

# Text Analytics of the Qur'ān using Statistical and Large Language Models

by

**Al-Ahmadgaid Bahauddin Asaad**

Submitted to the *Institute of Islamic Studies*  
in partial fulfillment of the requirements for the degree of  
**Master of Arts in Islamic Studies**



**UNIVERSITY OF THE PHILIPPINES DILIMAN**  
Diliman, Quezon City

May 2025

## Dedication

إلى والدي أeralح وعلى شعب فلسطين

*To my late father and to the people of Palestine*

أَفَلَا يَتَدَبَّرُونَ الْقُرْآنَ وَلَوْكَانَ مِنْ عِنْدِغَيْرِ اللَّهِ لَوَجَدُوا فِيهِ أَخْتِلَافًا كَثِيرًا

*Will they not think about this Qur'an? If it had been  
from anyone other than God, they would have found  
much inconsistency in it.*

QUR'AN 4:82

*The servants of the Lord of Mercy are those who walk  
humbly on the earth, and who, when aggressive people  
address them, reply, with words of Peace.*

QUR'AN 25:63

## **Abstract**

The interest of the paper is to provide a comprehensive text analytics of the Qur'an. This is by utilizing Statistical and Machine Learning methods to computationally analyze the said scripture. Specifically, the following procedures have been done: descriptive analyses of the structure of the Qur'an, morphological analyses of the Qur'an. Further, the data used for the Qur'anic Arabic corpus is the one in QuranTree.jl by Asaad (2022). The computational software used is Julia programming language.

# Chapter 1

## Introduction

The use of scientific computing to studying the Qur'ān is still in its early stage in the fields of Islamic Studies and Statistical and Machine Learning applications. This study will indeed benefit not only researchers from Islamic Studies but also Statisticians and Machine Learning researchers who are into text analytics. Having said that, it is therefore necessary to provide context to audiences of these disciplines to provide background on the state of Qur'ānic studies and the increasing adoption of scientific methodology to studying humanities.

### 1.1 Background

The Qur'ān or *al-qur'ān* القرآن meaning *the recitation*, the holy book of Islam, is revered by 1.9 billion (according to 2020 projection of Cooperman et al. (2011, p. 13)) Muslims across the globe as the literal words of God. Muslims believed that the Qur'ān was gradually revealed (Qur'ān 25:32) to Prophet Muhammad ﷺ through angel ḡibrīl جبريل or Gabriel (Qur'ān 2:97). The Qur'ān contains 77,429 Arabic words in total, which covers only 56 percent of the Greek New Testament which has 138,020 words in total (Sinai, 2017, p. 11).

The Qur'ān is divided into *sūrahs* سور which are the equivalent of chapters, each containing *ayāt* آيات (meaning *signs*), which are the equivalent of verses. The *sūrahs* سور are not arranged in chronological order as in the Bible's books and chapters, but rather arranged in monotonically decreasing length of number of verses after the first *sūrah* سورة (see Figure 1.1). The *sūrah* سورة of the Qur'ān can be categorized into two types: the *makkīyya* مكية (Meccan) and *madaniyya* مدانية (Medinan). The categories refer to the geographical location of where the *sūrah* سورة was revealed. Figure 1.1 shows the grouping of the *sūrahs* سور. Note that some of the *sūrahs* سور have mixed geographical locations<sup>1</sup>, that is, a few of the *ayāt* آيات in it were revealed in

---

<sup>1</sup>see list of the location in [https://tanzil.net/docs/revelation\\_order](https://tanzil.net/docs/revelation_order)

other geographical location apart from the geographical location of the rest of the *ayāt* آيات. Therefore, the categorization in Figure 1.1 highlights the geographical location of the majority of the *ayāt* آيات in the *sūrah* سورة.

The Qur'ān (meaning *the recitation*) was revealed *orally* by angel *ġibrīl* جبريل to Prophet Muhammad ﷺ and passed onto other believers through oral tradition (reciting the Qur'ān to students repeatedly so as to memorize it, instead of writing it down and let the believers read it and memorize it). Memorizing 77,429 Arabic words of the Qur'ān through oral transmission can be a difficult task, but what aids this memorization is the rhythm feature of the Qur'ān. According to one Orientalist, Sinai (2017), "rhyme, however, or rather a periodically recurrent word-final assonance, is a feature of the Qur'ān throughout, and it naturally partitions the

