PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
MOHAMED BOUDIAF UNIVERSITY - M'SILA

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
COMPUTER SCIENCE DEPARTMENT
N° :……

1985
جامعة محمد بوضياف - المسيلة
Université Mohamed Boudiaf - M'sila

Domain: Mathematics and Computer Science
Branch: Computer Science
Specialty: SIGL

A RESEARCH STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER

By: CHIKEUR RIHAB

TOPIC

STORY CLASSIFICATION IN THE HOLY QURAN

Before a Jury composed of:

Dr. …… University of M'sila Supervisor

Dr. ……. University of M'sila Reporter

Dr. ……. University of M'sila Examiner

Academic Year: 2020 /2021

DEDICATION:

In the name of God, the most merciful,

the most merciful and his prophet Muhammad.

«صلى الله عليه و سلم»

I dedicate to my parents that God protect them, as Almighty God said:

{ و قضى ربك الا تعبدو الا اياه و بالوالدين احسانا} الاسراء (23-25)

My father "Farid" Who lit the first candle in my life since my childhood, My sanctuary in this life, may God protect him and prolong his life.

My mother "Hamida" my angel the most beautiful mother in the world of her hugs

The warm friend who has always supported me and wished me much happiness.

To my sisters" Hana, Balsem" and "Sirin" and her husband Fateh and their dear children "Ossama and Tawba" ,

To my brothers "Ayman and Haitham"

To my uncles, aunts, grandfather and grandmother, may God save them,

And to all my friends and colleagues inside and outside the school .

Acknowledgement:

First of all, I thank God the Most High who gave me the courage and the will to achieve this modest work.

"(THE ONE WHO DOESN'T THANK PEOPLE, DOES NOT THANK ALLAH.)"

[Authentic Hadith]

I would like to thank my supervisor Ms. Bentrcia Rahima for the great honor she granted me

By giving me the subject of this graduation thesis. I had the honor and the privilege Work with his help and take advantage of his human and professional capacities With her vast experience, she guided me in this work, Detail with Friendly and dynamic. May this work satisfy you humble Critics, to show my gratitude and appreciation for the help and The advice you gave me, besides knowing that it instilled in me.

I thank all my professors at M'sila University.

My thanks also go to the members of the jury for agreeing to judge my job.

I warmly thank all my family, especially my parents, for always having me supported during my studies. May they find here the fruit of their patience and support permanent that they lavished me to face all the difficult moments.

I would also like to thank all of my colleagues, without exception, whom the university introduced me to, especially my friend Seliman khadi Cheyma, and I would like to thank her all for her help, advice, and everything she gave me.

Thank you for all those who have helped me from near or far to carry out this work.

الملخص:

أولت الأبحاث اهتماما كبيرا لمعالجة البيانات النصية بجميع اللغات و لكن تصنيف النصوص العربية أقل عددا من غيرها و خاصة أنظمة المعلومات التي تعتمد على النص العربي في القرآن الكريم ,هذا الهدف دفعنا للقيام بهذا البحث الذي يستخدم القران الكريم كذخيرة لغوية (حيث أخذنا عينة منه وهي سورة الاعراف) و يستغل تقنيات تعدين النصوص ,و منه فإن مخطط منهجية هذه الدراسةيتكون من : المعالجة المسبقة والنمذجة ، واستخراج المصطلح ،اختزال الأبعاد ومصطلحات الوزن (TF *IDF) ، تطبيق خوارزمية NNتقييم أداء نموذج التصنيف باستخدام قيم الدقة و الاسترجاع

الكلمات المفتاحية : التنقيب عن النصوص ،خوارزميةNN، التنقيب عن بيانات الشبكات العصبية

Abstract:

Research has paid great attention to processing textual data in all languages, but the classification of Arabic texts is less in number than others, especially information systems that depend on the Arabic text in the Holy Qur'an. This goal prompted us to do this research, which uses the Holy Qur'an as a linguistic repertoire (where we took a sample of it, which is Surah Al-Araf) and exploits the techniques of text mining, and from that, the outline of the methodology of this study consists of: pre-processing, modeling, term extraction, Dimensional abbreviation and weight terminology
(TF *IDF). The application of the NN algorithm to evaluate the performance of the classification model using accuracy and recall values.

Keywords: text mining, neural networks, data mining.

Résumé :

La recherche a accordé une grande attention au traitement des données textuelles dans toutes les langues, mais la classification des textes arabes est moins nombreuse que les autres, en particulier les systèmes d'information qui dépendent du texte arabe du Saint Coran. Cet objectif nous a poussé à faire cette recherche, qui utilise le Saint Coran comme répertoire linguistique (où nous en avons pris un échantillon, qui est la sourate Al-Araf) et exploite les techniques de text mining, et à partir de là, le contour de la méthodologie de cette étude se compose de: pré-traitement, modélisation, extraction de termes, abréviation dimensionnelle et terminologie de poids
(TF * IDF). L'application de l'algorithme NN pour évaluer les performances du modèle de classification à l'aide de valeurs précises et de rappel.

Mots-clés : text mining, réseaux de neurones, data mining.

TABLE OF CONTENTS:

LIST OF TABLES:

LIST OF FIGURES:

GENERAL INTRODUCTION

GENERAL INTRODUCTION

The worldwide development of the technical field and the rapid development of networks and the Internet have led to the inflation of the volume of data, or so-called big data, and research has paid great attention to the processing of textual data.

When we talked about the amount of information, we talked about information mining (data mining) of different types of data to extract knowledge.

There are a variety of life applications. Several tools are available that support various algorithms.

Many text classification studies focussed on French text and English text, the classification of Arabic text is less famous and more challenging than the other languages; for this purpose, we have worked on this research which concerns the Holy Quran and the stories of the prophets.

The Quran and Sunnah are the two main sources of Islamic theology, Quran is the word of God forwarded to our prophet Muhammad, (God bless him and grant him peace). When searching for the classification of prophets stories verses, we found that this topic is poorly investigated.

Our objectif is to study the classification of the stories of the prophets in the Qur'an. Since the Holy Qur'an contains many verses that talk about the stories of prophets, we have limited our study to only one surah which contains the stories of a number of prophets and their people. This surah is Surat Al-A'raf and it talked about the stories of 08 prophets. In order to achieve this goal, we extracted the verses of this surah from the Qur'an and analyzed and treated them.

Next, we applied a supervised classification method, namely the Artificial Neural Network ANN to classify each verse to its related prophet class.

The proposed structure of the dissertation can be presented as follows:

An introduction describing the field of research and our problem statement.

• **Chapter 1:** we introduce briefly general concepts related to the domain of:

Data Mining and Text Mining by giving some definitions and their main tasks.

• **Chapter 2**: In this chapter, we have talked about Quran and the verses that talk about the stories of the prophets. We have presented the process of classifying and handling these texts, as well as the difficulties associated with this process.

• **Chapter 3:** this chapter presents the tasks of classification. We have described the algorithms most used in learning textual data and we have demonstrated the different techniques of preprocessing the text before it can be classified.

• **Chapter 4:** Implementation and realization.

Finally, a general conclusion which summarizes the results obtained, the tools used, the difficulties encountered, as well as the future directions.

CHAPTER 01:

DATA MINING AND TEXT MINING

INTRODUCTION

Today, billions of data are collected every day on a global scale. Indeed, the low cost of machines in storage and energy encourages the company to accumulate more and more information. However, despite the amount of data, the number of people treated continues to increase, and experts in the field said that the data collected around the world is doubling every 20 months. Business until then He cannot convert his data into directly usable knowledge. In this chapter we will briefly summarize the definition of data mining (According to MIT, it is one of the 10 emerging technologies that will "change the world" in the 21st century), and its process. we also explain the different tasks of data mining

1.1. DATA MINING

1.1.1. Definition of Data Mining

- Knowledge Discovery is a set of techniques and tools used to work with unstructured documents and written text.
- Extracting meaning from unstructured documents and determining meaning from text.
- No need to read everything to uncover hidden information or making the right decision automatically.
- The non-trivial extraction, from data, of implicit, a priori unknown and potentially useful information (Piatetsky-Shapiro).

1.1.2. Data Mining tasks

The main task of data mining is classification which has the main job of assigning each record in a database to one of the predefined classes. Another is clustering which works in the way it finds groups of records instead of just one record that are close to each other according to user-defined metrics.

The next task is the association which defines involvement rules based on what record subset the attributes can be defined. Data mining is the main

step to achieve knowledge discovery. Normally for data preprocessing goes through various processes such as data cleaning, data integration, data selection and data transformation and after before it is prepared for the mining task. **[1]**

The following figure represents the different stages of the knowledge extraction process:

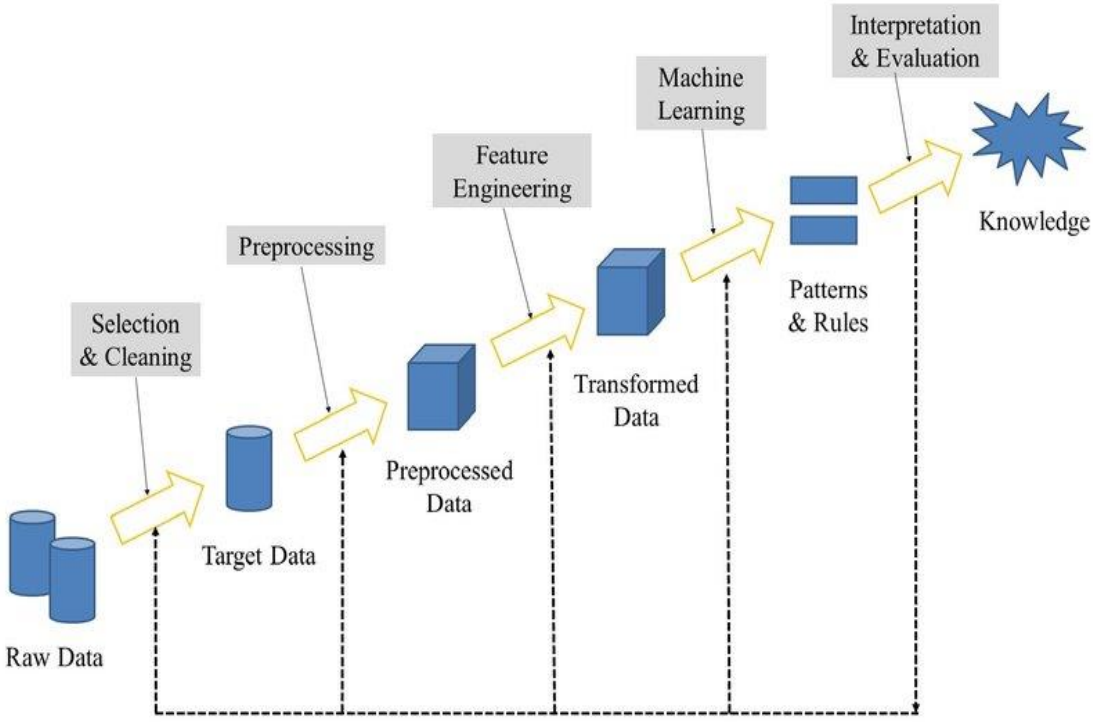

Figure 0.1: Knowledge extraction process from a database **[2]**

1.2. TEXT MINING

Text mining combines all data management and data mining techniques to enable the processing of specific data (i.e. text data). So-called textual data refers, for example, to a textual corpus, answers to unresolved questions in questionnaires. One of the main aspects of text mining is to convert this unstructured textual data (if this is the language used) into data that can be used by conventional data mining algorithms. It is enough to convert the plain text into a table of data, essential for the person in charge of the

analysis. It is the question of deploying the statistical method that best responds to a given problem.

