involves the utilization of an Arabizi Converter. This essential part enables the conversion of words written in Latin letters within tweets to Arabic script, allowing us to apply stemmers and other normalization tools.
- **Letter Normalization:** In Arabic, letters can take on various forms based on their position within a word. The process of letter normalization aims to unify these forms and replace them with a standardized representation. For instance, the different variations of the letter "alef" (,إ
آ ,أ) can be transformed into a single form (ا). This unification simplifies text analysis and ensures consistency in representing Arabic letters throughout the text.
- **Remove diacritics:** In Arabic text processing, the removal of diacritics plays a significant role. Diacritics are the small markings placed above or below letters to indicate vowel sounds and pronunciation. While diacritics provide valuable linguistic information, removing them can simplify the text and enhance processing efficiency. Additionally, the removal of the tatweel character which has no impact on the meaning of words is performed.
- **Remove URL and hashtags:** In our data preprocessing, we apply the removal of URLs and the elimination of hashtags. URLs as external links, do not directly impact the sentiment of a comment. Therefore, we remove them to focus on the textual content. Hashtags on the other hand play a crucial role in sentiment analysis by potentially influencing the polarity of a sentence. To maintain unbiased sentiment analysis, we remove hashtags and retain only the associated words, ensuring accurate analysis of the comment's sentiment without the interference of external links or hashtag biases.

- **Remove repeat character:** we remove repeated letters within words by reducing them to a single occurrence. This helps to reduce the dataset size and ensures consistency in lexical analysis. While repeated letters may convey stronger sentiment, treating words with different numbers of repeated letters as identical allows for accurate analysis.
- **Remove numbers and emoji:** During the data preprocessing phase, we apply a step to remove numbers and emojis from the text. Numbers are typically irrelevant to the sentiment analysis task and removing them helps us focus on the textual content. Although emojis can sometimes carry sentiment, we choose to remove them to maintain a text-based analysis approach. This ensures that our sentiment analysis model is primarily based on the textual information, enabling more accurate and reliable results.
- **Text tokenization:** We break down a sentence or a paragraph into individual tokens or words. By splitting the text into tokens, we create a structured representation that allows us to analyze and process the text more effectively. Tokenization is even more important in sentiment analysis than in other areas of NLP because sentiment information is often sparsely and unusually represented in phrases.
- **Remove stop words:** NLTK offers an extensive list of Arabic stop words covering prepositions, conjunctions, adverbs of space and time, and more. By tokenizing the text and comparing each word to the stop word list, we can effortlessly eliminate these non-essential words from the sentence, reducing noise and enhancing the accuracy of subsequent natural language processing tasks.
- **Stemming:** It aims to reduce words to their root form. It involves removing prefixes, suffixes, and other linguistic variations to extract the core meaning of a word. Arabic words often have complex morphological structures, and stemming plays a crucial role in standardizing them to their basic form. By applying algorithms like ISRI stemmers, we can simplify word representation, leading to improved information retrieval and text analysis. Stemming not only reduces the complexity of Arabic vocabulary but also enhances the accuracy and efficiency of language processing applications by treating different word forms as variants of the same root word.

4.7 Representation:

In our program, we employ TF-IDF (Term Frequency-Inverse Document Frequency) for machine learning algorithms and Embedding for deep learning algorithms.

4.7.1 TF-IDF: is a numerical representation that measures the importance of a term in a document based on its frequency in the document and polarity in the overall collection.

- TF= (Number of occurrences of the term in the document)/ (Total number of terms in the document)
- IDF=log (Total number of documents)/ (Number of documents containing the term)
- TF-IDF=TF*IDF

4.7.2 Embedding: is a technique that represents words as dense vectors, capturing their semantic relationships in a high-dimensional space.

The embedding process involves training model on a large corpus of text techniques like word2vec, Glove, or Bert.

4.8 Classification algorithms:

We employed a range of classification algorithms to process our dataset. These algorithms include both machine learning and deep learning algorithms, namely: Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Convolutional neural network (CNN), and Long Short-Term memory network (LSTM).

Indeed, predicting the best combination of representation and classification for optimal results is challenging without experimenting with different combinations. The quality of prediction heavily relies on the dataset and the nature of the texts.

4.9Experiments and results:

While using the supervised learning method, we divided the dataset into two parts, 80% for training and 20% for testing. The accuracy results are presented in Table 4 and figure 21.

Classifier	**Accuracy**
SVM	75.45%
NB	73.43%
DT	68.93%
RF	73.88%
CNN	76.40%
LSTM	**76.54%**

Table 4: Classification results.

