PriMera Scientific Engineering
Volume 2 Issue 2 February 2023
ISSN: 2834-2550

PriMera Scientific Publications

Extracting Semantic Relationship Between Fatiha Chapter (Sura) and the Holy Quran

Type: Review Article
Received: December 27, 2022
Published: January 07, 2023

Citation:
Ahmed Samir Ahmed Ibrahim El khadrawy., et al. "Extracting Semantic Relationship Between Fatiha Chapter (Sura) and the Holy Quran". PriMera Scientific Engineering 2.2 (2023): 02-12.

Ahmed Samir Ahmed Ibrahim El khadrawy[1,2]*, Safia Abbas[1,2], Yasser Kamal Omar[1,2] and Nady Hussain Abd Jawad[1,2]

[1]*Faculty of Computer and Information Sciences, Arab Academy for Science and Technology, Ain Shams University, Cairo, Egypt*

[2]*Faculty of Islamic and Arabic Studies, Al-Azhar University, Cairo, Egypt*

***Corresponding Author:** Ahmed Samir Ahmed Ibrahim El khadrawy, Faculty of Computer and Information Sciences, Arab Academy for Science and Technology, Ain Shams University, Cairo, Egypt; Faculty of Islamic and Arabic Studies, Al-Azhar University, Cairo, Egypt.

Abstract

There is a huge shortage of scientific research in the Arabic language, especially in natural language processing and relationships between Arabic documents and other specific documents. This shortage is also reflected in Arabic Books' introduction, Topics abstraction, and content summary engines. Furthermore, there are some good samples of Arabic words inside the Quran. The Quran is the holy book of Islam, that is divided into chapters (surah) and verses (ayat) of differing lengths and topics. This paper introduces a framework for both specialized researchers in Islamic studies as well as non-specialized researchers to find hidden relationships between one of the most important chapters of the Holy Quran which is Al Fatiha surah and the remaining chapters of the Holy Quran using Hierarchical Technique data modeling as an unsupervised learning technique. a new framework that can access tokens of the Holy Quran in different granule parts such as chapter (sura) part of the chapter (Aya) of the Holy Quran Sura, words, word roots, Aya roots, and Aya meaning in the Arabic language, Moreover, We had developed a lot of statistics related to Fatiha Sura and the holy Quran like (roots distinct for every Sura, Words Redundancy, Roots Redundancy, Matrix report by roots and every sura, Matrix report shows Percentage of roots similarity by every sura and whole Quran distinct roots, etc.). Furthermore, we enhance the search engine results by adding search by roots and Aya meaning for every Sura. And the results for sample queries show accuracy with more than 3% using meaning and roots compared to the text of the Holy Quran only.

Keywords: holy Quran text Analysis; Text Mining; Arabic Text Mining; 1N Form; 2Nform; Data Modeling; holy Quran Tafseer text

Introduction

Natural language processing is a branch of machine learning that is capable of understanding computer analysis as well as manipulating human language [1]. unfortunately, the Arabic language has a shortage in research that had been applied to Text analysis and NLP. Moreover, the Arabic language is not used by one country only, but by 26 countries across North Africa and the Middle East. Moreover, it is considered a native language speaking for more than 422 million people [2].

This huge shortage in Text analysis and NLP results in missing important parts in both Information Retrieval (IR) like search engines, Question Answering (QA), and Reports Summarization. As per mentioned before the holy Quran is written in the Arabic language and it is an area of interest for many people, especially Muslims so we need more investigations as well as more deep text analysis for the holy Quran, especially "Sura El Fatiha" as it is considered as the base of the holy Quran since The Prophet Mohamed (Peace upon him) called it "the base of Quran". In this paper, we will introduce a proposed model to extract Semantic Relationship between Fatiha Chapter (Sura) and the whole Holy Quran. Our proposed model consists of three phases: phase one focuses on Preprocessing Text and the ETL Process which can collect (the holy Quran Sura, Aya, words, word roots, Aya roots, and Aya meaning), while phase two focuses on text analysis which is capable to compare the matched words that have been extracted from phase one and data modeling for three objects (the holy Quran Text for every chapter, Word roots, and Aya explanations and meaning). Then in phase three, we implement the Information retrieval data access layer and visualize the information to show the measurements and calculations.

Text Analysis is to process unstructured such as textual information, extract meaningful information indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. As per claims of Arabic Language researchers that is the holy Quran is the most powerful book in the Arabic language in rhetoric and Arabic literature, so we proposed this paper to be a text analysis solution for all the Arabic language researchers.

Related Work

Background - Here is a highlighting of the main concepts related to the research Data Modeling, Text Mining, Natural Language Processing (NLP), extract Transform Loading (ETL), and the holy Quran history.

Text Mining

The purpose of Text Mining is to process unstructured textual information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. There are a lot of befits of text mining like increased researcher efficiency, unlocking hidden information and developing new knowledge, exploring new horizons, improved research and evidence-based, and improving the research process and quality [3].