Schematically, we can state:

TEXT MINING = LINGUISTIQUE + DATA MINING

The figure below shows the processing chain of representation, preparation and processing Organized Text Information:

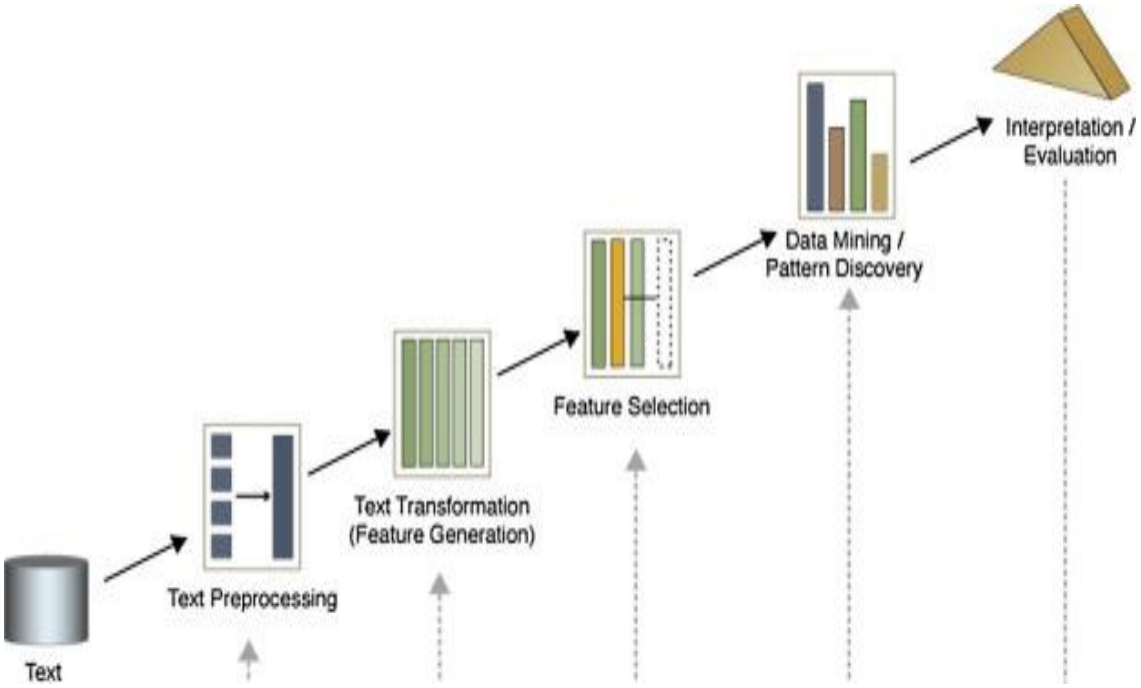

FIGURE 0.2: The Processing Chain For The Text Mining Process

1.2.1. OBJECTIVES OF TEXT MINING

In particular, textual data mining can be used in the following situations:

- Better understand the positioning of speeches, articles and articles Communication.
- To understand the repetitive topics related to the event, Company or competitor.
- Assess the weaknesses and strengths of the press review

.

- Compare text on the same topic to determine the main points
  Distinguish yourself in common or the opposite in style.
- Automatically create a website or email directory associated with it theme. Quantize text or part of text to extract text The most significant structures, such as automatic synthesis and
  Breakdown by subject.
- Establish links between the index and the documents used in the index.
- Establish rules for automatic classification of documents (classification
  Supervised or unsupervised).

1.3. TEXT CLASSIFICATION

Text Classification (TC), also known as text categorization, is the task of automatically assigning a set of documents into categories or classes or subjects from a predefined set. This task, which falls at the crossroads of research information (IR) and machine learning (ML), has witnessed a booming interest over the past decade from researchers and developers **[3]**

The purpose of automatic text categorization is to teach a machine to classify text into the correct category based on its content. Usually, the categories refer to the subjects of the texts, but for particular applications, they can take others **[4]** as shown in **figure 1.2**

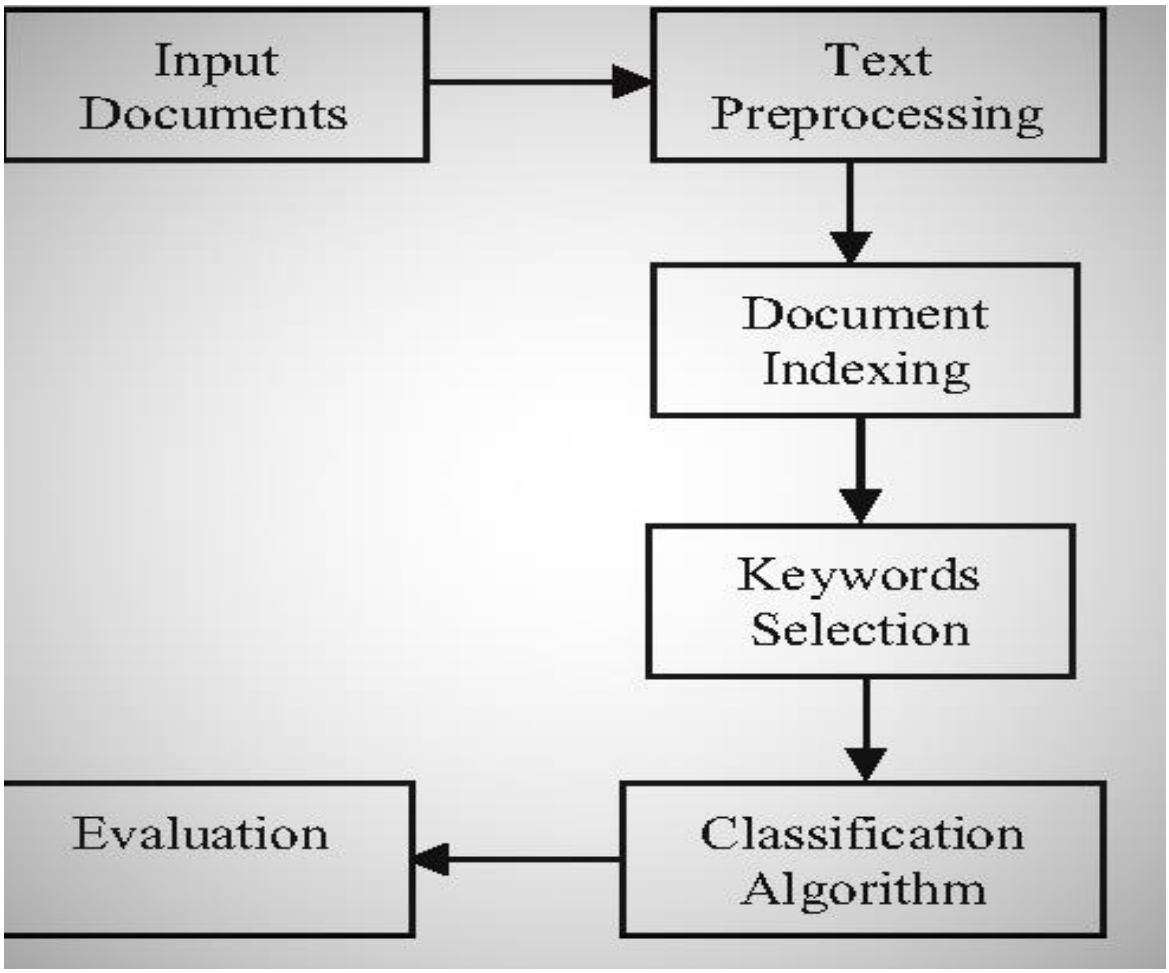

FIGURE 0.3: Text Categorization Process

Text classification requires the extraction of the characteristics of the text in digital form. Several functionalities and extraction methods are explored: word bag, reverse frequency-frequency term of the document, positive point mutual information and word2vec. Different modeling algorithms are used to classify texts by author and clustering algorithms are used to find clusters in the data.

Precision is used to evaluate modeling algorithms. The figure of merit proposed by (Levine & Domany, 2000) and the adjusted Rand index are used to assess the stability of the clusters. **[7]**

Table 1.1 lists the classification methods presented in the literature and their corresponding algorithms among these algorithms **[5]**

TABLE 1.1 : Classification Methods And Algorithms

Classification methods	Algorithms
– Classification and decision trees – Bayesian classifiers – Artificial neural network – K-nearest neighbor classifier – Regression – Classifiers based on association rules – Rough set	– ID3, C4.5 şi C5.0, CART, SPRINT, THAID, CHAID – Naive Bayes, Bayes Net – Single-Layer Perceptron, Multy-Layer Perceptron, RBFNetwork, SVM – K-NN, PEBLS – Linear Regression, Simple Logistic – RIPPER, CN2, Holte≠2s 1R, C4.5

1.3.1. APPLICATIONS OF TEXT CATEGORIZATION

The categorization of texts can be a support for different applications such as:

- Language identification.

- Recognition of writers and categorization of documents multimedia

- The labeling of documents,

- filtering (consisting in determining whether a document is relevant or not

(Binary decision)

- routing (consisting of assigning a document to one or more categories among n. **[8]**

1.3.2. TEXT CLASSIFICATION PROBLEMS

The text classification problem consists of several sub-problems that have been studied extensively in the literature such as document indexing, weighting assignment, document grouping, dimensionality reduction, determination threshold and type of classifiers ...

Several difficulties may oppose the process of categorizing texts, the main ones are as follows **[9]**:

A. Redundancy

Redundancy and synonymy allow the same concept to be expressed by different expressions, several ways of expressing the same thing. This difficulty is linked to the nature of the processed documents expressed in natural language unlike digital data.

B. The Ambiguity

Unlike digital data, textual data is semantically rich, because it is designed and reasoned by human thought. Because of ambiguity, words are sometimes bad descriptors; for example the word lawyer can designate the fruit, the lawyer, or even in the figurative sense, the person who defends a cause.

C. Spelling

A term can contain spelling or typing errors as it can be written in several ways or written with a capital letter. This will weigh on the quality of the results. Because if a term is spelled in two ways in the same document (Ghelizane, Relizane), the simple searching for this term with a single graphic form ignores the presence of the same term in other spellings.

D. Complexity Of The Learning Algorithm

A text is generally represented as a vector containing the number of appearances of the terms in this text. However, the number of texts that we are going to process is very large without forgetting the number of terms composing the same text so we can well imagine the size of the table (texts * terms) to be processed which will considerably complicate the task of classification by reducing system performance.

E. **Presence-Absence Of Terms**

The presence of a word in the text indicates a point that the author wanted to express, we therefore have a relation of implication between the word and the associated concept, although we know very well that there are several ways of expressing the same things, so the absence of a word does not necessarily imply that the concept associated with it is missing from the document. This careful thinking leads us to be careful about the use of learning techniques based on the exclusion of a particular word.

F. **Compound Words: Not Supporting Compound Words Like:**

like Arc-en-ceil, maybe, save-who-can, etc. Whose number is very large in all languages, and treat the word Arc -in-ceil for example by being 3 separate terms considerably reduces the performance of a classification system nevertheless the use of the technique of n-grams for the coding of texts considerably alleviates this problem of compound words.

1.4. ARABIC LANGUAGE

Various previous searches to classify tailor-made texts more on French and English texts, Arabic texts to be classified less frequently than other languages

Our research focuses on the stories of the prophets in the Holy Quran. Therefore, we had to look at the Arabic language, which is the language of the Holy Quran. The Arabic language is an inflectional language, it is not an analytical language. Derivation in Arabic is based on morphological patterns and the verb plays a more inflectional role than in other languages. In addition, Arabic words are made up of roots representing lexical and semantic connectors. Arabic offers the possibility of combining particles and pronouns affixed to words. In other words, Arabic allows a lot of freedom in the ordering of words in a sentence. Thus, the syntax of the sentence can change according to transformational mechanisms such as extra position, confrontation and omission, or

according to the syntactic replacement such as an agent name instead of a verb. **[6]** The research and development of the Arabic text still has a long way to go. Although

researchers have made great efforts, the complex morphological structure of the Arabic language poses challenges. Techniques must be demonstrated to make information retrieval effective for the Arabic language

The reason for this limitation of analytical studies to classify the verses of the Holy Quran and the Arabic language is that Arabic is a challenging language for a number of reasons **[16]:**

1) Arabic is one of the Semitic languages, which belong to the Afro-Asiatic language family an ancient language, spoken in the Middle East and North Africa, and it is written from right to left in a cursive way.

2) Arabic language has 28 consonants, and has no upper and lower case consonants as in English.

3) Arabic has a very complex morphology relative to the morphology of other languages such as: English.

4) Arabic language is a highly inflectional and derivational language which makes monophonically analysis a very complex and difficult task.

5) Arabic opinions are highly subjective to context domains, where you may face words that have different polarity categories in different contexts.

6) Arabic Internet users mostly used dialectal Arabic rather than using MSA(Modern Standard Arabic), where dialectal Arabic resources are scarce. The percentage of spelling mistakes within these Arabic opinions is high, and this represents an additional challenge.

CONCLUSION

In this chapter, we have presented the definition of Data mining and their tasks, the different stages of the knowledge extraction process and the definition of Text mining and their objectives. We also talked about text classification, and then we talked briefly about Arabic language

CHAPTER 02:

PROPHETS STORIES VERSES IN QURAN

INTRODUCTION

The Qur’an and the Sunnah are the two primary sources of Islamic philosophy. God, the Qur’an, was sent by Gabriel to the Messenger Muhammad the recognized in Arabic “The Qur’an is the word of God,” which is indicated by the message, not the law. Which implies all the comprehensive synthesis of humanity as a whole. God explained the rules that Islam follows with the most vital characteristics of this religion through the Qur’an, but unfortunately, there are not many studies on the classification of subjects of the Holy Qur’an, as it is classified as one of the most important critical works of the records of the Qur’anic universe. Usually, since the Heavenly Qur'an is composed in the Arabic language and the Arabic dialect is not dealt with, which leads to the development that dealing with dialects can be shared.

This chapter deals with concepts related to textual corpora in the Holy Qur’an and an overview of the stories of the prophets found in the studied surah.