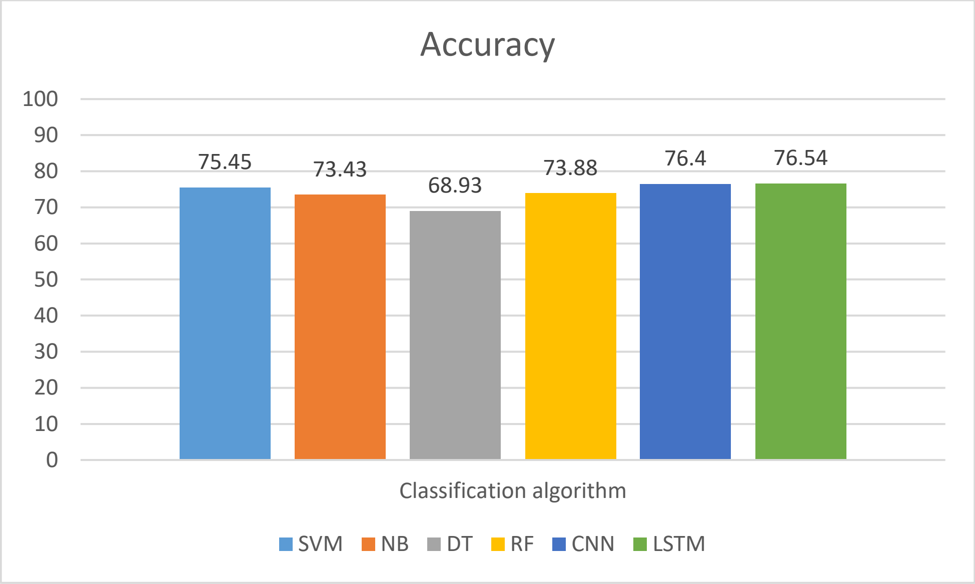


Figure 21: Classification results.

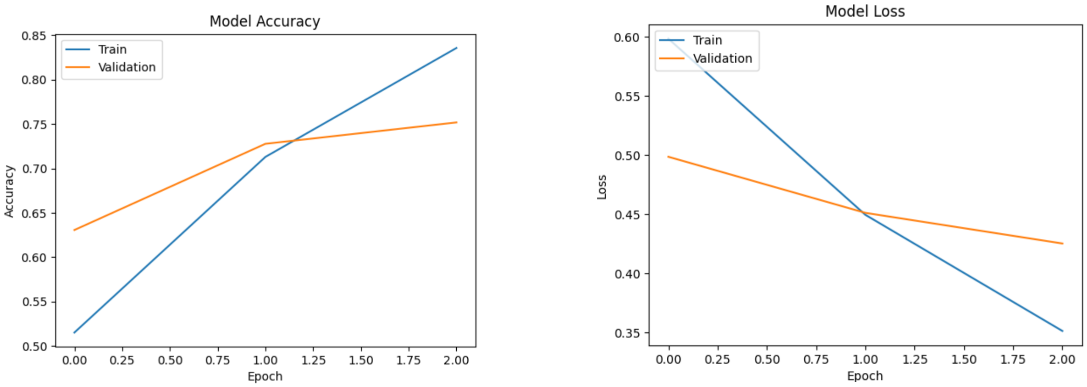


Figure 22: Accuracy and loss for the LSTM model.

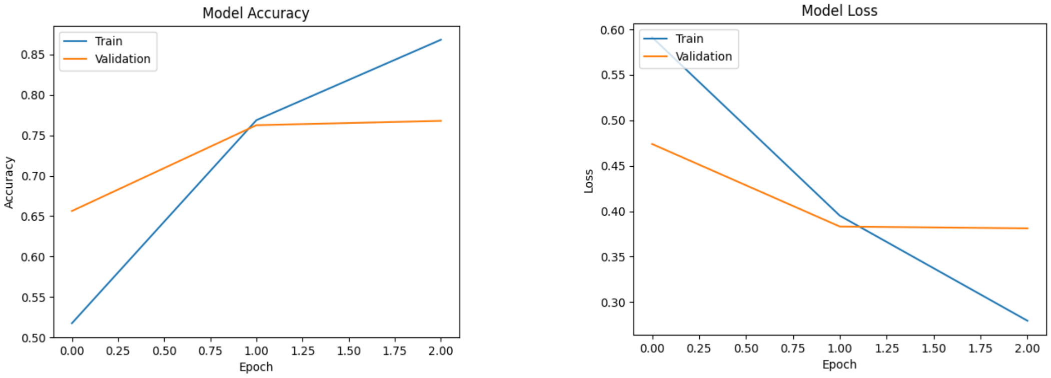


Figure 23: Accuracy and loss for the CNN model.

