order number: . . . . .

Thesis submitted to the

UNIVERSITY OF MOHAMED BOUDIAF – MSILA

FACULTY OF MATHMATICS AND COMPUTER SCIENCE
DEPARTEMENT OF COMPUTER SCIENCE

In partial fulfillment of the requirements for the degree of

Master in **Computer science**

By

Khelifi, Ayoub

Sahbi, Ramzi

Title of the thesis

DEEP LEARNING MODELS APPLIED TO ARABIC QURANIC TEXT

Under the supervision of

Madiha Halassa

Composition of the jury

Mourad Brik	University of Msila	President
Nour Elhouda Chalabi	University of Msila	Examiner

June, 2024

Dedications

نحمد االله عز وجل على منه وعونه لإتمام هذا البحث.

إلى من لا يضاهيهما أحد في الكون، إلى من أمرنا الله ببرّهما، إلى من بذلا الكثير، وقدّما ما لا يمكن أن يردّ، إليكما تلك الكلمات أمي وأبي الغاليان، أهدي لكما هذا البحث؛ فقد كنتما خير داعم لي طوال مسيرتي الدراسية.

إلى إخواننا وأخواتنا المجاهدين والمرابطين في غزة وفلسطين.

Khelifi Ayoub, Sahbi Ramzi

Acknowledgments

We are deeply indebted to our thesis supervisor Mrs. Madiha Hallassa whose unlimited steadfast support and inspirations have made this project a great success. In a very special way, we thank him for every support he has rendered unto us to see that we succeed in this challenging study.

Special thanks go to our friends and families who have contained the hectic moments and stress we have been through during the course of the research project.

We also thank Dr. Taha Zerrouki for his help.

Contents

List of figures

List of tables

Introduction

Understanding the topics covered in the verses of the Quran is a central interest for Islamic scholars, Quranic studies experts, and others. Traditional classification methods for Quranic verses can be simplified and enhanced using automated techniques such as Natural Language Processing (NLP) and Machine Learning (ML). This scientific report explores the application of deep learning algorithms to classify the Holy Quran verses and provides insights into the implementation, training, and evaluation of such systems.

Many textual classification studies focused on the French and English versions. The Arabic text classification is less well known and more difficult. For this purpose, we worked on this research, which concerns the application of natural language processing to the Holy Quran and its verses. The Quran is God's word referred to our Prophet Muhammad. When searching for the classification of the Holy Quran verses according to the subject matter of the verse, we discovered the lack of research and work on this subject. So, our goal is to design and develop a deep learning model to classify Quranic verses with high accuracy. Since the Holy Quran contains many verses talking about different subjects.

This work has 4 chapters:

The first chapter is mainly dedicated to the concepts of artificial intelligence, machine learning and deep learning as well as synthetic neural networks.

The second chapter, is about Arabic and its unique characteristics, which present a great challenge in the processing of natural language. We then presented a study of some aspects of the Holy Quran as well as text data. We also spoke about previous research in this area and its results.

In the third chapter, which provides an overview of the classification of texts, types and applications, and we have described the supervised classification algorithms.

Finally, the fourth chapter in which we presented the different techniques we used, the pre-processing processes on Quranic text for preparation, and then the part where we applied different deep learning models to our dataset, after which the results obtained and evaluated were presented.

Chapter1: Machine Learning and Deep Learning

1. Introduction

This chapter is devoted for talk about the Machine Learning and Deep Learning in general, by granting its definition, its main classes and the different approaches used. We will define the applications of deep learning in many domains, we will then present the types of neural networks, and their general architecture. This chapter's final point relates to the importance of deep learning and its challenges.

2. Artificial intelligence (AI)

Artificial Intelligence (AI) represents a dynamic branch of computer science dedicated to the creation of machines capable of performing tasks that conventionally require human intelligence. These artificially intelligent systems are made to simulate human brain functions, such as learning, reasoning, and decision-making, so that they can develop and become more capable via experience.

Artificial intelligence is a branch of computer science that includes machine learning and deep learning, and is frequently thrown around in tandem with machine learning or deep learning. These fields focus on creating artificial intelligence (AI) algorithms that can "learn" from existing data and gradually provide predictions or classifications that are more accurate. These algorithms are modelled after the decision-making processes of the human brain.

2.1. Definitions

A common definition of AI is that it is a technology that enables machines to imitate various complex human skills [1]

Artificial intelligence, or AI, is technology that enables computers and machines to simulate human intelligence and problem-solving capabilities.

3. Machine Learning

3.1. Definitions

Some people identify machine learning as follows:

3.1.1. Definition 1: Machine Learning is an application of Artificial Intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [4].

4. Deep Learning

4.1. Definitions

Several definitions of Deep Learning have emerged in recent years we cite in this context the following three definitions:

4.1.1. Definition 1:

Deep learning is a machine learning process using neural networks with several layers of hidden neurons. Since these algorithms have many parameters, they require a very large amount of data in order to be trained. [2]

4.1.2. Definition 2:

Deep learning is one of the main technologies of machine learning. With Deep Learning, we are talking about algorithms capable of mimicking the actions of the human brain thanks to artificial neural networks. The networks are composed of dozens or even hundreds of "layers" of neurons, each receiving and interpreting the information of the previous layer [3].

4.1.3. Definition 3:

One subfield of machine learning, known as deep learning (DL), is derived from the field of artificial intelligence (ML) It is a very recent area of research with a flashing popularity. This is due to its impressive performance on several problems long considered to be very difficult, as well as the very large number of techniques that greatly facilitate its use.

4.2. Rises and Historical Context

The evolution of deep learning has been a remarkable journey characterized by significant advancements and breakthroughs in artificial intelligence research. Initially rooted in the 1940s with the introduction of computational models inspired by neural networks, progress in the field stagnated during the AI winter of the 1970s and 1980s, as limitations in computing power and data availability hindered further development. However, a resurgence occurred in the late 1980s to the early 2000s, fueled by the introduction of backpropagation and exploration of architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This period laid the groundwork for the emergence of deep learning in the mid-2000s to the present, driven by breakthroughs in computational power, the availability of large-scale datasets, and advancements in optimization algorithms. Deep learning achieved widespread recognition with successes in image recognition competitions and breakthroughs in natural language processing and speech recognition. Looking ahead, ongoing research efforts are focused on addressing challenges and exploring new frontiers in efficiency, integration with

other AI techniques, and novel training paradigms, ensuring the continued evolution and impact of deep learning in shaping the future of technology and artificial intelligence [5]**.**

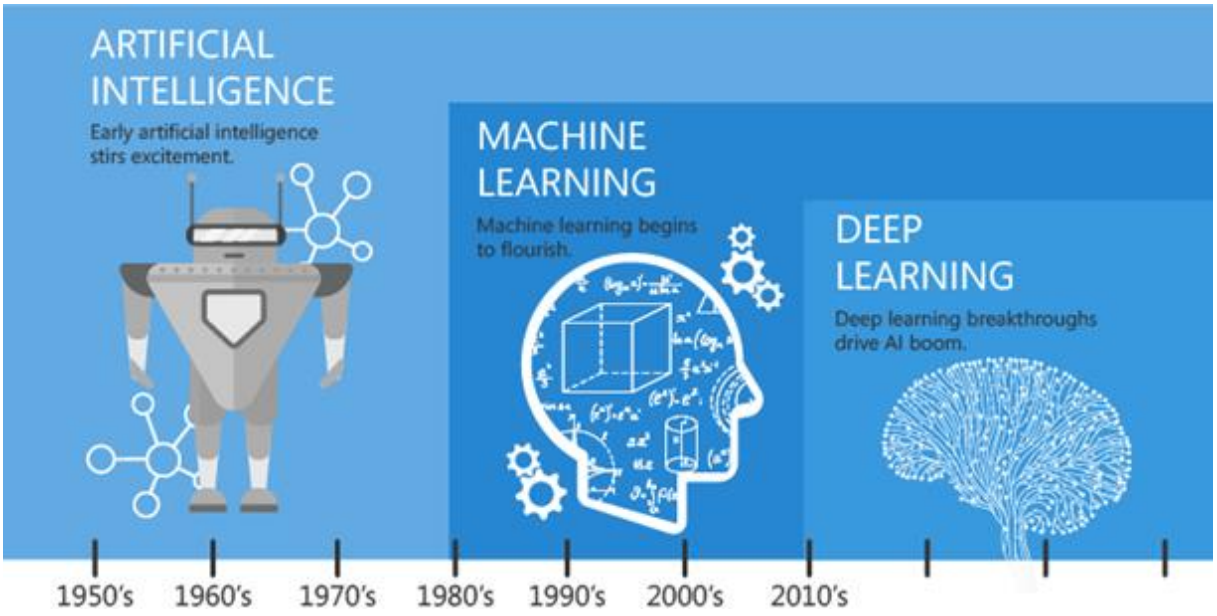

Fig. 1.1: The evolution and different phases of artificial intelligence (AI), machine learning, and deep learning.

4.3. The applications of deep learning

There are several uses of Deep learning, we cite in this context the following Applications:

4.3.1. Computer vision:

Computer vision is the computer's ability to extract information and insights from images and videos. Computers can use deep learning techniques to comprehend images in the same way that humans do. Computer vision has several applications, such as the following:

- Content moderation to automatically remove unsafe or inappropriate content from image and video archives.
- Facial recognition to identify faces and recognize attributes like open eyes, glasses, and facial hair.
- Image classification to identify brand logos, safety gear, and other image details [6].

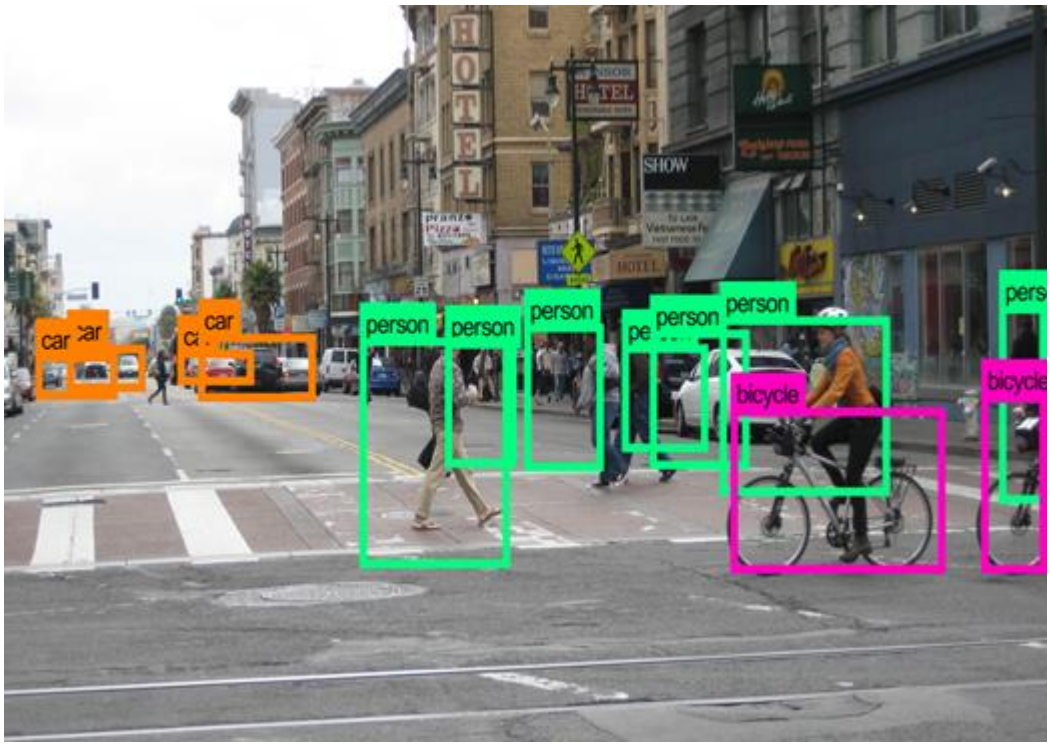

Fig. 1.2: The_Era-of-Computer-Vision-Is-Here.

4.3.2. Self-driving cars:

In self-driving cars, deep learning is used to create accurate models of the world around the car so that it can make driving decisions. These models are created by training a neural network on a large dataset of images and driving data. The neural network can then generalize from this data and make predictions about new data, such as what objects are in an image or what the car should do in each situation. Tesla is one popular example [7].

4.3.3. Healthcare:

The healthcare industry contends with inefficiencies, but deep learning plays a crucial role in streamlining the patient experience. KenSci, a company under the Advata umbrella, uses AI technology that learns from past performance data to predict how much space and what resources teams need to provide proper patient care. In addition, PathAI harnesses the predictive abilities of AI to garner more accurate data from drug research, clinical trials and patient diagnostics. Deep learning has also been proven to detect skin cancer through images, according to a National Centre for Biotechnology Report [8].

4.3.4. Robotics:

Robots can now automatically learn new skills and improve their existing ones caused by the widespread use of deep learning algorithms in the robotics industry. Similar to how people

learn, deep learning computers may also learn from data. It enables robots to perform better at a task without assistance from a human. Deep learning algorithms have made it possible for robots to communicate with people, recognize and handle things, and navigate through unfamiliar situations on their own. Here are some examples of how they are used in different robotic systems as, Object Detection and Recognition: Object detection and recognition are critical tasks in robotics that have become possible thanks to deep learning. By training neural networks with massive amounts of labelled data, robots can identify and classify objects in their environment with high accuracy [9].

4.3.5. Natural language processing (NLP):

Deep Learning algorithms have revolutionized Natural Language Processing, which enable autonomous meaning extraction from text, these algorithms have produced state-of-the-art results, in a variety of applications that we will mention some of them:

Fig. 1.3: Natural language processing.

- **Text Classification:**

This is one of the tasks we will be working on in our project, classification has several types:

- Topic labelling: classification of texts into predefined topics.
- Spam detection: The algorithms analyse and filter the email.

The importance of such categorization approach is to organize knowledge so that some specific treatments can be performed, including; information retrieval and efficient information extraction [10].

- **Sentiment analysis:**

is a process of identification and extraction of subjective information from a lot of data available on the web to determine the positive, negative, or neutral sentiment feelings of the public towards a particular topic or entity [11].

- **Chatbots:**

In the chatbots industry, deep learning creates chatbots that can understand and respond to human queries in natural language. It is one of the practical applications of deep learning. By using deep learning, chatbots can learn to recognize the intent of a user's utterance and generate an appropriate response. It allows chatbots to have conversations with users that feel natural and human-like [12].