1.1. DEFINITION OF THE HOLY QURAN

The pronouncement of the Qur’an is derived from the subject of the verb qura’a, meaning al-recitation; That is, the addition and the plural, which includes the saying: I have read the thing. It is the Qur’an; That is, she composed it and gathered some of it together, and the book of God Almighty was called the Qur’an. Because it includes in it stories and news, promises and promises, orders and prohibitions, and also collects verses and suras together

The Noble Qur’an is the word of God Almighty, revealed to His Prophet Muhammad - may God bless him and grant him peace -, miraculous in its wording, worshiped with its recitation, beginning with Surat al-Fatihah, and ending with Surat al-Nas, written in the Qur’an, and transmitted to us frequently

Muslims were interested in the Holy Qur’an and were keen on it to the point that they counted the number of its verses, its words, and even its letters, and their sayings were varied in their number **[10]**. Some of them conveyed consensus that the number

words of the Noble Qur'an is: 77437 words. As for its letters, it was said that they are: 321,000 letters, and it was said: 323015, The reason for the difference in the number of the letters of the Holy Qur'an is due to the disagreement in considering the Basmalah as a verse from the beginning. Every surah of the Qur'an or not to consider it, as well as the dispute over the letters of the tide and the like, is it one of the letters of the Holy Qur'an or not.

As for the number of verses of the Noble Qur'an, it is 6,200 verses, but their opinions were varied regarding more than that

As for the number of verses of the Noble Qur'an, it is 6,200 verses, Moreover, scholars found one hundred and fourteen (114) surahs that begin with Surat Al-Fatiha and end with Surat Al-Nas, and the number of its parties is 60.

2.2. DEFINITION OF THE STORY

The story is a very old global literary art, and it was found among most people and nations before Islam , especially among the Roman and Persian civilization , and the Holy Qur'an also contained many stories of previous nations, and it even addressed the Arabs in a narrative manner appropriate to their tendencies and natures based on their love of listening to stories and historical news.

The story is either a true realistic novel, either it is a fabricated story that aims to stimulate interest in portraying emotions and moral ideals, or the strangeness of its events and its language ,as it may be prose or poetry, as critics agree that there are specific elements of the story that must be available for its success, which are events ,and the characters , time ,place , and narration . Also, it can be said that the story is one of the literary arts that expresses the issues of daily life and their problems , and it meets the social and psychological needs of people by narrating events and facts**[11]**.

2.3. THE STORY IN THE HOLY QURAN

The Qur’an does not deal with the story from all its sides, as it is limited in most cases to mentioning one or more episodes of it and its failure to fulfill all the elements of the combined story of dialogue, people, time, place and complex

The story in the Holy Qur’an is considered the first committed story in Arab literature due to the style in which it was presented and its psychological and artistic influence, with the aim of it in terms of calling for monotheism.

What was revealed from the Qur’anic stories at the beginning of the da'wah was characterized by briefly presenting the events of the story by mentioning those who suffered torture without being subjected to the names of their prophets and the dialogue between them because the first purpose here is to warn the polytheists from stubbornness, denial and intimidation of what will befall them

2.4. ELEMENTS OF THE STORY IN THE HOLY QURAN

The familiar elements of the story, such as events, people, dialogue, spatial connection, temporal arrangement, and complex ... are not found all together in the Qur’anic story because the purpose is the style of presentation. The Qur’anic story is not concerned with mentioning the time and determining the date of the incident or its duration unless in its determination of the dimensions of the value of the incident itself, the chronology of the events, and the mention of the historical facts, the Qur’an has not adhered to it. Important The purpose of telling the stories:

- Warning and intimidation of the polytheists and reminding them of the consequences of denying them, and this is what the events element highlights.
- Confirmation of the Messenger and the believers on the shame, and this highlights the element of the people.
- Establishing an argument and persuading by telling the words of the opponent or introducing a personality to highlight the element of dialogue.

2.5. STORY REPETITION IN THE HOLY QURAN

Here we mention some aspects that could be an explanation of repetition of one story in the first Holy Quran:

That the repetition is due to the multiplicity of the religious purpose that results from the one story are multiple, so the story may come in a place to perform another purpose and so on

The second:

That the Noble Quran has token the story as basic to confirm some of the Islamic concepts of the Muslim community.

This was done by observing the external facts that the nation was experiencing and linking them to the reality of the story in terms of unity of purpose and content. This link between the Islamic concept in the story and the external lived experience of the Muslims may lead to a wrong understanding of the concept intended to be given to the nation ; So he understands his confinement within the story that context of the story that the story lived through and its special circumstances .

The one story in the Holy Quran is repeated in order to avoid this restriction and narrowing of the concept, and to confirm its comprehensiveness and breadth of all similar facts and events, so that it takes the characteristic of a moral or historical law that applies to all facts and events .

The third:

that repetition is the reason for the effectiveness of the story as a warning to the nation on the relationship of the external issue that it faces in the era of revelation or after it with the Islamic concept to derive from it its spirit and approach, so that the repetition of the opportunity is a statement of the stimulus when it is needed **[12].**

2.6. ANALYSIS OF THE STUDIED SURAH " SURAT AL-A'RAF "

Surat Al-A'raf is considered one of the longest Meccan suras, with 206 verses. It is the seventh chapter in the Qur'an, and it is the first of the Qur'anic surahs that have presented the stories of the prophets in detail since the beginning of

Adam. Creation until the end of creation. It could be a separator between Heaven and Hell, isolating its people and keeping them, and the owners of traditions are those whose sins are equal to their greatness, so their terrible deeds await them to enter Heaven, and their great deeds are avoided. Them enter Hell. The surah was defined by the gathering of the prophets, peace be upon them, and their position with their

members to serve as an excuse for the worshipers to persevere in hardship and tolerate harm.

The surah contained four main themes, and they are.

First (verses 11-25, 38-53): Introducing the beginning of man's creation, honoring him with the prostration of angels, the temptation of Satan to Adam and his wife, and dropping them out of Heaven, then their fate on the Day of Resurrection.

Second (verses 59-93, 103-157, 159-171):

Stories about a late stage in the lives of five of the prophets, namely: Noah, Aad, Thamud, Lot, and Shuaib. They brought their people with one message from their Lord, which is the Almighty saying: {O people Worship Allah, what you have of a god other than him), to show what the call ended in terms of the salvation of the believers and the destruction of the deniers, and what their earthly reward was and what will be in the hereafter, and the story of our master Moses in two stages, the first is his story with Pharaoh and the second with the people of bni Israel.

Third (verses 94-102, 180-206):

A statement that the world is a home of affliction, that there is an afterlife and an account of deeds, and that works result in a benefit or a loss, in this world and the hereafter.

Fourth (verses 1-10, 26-37, 54-58, 158, 172-179):

That God excused people and established the argument against them, and that disbelief and corruption are rooted in the owners of hell, and that he called them to consider the destruction of those who preceded them among the nations and that most of them are unbelievers And that he testified them against themselves by faith and that he is their Lord**.**

The number of words in Surat Al-A'raf is 3344, we extracted the lemmas(The lemmas shown below is split by part-of-speech and sorted by frequency. A
lemma groups word-forms that differ only by inflectional –as opposed to derivational-morphology, and do not vary in meaning. Each lemma is shown in Arabic and using Buckwalter transliteration. Verbs are shown in a seperate verb concordance) from it, processed it, deleted the duplicate words, and collected them in a text file to create a matrix that studies the stories of the prophets in Surat al-A'raf.

2.7. QURAN CORPUS

2.7.1. Annotated Arabic Corpora

The Quran contains over 77,000 words; it is divided into 114 chapters where each chapter is divided into verses, adding up to total of 6,243 verses. Some relevant corpora are created from the original text of the Holy Quran, namely **[13][26]:**

1. Quran Corpus of Haifa (Dror et al. 2004): This corpus has been built using an automatic morphological analysis for complete, it remains manually unverified and has multiple possible analyses for each word in the final published data set. Considering a random sample, the authors of the Haifa corpus estimate the final accuracy of annotation using an F-measure of 86%. Further, approximately 40% of the roots in Haifa's corpus are missing and the words lemmas are not given**[27].**
2. The Qur'anic Arabic Corpus (Dukes and Habash 2010): It is online-annotated corpus with multiple layers of annotation including morphological segmentation.

Part of speech tagging. Syntactic analysis using dependency grammar and a semantic ontology. Despite that this corpus in manually verified, it has some problems on the level of lemmas and roots, and has sufficient grammatical information: yet, the patterns are not given**[28].**

3. QurAna corpus (Sharaf and Atwell 2012): In this corpus, only the personal pronouns are tagged with antecedent information (over 24,500 pronouns). These antecedents are maintained as an ontological list of concepts. The Qur'anic Arabic Corpus was used to identify the targeted segments that contain pronouns, and for each pronoun, the starting and ending IDs of the text span that represents antecedents were recorded manually through forms developed using PHP scripting language**[29].**

4. QurSim corpus (Sharaf and Atwell 2012): It is an annotated corpus where semantically similar or related verses are linked together. With the help of domain experts. The authors adopt the same methodology of Ibn Kathir, a Muslim scholar who is known for his classic book of Quran commentary (or Tafsir in Arabic). In fact, the principle of this method is to link two verses if one of them was cited while commenting on the other. The size of the dataset is over 7,600pairs of related verses and the authors claimed that this dataset could be extended to over 13,500 pairs of related verses observing the commutative property of strongly related pairs**[30].**

5. The Boundary – Annotated Quran Corpus (Sawalha et al. 2014): Unlike the other Qur'anic corpora, the words in this corpus are tagged with prosodic and boundary annotation rather than morphological or syntactical annotation. It was built by gathering and tracking boundary stops from the "Tanzil Quran project", the part of speech tags from Tajwid (recitation) mark-up in the Quran**[31].**

6. Qurany: the Qur'anic text is augmented with an ontology or index of key concepts that were imported from "Mushaf al Tajweed", a recognized expert source which is compiled by Dr. Mohamed Habash, Director of the Islamic Studies Centre in Damascus, published by Dar Al-Maarifah in Syria and authenticated by Al-Azhar Islamic Research Academy in Egypt. The Qurany

allows users to search in the Holy Quran for abstract concept among nearly 1200 concepts and find the related verses to this concept; yet, the corpus includes 8 variant English translations**.**

7. Al-Mushaf corpus (Zeroual and Lakhouaja 2016): It is an enriched corpus with morph syntactical information. The process of building this corpus consists of a semi- automatic technique by using "Al-Khalil Morpho Sys", then manual processes. The corpus has 1770 roots, vowelled patterns for each stem and lemma, over 100 part of speech tags used and true lemma (1554 patterns**)[32].**

In our research we utilized the Arabic Qur'anic corpus (http://corpus.quran.com**)[33],** because it is an commented on phonetic source with different layers of commentaries, including Morphological division, portion marks, and linguistic analysis utilizing the rules of dependency.

2.7.2. Text description of « quran corpus »

The Qur'anic Arabic Corpus is an on-line annotated linguistic resource with multiple layers of annotation including morphological segmentation, part-of-speech tagging, syntactic analysis using dependency grammar "إعراب القرآن الكريم" and a semantic ontology. The motivation behind this work is to produce a resource that permits further analysis of the Quran, the 1,400 year old central sacred text of Islam. The 77,430 words of the Quran form a definite genre difficult to match to other texts of Arabic. Processing Qur'anic Arabic may be a unique challenge from a computational point of view, since it differs significantly from Modern Standard Arabic. This open source,includes the markers for the speech part of the Quran which represents an annotated linguistic resource shows the grammar, syntax and morphology of for each word of the Holy Quran,Segmentation morphological the formal representation of the Quran. Qur'anic Grammar using dependencies. The corpus provides three levels of analysis: morphological annotation, a syntactic Treebank and a semantic ontology.

The data in the corpus is written in Buckwalter Arabic transliteration scheme (Buckwalter scheme) and organized into four columns as follows:

1. **LOCATION:** consists of four parts: (chapter no: verse no: word no: part no).
2. **FORM:** consists of the main parts of the word.
3. **TAG:** includes the part-of-speech tag for each part of the word such as Noun, Verb, and Adjective, etc.