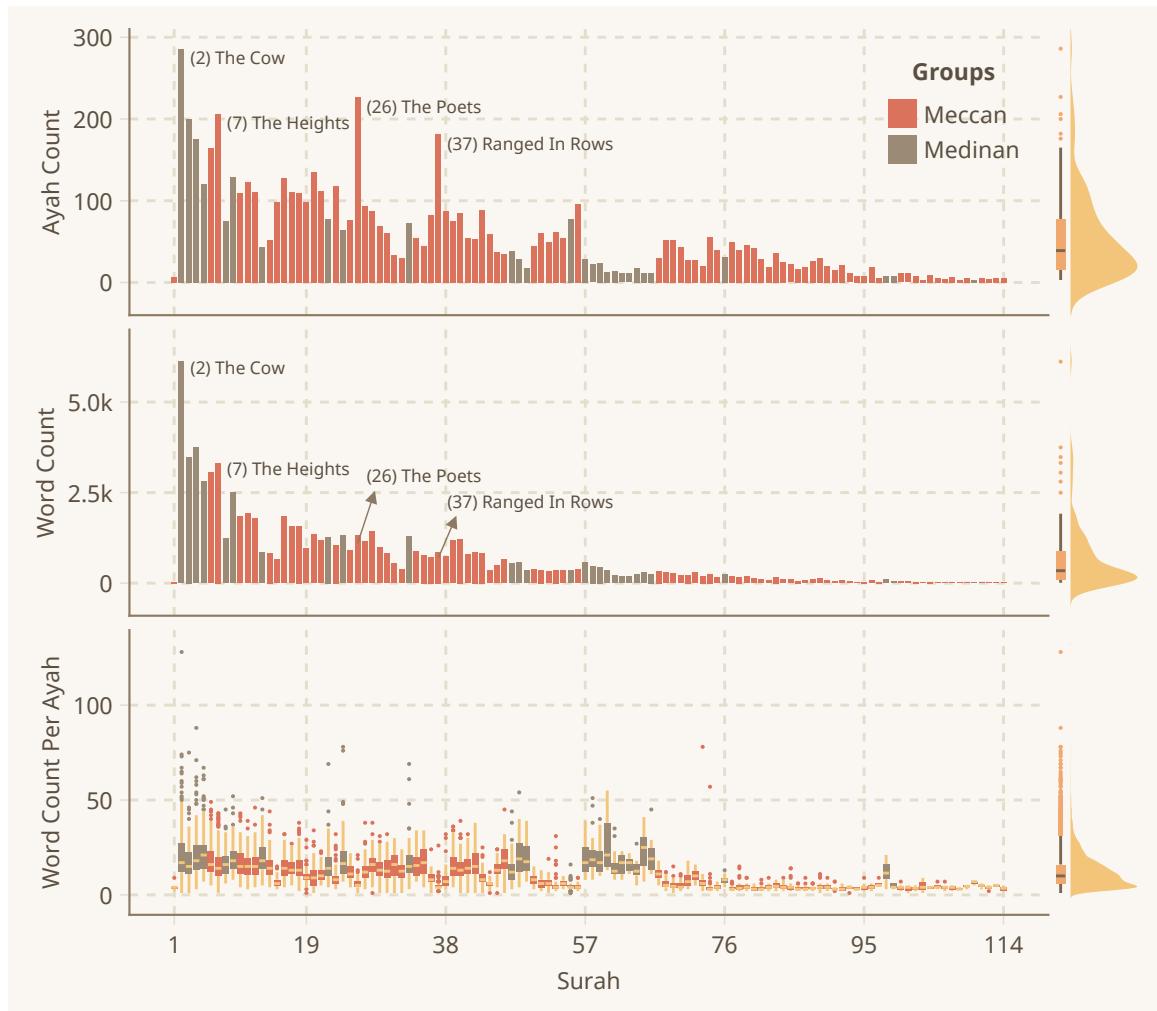


Figure 1.1: Statistics of the words and *ayāt* آيات (verses) of the Qur'ān

*sūrah* سُورَةٌ." Indeed, because of this feature, it makes it easy to memorize the entire Qur'ān, and the one who do so is called *hafiz* حَفِظْ meaning *one who remembers* or *keeper*. Qur'ān memorization contest is a common event in Muslim countries, the Philippine embassy has hosted one in 2022 in Saudi Arabia (Manila Bulletin, 2022).

According to the Muslim tradition, the oral transmission of passing the Qur'ān from a *hafiz* حَفِظْ to new believers was gradually put into writings as requested by the believers themselves. The idea was brought up after the battle of Yamama, where many of the Muslims who died were *qurrā'* قُرَّاءٌ (*the one who properly recite the Qur'ān*), and so fearing that their numbers will reduce in other battle fields, Umar ibn al-Khattab عمر بن الخطاب (who became the second caliph) suggested to the first caliph, Abū Bakr 'Abd Allāh ibn 'Abī Quhāfa, أبو بكر عبد الله بن أبي قحافة or short for Abū Bakr أبو بكر, to collect the Qur'ān into writing. Abū Bakr then authorized Zaid ibn Thabit زيد بن ثابت for the task. According to Zaid, he started collecting from the leafless stalks of the date-palm tree and from the pieces of leather and hides and from the stones, and from the chests of men (who had memorized the Qur'ān, i.e. the *hafiz* حَفِظْ)<sup>2</sup>. Long story short, the effort was finally codified by the third caliph, Uthman ibn Affan عُثْمَانُ بْنُ عَفَّانَ in the year 645 CE, which was then recopied and distributed to the different regional capitals of the early Islamic empire of that time. The rest of the copies outside this codification were then burned<sup>3</sup> down in order to have one standard Qur'ān. The Qur'ān nowadays is therefore assumed to be based on Uthmanic codex because of the story mentioned. That is, if indeed Uthman has ordered to burn other copies of the Qur'ān outside his codification, then what's left should only be based on his codex or archetype, and that should only be the inherited codex of the Muslims today.

The Qur'ān is believed by the Muslims to have been preserved since it was first recited by angel *gibril* جَبْرِيلْ or Gabriel to the Prophet ﷺ. Many orientalists had been skeptic about this claim, for example, John Wansbrough theorized that the Qur'ān was collected over a 200-year period (see Wansbrough, 2004, p. 101) after the death