- **Question Answering Systems:**

There is a database ready with "Q&A" duos, algorithms trained on them. Where the question is asked the model, the answer appears as a result.

5. Neural networks

Neural networks are the core architecture of deep learning algorithms, and deep learning and neural networks are closely intertwined. We'll examine the connection between deep learning and neural networks in this section, seeing how deep learning makes use of neural network topologies to accomplish amazing feats in artificial intelligence.

5.1. Basics of Neural Networks

Neural networks, at their core, are computational models designed to simulate the functioning of biological neural networks. Conceived in the 1950s, the early neural network models, known as perceptron’s, Perceptron is also known as an artificial neural network. It is mainly used to compute the logical gate like AND, OR, and NOR which has binary input and binary output.

laid the groundwork for modern artificial neural networks. A perceptron consists of interconnected nodes, or neurons, organized into layers, each performing simple computations. Despite their promising potential, early neural networks faced limitations in handling complex

tasks due to computational constraints and the absence of efficient training algorithms. However, the late 20th century witnessed renewed interest and advancements in neural network research, fuelled by breakthroughs in algorithmic techniques and computational resources. The development of backpropagation, an efficient training algorithm for multi-layer networks, marked a significant milestone, enabling neural networks to learn hierarchical representations of data. This resurgence in neural network research paved the way for the emergence of deep learning, characterized by the integration of multiple layers of neurons, thereby enhancing the network's capacity to learn and represent intricate patterns in data Neural Networks and Deep Learning [13].

5.2. Understanding the Mechanics of Neural Networks

These are some basic concepts about neural network:

- **Neurons:** Basic units of a neural network. Each neuron receives input, processes it, passes it through an activation function, and forwards the output to the next layer of neurons.
- **Weights and Biases:** Weighted neuronal circuits connect neurons. The network's performance can be optimized during training by adjusting these weights and biases.
- **Activation Function:** This function computes the weighted total and adds bias to it in order to decide whether or not to activate a neuron. Adding non-linearity to a neuron's output is the aim of the activation function.

Every neuron is connected with other neuron through a connection link. Each connection link is associated with a weight that has information about the input signal. This is the most useful information for neurons to solve a particular problem because the weight usually excites or inhibits the signal that is being communicated. Each neuron has an internal state, which is called an activation signal. Output signals, which are produced after combining the input signals and activation rule, may be sent to other units.

Here activation functions can be anything like sigmoid, tanh, Rule Based on the requirement we will be choosing the most appropriate nonlinear activation function to produce the better result. Now let us implement a single-layer perceptron.

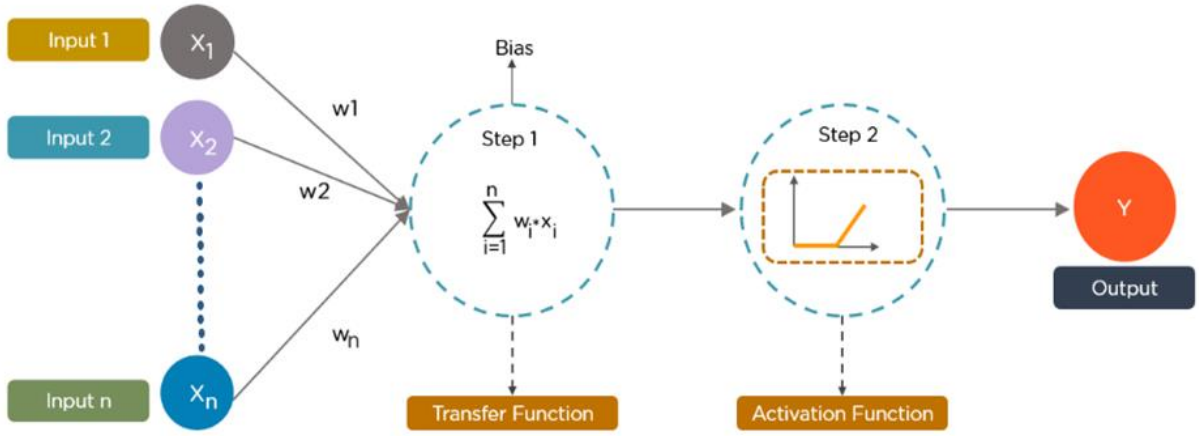

Fig. 1.4: Single-layer neural network.

5.3. Biological Neuron

A nerve cell (neuron) is a special biological cell that processes information. According to an estimation, there are huge number of neurons, approximately 1011 with numerous interconnections, approximately 10^{15} [14].

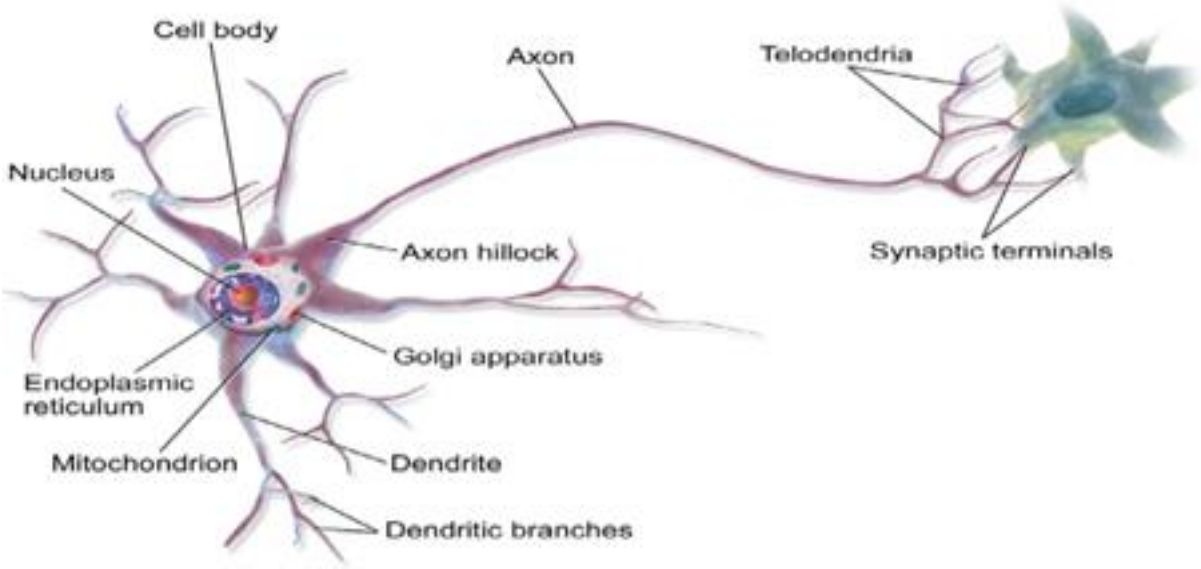

Fig. 1.5: Structure of a biological neuron.

Biological neuron has three basic functionalities

- Receive signal from outside.
- Process the signal and enhance whether we need to send information or not.

- Communicate the signal to the target cell which can be another neuron or gland.

In the same way, neural networks also work.

5.4. Neural Networks Models

There are many types of neural networks and each one has its own characteristics and application, some common types of neural networks include:

5.4.1. Feed-forward neural networks

This simple neural network variant passes data in a single direction through various processing nodes until the data reaches the output node. Feed-forward neural networks are designed to process large volumes of 'noisy' data and create 'clean' outputs. This type of neural network is also known as the multi-layer perceptron (MLPs) model.

A feed-forward neural network architecture includes the input layer, one or more hidden layers, and the output layer. Despite their alternate name, these models leverage sigmoid neurons rather than perceptron, thus allowing them to address nonlinear, real-world problems [15], commonly used for tasks like regression and classification.

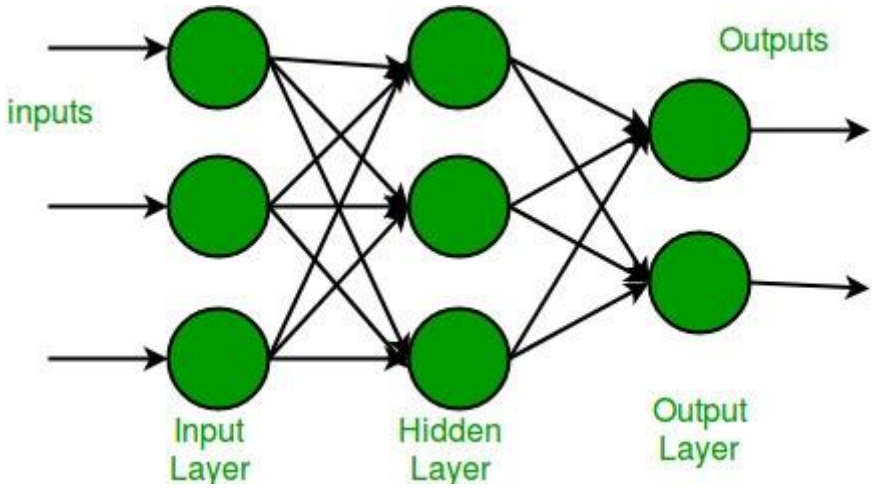

Fig. 1.6: Neural Network in Machine Learning.

5.4.2. Convolutional Neural Networks (CNNs)

The Convolutional Neural Network (CNN or ConvNet) is a popular discriminative deep learning architecture that learns directly from the input without the need for human feature extraction [16].

A convolutional neural network (CNN) is based on a multilayer perceptron variant. A CNN may have one or more convolutional layers. These levels might be fully linked or pooled. The

convolutional layer performs a convolutional operation on the input before sending the output to the next layer. Because of this convolutional process, the network may be considerably deeper while requiring many fewer parameters. Convolutional neural networks do exceptionally well in image and video recognition, natural language processing, and recommender systems as a result of this capacity. Convolutional neural networks do well in semantic parsing and paraphrase identification as well. They are also used in image categorization and signal processing [17].

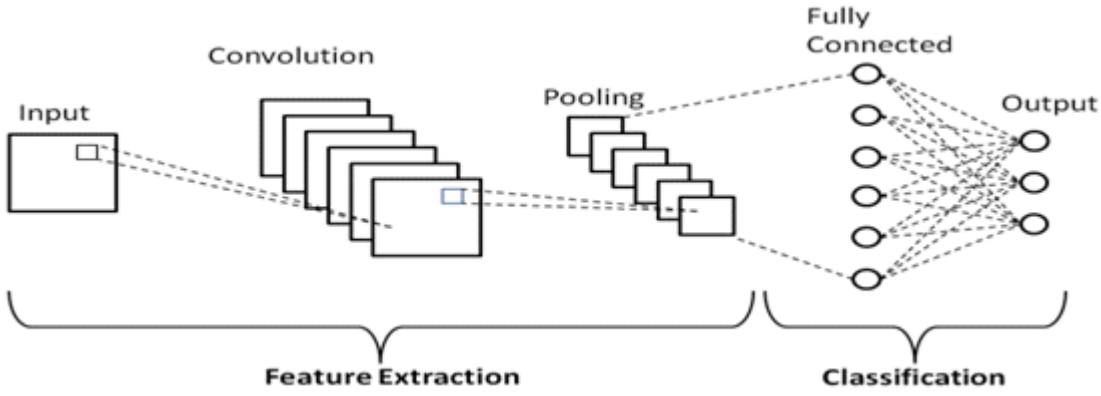

Fig. 1.7: Basic CNN Architecture.

- **The convolutional layer**

The convolutional layer, which is comprised of several convolution kernels, is responsible for exploring and filtering the training sample. The "weighted summation kernel layer" is regenerated together with the "weight matrix" of the fed data set by the convolutional layer, which is responsible for producing the "weight matrix" of the fed data set. In order to decrement the magnitude of the input, the filter makes use of integer values. The performance of convolutional kernels may be improved by adjusting three crucial hyperparameters: the filter 27 size, the zero padding, and the stride. Choosing the values that are ideal can help reduce the amount of complexity that the network has, which in turn can improve its accuracy.

- **Pooling layer**

Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to decrease the computational power required to process the data by reducing the dimensions. There are two types of pooling average pooling and max pooling.

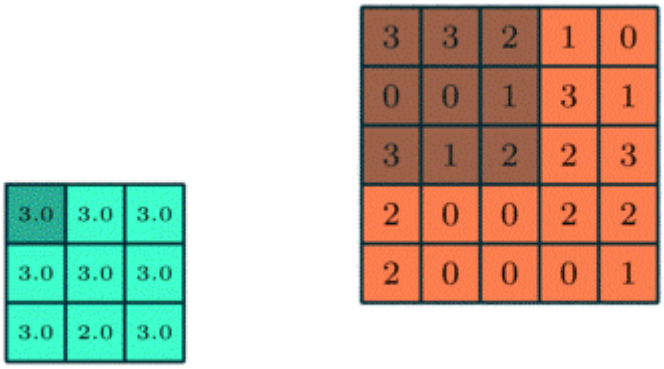

Fig. 1.8: Layers of Efficient B2.

So, what we do in Max Pooling is we find the maximum value of a pixel from a portion of the image covered by the kernel. Max Pooling also performs as a Noise Suppressant. It discards the noisy activations altogether and also performs de-noising along with dimensionality reduction.

On the other hand, Average Pooling returns the average of all the values from the portion of the image covered by the Kernel. Average Pooling simply performs dimensionality reduction as a noise suppressing mechanism. Hence, we can say that Max Pooling performs a lot better than Average Pooling [18].

5.4.3 Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other. Still, in cases when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus, RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is its Hidden state, which remembers some information about a sequence. The state is also referred to as Memory State since it remembers the previous input to the network. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. This reduces the complexity of parameters, unlike other neural networks [19].

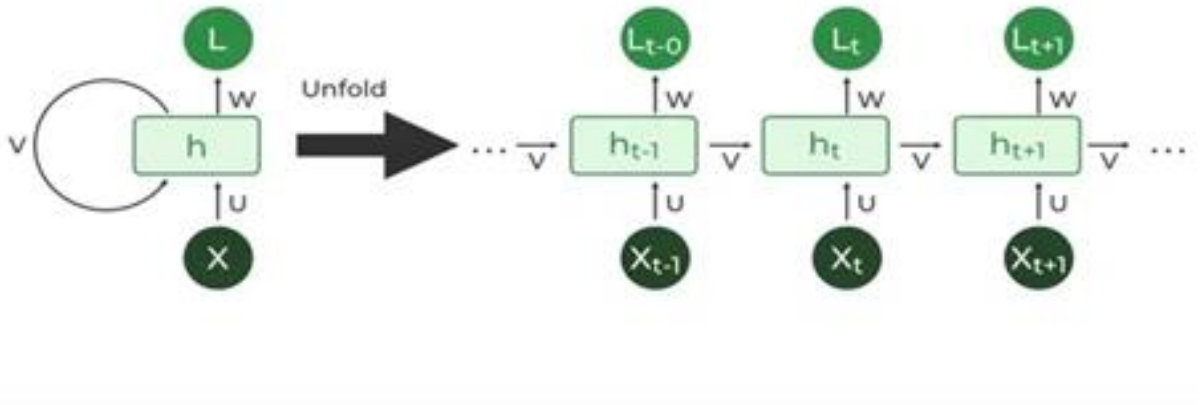

Fig. 1.9: The Architecture of a Basic RNN.