4. **FEATURES:** describes morphological features of the word such as Root, Stem, and Gender, etc. **[19]**

LOCATION	FORM	TAG	FEATURES
(7:1:1:1)	Al^m^S^	INL	STEM|POS:INL
(7:2:1:1)	kita`bN	N	STEM|POS:N|LEM:kita`b ROOT:ktb|M|INDEF|NOM
(7:2:2:1)	>unzila	V	STEM|POS:V|PERF|PASS|(IV)|LEM:>anzala|ROOT:nzl|3MS
(7:2:3:1)	<ilayo	P	STEM|POS:P|LEM:<ilaY`
(7:2:3:2)	ka	PRON	SUFFIX|PRON:2MS
(7:2:4:1)	fa	REM	PREFIX|f:REM+
(7:2:4:2)	laA	PRO	STEM|POS:PRO|LEM:laA
(7:2:5:1)	yakun	V	STEM|POS:V|IMPF|LEM:kaAna|ROOT:kwn|SP:kaAn|3MS|MOOD:JUS
(7:2:6:1)	fiY	P	STEM|POS:P|LEM:fiY
(7:2:7:1)	Sadori	N	STEM|POS:N|LEM:Sador|ROOT:Sdr|M|GEN
(7:2:7:2)	ka	PRON	SUFFIX|PRON:2MS
(7:2:8:1)	HarajN	N	STEM|POS:N|LEM:Haraj|ROOT:Hrj|M|INDEF|NOM
(7:2:9:1)	m~ino	P	STEM|POS:P|LEM:min
(7:2:9:2)	hu	PRON	SUFFIX|PRON:3MS
(7:2:10:1)	li	PRP	PREFIX|l:PRP+
(7:2:10:2)	tun*ira	V	STEM|POS:V|IMPF|(IV)|LEM:>an*ara|ROOT:n*r|2MS|MOOD:SUBJ
(7:2:11:1)	bi	P	PREFIX|bi+
(7:2:11:2)	hi.	PRON	STEM|POS:PRON|3MS
(7:2:12:1)	wa	CONJ	PREFIX|w:CONJ+
(7:2:12:2)	*ikoraY`	N	STEM|POS:N|LEM:*ikoraY`|ROOT:*kr|F|GEN
(7:2:13:1)	li	P	PREFIX|l:P+
(7:2:13:2)	lo	DET	PREFIX|Al+
(7:2:13:3)	mu&ominiyna	N	STEM|POS:N|ACT|PCPL|(IV)|LEM:mu&omin|ROOT:Amn|MP|GEN
(7:3:1:1)	{t~abiEu	V	STEM|POS:V|IMPV|(VIII)|LEM:{t~abaEa|ROOT:tbE|2MP

Figure 2.1 Verses of Al-A'raf chapter in Buckwalter transliteration scheme**[39]**

2.7.3. Morphological annotation

➢ Processing Qur'anic Arabic

Processing Quran Arabic is a unique challenge because the vocabulary and spelling are different from MSA. However, unlike most other Arabic scriptures, the Arabic Quran has the advantage of being completely unobtrusive. Each word of the Quran contains detailed diacritics (marks) on all letters that describe its exact vowel (**see Figure 2.2**). Using this information offers an advantage for automatic annotations over other forms of Arabic. **Figure 2.2** below shows an example of a word in the Arabic corpus of the Quran, displayed to website users displaying morphological annotations.

The three numbers at the top of the figure give the chapter number, the verse number and the word number. The Quran is divided into 114 chapters. Each chapter contains a

series of verse numbers. In this example, the comment is the fourth word of Chapter 21, Section 70.

The next line in **Figure 2.2** is linear segment-to-segment translation, followed by full translation and voice transcription. The displayed pronunciation is automatically extracted from morphological annotations and diacritics that already exist in the text. From a computer point of view, processing the Quran in Arabic is a unique challenge because the vocabulary and spelling are different from MSA. However,

unlike most other Arabic scriptures, the Arabic Quran has the advantage of being completely unobtrusive. Each word of the Quran contains detailed diacritics (marks) on all letters that describe its exact vowel (see Figure 2.2). Use this information to make automatic Notes compared to other forms of Arabic.

Figure 2.2 below shows an example word in the Qur'anic Arabic Corpus, as displayed to website users viewing the morphological annotation. The three numbers at the top of the figure give the chapter number, verse number and word number. The Quran is split into 114 chapters. Each chapter contains a sequence of numbered verses. In this example, the annotation corresponds to the fourth word of chapter 21, verse 70. The next line in Figure 1 is a segment-for-segment interlinear translation, followed by a full translation and a phonetic transcription. The pronunciation shown is derived automatically from the morphological annotation and diacritics already present in the text **[14]**.

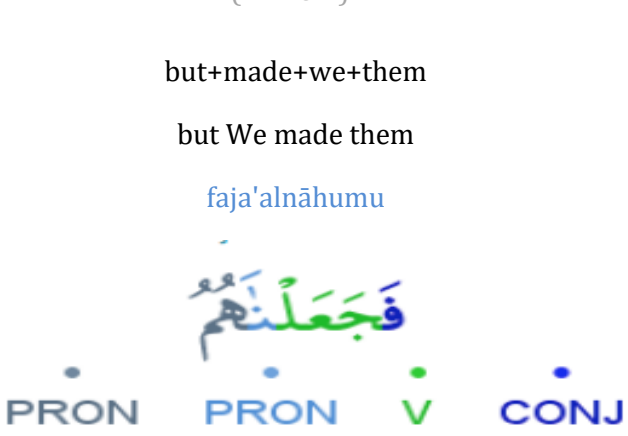

FIGURE 2.7: Morphological Segmentation Of A Fully Discretized Arabic Words In The Qur'anic Arabic Corpus

- **The qur'anic arabic corpus tagset :**

The Qur'anic Arabic Corpus differs from other related annotated Arabic corpora by adopting historical traditional Arabic grammar. Known as *i'rāb* "إعراب", this standardized grammar of the Quran has been developed and documented in detail for over 1,000 years – far longer than corresponding grammars for most other languages .In fact, traditional Arabic grammar is widely recognized as one of the origins of modern dependency grammar .Adopting this approach leads to morphological annotation which uses familiar terminology.

Using traditional grammar along with its standardized terminology also enables the morphological annotation to be verified against the many existing books and publications on Qur'anic grammar.**[34][35][36][37][38]**

Traditional Arabic grammar defines a detailed part-of-speech hierarchy which applies to both words and morphological segments. Fundamentally, a word may be classified as a verb, nominal, or a particle. The set of nominal include nouns, proper nouns, adjectives, subject pronouns and object pronouns. The particles include prepositions, conjunctions and interrogatives, as well as many others.

In the Qur'anic Arabic Corpus, initial automatic tagging was carried out using a modified version of the Buck Walter Arabic Morphological Analyzer (BAMA), adapted to the unique language of the Quran. BAMA defines its own tagset and segmentation scheme suitable for MSA **[39].** Since BAMA was used to perform the initial automatic analysis in the corpus, a mapping was required to convert to the desired Qur'anic Tagset. For the vast majority of words, this was a one-to-one process. However, in few cases, the Qur'anic Tagset was more detailed. For these words (such as the several types of particles), manual disambiguation was required.

TABLE 2.1: Buckwalter Transliteration With Examples

Characters			Examples			
Arabic	**Transliteration**	**Buckwalter**	**Arabic**	**Transliteration**	**Transcription**	**Gloss**
ء	‘	‘	سماء	samaA’	/samā’/	sky
آ	Ā	|	آمن	Āmana	/’āmana/	he believed
أ	Â	>	سأل	saÂala	/sa’ala/	he asked
ؤ	ŵ	&	مؤمن	muŵmin	/mu’min/	believer
إ	Ă	<	إنسان	Ăinsan	/’insan/	**man**
ئ	ŷ	}	سائل	saAŷil	/sā’il/	liquid
ا	A	A	كان	kaAna	/kāna/	he was
ب	b	b	بدا	badaA	/badā/	appear
ة	ћ	p	جنة	jan~ap	/janaћ/	heaven
ت	t	t	تحت	taHot	/taHot/	Below
ث	θ	v	ثلاثة† …	θalaAθaћ	/θalāθa/	three
ج	j	j	جنة ^	jan~ap	/janaћ/	heaven
ح	H	H	حكم	Hakama	/ Hakama /	judge
خ	x	x	خير Œ	xayor	/ xayor /	Well-being
د	d	d	دنيا	d~unoyaA	/ dunoyaA /	World
ذ	ð	*	ذلك	ða`lik	/ ða`lik /	That
ر	r	r	رجال	rijaAl	/ rijaAl /	men
ز	z	z	زينة	ziynaћ	/zīna/	decoration
س	s	s	سماء	samaA’	/samā’/	sky
ش	š	$	شريف	šariyf	/šarīf/	honest
ص	S	S	• ‹ صوت	Sawt	/Sawt/	sound
ض	D	D	–ضرير	Dala`lap	/ Dala`la/	blind
ط	T	T	طهارا	Tah~ara	/ Tahara /	Clean
ظ	Ď	Z	~ ™Ş ظلم	Ďulm	/Ďulm/	injustice
ع	ς	E	عمل	ςamal	/ςamal/	work
غ	γ	g	غافل	ga`fil	/ ga`fil /	oblivious
ف	f	f	~ ™ǂ f فاسق	faAsiq	/ faAsiq /	punk
ق	q	q	قادر	qaAdir	/qādir/	capable
ك	k	k	كريم	kariym	/karīm/	generous
ل	l	l	لكن	la`kin	/ la`kin /	But
م	m	m	مجرم	mujorim	/ mujorim /	criminal
ن	n	n	نور ›	nuwr	/nūr/	light
\	h	h	هل	hal	/hal/	Do you

و	w	w	وصل	waSl	/waSl/	receipt
ى	ý	Y	على	ςalaý	/ςala/	on
ي	y	y	دين	diyn	/tīn/	figs
◌َ	a	a	لَعَب	laEab	/dahana/	he played
◌ُ	u	u	لُعِب	luEib	/duhina/	it was played
◌ِ	i	i	لُعِب	luEib	/duhina/	it was play
◌ً	ã	F	كتاباً	kitaAbAã	/kitāban/	a book [nom.]
◌ٌ	ũ	N	كتابٌ	kitaAbũ	/kitābun/	a book [acc.]
◌ٍ	ĩ	K	كتابٍ	kitaAbĩ	/kitābin/	a book [gen.]
◌ّ†	~	~	كسّر	kas~ara	/kassara/	he smashed
◌ْ‡	.	o	مَسْجِد	mas.jid or masjid	/masjid/	mosque
_	_	§ -	مَسْـــجِد	mas._____jid	/masjid/	mosque

CONCLUSION

In this chapter we have studied and described the texts of the Holy Qur'an and the morphological explanation of its verses and how to treat its words. We also talked about the stories of the prophets that are included in the surah studied in the Holy Qur'an.

CHAPTER 03: TEXT PREPROCESSING AND CLASSIFICATION

INTRODUCTION

Classification is a very important process in data mining, it is about creating a model that can be applied to data when setting up our database (cleaning, filling, ...) linked to the treated subject and after collecting the data set, a preprocessing step must be carried out before classification, in the pre-treatment stage, it is necessary to go through several stages clean data and remove noise that can affect accuracy classification and this is what we will show in this chapter.

3.1. CLASSIFICATION

Classification is the most common task of Data Mining and which seems to be a human obligation. In order to understand our daily life, we are constantly classified, categorized and evaluated.

Classification consists of studying the characteristics of a new object in order to assign it a predefined class. The objects to be classified are generally records of a database, classification consists of updating each record in determining a class field. The classification task is characterized by a definition specific classes and a set of previously classified examples.**[14]**

The classification and structuring methods group together techniques which make it possible to reduce a data set of more or less large size to a smaller number of classes or general factors, thus making it easier to read and understand the initial data. The foundations of these methods have many points in common, particularly in terms of the objectives as well as the prior analysis and processing of the data.

Text classification is defined as an operation that identifies classes equivalence between text segments taking into account their informational content (words, n-gram, etc.).

3.2. TYPES OF CLASSIFICATION

There are two classification approaches, the classification supervised classes are known a priori, against unsupervised classification (clustering) classes are based on the structure of objects. Figure 3.1 shows the difference between supervised and unsupervised classification.

3.2.1. Supervised classification

The “classification” is a supervised method which consists in defining a function which assigns one or more classes to each data. In this approach it is assumed that an expert previously provides the labels for each data, the labels are membership classes. According to [Govaert, 2003]: "(the supervised classification (also called classification or inductive classification) aims to" learn "by example. It seeks to explain and predict the membership of documents in known classes a priori. Thus it is the set of techniques which aim to guess the membership of an individual to a class by using only the values which he takes) ”.

Discrimination (or supervised methods) can be based on probabilistic hypotheses (naive Bayes classifier, parametric methods) or on notions of proximity (closest neighbors) or even on searches in hypothesis spaces (trees of decision, neural networks).

The simplest purpose of a classification is to divide the sample into groups of homogeneous observations, each group being well differentiated from the others. Most of the time, however, this lens is more refined; we generally want to get sections inside main groups, then smaller subdivisions of these sections, and so on in short, we want to have a hierarchy, that is to say a series of “nested” partitions, more and more finer, on the initial set of observations.**[15]**

3.2.2. Unsupervised classification

This classification is also called “automatic classification”, “clustering” or even “regrouping”.

Unsupervised learning is learning without a supervisor. It is about extracting classes or groups of individuals with common characteristics. The quality of a classification method is measured by its ability to uncover some or all of the patterns hidden.

There are several families of unsupervised classification methods. The most common are:

- hierarchical classification;
- non-hierarchical classification, for example the k-means method;
- classification based on density;
- classification based on statistical / probabilistic models, for example a mixture of normal distributions.

3.3. Supervised Classification Vs Unsupervised Classification

The supervised classification consists in identifying the class of belonging of an object from of certain descriptive features. This approach allows the automatic assignment of documents in pre-existing classes.

The objective is to find a functional link, which is also called a model of prediction, between the texts to be classified and all the categories. prediction, it is necessary to have a set of previously labeled texts, called a learning set, from which the parameters of the prediction model are estimated the most efficient as possible, that is, which produces the fewest errors in prediction.

