Text Analytics of the Qur'ān using Bayesian Statistics and Large Language Models

Al-Ahmadgaid B. Asaad
https://www.al-asaad.com/

Institute of Islamic Studies
University of the Philippines Diliman

June 2024

Background and Motivation

The Qur'ān

- The Qur'ān or *al-qurʾān* اَلقُرآن meaning *the recitation*, the holy book of Islam, is revered by 1.9 billion (according to 2020 projection of [3, 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 *ǧi-brīl* جِبرِيل 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 [18, p. 11]

Background and Motivation

The Qur'ān

- The Qur'ān or *al-qurʾān* اَلقُرآن meaning *the recitation*, the holy book of Islam, is revered by 1.9 billion (according to 2020 projection of [3, 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 *ǧi-brīl* جِبرِيل 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 [18, p. 11]

Background and Motivation

The Qur'ān

- The Qur'ān or *al-qurʾān* اَلقُرآن meaning *the recitation*, the holy book of Islam, is revered by 1.9 billion (according to 2020 projection of [3, 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 *ǧi-brīl* جِبرِيل 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 [18, p. 11]

Background and Motivation

The Qur'ān

- The Qur'ān is divided into $s\bar{u}war$ سُوَر (*plural* of $s\bar{u}rah$ سُورَة) which are the equivalent of chapters, each containing $^{\supset}\bar{a}y\bar{a}t$ آيَات (*plural* of $^{\supset}\bar{a}yah$ آيَة meaning *signs*), which are the equivalent of verses.
- The $s\bar{u}war$ سُوَر 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 $s\bar{u}rah$ سُورَة (*see* Figure 1).
- The $s\bar{u}war$ سُوَر of the Qur'ān can be categorized into two types: the $makk\bar{\imath}yah$ مَكِّيّة (Meccan) and $madan\bar{\imath}yah$ مَدَنِيّة (Medinan).

Background and Motivation

The Qur'ān

- The Qur'ān is divided into *sūwar* سُوَر (*plural* of *sūrah* سُورَة) which are the equivalent of chapters, each containing *ʾāyāt* آيَات (*plural* of *ʾāyah* آيَة meaning *signs*), which are the equivalent of verses.
- The *sūwar* سُوَر 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* سُورَة (*see* Figure 1).
- The *sūwar* سُوَر of the Qur'ān can be categorized into two types: the *makkīyah* مَكِّيّة (Meccan) and *madanīyah* مَدَنِيّة (Medinan).

Background and Motivation

The Qur'ān

- The Qur'ān is divided into $sūwar$ سُوَر (*plural* of $sūrah$ سُورَة) which are the equivalent of chapters, each containing $ʾāyāt$ آيَات (*plural* of $ʾāyah$ آيَة meaning *signs*), which are the equivalent of verses.
- The $sūwar$ سُوَر 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 $sūrah$ سُورَة (*see* Figure 1).
- The $sūwar$ سُوَر of the Qur'ān can be categorized into two types: the $makkīyah$ مَكِّيّة (Meccan) and $madanīyah$ مَدَنِيّة (Medinan).

Background and Motivation

The Qur'ān

- The categories refer to the geographical location of where the *sūrah* سُورَة was revealed to Prophet Muhammad ﷺ. Figure 1 shows the groupings of the *sūwar* سُوَر.
- Note that some of the *sūwar* سُوَر have mixed geographical locations[a], that is, a few of the *ʾāyāt* آيَات in it were revealed in other geographical location apart from the geographical location of the rest of the *ʾāyāt* آيَات.
- Therefore, the categorization in Figure 1 highlights the geographical location of the majority of the *ʾāyāt* آيَات in the *sūrah* سُورَة.