<sup>2</sup>see <https://sunnah.com/bukhari:7191>

<sup>3</sup>see <https://sunnah.com/bukhari:4987>

of the Prophet ﷺ, instead of within a few years after the death of the Prophet ﷺ. However, recent findings through radiocarbon dating brings forward strong evidence of potential preservation of the whole Qur'ān, which the Muslims believed to be so. For example, the Birmingham Qur'ān manuscript discovered in 2015 is dated to be between 568 and 645 CE with 94.5% accuracy, making it among the oldest Qur'ānic manuscript in the world (see Birmingham University, 2015). Its predicted range of years intersects with the lifetime (570 to 632 CE) of the Prophet ﷺ. What is interesting is that the Birmingham Qur'ān is consistent with the Qur'ān today, word-by-word and letter-by-letter<sup>4</sup>, see Figure 1.2. This is indeed another evidence that the Qur'ān today was codified by Uthman since the discovery of the Birmingham Qur'ān manuscripts have confirm it. In addition to this, the Sana'a Palimpsest is also among the oldest Qur'ān radiocarbon dated to be between 578 CE and 669 CE with 95% accuracy (Sadeghi & Bergmann, 2010), which according to Sinai (2017), "neither does the edited portion of the Sana'a palimpsest offer evidence for additional or missing verses or for a divergent verse order within the *sūrahs* سور." Given these discoveries on the recent Quranic manuscripts, the claim of Wansbrough (2004, p. 101) is now untenable (see Sinai, 2014).

One likely reason as to why the scribes were able to preserve the Qur'ān in the two folios of the Birmingham Qur'ān manuscripts follows from the fact that the Qur'ān is firstly memorized before it was decided to be written. So that, the Topkapi manuscripts that contain 99% (only 23 verses missing)—dated around 701 to 750 CE (Karatay, 1962)<sup>5</sup>, that is, around 69 to 118 years after the death of the Prophet ﷺ—of the Qur'ān today was likely partly written from memory. This is possible since the Qur'ān possesses a rhythmic feature (that aids with memorization) that naturally divides its verses or *ayāt* آيات, and that the Qur'ān memorization competition is still held to this day (as in the example of Manila Bulletin, 2022), apart from

---

<sup>4</sup>The orthography of the Arabic letters in the early days had no diacritics and were written in its basic consonantal skeleton. That is, Arabic orthography and grammar were in their nascent when the Qur'ān was revealed, and therefore has to adjust and catch up, to capture and preserve the proper recitation of the Qur'an.

<sup>5</sup>see also <https://corpuscoranicum.de/en/manuscripts/1977/page/1-410?sura=1&verse=1>

the fact that it needs to be recited (any chapter after the first chapter of the Qur'ān according to the choice of the worshipper) every prayer from memory. Bottomline, there are many avenues for Qur'ān recitation from memory, and these have helped in its preservation. Moreover, since there are no significant evidence of insertion or malicious intention on addition or revision in all of the extant Qur'ānic manuscripts so far, some Orientalists came up with other theories of insertions on the basis of the literary style of the Qur'ān, see for example Sinai (2017, p. 92), where verse 102 of *sūra l-sāffāt* سُورَةِ الصَّافَاتِ or The Chapter of *Ranged in Rows* is theorized as addition because it is longer compared to other verses in the said chapter, refer to Sinai (2017, p. 92) for his other reasonings. Nonetheless, the Qur'ān is indeed stable based on the extant manuscripts.

Furthermore, the vastness of the early Islamic empire meant that different Mus-



Figure 1.2: 20th Century Qur'ān (left) in its fully featured orthographies vs Birmingham Qur'ān dated between 568 and 645 CE (right) in its basic consonantal skeleton. Image from Wikipedia (2015).

lim regional capitals have covered populations with different Arabic dialects, and so to accommodate these differences, Muslims believed that there were seven variant readings of the Uthmanic codex. Variant readings are defined as different pronunciations of the same word, in this case seven Uthamnic Qur'ān for seven different pronunciations. The *hadīt* حَدِيثٌ or *narration* comes from Ubayy ibn Ka'b who reported<sup>6</sup> that the Prophet ﷺ was near the tank of Banu Ghifar that ḡibrīl جِبْرِيلٌ or Gabriel came to him and said: "... Allah has commanded you to recite the Qur'ān to your people in *seven dialects*, and in whichever dialect they would recite, they would be right." Recent work of Sidky (2020) shows that the material evidence on the regional variants is in remarkable agreement with well-attested written variants documented in the traditional Muslim literature.

Muslim and non-Muslim scholars alike have been extremely interested in understanding the unique literary characteristics of the Qur'ān. As mentioned earlier, unlike other books like the Bible (arranged in chronological order), the Qur'ān does not follow any obvious organization. In addition to this, a *sūrah* سُورَةٌ does not fit the exact definition of a chapter. Indeed, the name attached to a *sūrah* سُورَةٌ is often decided as the unique entity mentioned in the said *sūrah* سُورَةٌ, its main purpose is to help early Muslims distinguish which *sūrah* سُورَةٌ they are talking about, this is contrary to the chapter name where the associated name is obviously the main topic of the chapter. Further, as described by Sinai (2017), "... the compositional unity of the long surahs located at the beginning of the corpus is anything but obvious: at least at first sight, they can appear a flit back and forth between different topics in a largely haphazard manner. This impression is not limited to Western readers: even pre-modern Muslim scholars have often approached their scripture as a quarry of unconnected verses and groups of verses that bear little intrinsic relation to what precedes and follows." It wasn't until Neuwirth (2007), that the compositional unity of the Qur'ān can be observed in tighter literary unities, as Neuwirth (2007) showed that the many of these texts display a tripartite structure and are often constructive

---

<sup>6</sup>source: <https://sunnah.com/muslim:821a>

around a narrative middle part (Sinai, 2017). Samples of the organizational style of the Qur’ān was shown in Sinai (2017, p. 88).