Not at all like the unsupervised classification where the computer must find itself groups of archives, the administered classification accept that there's as of now a classification of records. Usually the case, for case, of a library or a research. The objective is at that point to consequently classify a modern archive. It is therefore first to learn a demonstrate, or classifier, from a preparing set composed of couples (object, class). Unlike unsupervised classification, supervised classification can measure the significance of each word for classifying modern records. For illustration, a measure (information gain) calculates the regularity of a term. The more a word is connected to a

category and not to others, and more vitally: on the off chance that a modern record contains it, this word will be very discriminant. Numerous comparative measures have been created.

Finally, unlike the unsupervised classification, it is easy to evaluate the results here of a classification. Among the N examples of classified documents, part of the documents for training, and the rest for the test. During the test phase, we submit each document to the classification algorithm and we simply see if the machine find the right class. Of course, the result of this test is in no way guaranteed when the machine will have to classify new documents! (passing the test is necessary, but not sufficient) **[18]**

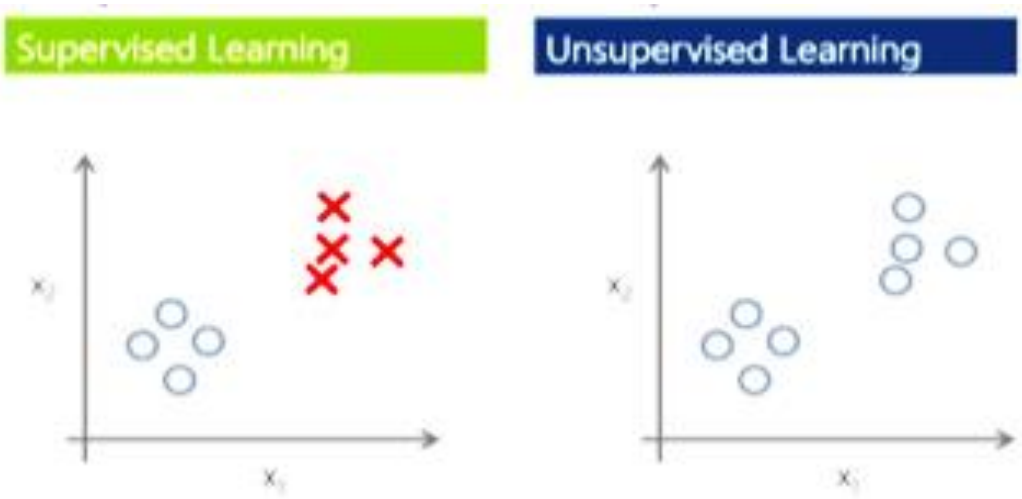

FIGURE 3.1: Supervised **Vs** Unsupervised Learning

3.4. THE STAGES OF A CLASSIFICATION

1. Choice of data.

2. Calculation of the similarities between the n individuals from the initial data.

3. Choice of a classification and execution algorithm.

4. Interpretation of the results: - evaluation of the quality of the classification, - description of the classes obtained.

In the following figure, we illustrated the steps that we followed in classifying verses of the stories of the prophets from Surat al-A'raf:

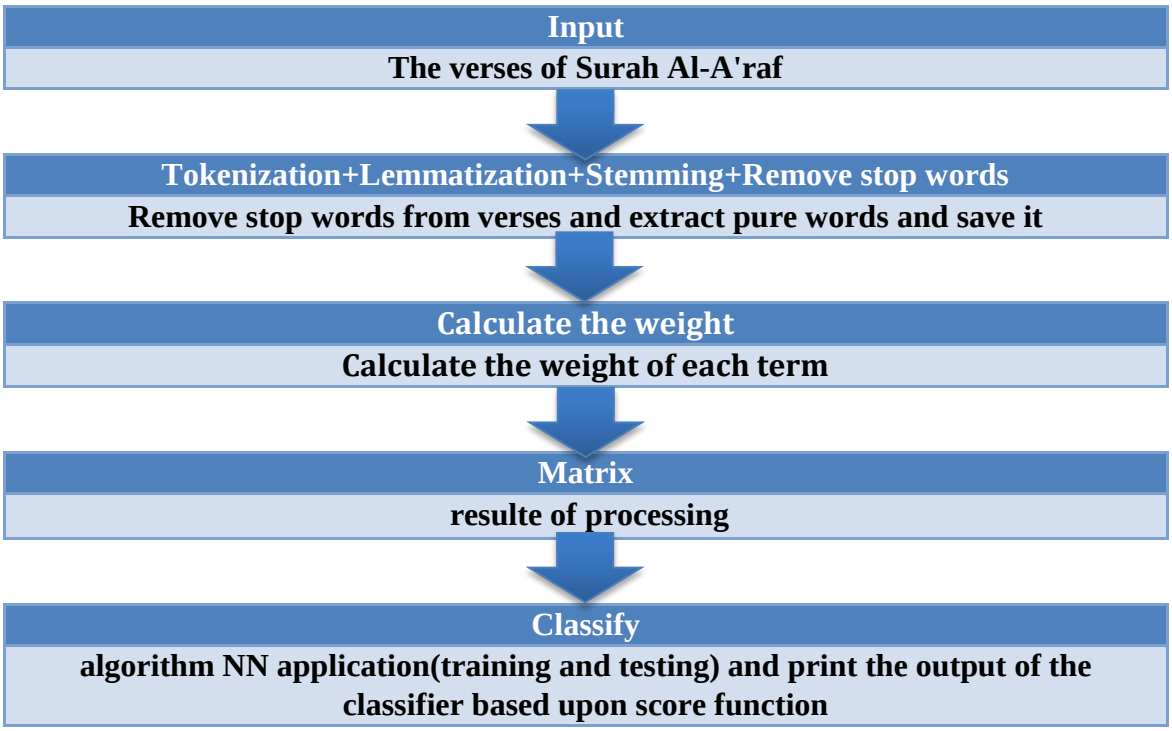

FIGURE 3.2: Classification Flow Diagram

3.5. PREPROCESSING

Textual data counting the specific frame of complex information. They are not explicitly delimited, organized and semantically labeled. Thus these data require earlier processing. Generally, the objective of preprocessing is to (Li, 2019) down look space. The preprocessing of the Arabic content could be a troublesome and critical step. **[17]**

It can have a positive or negative affect on the exactness of any data system, On the other side, Quran mining occupies a large area in text mining although very few approaches have been developed for Quranic Arabic due to the depth of knowledge needed in this field and the challenges related to Arabic script ,recovery. And thus the change of the pretreatment step leads to necessarily to the enhancement of any recuperation data framework exceptionally strongly. Preprocessing contains numerous sub-processes and each contains a particular work for prepare the information to be in ideal shape so the result can be moved forward. The system proposed centers on the taking after preprocessing steps:

- Tokenization.
- Lemmatization.
- Stemming.
- The removal of stop words.

The figure 3.3 shows the various operations performed during preprocessing:

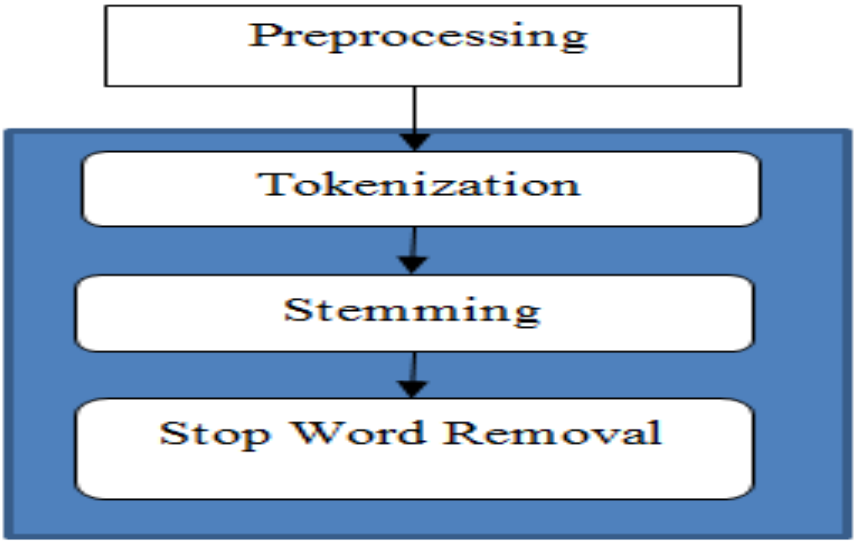

FIGURE 3.5: Preprocessing Operations

3.5.1. Tokenization

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens , perhaps at the same time throwing away certain characters, such as punctuation.

These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type/token distinction. A token is an instance of a sequence of

characters in some particular document that are grouped together as a useful semantic unit for processing. A type is the class of all tokens containing the same character sequence.

3.5.2. Lemmatization

Is the process where we take individual tokens from a sentence and we try to reduce them to their base form. The process that makes this possible is having a vocabulary and performing morphological analysis to remove inflectional endings.
The output of the lemmatization process is the lemma or the base form of the word. For instance, a lemmatization process reduces the inflections, "am", "are", and "is", to the base form, "be".

Lemmatization is useful for text flattening for text or search engine classification tasks, and for a variety of other NLP tasks such as verse classification that we are doing in this research. It is especially important when dealing with complex languages such as Arabic and Spanish.

3.5.3. Stemming

Stemming or deuffixation is just a simpler version of lemmatization where we are interested in stripping the suffix at the end of the word. When stemming we are interesting in reducing the inflected or derived word to it's base form.

Which to associate several words with the same root, that is to say to remove the suffixes of the words to keep only the root part, using simple rule-based algorithms string replacement to remove the most used suffixes.

Stemming example:

Word	الكاتب	الكتاب	المكتبة
Stem	كتب		

3.5.4. The removal of stop words

These words ought to be removed from the representation of texts for two reasons:

• From a linguistic point of see, these words contain exceptionally small information. The presence or nonappearance of these words does not offer assistance to guess the meaning of a content. For for this reason, they are commonly alluded to as "stop words".

• from a measurable point of view, these words are found in all the texts without no separation and are of no offer assistance in classification.

3.6. Weighting or calculate weight

Term weighting measures the importance of a term in a document. This significance is often calculated from statistical considerations and interpretations.

To compare the terms in a more productive way we utilize a system of weighting, a common illustration of the utilize of machine learning is the calculation of weights

based on word frequency. The TF.IDF estimation (Term Frequency, Inverse Document Frequency), commonly utilized in data recovery (IR), this measure is utilized to survey the significance of a term contained in a archive relative to a collection or a corpus.

The TF.IDF of a term i of a corpus is defined by:

$$TF.IDFi,j = TF.log_2 \frac{N}{n}$$

Where TF is the frequency of a term i in a document j;

N is the total number of documents of the corpus and n the number of documents in which the term i appears.

CONCLUSION

To process the text data, various algorithms need to be applied, there is a file a set of primary text filtering techniques and processes should be used of all the unnecessary and repetitive words, keep only informative text it is useful for the classification process regarding the subject we are studying, and the different stages of treatment.It is discussed in this chapter.

CHAPTER 04:

Implementation and Realization

INTRODUCTION

In order to find the best solutions to the problem that we talked about in our research, we have studied Surat Al-A'raf, analyzed it using C# language and applied it in Visual Studio 2017, as for studying the classification of verses at the expense of the stories of the prophets, we used Matlab, in this chapter we will explain the mechanism and tools necessary for this the work.

4.1. Development environment and tools

Choosing the right programming environment is extremely important to Project development. This is done according to several factors: Strength assembly, ease of use, availability of many features, Communication with the other environment...etc.

The tool we adopted is C# in Visual Studio 2017, we made this option to create the information matrix for this research and study it using Matlab.2009 because it contains a set of neural network tools and deep learning algorithms to classify.

4.1.1. Visual Studio 2017

The development environment used is visual studio because it has many strengths which are at the origin of its enormous success, the main ones being:

• An integrated development environment (IDE).

• In addition to C #, Visual Studio supports 36 different programming languages and allows the code editor and debugger to support (to varying degrees) almost any programming language, provided a specific service to a language exists. Built-in languages include C, C ++ / CLI, Visual Basic.NET, F #, JavaScript, TypeScript, XML, XSLT, HTML, and CSS. Support for other languages such as Python, Ruby, Node.js ..

• Visual Studio Tools allows you to test and debug applications by introducing error handling techniques as well as application monitoring concepts such as tracking, interacting with event logs, and using performance counters.

4.1.2. Language C#

We chose C# because it is an object-oriented programming language derived from C and C ++, similar to Java, easy to use, and used for developing web applications, as well as desktop applications, web services, commands, UI elements, or class libraries.

C# is a very powerful advanced language, yet it is characterized by complete ease in handling and very many uses and Compared to Java, C# C# generates files with more extensions than Java, such as exe, dll. That's why we worked it out in our research.

4.1.3. MATLAB

MATLAB is one of the high-performing languages and is generally used for the purpose of engineering calculations. Through integrated visualization and programming, MATLAB creates an easy-to-use environment in which math can be understood Symbols are used to express problems and solutions. This language is generally for the development of algorithms, calculations and mathematics. I decided to use Matlab because the framework itself should work on the desktop. Two Matlab toolkits are used: parallel computing and Neural Network Toolkit. The toolkit includes deep learning algorithms for the classification and functions of mission images. In order to speed up the learning of large datasets, you can decentralize computations and data to multicore processors and clusters using Parallel Computing Toolbox**[24].**

4.2. Implementation And Experimental Results

The database for this research consists of 206 verses of Surat Al-A'raf in the Holy Qur'an and data was divided into two text document sets (training and testing). The training set is (70%) Ayat (verses) and for testing set (30%) Ayat (verses). We choose only seven categories of the texts shown in the **table 4.1**.