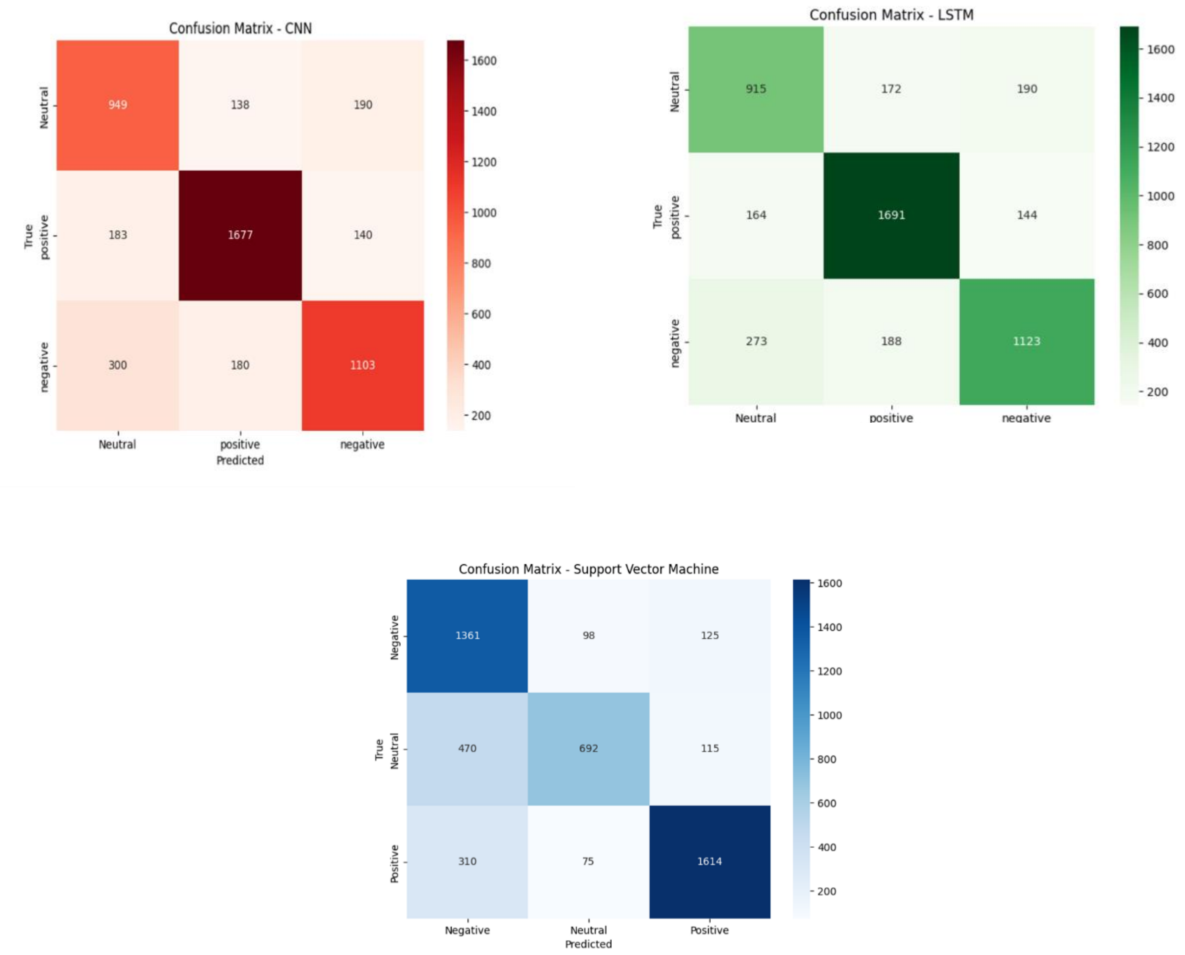


Figure 24: Confusion matrix of the best results.

4.10 Discussion:

Based on the given results, both the LSTM and CNN algorithms achieved the highest accuracies of 76.54% and 76.40% respectively, making them the top performers among the tested algorithms. SVM also demonstrated a commendable accuracy of 75.45%, making it another strong performer. On the other hand, RF, NB, and DT achieved slightly lower accuracies of 73.88%, 73.43%, and 68.93% respectively.

4.11 Presentation of our platform:

In this section, we will present some interfaces of the application, and we will implement the sentiment analysis algorithms, which consist of six algorithms.

4.11.1 Software development and libraries:

4.11.1.1 Python: is a popular high-level, interpreted programming language [53] that is often used for a broad range of tasks, including web development, scientific computing, data analysis, artificial intelligence, and machine learning. Python's quick code execution made possible by its interpreter mechanism makes it ideal for rapid prototyping. It is known for its ability to adapt, working successfully under Windows, Mac and Linux.

Python's syntax is similar to English, which facilitates learning and use. Programmers are able to develop programs in Python with fewer lines than in other languages, which improves readability and reduces maintenance costs.

4.11.1.2 Flask: is a web application framework. It is designed to make getting started quick and easy, with the ability to adapt to complex applications. It started as a simple wrapper around Werkzeug and Jinja and has become one of the most popular Python web application frameworks [54]. It is used to connect the Python code with the interface for our application.

4.11.1.3 Pandas: is a fast, powerful and easy use open source library for data analysis and manipulation in Python. It aims to be a fundamental tool for practical data analysis and strives to become the most powerful and flexible option across all languages. With its versatile capabilities, Pandas is suitable for various types of data especially numerical tables and time series. It provides efficient data manipulation with features like a high-performance DataFrame object, tools for reading/writing data in different formats, intelligent handling of missing data, flexible reshaping/pivoting, label-based slicing, fancy indexing, and efficient merging/joining of datasets [55]. Pandas can be seamlessly integrated with other powerful libraries and Python toolkits, enhancing productivity and performance in data analysis tasks.

4.11.1.4 NLTK (Natural Language Toolkit): NLTK is a popular open-source platform and collection of Python modules designed for natural language processing (NLP). It serves as a leading platform for developing Python applications that work with human language data. With NLTK, users can access more than 50 corpora and lexical resources through user-friendly interfaces. The platform also includes a variety of text processing libraries for tasks such as classification, tokenization, stemming, tagging, parsing, and semantic reasoning. NLTK provides wrappers for industrial-strength NLP libraries and has an active discussion forum where users can share their experiences and get support [56].

4.11.1.5 Scikit-Learn: is a free library for machine learning that is available for free. It offers a wide range of methods for supervised and unsupervised learning, including dimensionality reduction, classification, regression, and clustering. The library may be easily connected with other libraries like Pandas and Seaborn. It was constructed utilizing various libraries known as NumPy and SciPy [57].

4.11.2 Home page: The main interface of the application allows users to choose between typing text or uploading an Excel file for analysis.