## 1.2 Rationale of the Study

Attempts at understanding the Qur’ān by Qur’ānic scholars were mostly done with the use of manual processes, that is, studying the scriptures by going through its content one-at-a-time manually. However, with the advent of computers, some researchers have started using it to aid in their study. The first known to have used computers for studying the Qur’ān was likely Rashad Khalifa in 1968<sup>7</sup>, where he studied the significance of the mysterious initials at the beginning of some *sūrāhs* سور. Rashad uploaded the Qur’ān into his computer by transliterating the Arabic letters and other Qur’ānic orthographies into Roman letters and symbols that the computer can easily parse. This approach of using computers to find new insights is more common in the field of science, and it was new to the field of Qur’ānic studies.

Indeed, to proceed with the use of scientific computing, the Qur’ān will be treated as the data that needs to be analyzed using what is called Natural Language Processing (NLP), a branch of Machine Learning (ML) that aims to understand natural languages, such as Arabic. To instruct the computer to do Statistical analyses, ML, or NLP, one needs to use a *software application* or a formal language called *programming language*. There are several programming languages that the computer can understand. The popular one for researchers in the field of sciences are Python, R, and sometimes Julia (Bezanson et al., 2017). These programming languages will be used to construct instructions for computer. Therefore, if the data is the Qur’ān, then there should be a way to interface with it using any of these programming languages. Or alternatively, there should be a way to upload it into the chosen programming language and encode the Arabic letters into something that can be easily parsed by the computer, like what Rashad Khalifa did. Having said that, there are indeed some programming languages with libraries or packages for

---

<sup>7</sup>see [https://www.masjidtucson.org/quran/miracle/a\\_profound\\_miracle\\_sura68nun133.html](https://www.masjidtucson.org/quran/miracle/a_profound_miracle_sura68nun133.html)

interfacing with the Qur’ān, and this is true for Python, R, and Julia. For this study, Julia programming language since its library for interfacing the Qur’ān using the QuranTree.jl<sup>8</sup> has more features compared to those available in R and Python (see Asaad, 2021). QuranTree.jl is based on Tanzil<sup>9</sup> for the Qur’ānic Arabic texts, and Dukes and Habash (2010) for morphological annotation, which both libraries from R and Python do not have in terms of morphological annotation from Dukes and Habash (2010).

Now that the programming language for this paper is identified, next is to understand how Statistics and Machine Learning can help in studying the Qur’ān. Statistics is a branch of Science that aims to study features or characteristics of data generated from a random phenomenon. The findings of Statistical analyses can then be used to make conclusions or predictions of the general population of the data or general characteristics of the data. Machine Learning or ML, on the other hand, is a branch of Artificial Intelligence that heavily intersects with Statistics, albeit with distinct differences as well. Both Statistics and ML aims to characterize data by learning its features, but ML research aims on complex models that are often inspired by simpler models from Statistics. Therefore, one can think of Statistics as one of the fundamentals of ML.

One of the goals of Statistics and Machine Learning is to summarize all of the learned features into a hypothesized equation or hypothesized mathematical formula whose parameters or weights are optimized to capture the most characteristics of the data. It is called hypothesized equation since the researcher is the one who decide which family of equation best describe the characteristics of the data. The said equation is called a *model*. Therefore, a model is a generic entity that can be fitted or molded to capture the features of the data. Mathematically and in general, a model can be written as:

$$\hat{y} := h(x|\theta), \quad (1.1)$$

---

<sup>8</sup>see <https://alstat.github.io/QuranTree.jl/stable/>

<sup>9</sup><https://tanzil.net/download/>

which is a mapping between the sets  $\mathcal{X}$  and  $\mathcal{Y}$  where  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$  through a hypothesized function  $h$ . The idea of modeling is to find the optimal value of  $\theta$  in such a way it will minimize the error between the actual data (represented by  $y$  below) and the predicted one  $\hat{y}$ :

$$\varepsilon := y - \hat{y} = y - h(x|\theta). \quad (1.2)$$

To help understand the concept of modeling, and relate it to the fashion industry, which the author assumes most readers are familiar with, a model in a fashion industry is responsible for representing the characteristics of the target customers (in Eq. 1.1 this is  $y$ ). Therefore, for a clothing company, they hire Asian models (in Eq. 1.1, this is  $\hat{y}$  or  $h(x|\theta)$ ) to target Asian customers (in Eq. 1.1 this is  $y$ ). So that, when these models wore the clothes sold by the said company, the potential customer will more or less be able to relate to the model, and be able to imagine themselves wearing that same clothes as well, which help them incline to buying the said clothing. The model, therefore, does not necessarily have the looks of every target Asian customers, but at least in terms of height, skin tone, hair, and other common Asian features, the model will likely have it, or at least the difference is more or less minimal (in Eq. 1.1, the difference is represented by  $\varepsilon := y - \hat{y}$ ). The question now is, what are the benefits that this model can bring to the clothing company? Well, the clothing company will be able to create products that are tailored to their Asian customers using the said model, since the company will have the right baseline measurements needed. Relating this analogy to the technical concept discussed earlier, you can think of the target customers as the real or actual data (in Eq. 1.1 this is  $y$ ), and the model as the same technical term use in Machine Learning and Statistics (in Eq. 1.1 this is  $h(x|\theta)$ ), but this time this technical model is expected to capture the characteristics of the real data analogous to fashion model that is expected to capture the characteristics of the target customer. This Statistical or Machine Learning model brings the following benefit: researchers will be able to study the real data by simply using the model to answer questions that are not

available in the sampled real data.