[a] *see* list of the location in `https://tanzil.net/docs/revelation_order`

Background and Motivation

The Qur'ān

- The categories refer to the geographical location of where the $s\bar{u}rah$ سُورَة was revealed to Prophet Muhammad ﷺ. Figure 1 shows the groupings of the $s\bar{u}war$ سُوَر.
- Note that some of the $s\bar{u}war$ سُوَر have mixed geographical locations[a], that is, a few of the $^{\flat}\bar{a}y\bar{a}t$ آيَات in it were revealed in other geographical location apart from the geographical location of the rest of the $^{\flat}\bar{a}y\bar{a}t$ آيَات.
- Therefore, the categorization in Figure 1 highlights the geographical location of the majority of the $^{\flat}\bar{a}y\bar{a}t$ آيَات in the $s\bar{u}rah$ سُورَة.

[a] *see* list of the location in `https://tanzil.net/docs/revelation_order`

Background and Motivation

The Qur'ān

- The categories refer to the geographical location of where the *sūrah* سُورَة was revealed to Prophet Muhammad ﷺ. Figure 1 shows the groupings of the *sūwar* سُوَر.
- Note that some of the *sūwar* سُوَر have mixed geographical locations[a], that is, a few of the *ʾāyāt* آيَات in it were revealed in other geographical location apart from the geographical location of the rest of the *ʾāyāt* آيَات.
- Therefore, the categorization in Figure 1 highlights the geographical location of the majority of the *ʾāyāt* آيَات in the *sūrah* سُورَة.