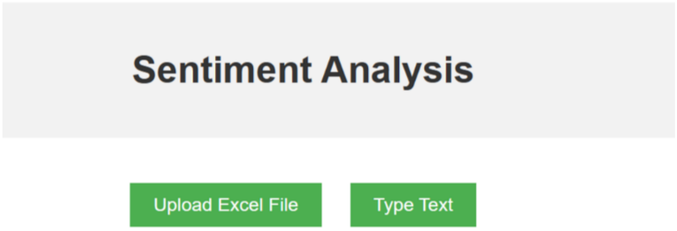
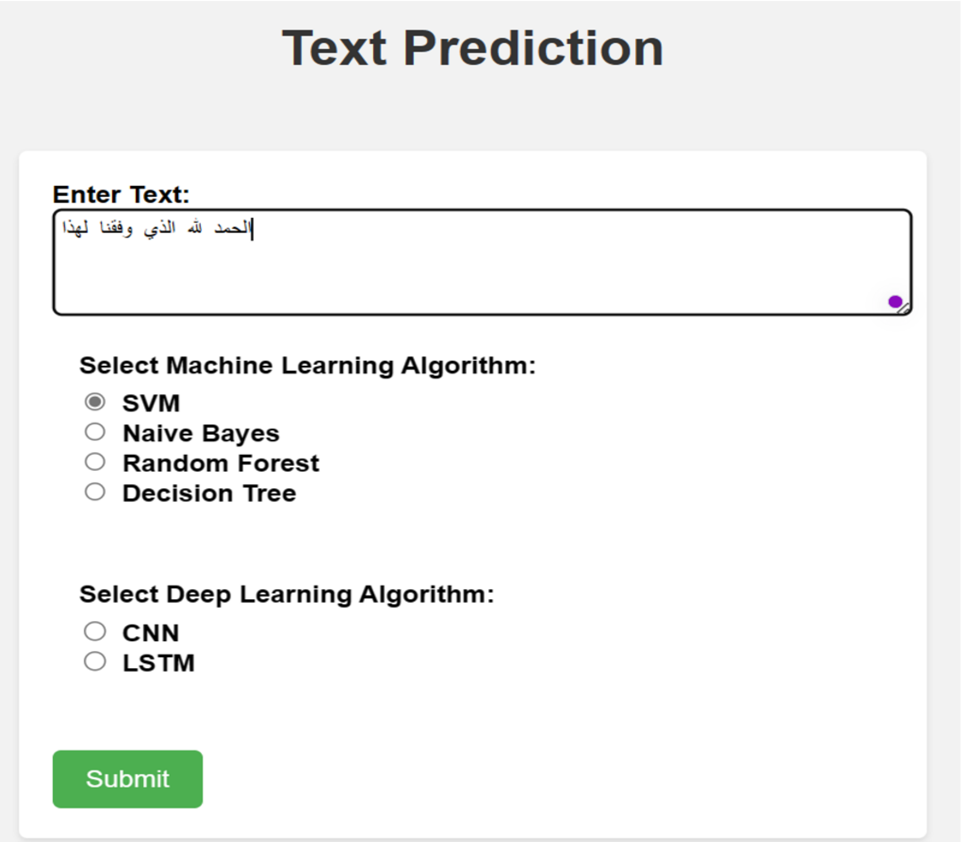


Figure 25: Home page.

4.11.3 Text prediction interface: the text prediction interface allows users to enter a text and choose an algorithm for text classification.



Text Prediction

Enter Text:

الحمد لله الذي وفقنا لهذا

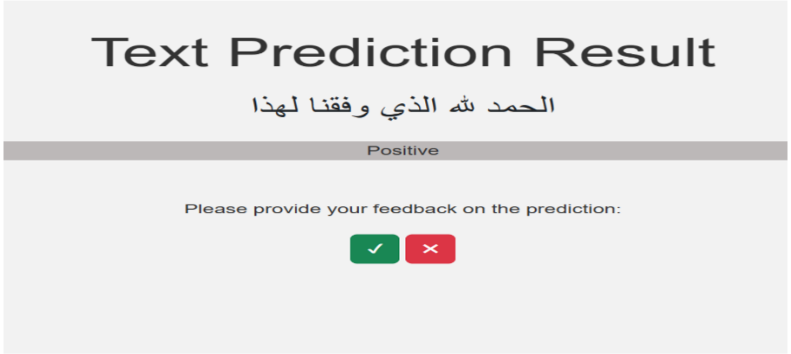
Select Machine Learning Algorithm:

- ● SVM
- ○ Naive Bayes
- ○ Random Forest
- ○ Decision Tree

Select Deep Learning Algorithm:

- ○ CNN
- ○ LSTM

Figure 26: Text prediction interface.



Text Prediction Result

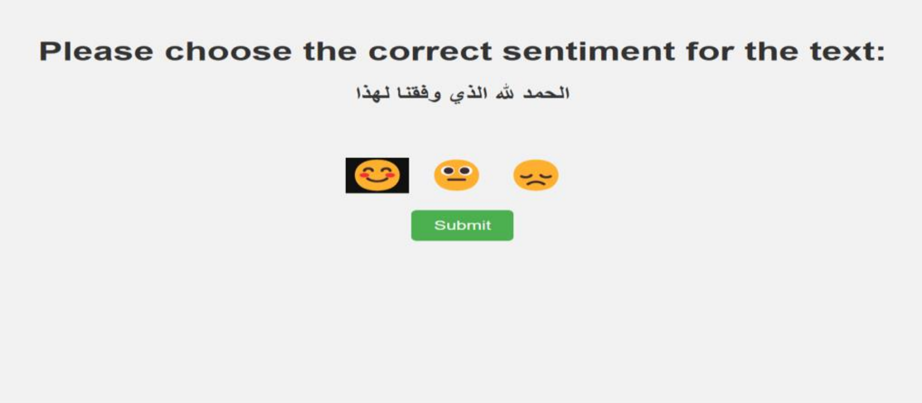
الحمد لله الذي وفقنا لهذا

Positive

Please provide your feedback on the prediction:

Figure 27: Text prediction result.