The architecture of a basic RNN consists of the following components:

- **Input Layer:**

This layer displays the input characteristics for every time interval in the series.

- **Recurrent Connection:**

which enables information to remain alive over various time steps, is an essential component of an RNN. Every time step, the output is generated and the hidden state is updated by combining the input from the previous time step with the current input.

- **Hidden State:**

The hidden state captures information about previous inputs in the sequence. It is updated at each time step based on the current input and the previous hidden state.

- **Output Layer:**

Using the current input and the hidden state as inputs, the output layer generates the output for the current time step.

RNNs have shown great success in many NLP tasks. At this point I should mention that the most commonly used type of RNNs are LSTMs, which are much better at capturing long-term dependencies than vanilla RNNs are. But don't worry, LSTMs are essentially the same thing as the RNN, they just have a different way of computing the hidden state.

5.4.3.1 Long Short-Term Memory Networks (LSTM)

Recurrent neural networks have an extension called Long Term Memory Networks (LSTM) that increases their memory. As a result, it is ideal for learning about important events occurring at widely spaced intervals.

due to the gradient dissipation problem (related to their gradient descent learning method), old information is easily forgotten. They say they have a short memory. The «Long Short-Term Memory» network (LSTM) is a network of short-term and long-term memory recurrent neurons that allows, thanks to its structure, to partially answer the gradient dissipation problem. This makes LSTM models good candidates for performing pattern recognition over time series[22].

LSTM is mainly proposed to solve the problem of gradient disappearance and gradient explosion during long sequence training. LSTM performs better than ordinary RNN in long sequence, LSTM also has some defects. Compared with ordinary RNN, LSTM requires more parameters, which increases the difficulty of training and may lead to the problem of over fitting.

There are three gates (input gate, forgetting gate and output gate) and memory cell in the cycle unit of LSTM. The input gate determines whether the input can be passed into the loop unit. If the output of the input gate is close to zero, it will block the value here and cannot enter the next level. The forgetting gate determines when to forget or retain the value stored in the memory unit. When the output of the forgetting gate is close to zero, the value originally remembered in the memory unit will be forgotten. The output gate can determine whether the input in the memory unit can be output. When the input gates, forgetting gate and output gate open and close depends on the learning process of LSTM. As shown in Figure below:

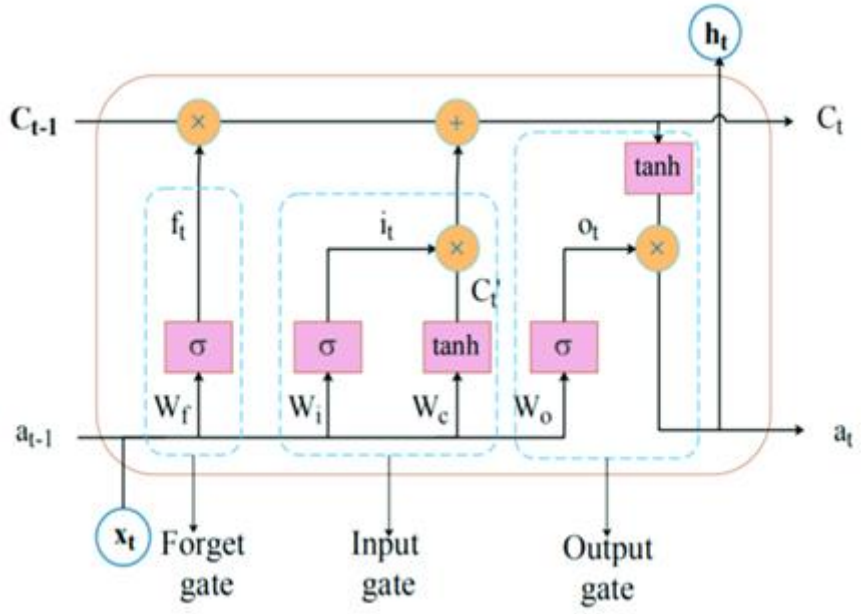

Fig. 1.10: Simplified cycle unit structure of LSTM.

5.4.3.2 Gated Recurrent Unit (GRU)

In order to solve the problem of vanishing graduation, the GRU model, a type of RNN, has been developed, we will learn about its most important characteristics and how it works:

Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) that was introduced by Cho et al. in 2014 as a simpler alternative to Long Short-Term Memory (LSTM) networks. Like LSTM, GRU can process sequential data such as text, speech, and time-series data.

The basic idea behind GRU is to use gating mechanisms to selectively update the hidden state of the network at each time step. The gating mechanisms are used to control the flow of information in and out of the network. The GRU has two gating mechanisms, called the reset gate and the update gate.

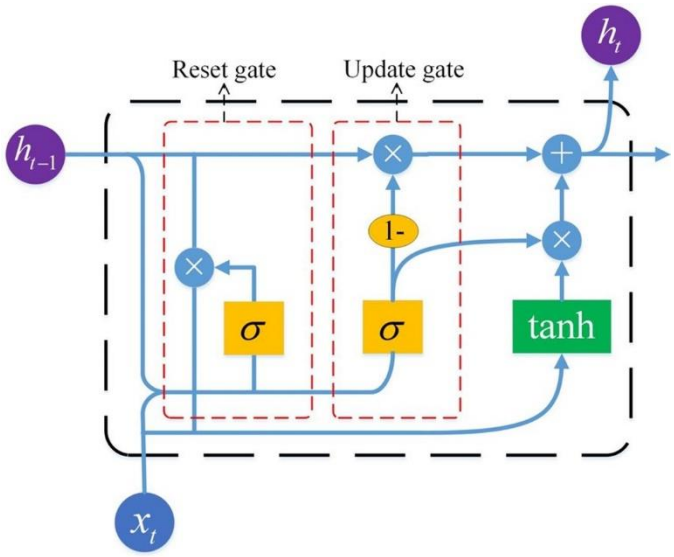

Fig. 1.11: Gated Recurrent Unit.

The reset gate determines how much of the previous hidden state should be forgotten, while the update gate determines how much of the new input should be used to update the hidden state. The output of the GRU is calculated based on the updated hidden state [20].

5.5 Importance of Deep Learning

The significance of Deep Learning is evident in its ability to tackle numerous challenges that are either too complex or impractical for traditional algorithms or human expertise. This technology excels in managing vast and intricate datasets, including those consisting of images, videos, audio, and text. Moreover, it is capable of learning from unstructured or unlabelled data, eliminating the need for human oversight or intervention to derive meaningful insights. Therefore, it plays an essential role in various aspects of everyday life. Here are several key reasons highlighting the importance of Deep Learning:

5.6.1 Maximum utilization of unstructured data

Research from Gartner revealed that a huge percentage of an organization's data is unstructured because the majority of it exists in different types of formats like pictures, texts etc. For the majority of machine learning algorithms, it's difficult to analyse unstructured data, which means it's remaining unutilized and this is exactly where deep learning becomes useful. You can use different data formats to train deep learning algorithms and still obtain insights which are relevant to the purpose of the training. For instance, you can use deep learning

algorithms to uncover any existing relations between industry analysis, social media chatter, and more to predict upcoming stock prices of a given organization [21].

5.6.2 Volatile data processing

Volatile data processing involves handling datasets that exhibit significant variations or fluctuations over time or across different instances. This process includes cleaning the data, extracting relevant features, selecting appropriate models, training and evaluating the models, dynamically updating them with new data, and monitoring their performance over time. By effectively managing volatile data, such as organizations can derive valuable insights to support decision-making and drive business outcomes or weather prediction .

5.6.3 Adaptability and Scalability

Deep learning models are highly adaptable. The models can be fine-tuned or adapted to new tasks with a limited amount of labelled data by employing and leveraging information acquired from previous tasks. This top advantage of deep learning comes in handy in application or use-case requirements where there is a dearth of labelled data [23].

5.7 The challenges and Limitation of deep learning

While deep learning is a relatively new technology and has many promising applications, there are also challenges and limitations that must be considered.

5.7.1 Ethics and fairness

The challenge of ethics and fairness in deep learning underscores the critical need to address biases, discrimination, and social implications embedded within these models. Deep learning systems learn patterns from vast and potentially biased datasets, which can perpetuate and amplify societal prejudices, leading to unfair or unjust outcomes. The ethical dilemma lies in the potential for these models to unintentionally marginalize certain groups or reinforce systemic disparities. As deep learning is increasingly integrated into decision-making processes across domains such as hiring, lending, and criminal justice, ensuring fairness and transparency becomes paramount. Striving for ethical deep learning involves not only detecting and mitigating biases but also establishing guidelines and standards that prioritize equitable treatment, encompassing a multidisciplinary effort to foster responsible AI innovation for the betterment of society [24].

5.7.2 Large quantities of high-quality data

Deep learning algorithms give better results when you train them on large amounts of high-quality data. Outliers or mistakes in your input dataset can significantly affect the deep learning process. For instance, in our animal image example, the deep learning model might classify an airplane as a turtle if non-animal images were accidentally introduced in the dataset.

To avoid such inaccuracies, you must clean and process large amounts of data before you can train deep learning models. The input data preprocessing requires large amounts of data storage capacity [3].

5.7.3 The hardware requirements

Deep learning algorithms are compute-intensive and require infrastructure with sufficient compute capacity to properly function. Otherwise, they take a long time to process results. The hardware demands of deep learning models also impose constraints. Multicore high-performance graphics processing units (GPUs) and similar processing units are necessary to enhance efficiency and reduce time consumption. These devices, however, are pricey and energy-intensive. RAM and a hard drive or solid-state drive with RAM as a foundation are additional hardware requirements.

5.7.4 Overfitting and Generalization Issues

Overfitting and generalization are fundamental challenges in deep learning, impacting model performance, reliability, and applicability to real-world scenarios.

- **Overfitting:**

occurs when a model learns the training data's noise and idiosyncrasies, rather than capturing underlying patterns and relationships. This phenomenon leads to inflated performance metrics on training data but poor generalization to unseen or new data, compromising the model's predictive accuracy and reliability. To mitigate overfitting, practitioners employ techniques such as regularization, dropout, early stopping, and data augmentation. These strategies constrain the model's capacity, introduce noise during training, and diversify the training samples, enhancing generalization and robustness across diverse datasets and environments

- **Generalization Issues:**

Generalization encompasses the model's ability to perform effectively on unseen data, encompassing diverse scenarios, variations, and conditions. Challenges arise when models trained on specific datasets, domains, or conditions fail to generalize to new contexts, exhibiting biases, inaccuracies, or unexpected behaviors [25].

6 Conclusion

In this chapter we review the main concepts of Deep learning and its applications. In particular, we discussed the fundamental components of neural networks, including neurons, weights, biases, and activation functions, and how these elements work together to process inputs through the network. At the end we presented some common neural network models especially in the field of natural language processing and many other areas.

Chapter2: Arabic language and Corpus of Quranic text (Quran)

1. Introduction

Arabic language holds a central position in Islamic culture, serving as the medium through which the Quran, the holy scripture of Islam, was revealed to Prophet Muhammad (peace be upon him) over fourteen centuries ago. Understanding the nuances of the Arabic language is paramount to comprehending the depth and richness of the Quranic text. In this chapter, we embark on an exploration of the Arabic language as it relates to the Quran, delving into its linguistic characteristics, historical significance, and unique features that contribute to the profoundness of Quranic discourse.

2. Arabic language

2.1. Particularity of the Arabic language

Arabic is one of the languages, often described as formally complex. Consisting of 28 letters (25 consonant letters and 3 long animatronic letters), the short animatronic letters are not represented by letters but by letters of formation, placed on or under the consonant letters. Monograms, in the sense that there are no small letters and large letters. Arabic texts are generally immobile, a great source of lexical ambiguity. Arabic is written from right to left with the specificity that the letters follow different forms depending on whether they are at the beginning, middle or end of the word. Table 1 shows the text of a few letters in the three spelling cases. However, it should be noted that some letters do not attach the letters that will succeed him as {ا، د، ذ، ر، ز، و}. [26]

The forms of the letters			
End	**Middle**	**Beginning**	**Isolated**
ـج	ـجـ	جـ	ج
ـص	ـصـ	صـ	ص
ـق	ـقـ	قـ	ق
ـي	ـيـ	يـ	ي

Table 2.1: State of transcription of Arabic letters

An Arabic word is written with consonants, vowels. Vowels are added above or below the letters. They are necessary for the correct reading and understanding of a

text, they allow to differentiate words with the same representation. Table 2.2 gives an example for words كتب and شعر. [29]

Lexical	**1st interpretation**		**2nd interpretation**		**3rd interpretation**	
كتب	**كَتَبَ**	**he wrote**	**كُتِبَ**	**it was written**	**كُتُبٌ**	**books**
شعر	**شَعَرَ**	**he felt**	**شِعْرٌ**	**poem**	**شَعْرٌ**	**hair**

Table 2.2: Ambiguity caused by absence of vowels for words كتب and شعر [26]

2.1.1. Structure of an Arabic word

In Arabic a word can designate an entire sentence thanks to its compound structure which is an agglutination of elements of grammar, the following representation schematizes a possible structure of a word. Note that reading and writing a word is from right to left.

Post fixed	Suffix	Schematic body	prefix	Antefix

Table 2.3: Structure of an Arabic word [30]

- Antefixes are prepositions or conjunctions at the beginning of words (question, future...).
- prefixes usually represented by a single letter, indicate the conjugating person of the verbs in present tense.
- Suffixes are the conjugation endings of verbs and masks which/plural/females for nouns including adverbals.
- post fixed are personal pronouns. [27]

Example: The word (أتتذكروننا)

This word expresses the sentence in English: "Do you remember us?" The segmentation of this word gives the following components:

Fixed post	Suffix	Schematic body	Prefix	Antefix
نا	ونـ	تذكر	تـ	أ
pronoun suffix complement of the name	verbal suffix expressing the plural	derived from the root	verbal prefix of the time of the unfulfilled	interrogation conjunction

Table 2.4: The segmentation of the word (أتتذكروننا) [27]

2.1.2. Word categories (The grammatical categories)

There are three categories for an Arabic word: (noun, verb and particle).