[a] *see* list of the location in `https://tanzil.net/docs/revelation_order`

Background and Motivation

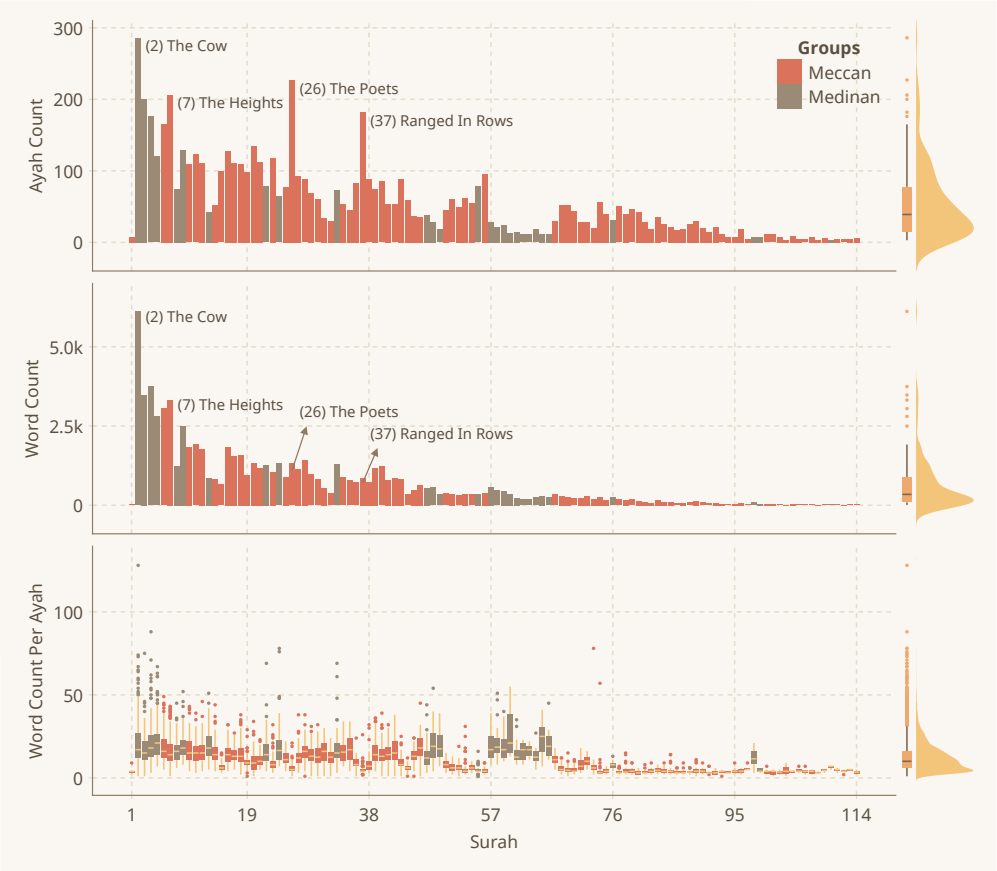

Figure: Statistics of the words and ʾāyāt آيَات (verses) of the Qur'ān

Background and Motivation

The Qur'ān

- Attempts at understanding the Qur'ān by Qur'ānic scholars were mostly done with the use of manual processes;
- 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[a], where he studied the significance of the mysterious initials at the beginning of some *sūwar* سُوَر.

[a] https://www.masjidtucson.org/quran/miracle/a_profound_miracle_sura68nun133.html

Background and Motivation

The Qur'ān

- Attempts at understanding the Qur'ān by Qur'ānic scholars were mostly done with the use of manual processes;
- 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[a], where he studied the significance of the mysterious initials at the beginning of some *sūwar* سُوَر.

[a] https://www.masjidtucson.org/quran/miracle/a_profound_miracle_sura68nun133.html

Background and Motivation

The Qur'ān

- Attempts at understanding the Qur'ān by Qur'ānic scholars were mostly done with the use of manual processes;
- 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[a], where he studied the significance of the mysterious initials at the beginning of some *sūwar* سُوَر.

[a] https://www.masjidtucson.org/quran/miracle/a_profound_miracle_sura68nun133.html

Background and Motivation

The Qur'ān

- 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 a scientific process called Natural Language Processing (NLP), a branch of Machine Learning (ML) that aims to understand natural languages, such as Arabic.

Background and Motivation

The Qur'ān

- 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 a scientific process called Natural Language Processing (NLP), a branch of Machine Learning (ML) that aims to understand natural languages, such as Arabic.

Background and Motivation

The Qur'ān

- To instruct the computer to do Statistical analyses or ML, one needs to use a *software application* or a formal[a] language called *programming language*.
- The popular one for researchers in the field of sciences are Python [21], R [12], and sometimes Julia [2].
- Therefore, if the data is the Qur'ān, then there should be a way to interface with it using any of these programming languages. There are indeed some programming languages with libraries or packages for interfacing with the Qur'ān, and this is true for Python, R, and Julia.

[a] "formal" because these languages were invented by man for particular purpose, in this case to communicate with computers

Background and Motivation

The Qur'ān

- To instruct the computer to do Statistical analyses or ML, one needs to use a *software application* or a formal[a] language called *programming language*.
- The popular one for researchers in the field of sciences are Python [21], R [12], and sometimes Julia [2].
- Therefore, if the data is the Qur'ān, then there should be a way to interface with it using any of these programming languages. There are indeed some programming languages with libraries or packages for interfacing with the Qur'ān, and this is true for Python, R, and Julia.

[a] "formal" because these languages were invented by man for particular purpose, in this case to communicate with computers

Background and Motivation

The Qur'ān

- To instruct the computer to do Statistical analyses or ML, one needs to use a *software application* or a formal[a] language called *programming language*.
- The popular one for researchers in the field of sciences are Python [21], R [12], and sometimes Julia [2].
- Therefore, if the data is the Qur'ān, then there should be a way to interface with it using any of these programming languages. There are indeed some programming languages with libraries or packages for interfacing with the Qur'ān, and this is true for Python, R, and Julia.

[a] "formal" because these languages were invented by man for particular purpose, in this case to communicate with computers

Background and Motivation

The Qur'ān

- For this study, the three programming languages will be used.
- The ruling is that Julia will be used for interfacing with the Qur'ān texts since its library for it has more features [1] compared to R and Python.
- The said Julia library is the QuranTree.jl[a]. QuranTree.jl is based on Tanzil[b] for the Qur'ānic Arabic texts, and [8] for morphological annotation, which both libraries from R and Python do not have in terms of morphological annotations from [8].