The feedback feature allows users to correct incorrect sentiment classifications.



Please choose the correct sentiment for the text:

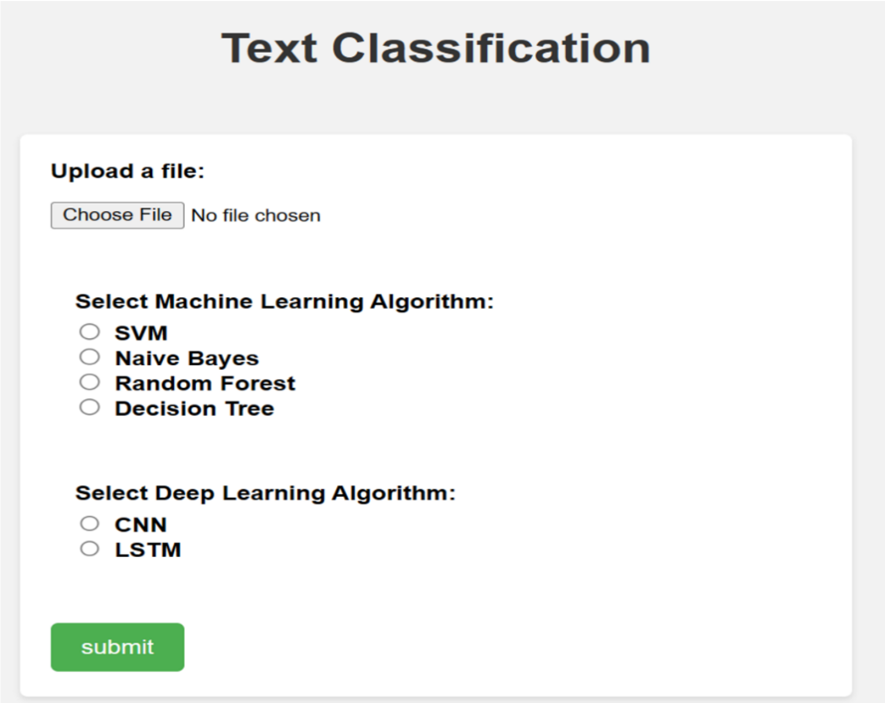
الحمد لله الذي وفقنا لهذا

😊 😐 😞

Submit

Figure 28: Correct the sentiment of the text.

4.11.4 Text classification interface: the text classification interface enables users to upload an Excel file and choose an algorithm for classification.



Text Classification

Upload a file:

Choose File No file chosen

Select Machine Learning Algorithm:

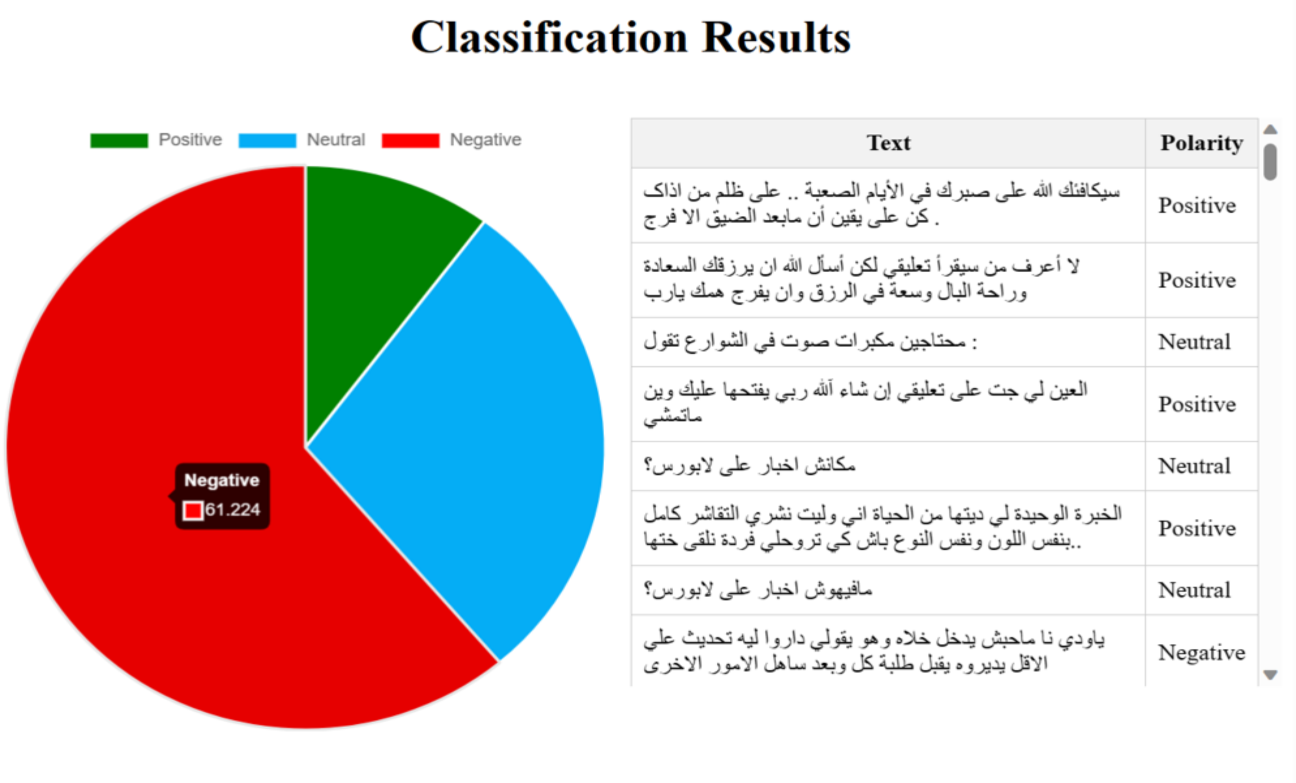
○ SVM
○ Naive Bayes
○ Random Forest
○ Decision Tree