TABLE 4.1: Categories Extracted From Surat Al-A'raf

Category Number	Category Name
00	No story found
01	Adam
02	Noah
03	Hud
04	Salih
05	Lut
06	Shoaib
07	Moses

This corpus contains 206 verses of Surat Al-A'raf in the Holy Qur'an written in using Buckwalter transliteration, grouped into 08 classes, the following table represents the different classes and the number of documents in each.

TABLE 4.2: Training Data Set Collection

Category	Total Documents
No story found	85
Adam	26
Noah	06
Hud	08
Saleh	07
Lut	05
Shoaib	09
Moses	60

4.2.1. Construction of the documents terms matrix

One of the basic text mining techniques for processing textual data is to convert each word in the text into a numerical value that represents the significance of the word in the set **[21].** We achieve this goal by building a matrix called the term

document matrix, where its rows are the extracted Qur'anic terms and its columns are the verses of the studied Qur'anic surah (Surat Al-A'raf**)[25].**

To construct this matrix, we follow these steps :

- Bag-of-words representation
- Dimensionality reduction(Removing Stop Words)
- Weighting

A. Bag-of-words representation

Converting a set of documents (verses of Sura Al-A'raf) into a data table This is done by extracting the sources of all the words of the studied surah, meaning that each word is deleted from it:

Stem prefixes:

• Remove Prefixes : **ال، وال، بال، كال، فال، لل، و**

Stem suffixes:

•Remove Suffixes : **ها، ان، ات، ون، ين، ية، ه، ي**

B. Dimensionality reduction (removing stop words)

Arabic dialect is exceptionally wealthy due to its lexicon and language structure. Consequently, a huge number of common and visit words existin Arabic writings andhave no critical semantic connection to the setting in which they happen and cannot be utilized alone to list and recognize records. Small considers were conducted to produce halt words for Arabic language **[22];[23]** .However, the coming about records of stopwords are not one or the other comprehensive nor exact sufficient since they don't consider vowelization in their approaches. In this work, we include an awfully imperative commitment by producing and approving physically a set of halt words for Arabic, vowelized Quran content. It incorporates relational words, conjunctions, things, and articles. Verbs with all their variations are not included since

Quran could be a sacrosanct content and verbs may hold vital semantics which we ought to not expel**.[19]**

examples of stop words that have the same letters but different vowelization and then different meanings:

لِمَا لَمَا لَمَا – ألّا أَلَا – أنّ أَنْ – بِكُمُ بِكمْ – أَفَمَنْ أَفَمِنْ

Moreover, each stop word is vowelized according to its position in the sentence such as these examples :

بَعْضُهُمْ ـبَعْضَهُ – بَعْضَهَا- بَعْضُهَا- بَعْضُكُمْ ـبَعْضَكُمْ – بَعْضِكُمْ ـبَعْضَ- بَعْضُ ـبَعْضِ ـبَعْضٍ - بَعْضًا

A sample of the extracted stop words from Arabic vowelized Quran is demonstrated in Figure :

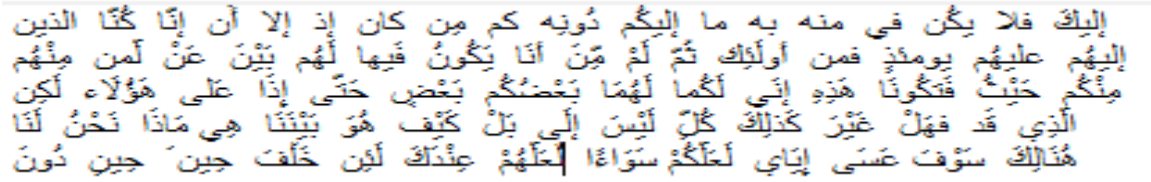
إِلَيكَ فلا يكُن في منه به ما إليكُم دُونِه كم مِن كان إذ إلا أن إنّا كُنّا الذين
إليهُم عليهُم يومئذٍ فمن أولئِك ثُمّ لَمْ مِّنَ أنا يَكُونُ فيها لَّهُم بَيْنَ عَنْ لّمن مِنْهُم
مِنْكُم حَيْثُ فَتَكُونَا هَذِهِ إِنِّي لَكُما لَهُمَا بَعْضَكُم بَعْضٍ حَتَّى إِذَا عَلَى هَؤُلَاء لَكِن
الَّذِي قَد فَهَلْ غَيْرَ كَذلِكَ كُلّ لَيْسَ إِلَى بَلْ كَيْف هُوَ بَيْنَنَا هِي مَاذَا نَحْنُ لَنَا
هُنَالِكَ سَوْفَ عَسَى إِيَاي لَعَلَّكُمْ سَوَاءً لَعَلَّهُمْ عِنْدَكَ لَئِن خَلْفَ حِينَ حِينٍ دُونَ

Figure 4.2: sample of vowelized stop words that exist in surah a'raf

C. Weighting

A specific weight in each document in the corpus. There is an efficient statistical method to calculate weights called Term Frequency Inverse Document Frequency (tf.idf) **[20]:**

$$TF.IDFi,j = TF.log_2 \frac{N}{n}$$

where **wi,j** is the weight of term iin document **j;**

tfi,j is the number of occurrences of term iin document **j**;

N is the total number of documents in the corpus; and **dfi** is the number of documents containing term **i**. A high weight in **tf.idf** is reached by a high term frequency in a given

document and a low document frequency of the term in the corpus; the weights hence tend to filter out common terms which are less discriminative.The result of this process is a matrix of **752** rows of unique words, namely terms, and **206** columns of verses from Surat Al-A'raf. The elements of this matrix are the calculated tf.idf weights of each term in a given verses.**[19]**

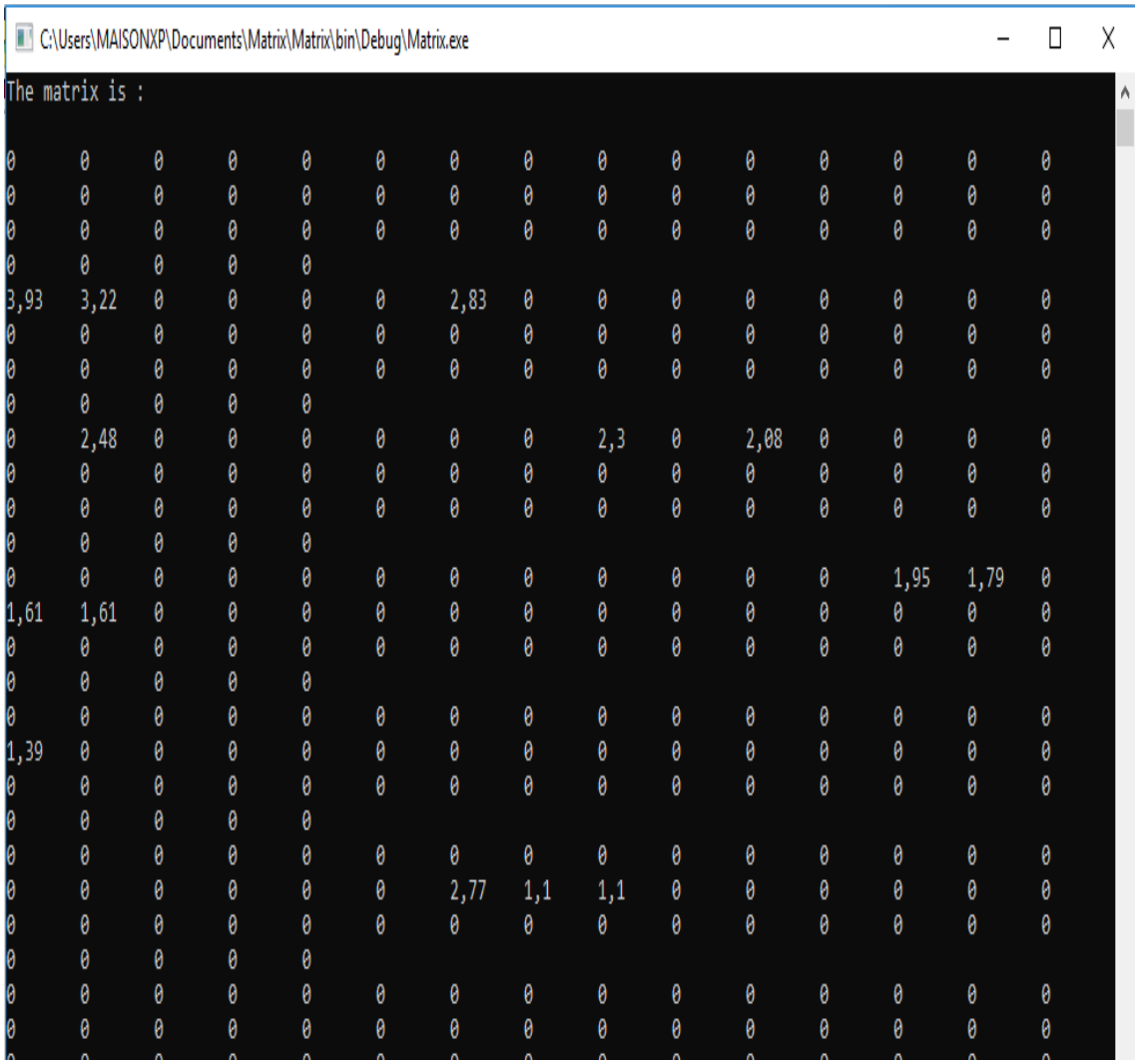

Figure 4.2: Preprocessing Results

4.3. ARTIFICIAL NEURAL NETWORKS:

Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements. We can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements.

Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output. Such a situation is shown below. There, the network is adjusted, based on a comparison of the output and the target, until the network output matches the target. Typically many such input/target pairs are used, in this supervised learning, to train a network**.[18]**

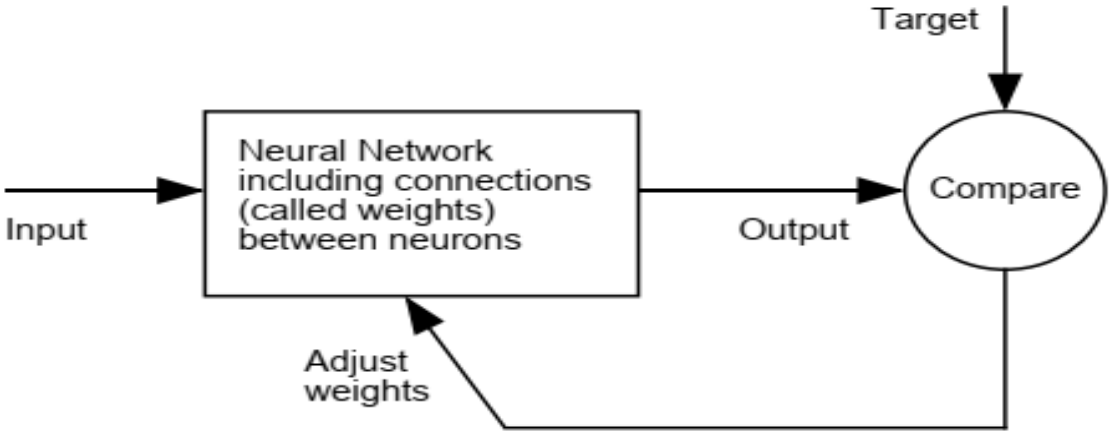

Figure 4.3: Structure of an Artificial Neural Network.

4.3.1. Typical Neural Network Design Process:

Each neural network application is unique, but network development typically follows the following steps:

- Access and preparation of your data
- Creation of the neural network
- Configuration of network inputs and outputs

- Adjustment of network parameters (weight and bias) to optimize performance
- Network learning
- Validation of network results
- Integration of the network into a production system

4.3.2. Artificial neural network architecture:

Ann is made of three layers namely input layer, output layer, and hidden layer/s. In our research, this input is the documents terms matrix that contains the weight of each word from Surat Al-A'raf(752 words), and each word has a vector component of its weight in each verse. As for the outputs, it is the group of categories that we want to classify the verses of Surat Al-A'raf, and they represent the prophets mentioned in this surah,number of hidden neurons 20 layers.

There must be a connection from the nodes in the input layer with the nodes in the hidden layer and from each hidden layer node with the nodes of the output layer. The input layer takes the data from the network. The hidden layer receives the raw information from the input layer and processes them. Then, the obtained value is sent to the output layer which will also process the information from the hidden layer and give the output. The interconnection of the nodes between the layers can be divided into two basic classes, namely the feedforward neural network and recurrent neural network. In the feedforward ann, the information movement from inputs to outputs is only in one direction. On the other hand, in the recurrent ann, some of the information moves in the opposite direction as well **[18].** Figure. 4.4 illustrates the feedforward ann and the recurrent ann architecture.