[a] https://alstat.github.io/QuranTree.jl/stable/
[b] https://tanzil.net/download/

Background and Motivation

The Qur'ān

- For this study, the three programming languages will be used.
- The ruling is that Julia will be used for interfacing with the Qur'ān texts since its library for it has more features [1] compared to R and Python.
- The said Julia library is the QuranTree.jl[a]. QuranTree.jl is based on Tanzil[b] for the Qur'ānic Arabic texts, and [8] for morphological annotation, which both libraries from R and Python do not have in terms of morphological annotations from [8].

[a] https://alstat.github.io/QuranTree.jl/stable/
[b] https://tanzil.net/download/

Background and Motivation

The Qur'ān

- For this study, the three programming languages will be used.
- The ruling is that Julia will be used for interfacing with the Qur'ān texts since its library for it has more features [1] compared to R and Python.
- The said Julia library is the QuranTree.jl[a]. QuranTree.jl is based on Tanzil[b] for the Qur'ānic Arabic texts, and [8] for morphological annotation, which both libraries from R and Python do not have in terms of morphological annotations from [8].

[a] `https://alstat.github.io/QuranTree.jl/stable/`
[b] `https://tanzil.net/download/`

Background and Motivation

The Qur'ān

- On the other hand, both R and Python will be used for libraries that are not available in Julia. In particular, R is known for niche statistical libraries since it was made for statistical computation, whereas Python is now popular for Deep Learning frameworks for complex modeling like TensorFlow[a] (a library made by Google[b]) and PyTorch[c] (a library made by Meta[d]).

[a] https://www.tensorflow.org/
[b] https://research.google/
[c] https://pytorch.org/
[d] https://ai.meta.com/

Background and Motivation

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 decisions, 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 researchers have been aiming 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.

Background and Motivation

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 decisions, 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 researchers have been aiming 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.

Objectives

The following are the general and specific objectives of this paper:

1. What are the structural characteristics of the Qur'an that can be extracted from its rich morphologies using statistical and large language models?
    1. What are the statistics of the Qur'ān's morphological features in terms of its parts of speech and selected entities like God's name and the prophets names mentioned?
    2. How do the rhythmic signatures of the Qur'ān of the verses looks like and what are statistical insights that can be extracted?
    3. What are the rhythmic signatures of the *ʾāyāt* آيَات of *makkīyah* مَكِّيَّة and *madanīyah* مَدَنِيَّة *sūwar* سُوَر?

Objectives

The following are the general and specific objectives of this paper:

1. What are the structural characteristics of the Qur'an that can be extracted from its rich morphologies using statistical and large language models?
    1. What are the statistics of the Qur'ān's morphological features in terms of its parts of speech and selected entities like God's name and the prophets names mentioned?
    2. How do the rhythmic signatures of the Qur'ān of the verses looks like and what are statistical insights that can be extracted?
    3. What are the rhythmic signatures of the *ʾāyāt* آيَات of *makkīyah* مَكِّيَّة and *madanīyah* مَدَنِيَّة *sūwar* سُوَر?

Objectives

The following are the general and specific objectives of this paper:

1. What are the structural characteristics of the Qur'an that can be extracted from its rich morphologies using statistical and large language models?
    1. What are the statistics of the Qur'ān's morphological features in terms of its parts of speech and selected entities like God's name and the prophets names mentioned?
    2. How do the rhythmic signatures of the Qur'ān of the verses looks like and what are statistical insights that can be extracted?
    3. What are the rhythmic signatures of the *ʾāyāt* آيَات of *makkīyah* مَكِّيّة and *madanīyah* مَدَنِيّة *sūwar* سُوَر?