Select Deep Learning Algorithm:

○ CNN
○ LSTM

submit

Figure 29: Text classification interface.



Classification Results

Text	Polarity
سيكافئك الله على صبرك في الأيام الصعبة .. على ظلم من اذاک كن على يقين أن مابعد الضيق الا فرج .	Positive
لا أعرف من سيقرأ تعليقي لكن أسأل الله ان يرزقك السعادة وراحة البال وسعة في الرزق وان يفرج همك يارب	Positive
محتاجين مكبرات صوت في الشوارع تقول :	Neutral
العين لي جت على تعليقي إن شاء آلله ربي يفتحها عليك وين ماتمشي	Positive
مكانش اخبار على لابورس؟	Neutral
الخبرة الوحيدة لي ديتها من الحياة اني وليت نشري التقاشر كامل بنفس اللون ونفس النوع باش كي تروحلي فردة نلقى ختها..	Positive
مافيهوش اخبار على لابورس؟	Neutral
ياودي نا ماحبش يدخل خلاه وهو يقولي داروا ليه تحديث على الاقل يديروه يقبل طلبة كل وبعد ساهل الامور الاخرى	Negative

Figure 30: Text classification results.

4.12 Conclusion:

In this chapter, we presented the main tools used to develop the application. We performed sentiment analysis on a corpus consisting of 24329 texts in Algerian dialect, labeled as follows: 9894 positive texts, 7883 negative texts, and 6552 neutral texts. We utilized various machine learning and deep learning classifiers, including Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM), along with some key interfaces of our system. We found that the Long Short-Term Memory (LSTM) classifier achieved the highest accuracy of 76.54%.

General Conclusion:

The objective of this thesis is to detect polarities of Arabic texts. We presented a platform that enables the generation of a database comprising tweets written in both Arabic and Algerian dialectal Arabic, including Arabizi. The main functionality of the platform was to classify given Arabic tweets or Algerian Arabic dialectal tweets into positive, negative, or neutral polarities. To successfully conduct this study, we utilized a dataset consisting of 24329 comments, each labeled as positive, negative, or neutral.

Our study commenced with an exploration of machine learning concept, followed by a detailed discussion on opinion mining, focusing on existing studies in the field. Furthermore, we studied the Arabic language, Algerian dialect, and Arabizi.

We employed various machine learning algorithms, including Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF), as well as deep learning algorithms including Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM), as classifiers for our analysis. After thorough evaluation, we found that the Long Short-Term Memory (LSTM) classifier achieved the highest accuracy, reaching 76.54%.

Perspectives:

For future work, several possibilities can be considered, including:

- Expanding the size of sentiment databases by incorporating a wider range of words from standard Arabic and Algerian dialects.
- Exploring alternative classifiers and incorporating the N-gram concept in the development of this work.
- Developing an analysis system that incorporates multilingual sentiment analysis. Algerian websites often contain multiple languages, including Algerian dialect, French, Arabic, and English. By considering multilingual sentiment analysis, the system would be better equipped to handle the diverse linguistic content found on these websites.
- Extending the analysis to include mixed classes in addition to the positive, negative, and neutral classes. Mixed class sentiment analysis would enable a more nuanced understanding of text data, capturing situations where multiple sentiments coexist within a single document or sentence.

Bibliography:

[1] T. M. Mitchell, *Machine Learning*. in McGraw-Hill series in computer science. New York: McGraw-Hill, 1997.

[2] Massih-Reza Amini, *Apprentissage machine: De la théorie à la pratique. Concepts fondamentaux en Machine Learning*. 2015.

[3] “Machine learning, explained | MIT Sloan,” Jun. 07, 2023. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (accessed Jun. 08, 2023).

[4] “Machine Learning: What It is, Tutorial, Definition, Types - Javatpoint.” https://www.javatpoint.com/machine-learning (accessed Jun. 08, 2023).

[5] T. O. Ayodele, “Machine Learning Overview,” in *New Advances in Machine Learning*, IntechOpen, 2010. doi: 10.5772/9374.

[6] B. Mahesh, “Machine Learning Algorithms - A Review,” vol. 9, no. 1, 2018.

[7] “Supervised Machine learning - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/supervised-machine-learning (accessed Jun. 08, 2023).