In this paper, the aim is to make use of the Statistical or Machine Learning modeling to understand the characteristics of the Qur'ān in terms of the morphologies of its canonical text in its modern orthography. More specifically, the following are the general objectives:

### 1.3 Objectives

The objective of the paper is to answer the following research questions:

1. What are the thematic themes of the surah that can be extracted by a Large Language Model?
2. What other insights that can be extracted by a Statistical and Machine Learning models in terms of the Qur'ānic morphology?
3. How does this combination of Statistical, Machine Learning, and Artificial Intelligence (AI) with the Muslim's traditional literatures help in understanding the Qurān, especially with the advent of Generative AI?

### 1.4 Significance of the Study

While the Qur'ān has been extensively studied by Muslims and non-Muslims scholars alike, especially in the topic of Meccan and Medinan surahs, there is still a lot to uncover from the perspective of Computational Statistics. Hence, the significance of this study is that it brings forward new ways of extracting insights from the Qur'ān by leveraging Computations, Statistics, Machine Learning, and AI, that is still in its early stage in the field of Qur'ānic Studies. Therefore, this new perspective or process of studying the scripture not only aids the scholars of the Islamic Studies, but may also contribute indirectly to community development and policy makers who use Qur'ān as part of their decision making.

### 1.5 Scope of the Study

The paper will cover all chapters of the Qur'ān both for the Morphological and Topic Modeling analyses, but it will only present the results of *sūrahs* سُورَاتٍ with at

least 1000 words. The rest of the result will be part of the web application that can be used to query the Qur'ān. It will also not delve into the *tafsir* تفسير of each of the verses, but only when necessary for further context.

### 1.6 Thesis Organization

The paper is organized as follows: Chapter 2 will discuss the related literatures, Chapter 3 will discuss the methodology, Chapter 4 will present the results and discussions, and finally Chapter 5 will contain the conclusion and recommendation. The references and appendices are placed after the Chapter 5.

## Chapter 2

### Literature Review

As mentioned in the previous chapter, the earliest work in Qur'ānic studies using computer was likely the work of Rashad Khalifa in 1968<sup>1</sup>, which led to one of his book entitled 'The Computer Speaks: God's Message to the World' (see Khalifa (1981)). While the work of Rashad started at studying the mystery letters in the beginning of some *sūrahs* سور (for example Qur'ān 2:1, 3:1, 7:1, etc.), it quickly went on to cover what he calls other *mathematical miracles*, all of which are covered in Khalifa (1981). His findings led him to generalize the claim that God revealed His words through this mathematical patterns throughout the Qur'ān, and that those verses that were off and did not conform to this discovered mathematical patterns led him to extensive investigation of the said verses, and concluded that those could be or surely be an insertion that should not have been in the Qur'ān in the first place. There are two verses that were off according to Rashad, and he called these verses as *false verses*<sup>2</sup>, these are the last two *ayāt* آيات of *sūra l-tāwba* سورة التوبة or The Chapter of *Repentance*. These two verses were removed in Rashad's Qur'ān translation<sup>3</sup>. Rashad believed so much on his findings that he self-proclaimed himself to be a messenger<sup>4</sup> with this new findings and that the Qur'ān nowadays should conform to his found mathematical patterns. This self-proclamation led to his assassination.

Fast forward to 20th century, among the pioneers to creating a stemmer system for the Qur'ān is the work of Thabet (2004). A stemmer system is a system for trimming inflected words into its basic form, which grammatically mean its the root form. For example, in English language the root word for *computational*, *computer*, *computation*, and *computerize* is *compute*. Therefore, from the root of the word forms different stems representing the different words. Hence, the idea of stemming is to

<sup>1</sup>[https://www.masjidtucson.org/quran/miracle/a\\_profound\\_miracle\\_sura68nun133.html](https://www.masjidtucson.org/quran/miracle/a_profound_miracle_sura68nun133.html)

<sup>2</sup><https://submission.org/App24.html>

<sup>3</sup><https://www.masjidtucson.org/quran/frames/>

<sup>4</sup>[https://www.masjidtucson.org/submission/faq/rashad\\_khalifa\\_summary.html](https://www.masjidtucson.org/submission/faq/rashad_khalifa_summary.html)

trim these words into its basic form, so that it would be easy to do word clustering or grouping through word similarity. According to Thabet (2004), the rich morphology of the Qur'ānic language or the Classical Arabic makes it even more difficult to do word stemming. Moving on, Thabet (2005) builds on this stemming system, and used it for tokenization of the Qur'ānic words. Tokenization is the process of listing all of the words in a sentence, and Thabet (2005) makes use of Thabet (2004) to further cleanse the noise from these tokens brought by the morphological variants of the Classical Arabic words. After cleansing the data, Thabet (2005) makes use of a statistical methodology for clustering or grouping the chapters of the Qur'ān, in particular the statistical methodology used is the Agglomerative Hierarchical Clustering based on the Euclidean distance of the adjusted word frequency of the *sūrah* سورة.

The work of Thabet (2004) and Thabet (2005) makes use of the Qur'ān corpus transliterated to Roman letters and symbols. Indeed, with interest growing on studying the Qur'ān from the lense of Data Analysis and Natural Language Processing, more work have been put in place into creating digital corpi of the Qur'ān that captures the different aspects of its linguistic styles. Hence, a series of work by Sharaf and Atwell led to the following publications: Sharaf and Atwell (2009) studied knowledge representation of the Qur'ānic verb valences using FrameNet frames, the output of which is a lexical database of the corpus of Qur'ān verbs. Further, the work of Sharaf and Atwell (2012a) came up with corpus for the annotations of the Qur'ānic pronouns, the authors named it as QurAna. Building on this work, Sharaf and Atwell (2012b) came up with a corpus for studying Qur'ānic relatedness based on the commentary of Ibn Kathir ابن كثیر, the authors named this corpus as QurSim.