- **The verb:**

We can classify Arabic verbs according to several criteria: According to the number and nature of the consonants of their roots, and according to their models.

By classifying verbs according to the number of consonants of the root, we will have either triliterous verbs that have three consonants, or quadrilateral verbs, few in number, that have four consonants. Depending on the model and number of consonants that make up the verbal structure, we have either verb nudes (مجرد) that are composed only by the consonants of their roots and short vowels, or verbs augmented or derived (مزيد) which are derived from three consonants of the root by modification of the vowels, by repetition of the second letter of the root, by addition and even by intercalation of affixes.

Verb conjugation depends on several factors:

- Time (completed, uncomplicated).
- The number of the subject (singular, duel, plural).
- The gender of the subject (male, female).
- The person (first, second and third)
- The mode (active, passive). [28]

- **The particle:**

The particles are invariable and limited lemmas. They indicate the articulation of the sentence. They are classified according to their semantic field and their function in the sentence; there are several types:

- **Preposition:** example (عن، ب، ك، ل، حتى).
- **Coordination particles:** example (أو، ثم، ف، و).
- **Interrogative particles:** example (أ، هل، ما).
- **Affirmation particles:** example (بلى، أجل، نعم).
- **Particles of negation:** example (لم، لن، لا).
- **Distinguishing particles:** example (أي).
- **Relative particles:** example (ما).
- **Future particles:** example (ف، سوف، لن).
- **Conditional particles:** example (لو، إن).

- **The name:**

Arabic nouns include nouns, adjectives and pronouns, as well as other invariable nouns. Nouns and adjectives are created by taking for origin sometimes a verbal type, sometimes a nominal type. We can distinguish in (table 2.5) two classes of names: the first group conjugable or semi-conjugable names which can have the form duelle, plural, etc. the second-class group non- conjugable names conjugals that keep the same form regardless of the context. Conjugable names are either primitive names, which escape any derivation such as «غَزالَ» (gazelle), or derivative names that are formed from a root such as "مكتبة" (library) of the root "كتب". [28]

Category	Derivation	Conjugation	Sub-category	Example
Name	**Irregular nel derivation**	**Not Conjugable**	Adverb	أين، حيث، قبل
			Voice name	كخ، نخ
			Verb name	هيهات، آه، أف
			Personal pronoun (affixed or isolated)	هو، أنا، ت، تن
			Interrogative pronoun	كيف، متى، ما
			Conditional pronoun	من، إذا
			Allusive pronoun	كم، أي
		Conjugable	Relative pronoun	الذي، التي
			Number name	ثلاثة، واحد، خمسة
			Demonstrative pronoun	هذا، هذه
			Proper noun	محمد، هند، صحراء
			Common name	قلم، أرنب، رجل
	Regular nel derivation	**Conjugable**	masdar	كتابة، القتل
			Active participle	قاتل، شارب
			Passive participle	مكتوب، مضروب
			Name of a time	جلسة، ضربة
			manner name	نظرة، جلسة
			name of time	مغرب
			name of place	مكتب، مقبرة
			instrument name	مطرقة، مسمار
			Adjective	حسن، جميل، بطل
			Elatif	أحسن، أفضل
			Diminutive name	كتيب، شويعر
			Relationship Name	تونسي، مصري
			intensive	قتال، غواص

Table 2.5: Classification of sub-categories of names [28]

2.1.3. Morphology of Arabic language

Arabic lexicon includes three categories of words: verbs, nouns and particles. Verbs and nouns are most often derived from a root with three radical consonants. A family of words can thus be generated from the same semantic concept from a single root using different schemes. This phenomenon is characteristic of Arabic morphology. It is therefore said that Arabic is a language with real roots from which the Arabic lexicon is deduced according to schemes that are additions and manipulations of the root. The essential elements of the morphology of the Arabic language are:

- **The scheme:**

The schema is a word composed of three consonants ف (f), ع (a) and ل (l), which are vocalized and can be increased by other letters (prefix, suffix and infix). The schema plays a very important role in the process of generating derived forms from a root. This generation process consists in replacing the root of the schema by the consonants of the root in question, while keeping the same vowels and the same letters raised while respecting the same order of the consonants, in other words, the scheme can be considered as a mold on which the root flows. [29]

- **The root**

The roots are at the origin of most Arabic words. They are verbs formed from three to five consonant letters. They are around 10,000 roots of which the vast majority (85%) is trilateral. The remains are quadrilateral roots. A root defines the fundamental meaning of derived words by using different diacritics and affixes with the letters of the root to create the inflection of meaning.

Example, the root (كتب) (he wrote) to the basic meaning “write”. Several words are derived from this root, combining it in several forms (present, imperfect, simple past, simple future, etc.).

There are also additional forms such as verbal names: [33]

	the root (كتب) (to write)				
Verbs	كتب	He wrote		يكتب	He writes
	كتبا	They wrote		يكتبون	They write
	كتبت	She wrote		تكتب	You write
	تكتبون	You write		نكتب	We write
Names	كاتب	Writer		كتابة	Writing
	كتاب	Book		مكتوب	Written
	مكتب	Desk		اكتتاب	Registration

Table 2.6: Some derivations of the root (كتب) [30]

- **The stems**

A Stem is the derivation obtained from a given root according to a model. The Stem corresponds to a schema if and only if it has the same number of letters and the same letters in the same positions. An exception is given to consonants ف (f), ع (å), ل (l) which are the letters of the base root (fål, to do). For example, there is: مكاتب (mkatb, offices), it is obtained from the root كتب (ktb, he wrote) according to the scheme مفاعل (mfaål). Not all stems produced are usable. [30]

Root	Model	Stem	Usable
(أكل, akl , he ate)	(مفعول, mfåwul)	(مفعول, mfåwul , edible)	Yes
(لعب, låb , he played)	(أفعلاء, afåla')	(ألعباء, alåba')	No

Table 2.7: An example of stem generation

- **Diacritics**

Diacritical signs are signs added above or above Arabic letters to signify the pronunciation of the word, this phonological role also influences the meaning of this word.

Three of these symbols are transcribed as follows:

- The fetha [a] is symbolized by a small line on the consonant (ba).
- The damma [u] is symbolized by a hook above the consonant (bu).
- The kasra [i] is symbolized by a small line on the consonant (bi).
- A small circle symbolizing the soukoun and affixed to a consonant when it is not linked to any vowel (baada). [31]

- **Affixes**

Affixes are letters that are added to the beginning (prefixes) or end of Arabic words (suffixes). In general, they are used to match words with syntactic elements. They mark the verbal aspect, the mode, the transitive properties, etc. They are around 150.

Prefixes depend on the words they attach to. Indeed, most Arabic words begin with the prefix (ALTARIF), al altaåryif, the definition article" which is used as a declarative term. For this, there are three types of prefixes. First, the nominal prefixes that are reserved for nouns and adjectives. Secondly, verbal prefixes that are reserved for verbs. And third, general prefixes that are used regardless of type of words. [32]

- **Derived words**

According to traditional grammar, the Arabic lexicon includes three categories of words verbs, nouns, and particles. Apart from the proper words, the words of the first two categories are derived from a root. They are called regular words or derivative words. Example: the word (أتطلبون) [33]

Derived word	Prefix	Radical	Suffix	root
أتطلبون	أ	تطلب	ون	طلب

Table 2.8: The derived word (أتطلبون) [36]

2.2. Some Problems of the Arabic language

Given its peculiarities, the Arabic language, faces a number of problems in without treatment, among the latter are cited vowels, agglutination and root extraction.

2.2.1. Vowels

The absence of vowels is very often a great source of ambiguity for morphological, syntactic, semantic and even pragmatic analysis. The majority of written Arabic texts, with the exception of holy texts and some educational works, are without vowels.

This ambiguity lies in the fact that 74% of the words that make up the Arabic vocabulary, accept more than one lexical vowel, and 89.9% of the names that constitute it accept more than one casual vowel. The proportion of ambiguous words is 90.5% if the counts concern their global vowels (lexical and casual, the casual ones in case of the name being: الأشكال). Example: the word "شهد". [26]

شُهْدٌ	Honey (beeswax)
شَهِدَ	Inform, affirm, was present, saw
شَهُدَ	Made a statement
شُهَدُ	
شهد	
شَهْد	Proper name female, plant

Table 2.9: The different voyellations of the word "شهد" [26]

2.2.2. Agglutination of words

Much of the Arabic words are generated by agglutinating proclitics and enclitics to a radical. To determine a name, for example, we add «ال = al», as in the word «الشمس = The sun». Personal pronouns can be related to the nouns «آياته = its signs», as to the verbs «أنزله = it has revealed». The particles with the names «كالمجرمين = on the same footing as the criminals», the coordination conjunctions with the verbs «فتولى = and he withdrew». The problem, in the context of the automatic processing of Arabic, is to be able to decompose the word into its different parts. [34]

2.2.3. The extraction of the root

In order to obtain the root of a word, one must first know the schema by which it was derived, delete the flexional elements (antefixes, prefixes, suffixes, post fixed) that are attached to it. Usually prefix and suffix tables are used. The agglutinative nature of Arabic makes this task quite difficult. This difficulty is even greater when it comes to non-vowel texts. Morphological analysis will therefore have to cut out the word and identify prefixes such as the conjunctions «و = and» and «ك = then», prepositions such as «ب = with» and «لـ = for», the defined article «ال = the» and suffixes of possessive pronouns «ـه = to him, ـها = to her, ـهم = to them, ـهن = to them» etc. The morphological analysis phase determines a possible pattern. Prefixes and suffixes are found by gradually removing prefixes and suffixes and trying to match all the roots produced by a schema in order to find the root. [27]

2.3. Some NLP Challenges in Arabic Language

we will focus on some of the major challenges that Natural Language Processing (NLP) faces when dealing with the Arabic language, which is spoken by more than 400 million people around the world. Arabic is a Semitic language that has a rich and complex morphology, syntax, and orthography. It also has a high degree of dialectal variation and diglossia. These features pose difficulties for NLP tasks such as

tokenization, segmentation, stemming, lemmatization, part-of-speech tagging, parsing, and more:

2.3.1. Complex Morphology:

Arabic is a highly inflectional language, which means that words can take many forms to express different grammatical functions. This complexity makes tasks such as tokenization, part-of-speech tagging, and lemmatization particularly challenging.

2.3.2. Diacritics:

Arabic script includes diacritics that can change the meaning of words, but these are often omitted in written text. This omission leads to ambiguity and difficulty in accurately understanding and processing the text.

2.3.3. Ambiguity:

Arabic exhibits a high degree of lexical and syntactic ambiguity. Words can have multiple meanings depending on the context, and the flexible word order adds to the complexity.

2.3.4. Dialectal Variation:

Arabic has many dialects that vary significantly from the Modern Standard Arabic (MSA) used in formal texts. These dialects can differ in vocabulary, grammar, and pronunciation, complicating NLP tasks.

2.3.5. Named Entity Recognition (NER):

Identifying proper nouns and entities in Arabic is challenging due to the lack of capitalization and the use of compound names. Additionally, entities often need to be disambiguated from common words.

These challenges highlight the complexities involved in developing effective NLP tools and applications for the Arabic language.

3. Corpus of Quranic text (Quran)

3.1. Quran

The Quran is for Muslims the verbatim Word of God, revealed during the twenty-three-year period of the prophetic mission of the Prophet Muhammad through the agency of the Archangel Gabriel. the meaning, the language, and every word and letter in the Quran, its sound when recited, and its text written upon various physical surfaces

are all considered sacred. the Quran was an oral revelation in Arabic first heard by the Prophet and later written down in the Arabic alphabet in a book consisting of 114 surahs (chapters) and over 6,200 verses (ayat), arranged according to an order that was also revealed. Considered the Book (al-Kitab) by all Muslims, it has many names, such as al-Furqan (“the Criterion”) and al-Huda (“the Guide”), but its most commonly used name is al-Quran, which means “the Recitation.” Known in English as the Quran (also Koran), it is the central theophany of Islam and the basic source and root of all that is authentically Islamic, from metaphysics, angelology, and cosmology to law and ethics, from the various arts and sciences to social structures, economics, and even political thought. the Quran is the constant companion of Muslims in the journey of life, as for the Quran as a book, it is found in nearly every Muslim home and is carried in various forms and sizes by men and women. [35]

3.1.1. Sound chain of transmission

This is the main condition for accepting a given reading, and has never been disputed as such. A sound chain of transmission in this context means that a given reading is referred back to the Prophet in the sense that he personally read it, or approved of it on hearing it from another reader. Readings are classified from the point of view of Sanad into four categories:

- Qira'a: This is a reading that is attributed to one of the seven readers chosen by Ibn Mujahid, (e.g. the reading of Nafi).
- Riwaya: This is a reading that is attributed to one of those who transmitted from one of the seven readers, (e.g. the Riwaya of Qalun from Nafi).
- Tariq: This is when a reading that is attributed to a transmitter of the generation following, (e.g. the Tariq of Abu Nashit from Qalun).
- Wajh: This is a preferred reading by a Qari when he has more than one of his disposal, all of which fulfill the necessary conditions. [36]

3.1.2. Quran readings (Qira'at)

There are 7 Mutawatir Qira'at and 3 Mashhur ones.

The Mutawatir are:

- Abd Allah b. Kathir al-Dari: His reading was popular in Mecca and was transmitted mainly by Qunbul and al-Bazzi.

- Nafi b. Abi Nuaym: His reading was very popular in Medina, and was transmitted mainly by Qalun and Warsh.
- Abd Allah b. Amir al-Yahsubi: His reading predominated in Syria, and was transmitted mainly by Hisham and Ibn Dhakwan.
- Abu Amr b. al-Ala: His reading was popular in Basra, and was transmitted mainly by al-Duri and al-Susi.
- Hamza b Habib al-Zayyat: His reading was transmitted mainly by Khalaf b. Hisham and Khallad b. Khalid.
- Asim b. Abi al-Najud: His reading was transmitted mainly by Abu Bakr Ibn Ayyash and Hafs b. Sulayman.
- Abu al-Hasan Ali b. Hamza al-Kisa'i: His reading was transmitted by Abu al-Harith al-Marruzi and Abu Umar Al-Duri, who was also the transmitter of Abu Amr b. al-Ala. The readings of these three readers attained prominence in Kufa.