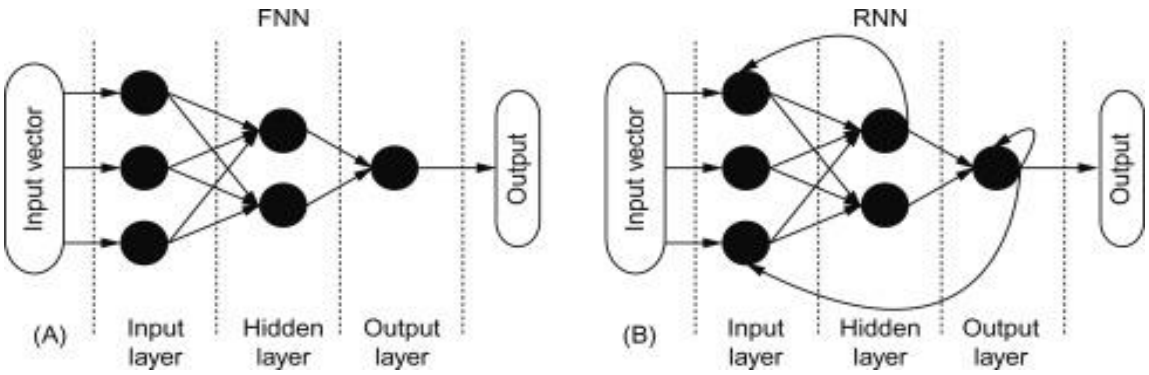

Figure. 4.4 the feedforward ann and the recurrent ann architecture.

4.3.3. Training Data Phase :

A system that maps an input to an output needs training to do this in a useful way. Just as people need to be trained to perform tasks, machine learning systems need to be trained. Training is accomplished by giving the system an input and the corresponding output and modifying the structure (models or data) in the learning machine . then the system should be able to produce correct outputs when new inputs are introduced.

4.3.4. Testing Data Phase:

A new document given the categorization model must predict the correct category label based on previous training.

All the pre-processing techniques are used, significant terms are obtained. TF*IDF calculations helps to create document vectors..

- **Precision:**

Accuracy is a measure of the system's ability to find valid documents. She gives the percentage of correct answers among the results obtained.

Precision i = the number of documents correctly / Attributes to class i

the number of documents classified by the system

It is theoretically possible to have a 100% reminder by returning the list of all documents from the base, but the precision will be poor and it will be difficult for the user to manage all of the results returned. If only one relevant result is returned, the precision is excellent. But the recall will be bad.

Accuracy testing and evaluation

We tried our program with 50 combinations of words in each verse

In the test group, we recorded the results in the following table:

True	**87 ,5%**	**52,9%**
False	12 ,5%	47,1%
Precision	87 ,5%	52,9%

Total Precision: (87,5+52,9)/2= 70,2%

- **Recall:**

Recall quantifies the number of positive class predictions made out of all positive examples in the dataset

Recall = TruePositives / (TruePositives + FalseNegatives)