Moving on, an unpublished work by Nassourou (2011) used a Support Vector Machine (SVM) to predict the classification of the place of revelation of the Qur'ānic *sūrahs* سور. The idea was to use *sūrahs* سور with well attested place of revalation as the training set, and then train an SVM to predict the remainings of the chapters.

Furthermore, the work of Shahzadi, ur Rahman, and Sawar (2012) developed a simple classifier based on the frequency distribution of words in *ayāt* آیات and the corresponding words in *sūrahs* سور.

Moving on, the work of

# Chapter 3

## Methodology

This chapter will discuss the methodologies that will be used to address the objectives of the paper. In particular, the chapter is organized as follows: Section 3.1 will discuss the Descriptive Statistics, Section 3.2 will discuss Probability distributions, Section 3.3 will discuss the concept of Histogram, Section 3.4 will discuss the concept of Topic Modeling, Section 3.5 will discuss the concept of Large Language Models, and finally Section 3.6 will discuss how to implement this in Julia programming language.

Further, the topics discussed here may be self-explanatory for Statisticians, ML researchers or those with Mathematical background. However, for the benefit of readers coming from humanities background, the paper will present the methodology as follows: mathematical formulas are formalized through Definition, Proposition, and Corollary for the sake of brevity, but immediate to these will be an explanation hoping to be simple enough for non-Statistician readers.

### 3.1 Descriptive Statistics

Among the basic statistical methodologies for summarizing information or data is what is known as Descriptive Statistics. From its name, these statistics are meant to convey simple description of the data. For example, *mean*, and *variance* are common statistics used for describing data. The formulas for these statistics are given in the following definitions:

**Definition 3.1.1** (Mean). Let  $x_i, i \in \{1, \dots, n\}$  where  $n \in \mathbb{N}$ , then the *mean* of  $x_i$ s is defined as follows:

$$\bar{x} = \sum_{i=1}^n x_i, \quad \text{where } x_i \in \mathbb{R}. \quad (3.1)$$

**Definition 3.1.2** (Variance). Let  $x_i, i \in \{1, \dots, n\}$  where  $n \in \mathbb{N}$  and let  $\bar{x}$  be the

mean defined in Defn. 3.1.1 then the *variance* of  $x_i$ s is defined as follows:

$$\text{Var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \text{where } x_i \in \mathbb{R}. \quad (3.2)$$

The mean is simply the average of the data points, while the variance is a single number that measures or summarizes the distances of the data points from the mean.

### 3.2 Probability Distribution

Among the basic tool to understanding patterns in the data is through the use of Probability distribution. Probability distribution is a model (*see* the discussion in Chapter 1 for understanding model). Therefore, Probability distribution is a mathematical formula design to characterize or describe the patterns a data point, mathematically represented as a variable (think of this like a placeholder of the data point of the said data). Modeling one variable makes it a univariate model or specifically univariate Probability distribution. Modeling more than one variable makes it a multivariate model or multivariate Probability distribution. Formally, the following are the definitions.

**Definition 3.2.1** (Probability Measure). Let  $\Omega$  be the *sample space*,  $\mathfrak{F}$  be the  $\sigma$ -algebra,  $\mathbb{P}$  be the probability measure, then the probability of a set  $\mathcal{A}, \mathcal{A} \in \mathfrak{F}$ , on the probability space  $(\Omega, \mathfrak{F}, \mathbb{P})$ , denoted by  $\mathbb{P}(\mathcal{A})$ , satisfies the following properties:

1. Non-negativity: For any set  $\mathcal{A}, \mathbb{P}(\mathcal{A}) \geq 0$
2. Normalization:  $\mathbb{P}(\Omega) = 1$
3. Countable additivity: For any sequence of disjoint sets  $\mathcal{A}_1, \mathcal{A}_2, \dots$  in  $\mathfrak{F}$ , such that the union  $\bigcup_{i \in \mathbb{N}} \mathcal{A}_i$  is also in  $\mathfrak{F}$ , we have:  $\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} \mathcal{A}_i\right) = \sum_{i \in \mathbb{N}} \mathbb{P}(\mathcal{A}_i)$

Basically, the idea of Defn. 3.2.1 is to define the concept of "measure." If someone is asked to measure something, the expectation is that that something satisfies the following assumptions: is that  $\mathbb{P}$  is a function that can be thought of like a machine. This machine only returns a positive number for every given collection of

data points. Further, this function or machine sums up to 1 if the whole collection of data points is measured. Finally, the probability measure of the union of the collection of these data points, is equal to the sum of the probability of each of the collection of the said data points.

**Definition 3.2.2** (Univariate Distribution).

### 3.3 Statistical Graphs

Graphs or plots are data visualization tools that are useful for exploratory data analysis apart from the Descriptive Statistics discussed above. It supplements the Descriptive Statistics findings through shapes visualized in the graphs. Among the popular statistical graphs is the bar graph. An example of this is given in Figure 1.1. Other graphs used are the box plots and the density plots which is also in Figure 1.1

#### 3.3.1 Box, Density, and Histogram Plots

While most statistical plots are easy to understand like bar graphs and scatter plots, others like Box, Density, and Histogram plots may not be easy to comprehend for someone with no Statistical background. This section will discuss how it is interpreted. Let's use Figure 1.1, Figure 3.1 for easy reference in this section.

As shown in Figure 3.1, both the Box and Density plots are tied to each other. This is indeed the case because both are describing the same information but presented in different style of visualization. In fact, Histogram is also used to describe the same information as the Box and Density, and the three are therefore related. So much so, that the three can be put into one graph. As to how to interpret these, readers are referred to for further details [to add reference].