Objectives

The following are the general and specific objectives of this paper:

1. What are the structural characteristics of the Qur'an that can be extracted from its rich morphologies using statistical and large language models?
    1. What are the statistics of the Qur'ān's morphological features in terms of its parts of speech and selected entities like God's name and the prophets names mentioned?
    2. How do the rhythmic signatures of the Qur'ān of the verses looks like and what are statistical insights that can be extracted?
    3. What are the rhythmic signatures of the $^{\supset}āyāt$ آيَات of $makkīyah$ مَكِّيَّة and $madanīyah$ مَدَنِيَّة $sūwar$ سُوَر?

Objectives

2. What other insights that can be extracted from the semantics of the Qur'an's texts using statistical and large language models?
    1. How does the theory of *concentrism* be formulated statistically, and what are the insights from the statistical and large language models on this?
    2. How do the $s\bar{u}war$ سُوَر are organized in terms of the topics? What are the themes that can be extracted for each of the surahs?
    3. How do these extracted themes compare to the summaries of Abdel Haleem's English translation of the Qur'an?
3. How does these combinations of statistical, machine learning, and artificial intelligence with the Muslim's traditional literatures help in understanding the Qur'ān, especially with the advent of Generative AI?

Objectives

2. What other insights that can be extracted from the semantics of the Qur'an's texts using statistical and large language models?
    1. How does the theory of *concentrism* be formulated statistically, and what are the insights from the statistical and large language models on this?
    2. How do the *sūwar* سُوَر are organized in terms of the topics? What are the themes that can be extracted for each of the surahs?
    3. How do these extracted themes compare to the summaries of Abdel Haleem's English translation of the Qur'an?
3. How does these combinations of statistical, machine learning, and artificial intelligence with the Muslim's traditional literatures help in understanding the Qur'ān, especially with the advent of Generative AI?

Objectives

2. What other insights that can be extracted from the semantics of the Qur'an's texts using statistical and large language models?
    1. How does the theory of *concentrism* be formulated statistically, and what are the insights from the statistical and large language models on this?
    2. How do the $s\bar{u}war$ سُوَر are organized in terms of the topics? What are the themes that can be extracted for each of the surahs?
    3. How do these extracted themes compare to the summaries of Abdel Haleem's English translation of the Qur'an?
3. How does these combinations of statistical, machine learning, and artificial intelligence with the Muslim's traditional literatures help in understanding the Qur'ān, especially with the advent of Generative AI?

Objectives

2. What other insights that can be extracted from the semantics of the Qur'an's texts using statistical and large language models?
    1. How does the theory of *concentrism* be formulated statistically, and what are the insights from the statistical and large language models on this?
    2. How do the $s\bar{u}war$ سُوَر are organized in terms of the topics? What are the themes that can be extracted for each of the surahs?
    3. How do these extracted themes compare to the summaries of Abdel
    Haleem's English translation of the Qur'an?
3. How does these combinations of statistical, machine learning, and artificial intelligence with the Muslim's traditional literatures help in understanding the Qur'ān, especially with the advent of Generative AI?

Objectives

2. What other insights that can be extracted from the semantics of the Qur'an's texts using statistical and large language models?
    1. How does the theory of *concentrism* be formulated statistically, and what are the insights from the statistical and large language models on this?
    2. How do the $s\bar{u}war$ سُوَر are organized in terms of the topics? What are the themes that can be extracted for each of the surahs?
    3. How do these extracted themes compare to the summaries of Abdel
    Haleem's English translation of the Qur'an?
3. How does these combinations of statistical, machine learning, and artificial intelligence with the Muslim's traditional literatures help in understanding the Qur'ān, especially with the advent of Generative AI?

Significance of the Study

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.
- This is especially true for Islamic Studies researcher, which the author hopes to benefit and get interested in Islamicate Digital Humanities, a new field which aims to take advantage of the scientific computations for studying Islamic texts, which the author hopes to have in any Islamic institute.
- Further, since this paper combines the several fields (Islamic Studies, Statistics, and Machine Learning), the results will also contain mathematical theories that is hoped to advance the field of Statistics and Machine Learning as well.