The Mashhur are:

Besides these seven readings, there were a number of other readings of which the authenticity is not disputed. Ibn al-Jazari acknowledged ten readings known as Qira'at al-ashara, which in addition to the above-mentioned seven readings, comprise those of the three following readers:

- Yaqub b. Ishaq al-Hadrami: His reading was popular in Basra.
- Khalaf b. Hisham: who was also the transmitter of the reading of Hamza. His reading was popular in Kufa.
- Abu Jafar Yazid b. al-Qaqa: His reading was well known in Medina. [36]

3.2. Corpora

3.2.1. Definition

A corpus is a set of documents, artistic or not (texts, images, videos etc.), grouped in a precise perspective, they are collected in electronic format, corpora can be used in several fields: literary, linguistic, scientific studies, etc. [29]

3.2.2. The corpus in literature

In literature the corpus includes a set of texts with a common aim. A corpus may consist of different documents (table, excerpt of text...) and these various documents have one thing in common. In general, this is the theme that reflects their resemblance. It takes a particular technique to decipher it. [29]

3.2.3. The corpus in linguistics

The branch of linguistics, which is concerned with corpora, is called corpus linguistics. It is linked to the development of computer systems, in particular to the constitution of textual databases. We speak of corpora to designate the normative aspect of language: its structure and its code in particular. In order to make corpora more useful for linguistic research, they are often subjected to a process known as annotation. [29]

3.2.4. The corpus in science

Corpora are indispensable and valuable tools in natural language processing. They make it possible to extract a set of useful information for statistical processing. From an informative point of view, they make it possible to extract trends and in particular to build sets of n-grams. From a methodological point of view, they provide the necessary objectivity for scientific validation in natural language processing. The information is no longer empirical, it is verified by the corpus. It is therefore possible to rely on well-trained corpora to formulate and verify scientific hypotheses. [29]

3.2.5. The Applications of Corpora

- Lexicography (help to build dictionaries).
- Language learning.
- Sociolinguistic studies.
- Linguistics: (the study of vocabulary, grammar, evolution of language or meanings of words).
- Computer linguistics (NLP), train or test textual analysis tools.
- Terminology, translation, technical writing.
- Analyze the characteristics of translated texts.
- translation aid. [29]

4. Literature review

The importance of the Quran among Muslims allows it to be one of the most important areas of software development of all kinds that exist around it. Many efforts have been devoted to the service of Quran to facilitate access to God's words. Despite these efforts, we note that there is a lack of classification systems of Quran verses.

4.1. The work of Masnizah Mohd, Faizan Qamar, Idris Al-Sheikh and Ramzi Salah (2021)

The document introduces a Quranic optical character recognition (OCR) system using deep learning models. Six deep learning models are developed to study the effect of different input and output representations, as well as the accuracy and performance of the models. The study compares long short-term memory (LSTM) and gated recurrent unit (GRU). A new Quranic OCR dataset is created based on the most famous printed version of the Holy Quran (Mushaf Al-Madinah), and the experiments achieve better performance in word recognition rate (WRR) and character recognition rate (CRR). The study also compares LSTM and GRU in the Arabic text recognition domain, showing that the proposed system achieves an accuracy of 98% on the validation data, with a WRR of 95% and a CRR of 99% in the test dataset. [46]

4.2. The work of Zineb Touati-Hamad, Mohamed Ridda Laouar, Issam Bendib and Saqib Hakak (2022)

The document presents a study on the authentication of Arabic Quranic verses using deep learning and word embeddings. The authors propose a new approach based on deep learning and word embeddings to automatically classify Quranic and Arabic texts. They evaluate the proposed approach using different word embeddings models and two popular classifiers, achieving an accuracy of 98.33%. The study highlights the potential of deep learning techniques in distinguishing Quranic verses from regular Arabic text and discusses the challenges and motivations for this research. The results show that the combination of convolutional neural network (CNN) and long short-term memory (LSTM) outperforms traditional methods in classifying Quranic verses. The study concludes by discussing future work and the implications of using deep learning models to identify Quranic verses in Arabic textual content. [38]

4.3. The work of Mohamed Galal, Magda M. Madbouly and Adel El-Zoghby (2019)

A new algorithm called GStem was introduced to group similar Arabic words based on extra Arabic letters. Deep neural networks, like CNN, were successfully used for Arabic text classification. Performance of Arabic text classification was improved by applying linguistic processing techniques such as normalization and stemming. Word2Vec was used to convert texts into two-dimensional arrays for representation in training models. Results showed that SVM outperformed NB and KNN in Arabic text classification. [39]

4.4. The work of Mohamed G. Mahdi, Ahmed Sleem and Ibrahim Elhenawy (2024)

The documents include a dataset with a distinctive set of 113,284 Arabic words and the MMAC dataset containing 282,593 unique words and 66,725 images of Arabic words. A study on Arabic character recognition techniques using artificial intelligence was presented. Previous research was reviewed, analyzing key trends and challenges in Arabic character recognition. Test accuracy of up to 92.88% was achieved using CNN models, transfer learning techniques, and genetic algorithms. [40]

4.5. The work of Suhaib Kh. Hamed and Mohd Juzaiddin Ab Aziz (2018)

The article discusses the optimization of the string to word vector filter for classifying verses by employing feature reduction methods and text preprocessing techniques. It highlights the use of the Term Frequency (TF) technique to rank and select high scoring features, emphasizing the importance of terms in verses classification. The application of tokenization, lowercasing, and stemming enhances classification accuracy by improving content understanding. Additionally, handling stop words by removing irrelevant features aims to increase the effectiveness of the neural network classifier, ultimately leading to a more efficient learning process for classifying verses. [41]

4.6. The work of Suhaib Kh. Hamed and Mohd Juzaiddin Ab Aziz (2016)

The study centers on creating a Question Answering System (QAS) for the Holy Quran, utilizing WordNet, Islamic terminology, and Neural Network categorization. The process involves utilizing N-gram method, document categorization, and Neural Network classifier to improve the precision in retrieving verses associated with Fasting and Pilgrimage themes. The outcomes indicate enhanced effectiveness in retrieving information from the Quran, with high precision and recall values resulting in an F-score of around 87%. [42]

4.7. The work of Abdullahi O. Adeleke, Noor Azah Samsudin, Aida Mustapha and Nazri Mohd Nawi (2018)

The article presents a Group-Based Feature Selection Approach for improving the classification of Holy Quran verses. The study aims to automatically label Quranic verses into categories related to fundamental aspects of Islam such as 'Iman' (profession of faith), 'Ibadah' (worship), and 'Akhlak' (etiquettes). The dataset consists of 451 verses from chapter two and six of the Holy Quran, normalized using the TF-IDF method. Five feature selection algorithms were experimented with, and four classification

algorithms (k-NN, LibSVM, NB, J48) were implemented for the labeling task. The classification experiments used 10-fold cross-validation and evaluated accuracy and AUC metrics. The proposed approach showed promising results in classifying Quranic verses, providing a valuable method for automated labeling and analysis in Quranic studies. [43]

4.8. The work of Aqsa Noor and Ahmad Ali (2021)

The text investigates using deep learning methods like BERT word embedding, LSTM, and GRU for classifying imbalanced multiclass Quranic verses. The dataset contains translations of Quranic verses in English classified into six groups, showing a noticeable imbalance in class distribution. The research tackles this disparity by utilizing methods such as tf-idf vectorization, n-gram strategies, and Synthetic Minority Over-sampling Technique (SMOTE). The model benefits from BERT word embedding to gain a contextual understanding of Quranic verses by taking into account the meaning and context of the words. The paper emphasizes the significance of maintaining a balanced dataset and comprehending the subtleties of the text in order to attain precise classification outcomes. The efficiency of deep learning classifiers with LSTM, GRU, and fine-tuned BERT models has been proven through experimental results in achieving high accuracy and F1-scores for categorizing Quranic verses according to their themes and topics.

In general, the research demonstrates how deep learning techniques, specifically BERT word embedding, can effectively address skewed data and uncover valuable information from religious texts such as the Quran. [45]

However, the scope of these works is limited and has focused on recognition, authentication and preserving the integrity of the Digital Quran content and different aspects of the Arabic language such as abuse detection, mood detection, and opinion mining aspects.

5. Conclusion

In conclusion, this chapter serves as a foundational exploration of the Arabic language and its inseparable relationship with the Quranic text. By delving into the linguistic characteristics and unique features of Arabic, we lay the groundwork for a deeper understanding and appreciation of the Quran's divine message. From this study we found the need for a system of classification of Quranic verses, which will facilitate understanding and meditation on God's words and meanings.

Chapter3: Text Classification

1. Introduction

Text classification is a generic task that involves assigning one or more categories, from a predefined list, or not to a document. Currently, the classification of texts is a very active field of research and the automation of this operation has become a challenge for the scientific community, the work has evolved considerably over the last twenty years and several models have emerged such as filtering (supervised classification bi-class), routing (supervised classification multi-class) or ordered classification (ranking of texts in order of relevance for each category).

2. Classification Definition

"The only way to make an informative and natural method is to put together the things that are similar and separate those that are different from each other."

The classification process seeks to highlight the implicit dependencies that exist between objects, classes between them, classes, and instances. Classification covers the processes of recognizing the class of an object, and the possible insertion of a class in a hierarchy. This mode of reasoning makes it possible to recognize an object by identifying its characteristics, relative to the hierarchy studied. Classification involves a membership decision-making process. [46]

2.1. Why automate classification?

An automatic document classification method is not only possible, but a viable solution for many problems. It frees people from dealing with unorganized piles of paper. Using a scanner equipped with the necessary features, automatic document classification allows the user to quickly sort separate pages and letters and faxes in case you need them. More and more organized, without the use of many people will save you money, time and space [47].

2.2. Bi-class and multi-class classification

2.2.1. Binary Classification

Binary classification corresponds to the filtering. This is a problem for which the classification system answers the question: "Does the text belong to category C or not (i.e. or its complementary category --C?" (For example, is a document allowed to children or not). However, when it comes to performing a multi-class classification that

allows the document to be transmitted to the most appropriate category(s), we are talking about routing. This multi-class classification, as the case may be, may or may not be disjointed.

2.2.2. Multi-Class Classification

The multi-class classification is the context of classification into a number of classes greater than one and for which a text is assigned to one and only one class. A disjoint multi-class classification system answers the question "To which class (singular) does the document belong?".

2.2.3. Multi-label classification

In a multi-label classification system, one can associate a text to one or more classes or to no class. The system answers the question, "To which classes (plural) does the document belong?", This is the most general case of classification. [48]

3. The types of automatic classification

In the field of automatic classification, there are two types of approach: supervised classification and unsupervised classification. These two methods differ in how classes are generated. In the case of unsupervised classification, document groups (classes) are automatically calculated by the machine, while they are, in the supervised approach defined by an expert.

However, there are other types of classification that rely on other types of learning methods such as "semi-supervised learning" and "reinforcement learning." Indeed, semi-supervised learning is a good compromise between the two types of "supervised" and "unsupervised" learning because it allows a large amount of data to be processed without having to label them all, and it benefits from the two types mentioned. While reinforcement learning is widely used in the case of interactive learning.

3.1. Unsupervised classification

When clustering, objects are grouped into disjointed homogeneous classes. To highlight the document sets, the internal homogeneity of the classes and the dispersion between them must be maximized. The two main methods of clustering are: hierarchical methods and non-hierarchical methods.

3.2. Supervised classification

In this type of classification, classes are predefined with a description of the documents. When a new document arrives, we compare it with the description of each class and put it in the one that most resembles it. Several techniques are used, we can mention K-nearest neighbors (KNN), decision tree, Naive Bayes, Support vector machine [47] .

4. Supervised learning algorithms

In a simple way, the goal of the algorithm is to find out why each sample document has been filed in this or that class, in order to predict the class of new documents to be filed in the future. Below are some commonly used supervised learning algorithms for automatic categorization of text:

4.1. K-Nearest Neighbor (KNN) Algorithm

KNN (KNN for K Nearest Neighbors) has proven its effectiveness against textual data processing. The learning phase is to store the labeled examples. The classification of new texts is done by calculating the distance between the vector representation of the document and that of each example of the corpus. The nearest K elements are selected and the document is assigned to the majority class (the weight of each example in the vote being possibly weighted by its distance). [49]

4.2. Decision tree

A decision tree represents the studied objects in a tree form, according to a hierarchy of attributes determined by an entropy calculation. These methods are popular for the summary presentation of the data they provide, as well as for the clarity of the explanations for the decision rendered. [47]

4.3. Naïve Bayes (or Simple Bayes)

The Naïve Bayes (NB) algorithm, is another well-known method in learning, it is also used in the categorization of documents. It is based on a probabilistic model, which aims to estimate the conditional probability of a category knowing a document and assigns the most likely category(s) to the document. The naive part of this model is the hypothesis of word independence, that is, the conditional probability of a word knowing a category is assumed to be independent of this probability for other words. This hypothesis makes categorization by NB more effective than the exponential complexity of non-naive Bayesian approaches that use word combinations as preachers. [50]

4.4. Neural networks

A neural network (or Artificial Neural Network) is a computational model whose design is very schematically inspired by the functioning of real neurons (human or not). Neural networks are generally optimized by statistical learning methods thanks to their ability to classify and generalize, such as the automatic classification of postal codes or the decision-making regarding a stock market purchase based on price movements. They enrich with a set of paradigms to generate large functional, flexible and partially structured spaces. [51]

4.5. Vector support machines (or SVM)

This technique initiated by Vapnik attempts to linearly separate positive and negative examples in the set of examples. Each example must be represented by a vector of dimension n. The method then looks for the hyperplane that separates the positive examples from the negative examples, ensuring that the margin between the nearest positive and negative is maximum. Intuitively, this ensures a good level of generalization because new examples may not be too similar to those used to find the hyperplane but still be located frankly on one side or the other of the border. The effectiveness of SVM is superior to that of all other methods on text classification. Its effectiveness is also very good for pattern recognition. Another interest is the selection of Carrier Vectors that represent the discriminating vectors through which the hyperplane is determined. The examples used when searching for the hyperplane are no longer useful and only these support vectors are used to classify a new case. This makes it a very fast method. [47]

5. Classification Applications

Classification is in practice applied in most domains of the real world. We find it as an example in:

- The Web, for classifying documents according to their subjects and filtering spam (spam/non spam);
- The medical sector, for the classification of patients according to their diseases;

Bioinformatics, for the classification of genes when a large number of genes can show similar behaviors.