### 3.4 Topic Modeling

As presented in Chapter 1, the first objective is to extract the thematic themes of *sūrahs* سُورَاتْ with at least 1000 words. In Statistics and Machine Learning, this task is called Topic Modeling. There are several ways to do this, but the popular methodology is to use the Latent Dirichlet Allocation (LDA) discussed in the next section.

### 3.4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a Statistical methodology that is based on Bayesian inference (Bayes, 1763; Laplace, 1986). It is a generative probabilisitic model for collection of discrete data such as text corpora (Blei, Ng, & Jordan, 2003). The idea of behind this methodology is to

## 3.5 Large Language Models

Generative Artificial Intelligence or GenAI for short has been making waves on its effectiveness to generate texts, images, audio, video, etc. It has elevated humanity to a new level of capability. However, behind this amazing capabilities is that GenAI is by design a mathematical formula that are called *model*. There are several types of

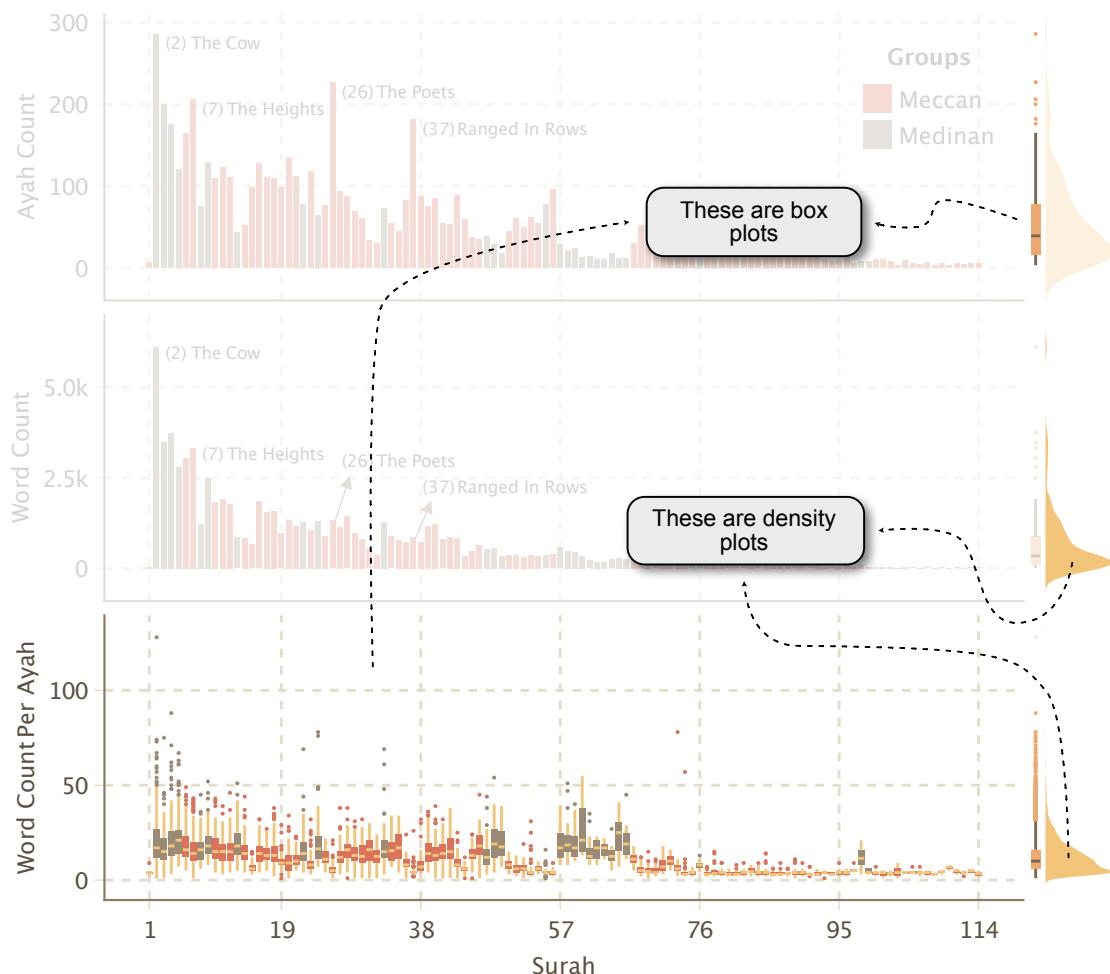


Figure 3.1: Statistics of the words and *ayāt* (verses) of the Qur'ān

*models*, and one of those is the Large Language Model (LLM). The following section will discuss what LLM is and its mathematical formulation.

### 3.5.1 Retrieval-Augmented Generation

The problem with LLM is that it was only trained on huge but limited data, and is therefore not able to infer what should be the context when asked.

### 3.6 Julia Code Setup

This section will discuss the coding setup

### 3.7 Web Application using SvelteKit and Julia

# **Chapter 4**

## **Results and Discussions**

The chapter is organized as follows: Section 4.4 will discuss the results for thematic analysis using Large Language Model with a comparison to one Statistical model that to emphasize the power of the LLM. Further, Section 4.5 will discuss the Information Retrieval result.

### **4.1 Descriptive Statistics**

In this section, the count statistics for the verses and words of the Qur'an. Figure 1.1 shows the count of *ayāt* آيات (verses) and word of the Qur'an.