Significance of the Study

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.
- This is especially true for Islamic Studies researcher, which the author hopes to benefit and get interested in Islamicate Digital Humanities, a new field which aims to take advantage of the scientific computations for studying Islamic texts, which the author hopes to have in any Islamic institute.
- Further, since this paper combines the several fields (Islamic Studies, Statistics, and Machine Learning), the results will also contain mathematical theories that is hoped to advance the field of Statistics and Machine Learning as well.

Significance of the Study

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.
- This is especially true for Islamic Studies researcher, which the author hopes to benefit and get interested in Islamicate Digital Humanities, a new field which aims to take advantage of the scientific computations for studying Islamic texts, which the author hopes to have in any Islamic institute.
- Further, since this paper combines the several fields (Islamic Studies, Statistics, and Machine Learning), the results will also contain mathematical theories that is hoped to advance the field of Statistics and Machine Learning as well.

Significance of the Study

The significance of this study is that:

- With that said, the author would like to also emphasize the opportunities that scientific methodologies can bring to studying Islamic studies, especially for Muslim researchers who are in the field of science.
- Finally, this new perspective or process of studying the scripture not only aids the scholars of the Islamic Studies, Statistics, and Machine Learning, but may also contribute indirectly to community development and policy makers who use Qur'ān as part of their decision making.

Significance of the Study

The significance of this study is that:

- With that said, the author would like to also emphasize the opportunities that scientific methodologies can bring to studying Islamic studies, especially for Muslim researchers who are in the field of science.
- Finally, this new perspective or process of studying the scripture not only aids the scholars of the Islamic Studies, Statistics, and Machine Learning, but may also contribute indirectly to community development and policy makers who use Qur'ān as part of their decision making.

Scope of the Study

The significance of this study is that:

- The paper will cover all chapters of the Qur'ān as much as possible, except maybe for cases like thematic modeling on very short $sūwar$ سُوَر, since topics on these $sūwar$ سُوَر may be obvious already or easier to see due to very short number of $ʾayah$ أَيَة.
- However, for cases where the analyses is not at the level of $sūrah$ سُورَة, but rather on the level of the Qur'ān as a whole, then all of the $sūwar$ سُوَر will be used.

Mathematical Sections

- Like any humanities studies, Mathematics is mostly concern with understanding facts about objects that is being studied.
- For Islamic studies, these objects can be physical like Qur'ān and other Islamic texts, or metaphysical like understanding the purpose of life.
- For Mathematics, the objects can be explicit or abstract as well, but it does revolve heavily on numbers and logics, and like other domain it studies facts about these objects.
- With that said, any object being studied in Mathematics are presented as Definitions, a formal way of defining terms or objects

Mathematical Sections

Mean

Let $x_i, i \in \{1, \cdots, n\}$ where $n \in \mathbb{N}$, then the *mean* of x_is is defined as follows:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \text{where } x_i \in \mathbb{R}. \tag{1}$$

- Verified facts that are of significant findings are sectioned as Theorem in mathematics, whereas those that are proposed are called Proposition.
- Other small results or facts supporting the Proposition are sectioned as Corollary.

Review of Related Literatures

- The earliest paper on Qur'ānic studies using computer was likely the work of Rashad Khalifa in 1968[1], which led to one of his book entitled 'The Computer Speaks: God's Message to the World' (*see* [11]).
- Rashad started at studying the mystery letters in the beginning of some *sūrahs* سُور (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 [11].
- Rashad was able to do this by transliterating the Qur'ān into Alpha-Numeric typesets for easy parsing of the computers back then.

Review of Related Literatures

- Among the pioneers in computational applications for the Qur'ān is the work of [19], who built a stemmer system for the Qur'ān.
- For example, in English language the root word for *computational*, *computer*, *computation*, and *computerize* is *compute*.
- According to [19], the rich morphology of the Qur'ānic language or the Classical Arabic makes it even more difficult to do word stemming.