- Marketing, for the classification of companies according to their productions. [47]
- …

6. Algorithm performance metrics

6.1. Recall:

Recall measures the proportion of correct positive predictions out of all actual positive instances. Recall is concerned with the correct prediction of actual positives and is crucial when the consequences of false negatives are significant.

6.2. Precision:

is a metric that measures how often a machine learning model correctly predicts the positive class. You can calculate precision by dividing the number of correct positive predictions (true positives) by the total number of instances the model predicted as positive (both true and false positives). [52]

6.3. F1 score:

is a metric in machine learning that quantifies a model's precision. It merges the model's precision and recall scores,

6.4. Accuracy:

The most straightforward way to measure a classifier’s performance is using the Accuracy metric. Here, is the ratio of the number of correct predictions to the total number of input sample. [53]

6.5. Hamming Loss

It reports how many times on average, the relevance of an example to a class label is incorrectly predicted. Therefore, hamming loss takes into account the prediction error (an incorrect label is predicted) and missing error (a relevant label not predicted), normalized over total number of classes and total number of examples. [44]

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

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

Fig. 3.1: Algorithm performance metrics

6.6. Confusion matrix:

A matrix that presents the evaluation of a machine learning model on a specific dataset. It is a tool for showing the count of correct and incorrect predictions made by the model. Commonly, it is utilized to assess the effectiveness of classification models that strive to forecast a specific category for every given input.

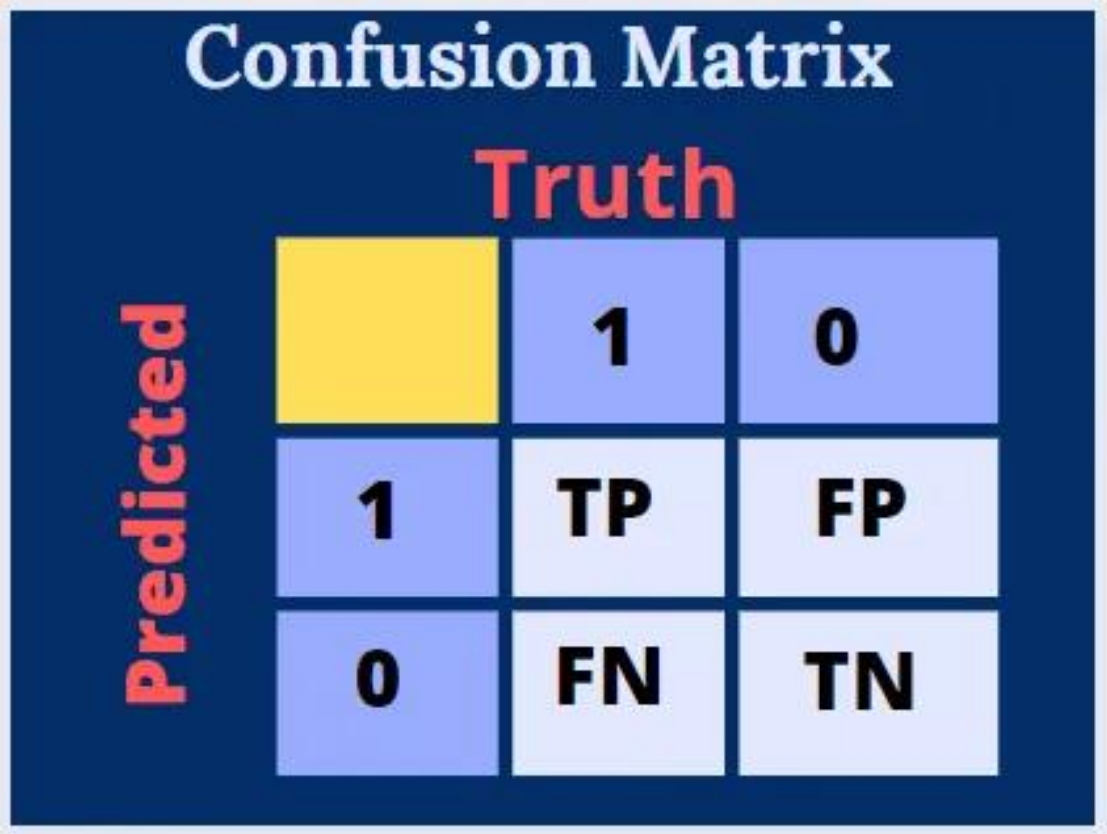

Fig. 3.2: Confusion Matrix [54]

And this is the metrics derived from the Confusion Matrix:

- **True Positive (TP):** An instance for which both predicted and actual values are positive
- **False Positive (FP):** An instance for which predicted value is positive but actual value is negative
- **False Negative (FN):** An example where the actual value is positive but the predicted value is negative
- **True Negative (TN):** A situation in which both predicted and real values are negative

7. Conclusion

In this chapter we had talked about Classification, its definition, its types and its applications..., we concluded that text classification is very useful and important in many domains. So how do we apply the classification of texts to the Holy Quran verses?

Chapter4: Implementation and Results

1. Introduction

The objective of this chapter is to present steps for the implementation of the proposed approach within the framework of infusing classification of verses of Quran using deep learning models. We begin by presenting the resources, language and development environment we have used. Then the steps to achieve the model and end with the tests conducted.

2. Programming Languages and Libraries used

Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs. Colab is especially well suited to machine learning, data science, and education. [56]

Python is an interpreted high-level general-purpose programming language, python's design philosophy emphasizes code readability with its notable use of significant indentation. [57]

Library	Description
NumPy	NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices)
Pandas	Pandas is a library for data manipulation and analysis. It provides data structures like DataFrame and Series, which are highly efficient for working with structured data.
Os	This module provides a way of using operating system-dependent functionality, such as reading or writing files, manipulating paths, etc
JSON	JavaScript Object Notation (JSON) is a standardized format commonly used to transfer data as text that can be sent over a network. It's used by lots of APIs and Databases, and it's easy for both humans and machines to read.
Gensim	Gensim is a popular open-source library in Python that is used for topic modelling and natural language processing tasks. It is designed

	to automatically extract key themes and concepts from large sets of text data
Re	A Regular Expression or RegEx is a special sequence of characters that uses a search pattern to find a string or set of strings, It can detect the presence or absence of a text by matching it with a particular pattern and also can split a pattern into one or more sub-patterns.
Keras	Keras is an open-source high-level Neural Network library, which is written in Python is capable enough to run on Theano, TensorFlow, or CNTK. It was developed by one of the Google engineers, Francois Chollet. It is made user-friendly, extensible, and modular for facilitating faster experimentation with deep neural networks. It not only supports Convolutional Networks and Recurrent Networks individually but also their combination.
Matplotlib	Matplotlib is a powerful plotting library in Python used for creating static, animated, and interactive visualizations. Matplotlib's primary purpose is to provide users with the tools and functionality to represent data graphically, making it easier to analyze and understand
Seaborn	Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.
Sklearn	Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.
TensorFlow	The library TensorFlow is a popular open-source machine learning framework used for building and training neural networks.
Pyarabic	A specific Arabic language library for Python, provides basic functions to manipulate Arabic letters and text, like detecting Arabic letters, Arabic letters groups and characteristics, remove diacritics etc.

Table 4.1: The libraries used in the project

3. The architecture of the Proposed System

Fig. 4.1: The main steps of project

4. Data set:

The dataset is primarily generated based on the information obtained from the "Quran by Subject" website [58], and it is enhanced by extra data gathered from the original Arabic version of the Quran [59].

The Quran by Subject offered a text document containing details on numerous Quran verses organized and labelled according to their relevant subtopics. To be more precise, there are numerous verses stored and written in the text file(quran_topics.text) under their respective subtopics.

4.1. Imports libraries

We have included the essential libraries needed for our project.

```
import os
import numpy as np
import json
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns
import gensim.downloader as api
import tensorflow as tf
from sklearn.preprocessing import LabelEncoder
from gensim.models import KeyedVectors
# mlp for multi-label classification
from numpy import mean
from numpy import std

from sklearn.datasets import make_multilabel_classification
from sklearn.model_selection import RepeatedKFold
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.metrics import  multilabel_confusion_matrix, hamming_loss
from keras.models import Sequential
from keras.layers import Embedding, Conv1D, GlobalMaxPooling1D, Dropout, Dense,GRU
from keras.layers import Bidirectional, LSTM
```

Fig. 4.2: Libraries Imports

4.2. Download the Main Files of Quran

Here the Quran file was downloaded from the drive and linked to the editor coolab:

```
!pip install gdown -U --no-cache-dir
import gdown
url = 'https://drive.google.com/drive/folders/1WFtHaE3gWronMdT_jbbgcFQdVQoAhuMQ'
gdown.download_folder(url)

# All Quran files are located in the main directory.
quran_data_dir = 'Quran Dataset'
```

Fig. 4.3: Main Files of Quran Dataset

Result:

```
  Attempting uninstall: gdown
    Found existing installation: gdown 5.1.0
    Uninstalling gdown-5.1.0:
      Successfully uninstalled gdown-5.1.0
Successfully installed gdown-5.2.0
Retrieving folder contents
Processing file 1Hsu8-AeWr_Y4MNEKgQZHlzTH2yY1Dy8l quran_original.csv
Processing file 1mfxU3P6hMH__C3hb6jczTn2UKpR5l38G quran_topics.txt
Processing file 1z4vS38ZbQ_RVfzOXK2Y5nfrR_noF_pws souras_topics.json
Retrieving folder contents completed
Building directory structure
Building directory structure completed
Downloading...
From: https://drive.google.com/uc?id=1Hsu8-AeWr_Y4MNEKgQZHlzTH2yY1Dy8l
To: /content/Quran Dataset/quran_original.csv
100%|██████████| 1.36M/1.36M [00:00<00:00, 13.7MB/s]
Downloading...
From: https://drive.google.com/uc?id=1mfxU3P6hMH__C3hb6jczTn2UKpR5l38G
To: /content/Quran Dataset/quran_topics.txt
```

Fig. 4.4: Downloading files

5. Preparing and Preprocessing Dataset

The Quran database contains three different file types in the following steps there will be a process of processing the Quranic texts and collecting them in a table containing the verses and their classifications into 12 topics which are: “allah”, “disbelief”, “evils”, “faith”, “mohammad”, “quran”, “rulings”, “stories”, “struggle”, “universe”, “unseen” and “worship”.

5.1. Add the Soura names to Quran Original CSV

add the names of the suras (chapters) to your Quran dataset using a JSON file “souras_topics.json” containing sura information, This loads the original Quran CSV file, reads the Soura names from a JSON file, and adds the Soura names to the Quran original Data Frame.

	snum	sname	anum	aya
0	1	الفاتحة	1	بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ
1	1	الفاتحة	2	الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ
2	1	الفاتحة	3	الرَّحْمَٰنِ الرَّحِيمِ
3	1	الفاتحة	4	مَالِكِ يَوْمِ الدِّينِ
4	1	الفاتحة	5	إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ
...	...	...	...	...
6231	114	الناس	2	مَلِكِ النَّاسِ
6232	114	الناس	3	إِلَٰهِ النَّاسِ
6233	114	الناس	4	مِنْ شَرِّ الْوَسْوَاسِ الْخَنَّاسِ
6234	114	الناس	5	الَّذِي يُوَسْوِسُ فِي صُدُورِ النَّاسِ
6235	114	الناس	6	مِنَ الْجِنَّةِ وَالنَّاسِ

6236 rows × 4 columns

Fig. 4.5: Chapter names

5.2. Cleaning CSV File with subtopics

We used advanced regular expression (RegEx) programming techniques to computationally analyse the text file with following steps:

- The regex pattern rx_topic is utilized for extracting Quran topics and their contents.
- rx_soura is employed for extracting details related to chapters (suras) and verses within each topic.
- strip to remove empty lines.

- Remove undesired text: the str.replace() function is utilized, specifically for text that commences with "القرآن التي تتكلم عن:".
- remove '()' and insides.
- removes any characters from anum column but not the numbers
- Organizing the Data into a Data Frame.
- Saving the Data to a CSV File.

	topic	sname	ayat	aya	anum
0	الدعاء من القران و الاستجابه	الفاتحة	بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيمِ الْحَمْد...	بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيمِ	1
1	الدعاء من القران و الاستجابه	الفاتحة	بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيمِ الْحَمْد...	الْحَمْدُ لِلهِ رَبِّ الْعَالَمِينَ	2
2	الدعاء من القران و الاستجابه	الفاتحة	بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيمِ الْحَمْد...	الرَّحْمَنِ الرَّحِيمِ	3
3	الدعاء من القران و الاستجابه	الفاتحة	بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيمِ الْحَمْد...	مَالِكِ يَوْمِ الدِّينِ	4
4	الدعاء من القران و الاستجابه	الفاتحة	بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيمِ الْحَمْد...	إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ	5
...	...	...	...	...	...
7060	الربا	آل عمران	... يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَأْكُلُوا	...وَلِيُمَحِّصَ اللهُ الَّذِينَ آمَنُوا وَيَم	141
7061	الربا	آل عمران	... يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَأْكُلُوا	...أَمْ حَسِبْتُمْ أَنْ تَدْخُلُوا الْجَنَّةَ وَ	142
7062	الربا	الروم	...أَوَلَمْ يَرَوْا أَنَّ اللهَ يَبْسُطُ الرِّزْ	...أَوَلَمْ يَرَوْا أَنَّ اللهَ يَبْسُطُ الرِّزْ	37
7063	الربا	الروم	...أَوَلَمْ يَرَوْا أَنَّ اللهَ يَبْسُطُ الرِّزْ	...فَآتِ ذَا الْقُرْبَى حَقَّهُ وَالْمِسْكِينَ وَ	38
7064	الربا	الروم	...أَوَلَمْ يَرَوْا أَنَّ اللهَ يَبْسُطُ الرِّزْ	...وَمَا آتَيْتُمْ مِنْ رِبًا لِيَرْبُوَ فِي أَم	39

7065 rows × 5 columns

Fig. 4.6: Quran Subtopics

5.3. Remove diacritics

The code utilizes the strip diacritics function from pyarabic.araby library to eliminate diacritics from the text in the 'aya' column.

```
#Remove diacritics
!pip install pyarabic
import pyarabic.araby as araby

df_quran_original_diac['aya']=df_quran_original_diac['aya'].apply(lambda aya :
araby.strip_diacritics(aya))
sentences = [aya.strip(' ').split(' ') for aya in df_quran_original_diac['aya']]
build_bin('quran_original_clean',sentences,0, 'الرحيم')
build_bin('quran_original_clean',sentences,1, 'الرحيم')
```

Fig. 4.7: Remove diacritics

5.4. Collect Subtopics

This prepares the final dataset by categorizing 167 subtopics into 12 major topics and saving it to a CSV file, 167 subtopics are condensed into 12 main topics. Each row in the CSV file corresponds to a labelled verse, with a total of 6,246 individual rows in the file. One verse may be found multiple times if it covers multiple topics.