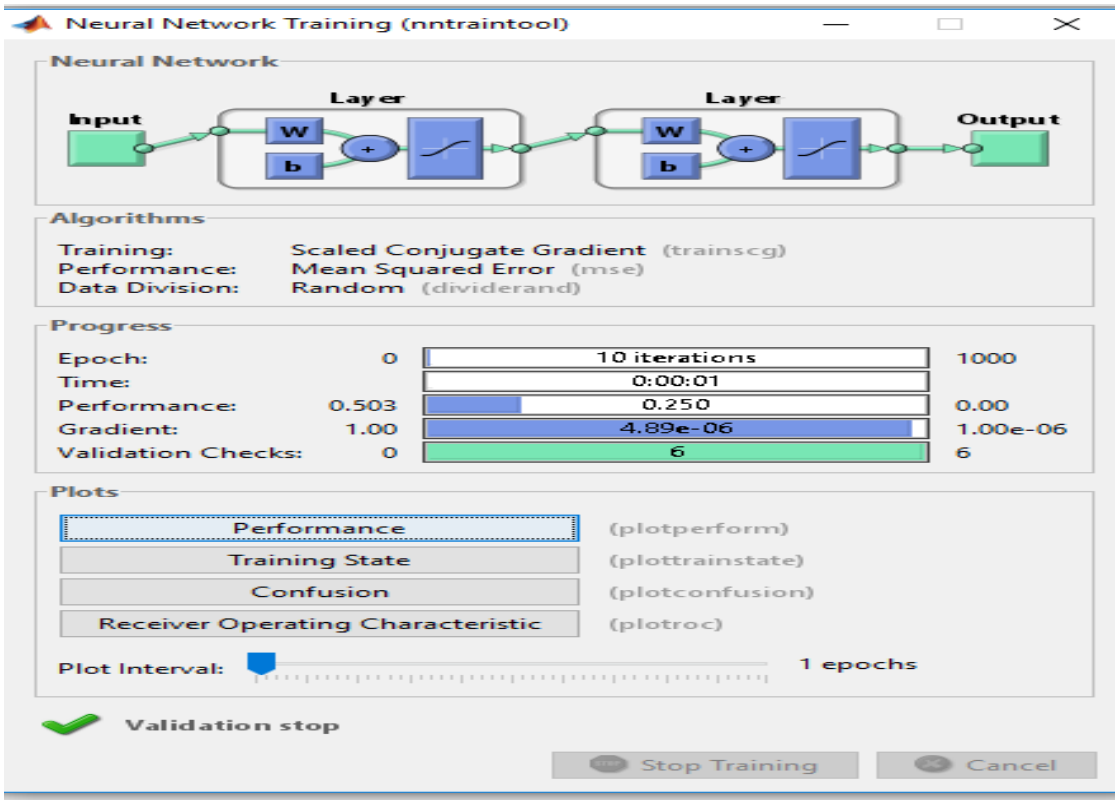

Figure 4.4: Neural Network Training

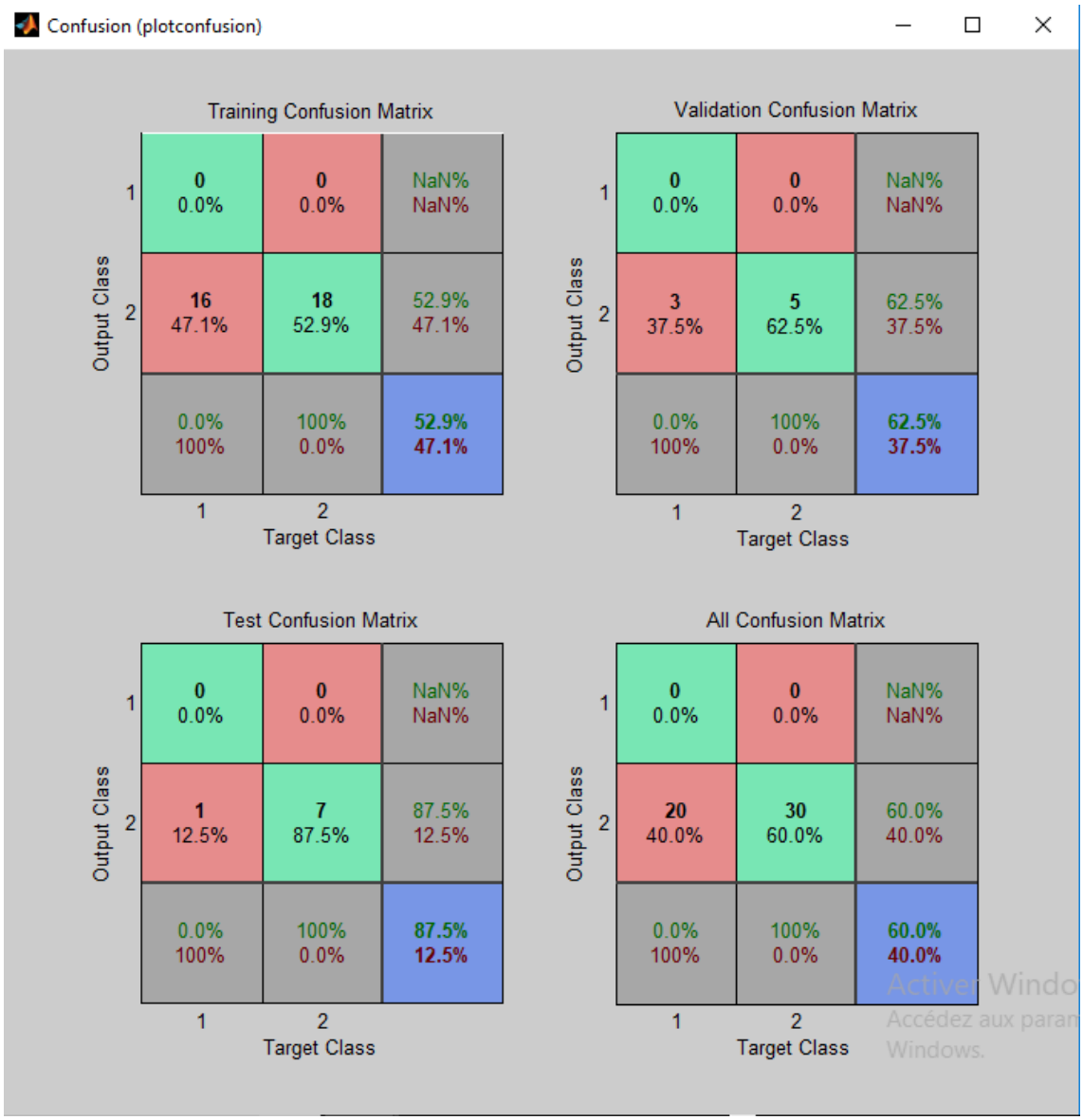

Figure 4.5: Confusion matrix

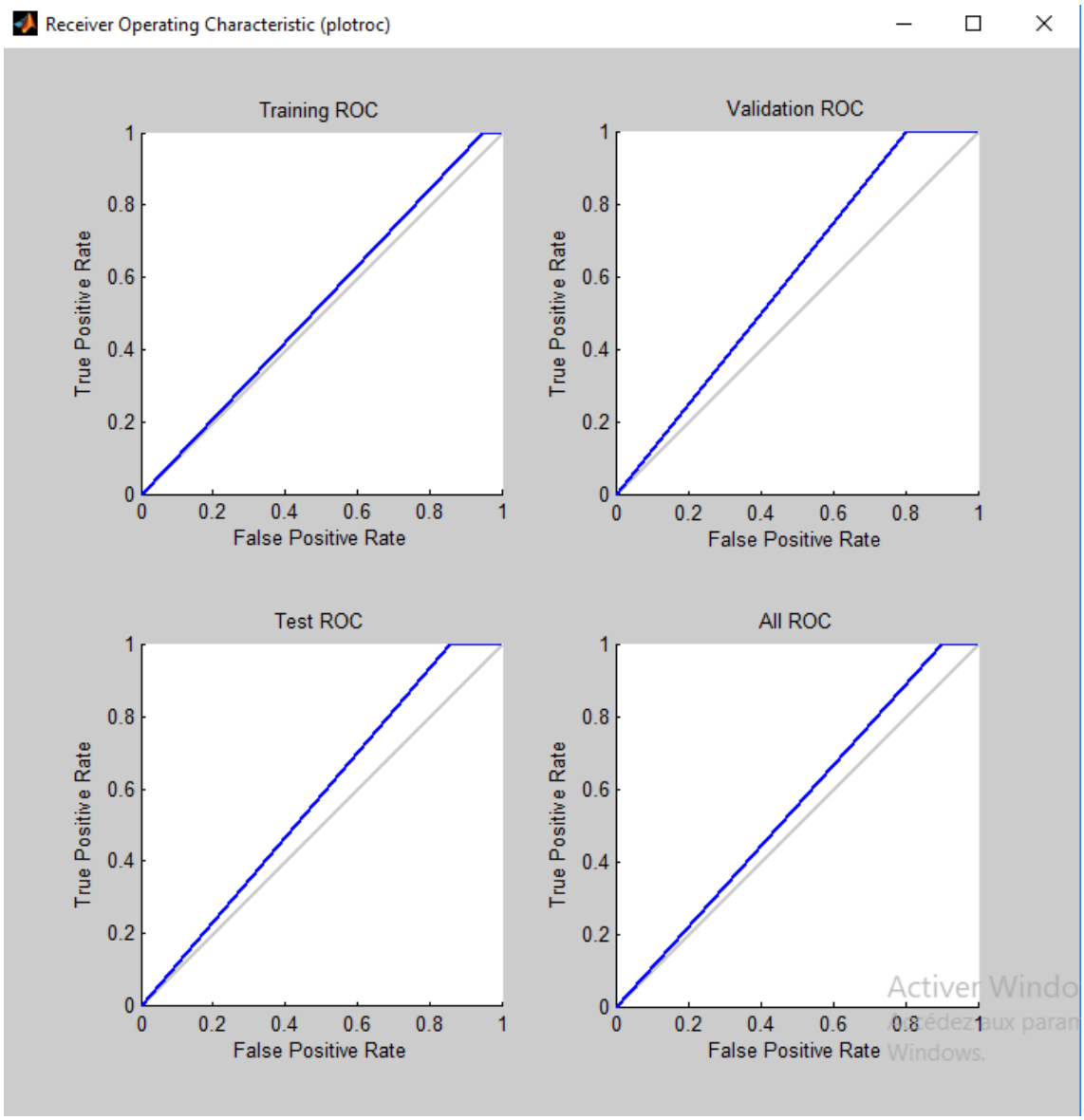

Figure 4.6: Receiver Operating Characteristic

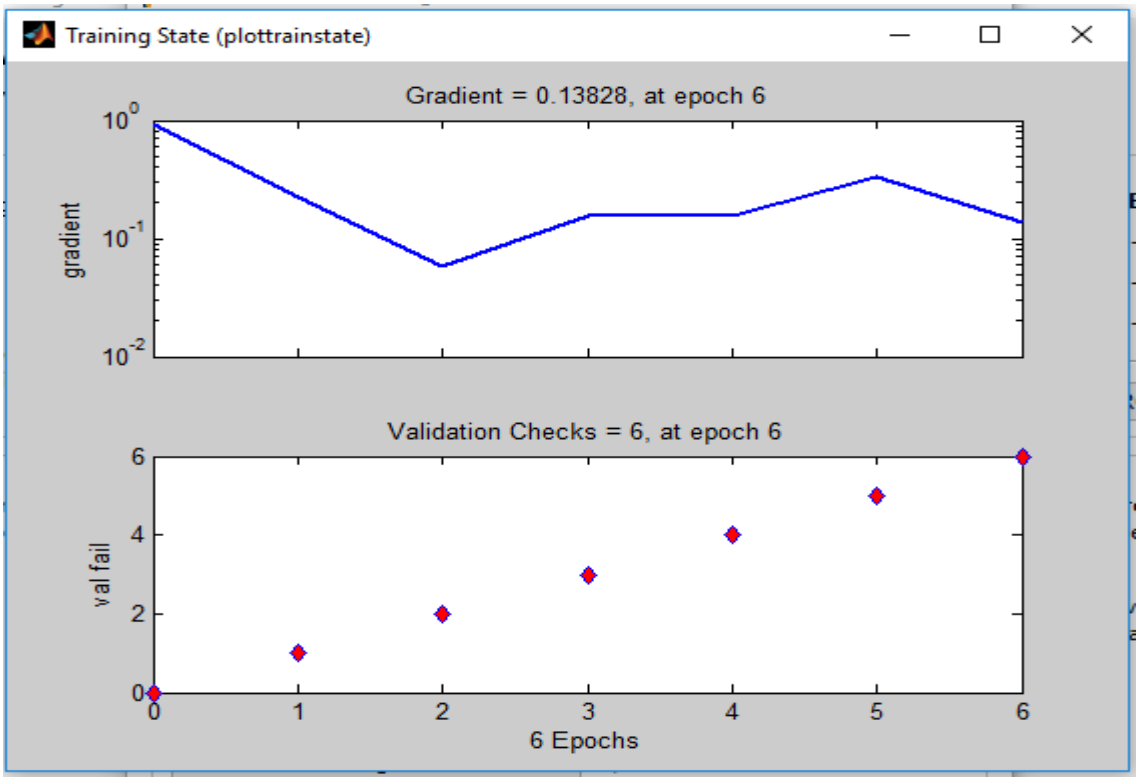

Figure 4.7: Training State

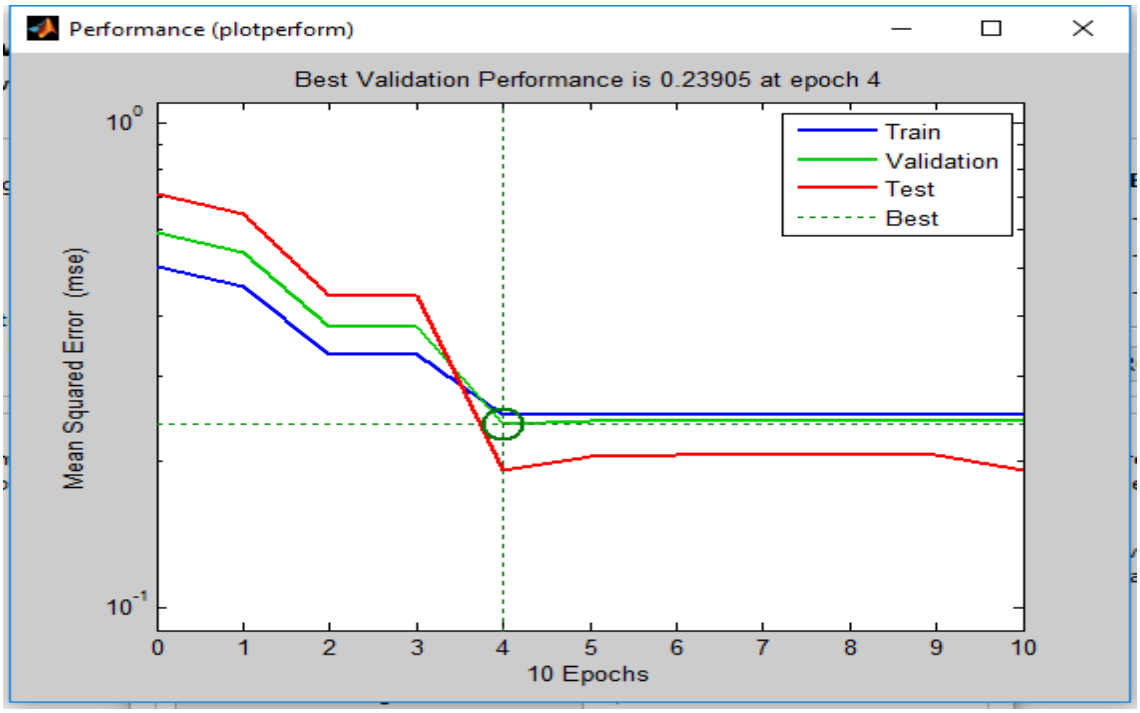

Figure 4.8: Performance

GENERAL CONCLUSION

GENERAL CONCLUSION

Since the Qur'an is the heavenly book for Muslims and is the most important source for understanding their religion, there are a huge number of data frameworks that draw on Arabic Qur'anic content to provide accurate and nearly comprehensive information about the Qur'an to the world. However, the development of such frameworks remains a challenging task due to the nature of Arabic authorship, the semantic uncertainty of words, the lack of origins and devices that support the Arabic language, and the religious nature of the Qur'an's content that needs careful excavation.

In this study, we describe the design and successful implementation of classifying texts (Quranic verses) from Surat Al-A'raf from the Holy Qur'an in the formulation of the stories of the prophets that we represented in 08 classes, using the NN algorithm and it was applied using Matlab. This paper deals with a limited number of text files (206 documents) in the training set and the test set.

for the prospects for this work. We suggest to improve this study the following points:

- Collect a larger group of Quranic verses or study the stories of the prophets in the entire Quran.

- Application of other classification algorithms (KNN, etc.).

- Combine unsupervised classification techniques (K-mean...) with classification techniques to achieve the best results.

- Improving classification performance using selection techniques

Features (Feature selection)

- Add more classifiers, more parameters and more functionality.
- Add the analysis using the Mixed class.

REFERENCES

[35] A. Nadwi ,Vocabulary of the Holy Quran. Millat Book Centre,2006.
[36] A. M. Omar, Dictionary of the Holy Quran. Noor Foundation International,2005.

[12] Al-Barez, H. Fadel, *Dialiktik alksah fi alns alkra'ni,* 2020.

[10] Al-Qaisi, A. Mohsen, *the development of the study of the concept of the Holy Qur'an,* 2018*.*

[29] A.Sharaf, E.Atwell, QurAna: Corpus of the Quran annotated with Pronominal Anaphora. The 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 130-137, Istanbul, Turkey, 2012.
[30] A.Sharaf, E.Atwell, QurSim: A corpus for evaluation of relatedness in short texts. The 8th International Conference on Language Resources and Evaluation (LREC 2012),pp. 2295-2302, Istanbul, Turkey, 2012.

[38] A. R. Siddiqui, Quranic Keywords: A Reference Guide. The Islamic Foundation,2008.

[23] B .Alhadidi, M. Alwedyan, Hybrid Stop-Word Removal Technique for Arabic. *Egyptian Computer Science Journal*, 2008, 35-38.

[18] Beale, H. Demuth, Neural Network Toolbox For Use with MATLAB, 2004, October ,

p. 6 / 846.

[14] C.Djazia, *Une plateforme orientée agent pour le data mining, En vue de l'obtention.* Université HADJ LAKHDAR – BATNA, 2009.

[5] C.OPREA, Performance Evaluation Of The Data Mining Classification, 2014.

[33] Corpus Quran, http://corpus.quran.com, 2021 06 02.
[34] E. Muhammad, From the Treasures of Arabic Morphology. Zam Zam Publishers,2007.
[20] G.Salton, M.McGill, Introduction to modern information retrieval. p. New York, 1983.

[22] I. El-Khair, Effect of Stop Words Elimination for Arabic Information Retrieval: A comparative Study. International journal of Computing & Information Sciences. pp. 119-133, 2006.

[32] I.Zeroual,A.Lakhouaja,Al-Mus'haf Corpus: A New Quranic Corpus rich in Morphosyntactical Information and accurate Part of Speech tagging, 2016 December 05.

[37] J.N. Rafai, Basic Quranic Arabic Grammar.Ta-Ha Publishers Ltd,1998.

[27] J. Dror, D. Shaharabani, R. Talmon , S. Wintner, Morphological Analysis of the Qur'an. Literary and Linguistic Computing, 19(4):431-452, 2004.

[28] K.Dukes, N. Habash, Morphological annotation of Quranic Arabic. The 7th International Conference on Language Resources and Evaluation (LREC), pp. 2530-2536, Valletta, Malta, 2010.

[39] K.Dukes, Arabic Corpus (morphology, version 0.4), 2011

[13] K. Shaalan, A. Hassanien, F.Tolba . *Intelligent Natural Language Processing: Trends and Applications.*

[2] L. Fatma, Classification des textes prophétiques. *Classification des textes prophétiques*. DEPARTEMENT D'INFORMATIQUE Université de M'sila, 2016.

[11] L.Rachid, *Basma Al-Nimri's story art .* Alaan Publishing ,2015.

[24] M.Chahrazed, Data classification using deep learning. Msila, mathematics and informatics- computre sience, 2018, 06 23.

[1] M.Goel, G. Shivani, Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity. *Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity*, p. 113, 2015, march 18.

[9] M.Hocine, classification automatique de textes Orienté Agent , faculté des sciences – algerie,2010-2011.

[7] M.Li, Modeling versus Clustering for Text Classification, 2019, June 24.

[15] M. Roux, *ALGORITHMES DE CLASSIFICATION.* Université Paul Cézanne ,Juin 2006.

[6] M.Sanan, M.Rammal , K. Zreik, L'accès Multilingue à l'information scientifique et technologique : limitations des moteurs de recherche en langue Arabe.

[31] M.Sawalha, C.Brierley, E. Atwell, Automatically generated, phonemic Arabic-IPA pronunciation tiers for the Boundary Annotated Qur'an Dataset for Machine Learning (version 2.0). LRE-REL2, 42, 2014.

[3] N.Pio, MercurioWeb SNC, F. Sebastiani,A. Sperduti, Discretizing Continuous Attributes in AdaBoost for Text Categorization. pp. 320-334, 2003.

[8] O. Choayb, Classification automatique de textes. *Classification automatique de textes*. Département d'Informatique UNIVERSITE DE M'SILA , 2014.

[19] R.BENTRCIA, Text Mining and Analytics for Extracting/ Discovering. *Doctorat 3ème cycle LMD*. Batna, Department of Computer Science, College of Math and Computer science, Batna 2 University, 2017, October.

[25] R.Bentrcia, S. Zidat, and F. Marir, Extracting Semantic Relations from the Quranic Arabic Based on Arabic Conjunctive Patterns, Journal of King Saud University - Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2017.09.004

[26] R.Bentrcia, S.Zidat, and F. Marir, An analytical study on the holy Quran based on the order of words in Arabic AND conjunction. The Malaysian Journal of Computer Science, vol. 31, no. 1, 2018.

[17] S.ABDELOUAHAB, Processus de classification supervisée de textes arabes par la méthode K PPV Application aux articles de presse, Mémoire de Master, Université de M'sila, 2011-2012.

[4] S.RÉHEL, *CATÉGORISATION AUTOMATIQUE DE TEXTES*, 2005, JANVIER.

[21] S.Weiss, N. Indurkhya, , T.Zhang, , F. Damerau, Text Mining Predictive Methods for Analyzing Unstructured Information. p. New York , 2005.

[16] W. Al-Harbi, A. Emam , EFFECT OF SAUDI DIALECT PREPROCESSING ON ARABIC SENTIMENT ANALYSIS. *EFFECT OF SAUDI DIALECT PREPROCESSING ON ARABIC SENTIMENT ANALYSIS*, 2015, December.