Review of Related Literatures

- Among the pioneers in computational applications for the Qur'ān is the work of [19], who built a stemmer system for the Qur'ān.
- For example, in English language the root word for *computational*, *computer*, *computation*, and *computerize* is *compute*.
- According to [19], the rich morphology of the Qur'ānic language or the Classical Arabic makes it even more difficult to do word stemming.

Review of Related Literatures

- Moving on, [20] builds on top of this stemming system, and used it for tokenization of the Qur'ānic words, and building a statistical methodology for clustering or grouping the chapters of the Qur'ān, in particular [20] used a Agglomerative Hierarchical Clustering based on the Euclidean distance of the adjusted word frequency of a $s\bar{u}rah$ سُورَة.
- With the growing interests on studying the Qur'ān from the lense of Data Analysis and Natural Language Processing, resulted into creating digital corpi of the Qur'ān.
- A series of work by Sharaf and Atwell led to the following publications: [14] studied knowledge representation of the Qur'ān's verb valences using FrameNet frames, the output of which is a lexical database as a corpus of Qur'ān's verbs.

Review of Related Literatures

- Further, the work of [15] came up with corpus for the annotations of the Qur'ānic pronouns, the authors named it as QurAna.
- Building on this work, [16] came up with a corpus for studying Qur'ānic relatedness based on the commentary of Ibn Kathir اِبن كَثِر, the authors named this corpus as QurSim.
- Apart from Sharaf and Atwell, Dukes and Habash was also working on this, but specifically on the morphological annotations.
- For example, [8], which also led to other publications [6, 5, 7] related to this.

Review of Related Literatures

- With the establishment of a morphological annotated corpus for the Qur'ān, the hoped was to have further analyses on the said scripture using statistical and machine learning methodologies.
- As such, the work of [9, 4] were among the first to do so, where they constructed a statistical parser through machine learning.
- This was then followed by siddiqui2013 who used the said corpus for topic modeling using Latent Dirichlet Allocation, the said study started with 114 $s\bar{u}war$ سُوَر, but after processing it went down to 24 $s\bar{u}war$ سُوَر after considering $s\bar{u}war$ سُوَر with 1000 or more words.dukes2010dependency related to this.

Review of Related Literatures

- This was mainly due to the very sparse document term matrix for the Term Frequency - Inverse Document Frequency[2] (TF-IDF) embedding if considering all of the 114 $s\bar{u}war$ سُوَر.
- Aside from this, the rest have used the corpus by [8] as part of benchmark for morphological analysis or for new lexicographic database, for example [13, 10].

With the establishment of a morphological annotated corpus for the Qur'ān, the hoped was to have further analyses on the said scripture using statistical and machine learning methodologies. As such, the work of [9, 4] were among the first to do so, where they constructed a statistical parser through machine learning. This was then followed by [17] who used the said corpus for topic modeling using Latent Dirichlet Allocation, the said study started with 114 $sūwar$ سُوَر, but after processing it went down to 24 $sūwar$ سُوَر after considering $sūwar$ سُوَر with 1000 or more words. This was mainly due to the very sparse document term matrix for the Term Frequency - Inverse Document Frequency[3] (TF-IDF) embedding if considering all of the 114 $sūwar$ سُوَر. Aside from this, the rest have used the corpus by [8] as part of benchmark for morphological analysis or for new lexicographic database, for

[3] *See* Section **??**

example [13, 10].

Thank you!

References I

[1] Al-Ahmadgaid B. Asaad. "QuranTree.jl: A Julia Package for Quranic Arabic Corpus". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, 2021, pp. 208–212. URL: https://aclanthology.org/2021.wanlp-1.22.

[2] Jeff Bezanson et al. "Julia: A fresh approach to numerical computing". In: *SIAM Review* 59.1 (2017), pp. 65–98. DOI: 10.1137/141000671. URL: https://epubs.siam.org/doi/10.1137/141000671.