#### **4.1.1 Verses**

### **4.2 Morphological Analysis**

### **4.3 Structural Analysis**

#### **4.3.1 Concentric Structure**

#### **4.3.2 Mathematical Structure**

### **4.4 Topic Modeling**

#### **4.4.1 Latent Dirichlet Allocation**

#### **4.4.2 OpenAI LLM**

### **4.5 Information Retrieval using RAG**

## **Chapter 5**

### **Conclusion and Recommendation**

## References

- Asaad, A.-A. B. (2021). QuranTree.jl: A Julia package for Quranic Arabic corpus. In *Proceedings of the sixth arabic natural language processing workshop* (pp. 208–212). Kyiv, Ukraine (Virtual): Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.wanlp-1.22>
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53, 370-418. Retrieved from <http://www.jstor.org/stable/105741>
- Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1), 65–98. Retrieved from <https://pubs.siam.org/doi/10.1137/141000671> doi: 10.1137/141000671
- Birmingham University. (2015). *Birmingham qur'an manuscript dated among the oldest in the world*. Birmingham University. (Available at: <https://www.birmingham.ac.uk/news-archive/2015/birmingham-quran-manuscript-dated-among-the-oldest-in-the-world> (Accessed: July 8th, 2023))
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
- Cooperman, A., O'Connell, E., & Stencel, S. (2011). *the future of the global muslim population* (Tech. Rep.). Pew Research Center.
- Dukes, K., & Habash, N. (2010, May). Morphological annotation of Quranic Arabic. In N. Calzolari et al. (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). Retrieved from [http://www.lrec-conf.org/proceedings/lrec2010/pdf/276\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/276_Paper.pdf)
- Karatay, F. E. (1962). *Topkapı sarayı müzesi kütüphanesi arapça yazmalar katalogu. kur'an, kur'an ilimleri, tefsirler no. 1 - 2171*. Küçükaydin Matbaası, İstanbul.
- Khalifa, R. (1981). *The computer speaks: God's message to the world*. Renaissance Production.

- Laplace, P. S. (1986, 08). Memoir on the probability of the causes of events. *Statist. Sci.*, 1(3), 364–378. Retrieved from <http://dx.doi.org/10.1214/ss/1177013621> doi: 10.1214/ss/1177013621
- Manila Bulletin. (2022). *PH embassy in riyadh hosts first asian qur'an memorization contest*. Manila Bulletin. (Available at: <https://mb.com.ph/2022/04/30/ph-embassy-in-riyadh-hosts-first-asian-quran-memorization-contest/>) (Accessed: July 8th, 2023))
- Nassourou, M. (2011). *Using machine learning algorithms for categorizing quranic chapters by major phases of prophet mohammad's messengership*.
- Neuwirth, A. (2007). *Studien zur komposition der mekkanischen suren*. Berlin, Boston: De Gruyter. Retrieved 2023-07-09, from <https://doi.org/10.1515/9783110920383> doi: doi:10.1515/9783110920383
- Sadeghi, B., & Bergmann, U. (2010). The codex of a companion of the prophet and the qurān of the prophet. *Arabica*, 57(4), 343 - 436. doi: \url{https://doi.org/10.1163/157005810X504518}
- Shahzadi, N., ur Rahman, A., & Sawar, M. J. (2012). Semantic network based classifier of holy quran. *International Journal of Computer Applications*, 39, 43-47. Retrieved from <https://api.semanticscholar.org/CorpusID:6539715>
- Sharaf, A., & Atwell, E. (2009). Knowledge representation of the quran through frame semantics: a corpus-based approach. In *Proceedings of the fifth corpus linguistics conference*. The Fifth Corpus Linguistics Conference. Retrieved from <https://api.semanticscholar.org/CorpusID:18278736>
- Sharaf, A., & Atwell, E. (2012a, May). QurAna: Corpus of the Quran annotated with pronominal anaphora. In N. Calzolari et al. (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 130–137). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from [http://www.lrec-conf.org/proceedings/lrec2012/pdf/123\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/123_Paper.pdf)
- Sharaf, A., & Atwell, E. (2012b, May). QurSim: A corpus for evaluation of

- relatedness in short texts. In N. Calzolari et al. (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 2295–2302). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from [http://www.lrec-conf.org/proceedings/lrec2012/pdf/190\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/190_Paper.pdf)
- Sidky, H. (2020). On the regionality of qur'anic codices. *Journal of International Quranic Studies Association*. doi: <http://dx.doi.org/10.5913/jiqsa.5.2020.a005>
- Sinai, N. (2014). When did the consonantal skeleton of the quran reach closure? part ii. *Bulletin of the School of Oriental and African Studies*, 77(3), 509–521. doi: 10.1017/S0041977X14000111
- Sinai, N. (2017). *The qur'an: A historical-critical introduction*. Edinburgh University Press Ltd.
- Thabet, N. (2004). Stemming the qur'an. In *Proceedings of the workshop on computational approaches to arabic script-based languages* (p. 85-88). USA: Association for Computational Linguistics.
- Thabet, N. (2005). Understanding the thematic structure of the qur'an: an exploratory multivariate approach. In *Proceedings of the acl student research workshop* (p. 7-12). USA: Association for Computational Linguistics.
- Wansbrough, J. (2004). *Quranic studies: Sources and methods of scriptural interpretation*. Prometheus Books.
- Wikipedia. (2015). *Comparison of a 20th-century edition of the quran (left) and the birmingham quran manuscript (right)*. Wikipedia. (Available at: [https://en.wikipedia.org/wiki/Birmingham\\_Quran\\_manuscript](https://en.wikipedia.org/wiki/Birmingham_Quran_manuscript) (Accessed: July 9th, 2023))