5.5. Merging topics for duplicate verses

One verse may be found multiple times if it covers multiple subtopics. Therefore, the CSV file undergoes a new processing where each verse is placed in a single row along with its related topics. Each verse is then categorized under a minimum of one of the 12 main topics, this indicates that the verses cover 12 main subjects and the total count of categorized verses is 4911.

5.6. Data representation

The neural networks and other deep learning models cannot interpret or process text input. Instead, the text needs to be translated into numerical values. The process of converting to this form is referred to as word embedding.

Word2Vec is a popular algorithm in the field of natural language processing (NLP) that aims to capture the semantic meaning of words and phrases in a numerical forma [60], there are two algorithms for Word2Vec which are Continuous Bag of Words (CBOW) and skip-gram.

The neural network requires verses to be in the form of fixed length vectors, so short sentences are filled with padding to match the length of the longest verse in the dataset, two functions have been defined for this representation:

5.7. Splitting the data

We used two techniques to split the database:

- train_test_split is a function from the sklearn library that splits the dataset into two groups, randomly selecting 75% of the data as a training group and 25% as a test group.

```
#Splitting the data
    X_train, X_test, y_train, y_test = train_test_split(X_seq, Y_seq, random_state=123)
    print("Size of test data: {:d}".format(len(X_test))) #Print Test Data Size
    print("Size of train data: {:d}".format(len(X_train))) #Print Train Data Size
```

Fig. 4.8: Train test split

6. Implementation of models and results

6.1. Models Building

These are three different neural network models (CNN, RNN, GRU) defined using Keras Sequential API:

```
models = [
    Sequential([ #CNN
        Embedding(len(E), embed_dim, input_length=max_seqlen, trainable=True,
weights=[E]),
        Conv1D(64, 5, activation='relu'), # num_filters=32, kernel_size=5
        GlobalMaxPooling1D(),
        Dense(num_labels*2, activation='relu'),
        Dropout(0.2),
        Dense(num_labels, activation='sigmoid')
        ]),

    Sequential([ #RNN
        Embedding(len(E), embed_dim, input_length=max_seqlen, trainable=True,
weights=[E]),
        Bidirectional(LSTM(64)),
        Dense(num_labels*2, activation='relu'),
        Dropout(0.2),
        Dense(num_labels, activation='sigmoid')
        ]),
```

Figure 4.9: Models building

6.2. Model summary

Figures displays a condensed overview of the (RNN, CNN, GRU) neural network and the attributes of its layers, receiving input of vectorized Quran verses with diacritized text.

During the creation of word sequences for the 4,911 verses in the Quan dataset with diacritized text, a vocabulary of 15,736 words was formed. Each verse vector has a length of 145, which is the maximum length for a verse. The unaccented text of Quran verses has a vocabulary consisting of 13,345 words, with each word being represented as a vector of 129 integer values.

6.3. Models Training and Evaluation:

After building models and defining word representation functions, we train and test models and then evaluate these models with test data and print evaluation metrics like accuracy, classification report, and hamming loss are computed.

During the training process, we set up the loss function, optimizer, we iterated over the dataset, calculated the loss, and updated the model's weights to improve its performance. The training was conducted for a specified number of 60 epochs.

```
def cv (X_seq, Y_seq, model, batch_size, epochs ):
    #splitting the data
    X_train, X_test, y_train, y_test = train_test_split(X_seq, Y_seq,
random_state=123)
    print("Size of test data: {:d}".format(len(X_test))) #Print Test Data Size
    print("Size of train data: {:d}".format(len(X_train))) #Print Train Data Size

    model.compile(optimizer='adam', loss='binary_crossentropy', metrics=
['accuracy'])
    tensorboard = tf.keras.callbacks.TensorBoard(log_dir=quran_logs_dir)
    model.fit(X_train, y_train, batch_size=batch_size, epochs=epochs,
              # validation_data=(X_test, y_test), # or
              validation_split=0.2, # 20% of the training used for validation
              verbose=True, callbacks=[tensorboard])
```

Figure 4.10: Implement the model with test train split

Since our project used 2 techniques: cross validation and train test split, and after training and testing on all models produced that k-fold cross-validation technique showed better results than other technology.

We start by presenting the results with diacritized text:

- **CNN model**

Now our model is trained and we have got 90.75% accuracy.

- **RNN model:**

Our model is trained and we have got 91.52% accuracy

- **GRU model:**

Our model is trained and we have got 87.61% accuracy

The results without diacritized text:

- **CNN model:**

```
13/13 [==============================] - 0s 16ms/step
Accuracy: 0.906536
```

Figure 4.11: Result of CNN model

- **RNN model:**

```
Fold 11:
13/13 [==============================] - 1s 8ms/step
Accuracy: 0.915089
```

Figure 4.12: Result of RNN model

The model RNN is trained and we have got the best result with: 91.50% accuracy.

In addition, we reviewed the results of accuracy and loss by the significance of the epochs and loss where the value of 60 for the hyperparameter era was set which represents the number of processing times performed in the entire dataset. The value of 64 magnitude of hyperparameter is assigned the size of the batch representing the number of samples fed on the neural network at a time. Figure 11 show the values of accuracy and loss, respectively, produced through the executive sequence of covenants when conducting an RNN model on the distorted text of the Qur'an verses. As shown in figure 11, accuracy was the highest for training samples and data set verification after approximately 32 epochs.

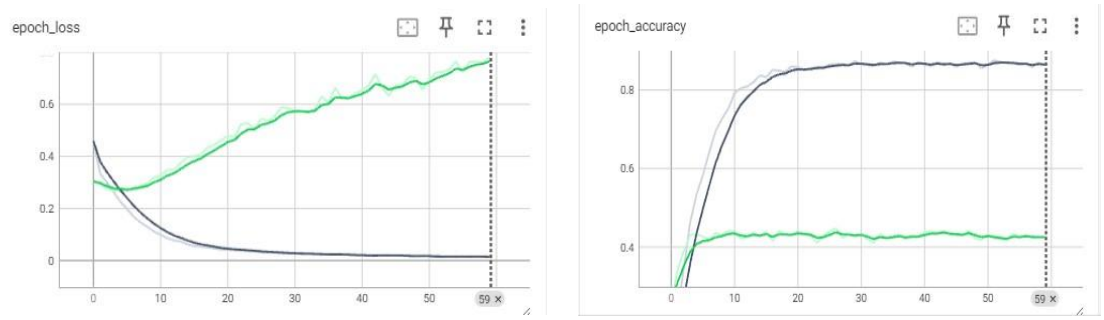

Figure 4.13: Accuracy and Loss per epoch

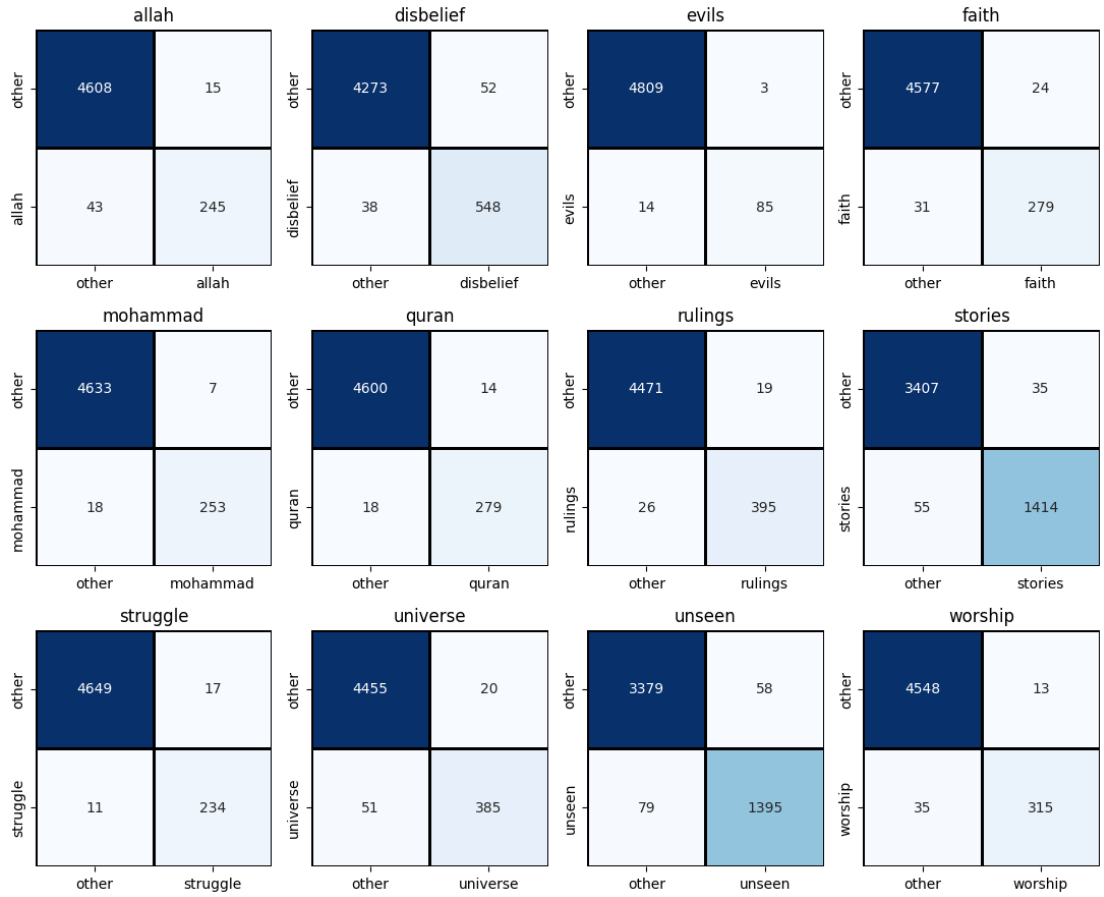

Figure 4.14: Confusion matrices for 12 classified topics of Quran verses.

6.4. Analyse the results and comparison

The details of evaluation results for each label (topic) using diacritized and undiacritized text with cross validation technique are provided through previous results of models.

The previous results of models present the evaluation outcomes for each label (topic) achieved through cross-validation technique using diacritized and undiacritized text. Through making comparisons When comparing the performance results among the different categories, it is clear that the RNN had the highest level of prediction for the "stories" category regardless of whether the text was diacritized or undiacritized. Furthermore, the CNN demonstrated the highest recall and F1-score results for the "stories" category with both types of text. It achieved the highest precision for the label "Mohammad" with diacritized text, GRU achieved less results than previous models in all 12 topics, for example in the theme "unseen" the precision value is equal to 0.93 while in CNN it was 0.96.

By analysing these results, it clear that the RNN and CNN classifiers mostly performed better with the “stories” label which represents the topic that has the largest number of labelled.

This Results with 12-fold cross validation:

	RNN		**CNN**		**GRU**	
Metric	**Diacritiz**	**Undiacritiz**	**Diacritiz**	**Undiacritiz**	**Diacritiz**	**Undiacritiz**
Accurac	0.915292	0.915089	0.907550	0.906536	0.876196	0.8625
Hammi	0.1198	0.01181	0.01115	0.1135	0.018411	0.01912

Table 4.2: Performance results of multi-label classification of Quran verses

7. Conclusion

In this chapter, we have presented the essence of our work which consists in creating a deep learning models to classify Quran verses in 12 topics, for the implementation, we chose three models (CNN, RNN, GRU) for multi label classification and our models produced highest result with cross validation technique.

Conclusion

In conclusion, this project successfully demonstrated the capabilities of deep learning and natural language processing techniques and models in classifying verses of the Holy Quran with high accuracy. The proposed deep learning models, which incorporates advanced neural network architectures and feature extraction techniques, has consistently outperformed traditional classification methods on various benchmark datasets. The results of this research provide valuable contributions to the field of Quranic text classification and open new horizons for exploring applications of deep learning in the field of Islamic studies.

The effectiveness of the proposed deep learning models can be attributed to several factors. First, the models ability to learn complex patterns and relationships within Quranic text data plays a crucial role in accurately classifying verses. Second, the use of advanced neural network architectures, such as adaptive neural networks (CNNs) and recurrent neural networks (RNNs) and Gated recurrent unit (GRU), allows the models to capture local and long-range dependencies in the text, which is essential for understanding the semantic nuances of Quranic verses. In addition, incorporating powerful feature extraction techniques, such as word embedding extraction, helps the models transform raw text data into meaningful representations that can be effectively processed by the neural network.

The contributions of this research and the results of this work provide valuable insights into applying deep learning techniques to analyze and process Quranic text data. The proposed model can be further extended to address other challenging tasks in the field of Quranic studies, such as sentiment analysis, topic modeling, and question answering.

As the field of deep learning continues to develop, more powerful and advanced techniques are expected to emerge, enhancing the capabilities of deep learning models to analyze and process Quranic text data. The research presented in this dissertation represents a starting point toward developing more advanced solutions.

Bibliography

[1] H. Sheikh, C. Prins, and E. Schrijvers, "Artificial Intelligence: Definition and Background," in *Mission AI: The New System Technology*, Cham, Springer International Publishing, 2023, pp. 15-41.

[2] cnil, “cnil,” [Online]. Available: https://www.cnil.fr/fr/definition/apprentissage-profond-deep-learning. [Accessed 13 may 2024].

[3] D. Scientest, “All about deep learning,” 2021.

[4] Alam, Azmir. What is Machine Learning? 2023. 10.5281/zenodo.8231580.

[5] Chat gpt.

[6] “aws.amazon,” [Online]. Available: https://aws.amazon.com/what-is/deep-learning/ [Accessed 13 may 2024].

[7] knowledgehut, “knowledgehut,” [Online]. Available: https://www.knowledgehut.com/blog/data-science/deep-learning-applications. [Accessed 14 may 2024].