[3] Alan Cooperman, Erin O'Connell, and Sandra Stencel. *the future of the global muslim population*. Tech. rep. Pew Research Center, 2011.

References II

[4] Kais Dukes. "Statistical parsing by machine learning from a classical Arabic treebank". In: *arXiv preprint arXiv:1510.07193* (2015).

[5] Kais Dukes, Eric Atwell, and Nizar Habash. "Supervised collaboration for syntactic annotation of Quranic Arabic". In: *Language resources and evaluation* 47 (2013), pp. 33–62.

[6] Kais Dukes, Eric Atwell, and Abdul-Baquee Sharaf. "Online visualization of traditional Quranic grammar using dependency graphs". In: *The Foundations of Arabic Linguistics Conference*. Citeseer. 2010.

References III

[7] Kais Dukes and Tim Buckwalter. "A dependency treebank of the Quran using traditional Arabic grammar". In: *2010 the 7th International Conference on Informatics and Systems (INFOS)*. IEEE. 2010, pp. 1–7.

[8] Kais Dukes and Nizar Habash. "Morphological Annotation of Quranic Arabic". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari et al. Valletta, Malta: European Language Resources Association (ELRA), May 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/276_Paper.pdf.

References IV

[9] Kais Dukes and Nizar Habash. "One-Step Statistical Parsing of Hybrid Dependency-Constituency Syntactic Representations". In: *Proceedings of the 12th International Conference on Parsing Technologies*. Ed. by Harry Bunt, Joakim Nivre, and Özlem Çetinoglu. Dublin, Ireland: Association for Computational Linguistics, Oct. 2011, pp. 92–103. URL: https://aclanthology.org/W11-2912.

References V

[10] Mustafa Jarrar and Tymaa Hasanain Hammouda. "Qabas: An Open-Source Arabic Lexicographic Database". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 13363–13370. URL: https://aclanthology.org/2024.lrec-main.1170.

[11] Rashad Khalifa. *The computer speaks: God's message to the world*. Renaissance Production, 1981.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

References VI

[13] Yasser Sabtan. "Morphological Analysis of the Glorious Qur'an: A Comparative Survey of Three Corpora". In: *Arab World English Journal (AWEJ)* 8.4 (2017).

[14] Abdul-Baquee Sharaf and Eric Atwell. "Knowledge representation of the Quran through frame semantics: a corpus-based approach". In: *Proceedings of the Fifth Corpus Linguistics Conference*. University of Liverpool, United Kingdom: The Fifth Corpus Linguistics Conference, 2009. URL: https://api.semanticscholar.org/CorpusID:18278736.

References VII

[15] Abdul-Baquee Sharaf and Eric Atwell. "QurAna: Corpus of the Quran annotated with Pronominal Anaphora". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 130–137. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/123_Paper.pdf.

References VIII

[16] Abdul-Baquee Sharaf and Eric Atwell. "QurSim: A corpus for evaluation of relatedness in short texts". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2295–2302. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/190_Paper.pdf.

[17] Muazzam Ahmed Siddiqui, Syed Muhammad Faraz, and Sohail Abdul Sattar. "Discovering the Thematic Structure of the Quran using Probabilistic Topic Model". In: *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*. 2013, pp. 234–239. DOI: 10.1109/NOORIC.2013.55.

References IX

[18] Nicolai Sinai. *The Qur'an: A historical-critical introduction*. Edinburgh University Press Ltd, 2017.

[19] Naglaa Thabet. "Stemming the Qur'an". In: *Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*. Semitic '04. Geneva, Switzerland: Association for Computational Linguistics, 2004, pp. 85–88.

[20] Naglaa Thabet. "Understanding the thematic structure of the Qur'an: an exploratory multivariate approach". In: *Proceedings of the ACL Student Research Workshop*. ACLstudent '05. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 7–12.

References X

[21] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.