[8] M. Birjali, M. Kasri, and A. Beni-Hssane, “A comprehensive survey on sentiment analysis: Approaches, challenges and trends,” *Knowl.-Based Syst.*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[9] R. Amiri, H. Mehrpouyan, L. Fridman, R. Mallik, A. Nallanathan, and D. Matolak, “A Machine Learning Approach for Power Allocation in HetNets Considering QoS,” Mar. 2018.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[11] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.

[12] “RECURRENT NEURAL NETWORKS,” 2001.

[13] W. Li, F. Qi, M. Tang, and Z. Yu, “Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification,” *Neurocomputing*, vol. 387, pp. 63–77, Apr. 2020, doi: 10.1016/j.neucom.2020.01.006.

[14] S. Zhou, Q. Chen, and X. Wang, “Fuzzy deep belief networks for semi-supervised sentiment classification,” *Neurocomputing*, vol. 131, pp. 312–322, May 2014, doi: 10.1016/j.neucom.2013.10.011.

[15] M. Bahi and M. Batouche, *Deep Learning for Ligand-Based Virtual Screening in Drug Discovery*. 2018, p. 5. doi: 10.1109/PAIS.2018.8598488.

[16] T. Rahmat, A. Ismail, and S. Aliman, “Chest X-ray Image Classification using Faster R-CNN,” *Malays. J. Comput.*, vol. 4, May 2019, doi: 10.24191/mjoc.v4i1.6095.

[17] “LSTM RNN in Tensorflow - Javatpoint.” https://www.javatpoint.com/long-short-term-memory-rnn-in-tensorflow (accessed Jun. 20, 2023).

[18] K. Buvaneshwaran, “Support Vector Machine(SVM) In Machine Learning - CopyAssignment,” Aug. 14, 2022. https://copyassignment.com/support-vector-machine-svm-in-machine-learning/ (accessed Jun. 08, 2023).

[19] T. A. Team, “Decision Trees Explained With a Practical Example – Towards AI.” https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53, https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53 (accessed Apr. 16, 2023).

[20] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, “Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 127, pp. 511–520, 2018, doi: 10.1016/j.procs.2018.01.150.

[21] B. YEGNANARAYANA, *ARTIFICIAL NEURAL NETWORKS*. PHI Learning Pvt. Ltd., 2009.

[22] K. L. Priddy and P. E. Keller, *Artificial Neural Networks: An Introduction*. SPIE Press, 2005.

[23] “opinion,” May 03, 2023. https://dictionary.cambridge.org/dictionary/english/opinion (accessed May 03, 2023).

[24] N. Mishra and C. K. Jha, “Classification of Opinion Mining Techniques,” *Int. J. Comput. Appl.*, vol. 56, no. 13, pp. 1–6, Oct. 2012, doi: 10.5120/8948-3122.

[25] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[26] S. Kausar, X. Huahu, W. Ahmad, M. Y. Shabir, and W. Ahmad, “A Sentiment Polarity Categorization Technique for Online Product Reviews,” *IEEE Access*, vol. 8, pp. 3594–3605, 2020, doi: 10.1109/ACCESS.2019.2963020.

[27] “What Is Opinion Mining & Why Is It Essential?,” *MonkeyLearn Blog*, Sep. 16, 2020. https://monkeylearn.com/blog/opinion-mining/ (accessed Apr. 01, 2023).

[28] V. M., J. Vala, and P. Balani, “A Survey on Sentiment Analysis Algorithms for Opinion Mining,” *Int. J. Comput. Appl.*, vol. 133, no. 9, pp. 7–11, Jan. 2016, doi: 10.5120/ijca2016907977.

[29] A. Assiri, A. Emam, and H. Aldossari, “Arabic Sentiment Analysis: A Survey,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 12, 2015, doi: 10.14569/IJACSA.2015.061211.

[30] T. Hardeniya and D. A. Borikar, “Dictionary Based Approach to Sentiment Analysis - A Review,” vol. 2, no. 5.

[31] M. Darwich, S. A. Mohd Noah, N. Omar, and N. A. Osman, “Corpus-Based Techniques for Sentiment Lexicon Generation: A Review,” *J. Digit. Inf. Manag.*, vol. 17, no. 5, p. 296, Oct. 2019, doi: 10.6025/jdim/2019/17/5/296-305.

[32] A. Sadia, F. Khan, and F. Bashir, “An Overview of Lexicon-Based Approach For Sentiment Analysis”.

[33] N. A. M. Razali *et al.*, “Opinion mining for national security: techniques, domain applications, challenges and research opportunities,” *J. Big Data*, vol. 8, no. 1, p. 150, Dec. 2021, doi: 10.1186/s40537-021-00536-5.

[34] A. Ghallab, A. Mohsen, and Y. Ali, “Arabic Sentiment Analysis: A Systematic Literature Review,” *Appl. Comput. Intell. Soft Comput.*, vol. 2020, pp. 1–21, Jan. 2020, doi: 10.1155/2020/7403128.

[35] R. M. Duwairi, R. Marji, N. Sha’ban, and S. Rushaidat, “Sentiment Analysis in Arabic tweets,” in *2014 5th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan: IEEE, Apr. 2014, pp. 1–6. doi: 10.1109/IACS.2014.6841964.

[36] M. Mataoui, O. Zelmati, and M. Boumechache, “A Proposed Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic,” *Res. Comput. Sci.*, vol. 110, no. 1, pp. 55–70, Dec. 2016, doi: 10.13053/rcs-110-1-5.