[8] builtin, “builtin,” artificial-intelligence, [Online]. Available: https://builtin.com/artificial-intelligence/deep-learning-applications.. [Accessed 15 may 2024].

[9] M Soori , B Arezoo , R Dastres, ” Artificial intelligence, machine learning and deep learning in advanced robotics, a review”2023 .vol.3,pp 54-70.

[10] A. M. Said Gadri, "Contextual Categorization of Documents Using a New Panoply of Similarity Metrics," in *International Conference on Advanced Technology & Science*, Antalya, Turkey, 2014.

[11] N. n. Khoudour Aya nor elhouda, “Application of ensemble Learning in visual question-answering,” 10 octobre 2023.

[12] knowledgehut, “knowledgehut,” [Online]. Available: https://www.knowledgehut.com/blog/data-science/deep-learning-applications. [Accessed 14 may 2024].

[13] A. K. Amey Thakur, "Fundamentals of Neural Networks," *Research in Applied Science & Engineering Technology*, vol. 9, no. VIII, 2021

[14] tutorialspoint, “tutorialspoint,” [Online]. Available: https://www.tutorialspoint.com/artificial_neural_network. [Accessed 17 may 2024].

[15] spiceworks, “spiceworks,” [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-a-neural-network/. [Accessed 17 may 2024].

[16] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, p. 420, 2021.

[17] A. K. Amey Thakur, "Fundamentals of Neural Networks," *Research in Applied Science & Engineering Technology*, vol. 9, no. VIII, 2021.

[18] Mandal, “analyticsvidhya,” 23 Feb 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/Introduction to Convolutional.

[19] geeksforgeeks,“geeksforgeeks,”[Online].Available:https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/. [Accessed 19 may 2024].

[20] geeksforgeeks, “geeksforgeeks,” [Online]. Available: https://www.geeksforgeeks.org/gated-recurrent-unit-networks/?ref=lbp. [Accessed 19 may 2024].

[21] magnimindacademy, “magnimindacademy,” [Online]. Available: https://magnimindacademy.com/blog/deep-learning-and-its-5-advantages/. [Accessed 19 may 2024].

[22] Eguerrand de Rautlin de la Roy ,Thomas R, Akka Z, Pierre B, Kamel H, Laurent M, "Détection d’ouvertures de fenêtres au moyen de réseaux de neurones artificiels LSTM" IBPSA France - Châlons ., May 2022.

[23] careerera, “careerera,” [Online]. Available: https://www.careerera.com/blog/advantages-and-disadvantages-of-deep-learning. [Accessed 20 may 2024].

[24] T. Talaei Khoei, H. Ould Slimane, and N. Kaabouch, "Deep learning: Systematic review, models, challenges, and research directions," *Neural Computing and Applications*, vol. 35, no. 31, pp. 23103-23124, 2023.

[25] A. A. Shan Masood, “A Comprehensive Overview of Modern Techniques and Applications Research,” 2024.

[26] S. Boulaknadel, "Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de Spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation," Doctoral Thesis, Faculty of sciences, Univ. Of Nantes, Nantes, 2008. Accessed on: May, 3, 2024. [Online]. Available: https://theses.hal.science/tel-00479982

[27] F. S. Douzidia, "Résumé automatique de texte arabe," Master's Thesis, Faculty of sciences, Univ. of Montreal, Montreal, 2004. Accessed on: April, 3, 2024. [Online]. Available: https://papyrus.bib.umontreal.ca/xmlui/handle/1866/16637

[28] Tahar DILEKH, "Implémentation d’un outil d’indexation et de recherche des textes en arabe," Master's Thesis, Faculty of sciences, Batna 2 Univ., Batna, 2011. Accessed on: April, 3, 2024. [Online]. Available: https://dspace.univ-batna2.dz/items/d5be6eca-65e2-4fba-8759-b97c766659ad/full

[29] A. KHEMAKHEM, "ArabicLDB: une base lexicale normalisée pour la langue arabe," Master's Thesis, Faculty of Economics and Management, Sfax Univ., Sfax, 2011. Accessed on: April, 3, 2024. [Online]. Available:

[30] A. ED-dariouache, "Etude et réalisation d’un analyseur morphologique de la langue

arabe," Master's Thesis, Faculty of Sciences and Technology Fez, Sidi Mohamed Ben Abdellah Univ., Morocco, 2015. Accessed on: April, 3, 2024. [Online]. Available: https://memoirepfe.fst-usmba.ac.ma/book/3068

[31] S. Baloul, " Développement d'un système automatique de synthèse de la parole à partir de texte arabe standard voyelle," Doctoral Thesis, Faculty of Sciences, Marie France Univ., France, 2003. Accessed on: April, 3, 2024. [Online]. Available: https://cyberdoc.univ-lemans.fr/theses/2003/2003LEMA1015.pdf

[32] Al Hajjar, Abd El Salam, "Extraction et gestion de l'information à partir des documents arabes," Doctoral Thesis, Faculty of Sciences, Paris 8 Univ., Saint-Denis, 2010. [Online]. Available: https://octaviana.fr/document/159212537#?c=&m=&s=&cv=

[33] K. Azeibar, "Un SRI Sémantique pour les traditions prophétiques," Master's Thesis, Faculty of Mathematics and Computer Science, M'sila Univ., M'sila, 2017. Accessed on: May, 1, 2024. [Online]. Available:

[34] S. Zaidi-Ayad, "Une plateforme pour la construction d'ontologie en arabe: Extraction des termes et des relations à partir de textes (Application sur le Saint Coran)," Doctoral Thesis, Faculty of Sciences, Badji Mokhtar - Annaba Univ., Annaba, 2013. Accessed on: May, 16, 2024. [Online]. Available: https://theses-algerie.com/9040629962574994/these-de-doctorat/universite-badji-mokhtar---annaba/une-plateforme-pour-la-construction-d-ontologie-en-arabe-extraction-des-termes-et-des-relations-%C3%A0-partir-de-textes-application-sur-le-saint-coran-

[35] Seyyed Hossein Nasr, Caner K. Dagli, Maria Massi Dakake, Joseph E. B. Lumbard " The Study Quran," HarperCollins Publishers, 2010.

[36] F. I. A. Fayyad, " The seven readings of the Qur'an: A CRITICAL STUDY OF THEIR LINGUISTIC DIFFERENCES," Doctoral Thesis, Department of Arabic and Islamic Studies, Glasgow Univ., M'sila, 1989. Accessed on: May, 17, 2024. [Online]. Available: https://www.gla.ac.uk/myglasgow/research/enlighten/theses/digitisation/

[37] M. Mohd, F. Qamar, I. Al-Sheikh, and R. Salah, "Quranic optical text recognition using deep learning models," *IEEE Access*, vol. 9, pp. 38318-38330, 2021.

[38] Z. Touati-Hamad, M. R. Laouar, I. Bendib, and S. Hakak, "Arabic Quran verses authentication using deep learning and word embeddings," *The International Arab Journal of Information Technology*, vol. 19, no. 4, pp. 681-688, 2022.

[39] M. Galal, M. M. Madbouly, and A. D. E. L. El-Zoghby, "Classifying Arabic text using deep learning," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 23, pp. 3412-3422, 2019.

[40] Mahdi, M. G., Sleem, A., & Elhenawy, I. (2024). Deep Learning Algorithms for Arabic Optical Character Recognition: A Survey. Multicriteria Algorithms with Applications, 2, 65-79.

[41] S. K. Hamed and M. J. Ab Aziz, "Classification of holy Quran translation using neural network technique," *Journal of Engineering and Applied Sciences*, vol. 13, no. 12, pp. 4468-4475, 2018.

[42] S. K. Hamed and M. J. Ab Aziz, "A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification," *Journal of Computer Science*, vol. 12, no. 3, pp. 169-177, 2016.

[43] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, "A group-based feature selection approach to improve classification of Holy Quran verses," in *Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018)*, Johor, Malaysia, Feb. 06-07, 2018, pp. 282-297. Springer International Publishing, 2018.

[44] "Metrics for Multilabel Classification", Jan. 25, 2020. Accessed on: Apr. 6, 2024. [Online]. Available: https://mmuratarat.github.io/2020-01-25/multilabel_classification_metrics

[45] A. Noor and A. Ali, "Multiclass imbalanced classification of Quranic verses using deep learning approach," in *Proc. 2021 4th International Conference on Computing & Information Sciences (ICCIS)*, Nov. 2021, pp. 1-6.

[46] A. El Akadi, "Contribution à la sélection de variables pertinentes en classification supervisée: Application à la sélection des gènes pour les puces à ADN et des caractéristiques faciales," Doctoral Thesis, Faculty of Sciences, Mohammed V Univ., Agdal, Rabat, 2012. Accessed on: May, 4, 2021. [Online]. Available: http://dspace.univ-setif.dz:8888/jspui/handle/123456789/35

[47] S. Behlouli, "Une approche sémantique pour la classification des traditions prophétiques," Master Thesis, Faculty of Mathematics and Computer Science, M'sila Univ., M'sila, 2016. Accessed on: May, 4, 2024. [Online]. Available: https://dspace.univ-msila.dz/items/04090e5a-fc28-4066-a613-8b4659714d30

[48] H. Matallah, "Classification Automatique de Textes Approche Orientée Agent," Magister's Thesis, Faculty of Sciences, University of Aboubekr Belkaid Tlemcen, Tlemcen, 2011. Accessed on: May, 4, 2021. [Online]. Available: http://dspace.univ-tlemcen.dz/handle/112/218

[49] F. Dahoumi, "Identification de la langue et catégorisation thématique des textes d'un corpus multilingue en utilisant les algorithmes: NB, SVM," Master's Thesis, Faculty of Mathematics and Computer Science, M'sila Univ., M'sila, 2013. Accessed on: May, 4, 2024. [Online]. Available: https://dspace.univ-msila.dz/items/184d1a1a-8264-496c-b42a-44a1d3c71524

[50] M. Baali, "Utilisation de la technique des n-gramme dans l'extraction des racines en langue arabe," Master's Thesis, Faculty of Mathematics and Computer Science, M'sila Univ., M'sila, 2015. Accessed on: May, 4, 2024. [Online]. Available: https://dspace.univ-msila.dz/items/1904851d-e7e5-4d59-a150-b50a1f944d55

[51] M. Helassa, "Extraction de connaissances à partir de données: application au hadith," Magister's Thesis, Faculty of Sciences, Ferhat Abbas Sétif Univ., Sétif, 2012. Accessed on: May, 4, 2024. [Online]. Available: https://www.pnst.cerist.dz/detail.php?id=64227

[52] "Classification Metrics: Accuracy, Precision, Recall," Aug. 19, 2022. Accessed on: May. 3, 2024.Available: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall.

[53] "What is the best metric (precision, recall, f1, and accuracy) to evaluate the machine learning model for imbalanced data?," Jul. 14, 2020. [Online]. Accessed on: May. 3, 2024. Available: https://www.researchgate.net/post/What_is_the_best_metric _precision_recall_f1_and_accuracy_to_evaluate_the_machine_learning_model_for _imbalanced_data.

[54] "Confusion Matrix," June. 30, 2022. [Online]. Accessed on: June. 4, 2024. Available: https://pub.towardsai.net/confusion-matrix-179b9c758b55

[55] Seyyed Hossein Nasr, Caner K. Dagli, Maria Massi Dakake, Joseph E. B. Lumbard, Mohammed Rustom " The Study Quran," Book, Faculty of Sciences, Marie France Univ., France, 2003. Accessed on: Apr. 3, 2024. [Online]. Available: https://cyberdoc.univ-lemans.fr/theses/2003/2003LEMA1015.pdf.

[56] A. S. Bedi, "A better way to use Google Cloud from Colab," Medium, Apr. 2, 2021. [Online]. Available: https://medium.com/google-colab/a-better-way-to-use-google-cloud-from-colab-bb93f88b5021. Accessed: Jun. 4, 2024.

[57] Kennesaw State University, "Quickstart: Python," Kennesaw State University, [Online].Available:https://research.kennesaw.edu/computing/resources/quickstart_python.php. Accessed: Jun. 4, 2024.

[58] Quran By Subject, "Quran By Subject," [Online]. Available: https://quranbysubject.com. Accessed: Jun. 4, 2024.

[59] Tanzil, "Browse, Search, and Study the Quran," Tanzil, [Online]. Available: https://tanzil.net. Accessed: Jun. 4, 2024.

[60] Dremio, "Word2Vec," Dremio, [Online]. Available: https://www.dremio.com/wiki/word2vec/. Accessed: Jun. 4, 2024.

[61] Martin Pollák , Peter Gabštur , Marek Kočiško, Prediction of Errors in the Field ofAdditive Manufacturing Technology ,Technical University of Košice, Faculty of Manufacturing Technologies,May 2024.

Abstract:

Deep learning has seen significant growth and development in natural language processing (NLP) for many languages, including Arabic. However, the unique characteristics of Arabic represent many challenges for deep learning models.

The aim of this research is to develop accurate and effective systems for classifying Arabic verses of the Holy Quran using recurrent neural networks (RNN) and convolutional neural networks (CNN).

The proposed model achieved promising results on the data set used, demonstrating its effectiveness in classifying the verses of the Holy Quran with high accuracy. This research contributes to the promotion of Quranic text classification techniques and opens new avenues for exploring deep learning applications in Islamic studies.

Key words: Arabic Language ; Holy Quran ; Deep Learning ; Natural Language Processing (NLP) ; CNN ; RNN

الملخص:

شهد التعلم العميق نموا وتطورا هائلين في معالجة اللغة الطبيعية (NLP) للعديد من اللغات، بما في ذلك اللغة العربية. ومع ذلك فإن الخصائص الفريدة للغة العربية تمثل العديد من التحديات لنماذج التعلم العميق.

الهدف من هذا البحث هو تطوير نظم دقيقة وفعالة لتصنيف آيات القرآن الكريم العربية وذلك باستخدام كل من الشبكات العصبية المتكررة (CNN) والشبكات العصبية التلافيفية (RNN).

حقق النموذج المقترح نتائج واعدة على مجموعة البيانات المستعملة، مما يدل على فعاليته في تصنيف آيات القرآن الكريم بدقة عالية. يساهم هذا البحث في تعزيز تقنيات تصنيف النص القرآني ويفتح آفاقا جديدة لاستكشاف تطبيقات التعلم العميق في مجال الدراسات الإسلامية.

الكلمات المفتاحية: اللغة العربية ؛ القرآن الكريم ؛ التعلم العميق ؛ معالجة اللغة الطبيعية