[37] M. A. Ayyoub, S. B. Essa, and I. Alsmadi, “Lexicon-based sentiment analysis of Arabic tweets,” *Int. J. Soc. Netw. Min.*, vol. 2, no. 2, p. 101, 2015, doi: 10.1504/IJSNM.2015.072280.

[38] A. Al-Thubaity, Q. Alqahtani, and A. Aljandal, “Sentiment lexicon for sentiment analysis of Saudi dialect tweets,” *Procedia Comput. Sci.*, vol. 142, pp. 301–307, 2018, doi: 10.1016/j.procs.2018.10.494.

[39] H. AL-Rubaiee, R. Qiu, K. Alomar, and D. Li, “Sentiment Analysis of Arabic Tweets in e-Learning,” *J. Comput. Sci.*, vol. 12, no. 11, pp. 553–563, Nov. 2016, doi: 10.3844/jcssp.2016.553.563 (accessed Apr. 16, 2023).

[40] A. Elouardighi, M. Maghfour, H. Hammia, and F.-Z. Aazi, “Analyse des sentiments à partir des commentaires Facebook publiés en Arabe standard ou dialectal marocain par une approche d’apprentissage automatique” (accessed Apr. 17, 2023).

[41] M. A. Sghaier, H. Abdellaoui, R. Ayadi, and M. Zrigui, “Analyse de sentiments et extraction des opinions pour les sites e-commerce : application sur la langue arabe” (accessed Apr. 16, 2023).

[42] H. K. Aldayel and A. M. Azmi, “Arabic tweets sentiment analysis – a hybrid scheme,” *J. Inf. Sci.*, vol. 42, no. 6, pp. 782–797, Dec. 2016, doi: 10.1177/0165551515610513 (accessed Apr. 16, 2023).

[43] Babbel.com and L. N. GmbH, “How Many People Speak Arabic Around The World, And Where?,” *Babbel Magazine*. https://www.babbel.com/en/magazine/how-many-people-speak-arabic (accessed Jun. 05, 2023).

[44] U. Nations, “Official Languages,” *United Nations*. https://www.un.org/en/our-work/official-languages (accessed Jun. 05, 2023).

[45] Quranic, “The Types Of Arabic And Their Differences | Quranic Arabic,” *Quranic Arabic For Busy People*, Sep. 29, 2021. https://www.getquranic.com/types-of-arabic-and-their-differences/ (accessed Apr. 16, 2023).

[46] “The Richest Language?” https://www.linkedin.com/pulse/richest-language-ali-moghnieh (accessed Apr. 16, 2023).

[47] Christina, “Arabic Language | Learn Million-Word Language | Lonet.Academy,” *Lonet.Academy Blog*, Jan. 25, 2019. https://lonet.academy/blog/learn-the-million-word-arabic-language/ (accessed Apr. 16, 2023).

[48] K. Meftouh, K. Smaili, and Nadjette Bouchemal, “A study of non-resourced language: the case of an Algerian dialect,” 2012, doi: 10.13140/RG.2.1.4881.1041 (accessed Apr. 16, 2023).

[49] R. Cotterell, A. Renduchintala, N. Saphra, and C. Callison-Burch, “An Algerian Arabic-French Code-Switched Corpus” (accessed Apr. 16, 2023).

[50] “The ultimate guide to the Arabizi language - Cudoo.” https://cudoo.com/blog/the-ultimate-guide-to-the-arabizi-language/ (accessed Jun. 04, 2023).

[51] M. Al-Badrashiny, R. Eskander, N. Habash, and O. Rambow, “Automatic Transliteration of Romanized Dialectal Arabic,” in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Ann Arbor, Michigan: Association for Computational Linguistics, 2014, pp. 30–38. doi: 10.3115/v1/W14-1604 (accessed Apr. 16, 2023).

[52] H. Mulki, H. Haddad, and I. Babaog, “Modern Trends in Arabic Sentiment Analysis: A Survey”.

[53] “What is Python? Executive Summary,” *Python.org*. https://www.python.org/doc/essays/blurb/ (accessed May 23, 2023).

[54] A. Ronacher, “Flask: A simple framework for building complex web applications.”

[55] “pandas - Python Data Analysis Library.” https://pandas.pydata.org/about/ (accessed May 23, 2023).

[56] “NLTK :: Natural Language Toolkit.” https://www.nltk.org/ (accessed May 23, 2023).

[57] Nik, “Introduction to Scikit-Learn (sklearn) in Python • datagy,” *datagy*, Jan. 05, 2022. https://datagy.io/python-scikit-learn-introduction/ (accessed May 23, 2023).

[58] “Algerian Dialect Review for sentiment analysis.” https://www.kaggle.com/datasets/djoughimehdi/algerian-dialect-review-for-sentiment-analysis (accessed Jun. 09, 2023).

[59] “Algerian Corpus (Algerian Dataset).” https://www.kaggle.com/datasets/massinissaissighid/algerian-corpus-algerian-dataset (accessed Jun. 09, 2023).

[60] “GitHub - kahinasassi/Algerian_dialect_dataset:” https://github.com/kahinasassi/Algerian_dialect_dataset (accessed Jun. 09, 2023).

[61] “GitHub - massiAp1/Algerian-oppinion-mining: Algerian text classification with Machine Learning, new corpus, new text preprocessing.” https://github.com/massiAp1/Algerian-oppinion-mining (accessed Jun. 09, 2023).