Natural language processing (NLP)

NLP is used to build machines and models that could understand and respond to a text, or voice data as well as respond to text or speech in the same way as humans do [4]. In the current proposed solution we use this concept to prepare the Quran word text like change text format, text steaming, and term frequency.

Data Modeling

Data Modelling is the process of analyzing the data objects and their relationship to the other objects. It is used to analyze, classify and simplify the data into useful information and enhance information retrieval.in section 4 we describe how we have used this concept to calculate the statistical information related to Fatiha Sura and every Chapter of the holy Quran.

Extract Transfor Loading (ETL)

ETL stands for extract, transform, and loading. It is the process that is used for extracting data from various sources, moreover, it can transform the data into a usable and trusted resource. Finally, it is capable to load it into a single data source then the system end-users will be able to access and solve business problems [5]. We used this concept to build ETL using the MSSQL integration service and prepared CSV files into a Table design structure.

The Holy Quran History

The holy Quran has 114 Sura (Chapters), every Sura has many Aya (Verses) by a total number of Aya 6236 [6]. Every Aya contains some words, the total numbers of words are 77779, and the total unique word is 14780. Quran words coming to us in 2022 by high accuracy transfer by Islamic researchers, auditors throw a practical approach including a series of readers from much millions of examples in Muslims prays and memorizing the Holy Quran schools and it was transmitted from generation to generation, from God to the Prophet Mohamed (Peace upon him) to us till now a day.

Related work - Several approaches have been introduced in previous papers for extracting the words and their roots/meaning from the Holy Quran, yet still, no proposed system shows the relation between Fatiha and the whole Holy Quran and Arabic word roots.

The proposed system by Halim Sayoud and et.al [7] had conducted some experiments for classifying the relationships between Hadith and Holly Quran, but the proposed system was missing to work on many other Hadith by many people as well as the system does not cover the relationship between Fatiha and other holy Quran chapters.

The approach by Rahima and et.al [8] introduced a proposed model which could classify the conjunctive patterns with two terms *AND & BETWEEN*, but unfortunately, they are missing the expert judgment for their results.

The proposed approach by Alhawarat and et.al [9] developed an showed the word cloud of Holly Qur`an that introduced the most frequent 100 words as well as measured the TF & TF-IDF. But the algorithms they had used is not efficient enough from the performance perspective. Moreover, their reports are not dynamic enough as they extract reports as they are used only type of word cloud.

The proposed approach by Sadi and et.al [10] introduced an ontology that classifies just the Quranic “Nature” Domain, by collecting each group of words with the same meaning in one category using SPARQ and OWL.

The proposed approach by Waseem Alromima and et.al [11] introduced an Ontology-Based Model for Arabic Lexicons that covers the knowledge of the Arabic language vocabulary associated with the Place Noun vocabulary mentioned in the Holy Quran using the Web Ontology Language (OWL), The ontology will be useful in the knowledge of the Islamic learning, linguistics researches, and Semantic Web applications.

The proposed approach by Khaled and et.al [12] introduced A Qur’anic Code for Representing the Holly Qur’an (Rasm Al- ’Uthmani), Holly Quran must be written correctly and precisely without any modification, even though some characters used in Quran do not have a corresponding Unicode representation so the representations are not the best way to represent the Holly Qur’an by using the Quranic code, they successfully to solved 5 problems out of 6, reached more than 65% reduction ratio, more searching capability, and standard way to store and present the Holly Qur’an on any electronic device but still have Problem Lengthening like (Tatweel) ـــــــــــــ still need to fix.

The previously mentioned papers introduced different good approaches related to the Arabic language and the holy Quran. Moreover, to the best of our knowledge, there is a lack of previous research papers on text analysis of the Fatiha Chapter compared to the whole holy Quran especially the Fatiha chapter is considered the first Sura in the Holy Quran and considered a summary and guidelines for the remaining chapters in the holy Quran.

Data Set

In this paper, the applied dataset was in three CSV files format as follows and all of the files authorized from King Fahd Glorious Qur'an Printing Complex [13]:

holy Quran in the Arabic language - CSV 1 (total number of columns: 16 and the total number of rows: 6232)

This is a sample of the CSV 1 file that contained the holy Quran words in the Arabic language for every chapter The first Row is columns as below:

(id,jozz,sora,sora_name_en,sora_name_ar,page,line_start,line_end,aya_no,aya_text,aya_text_emlaey).
Row Sample as below

(1,1,1,Al-Fātiḥah,الفَاتِحة",1,2,2,1,"بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ ١","بسم الله الرحمن الرحيم)

Quran text Roots for every chapter - CSV 2 (total number of columns: 3 and the total number of rows: 77432)

This is a sample of the CSV 2 file that contained the roots of the holy Quran words The first Row is columns as below:

(id, word, root).
Row Sample as below:

(1,بِسۡمِ , /سمي/)

Core Quran Text Meaning for every Chapter - CSV 3 (total number of columns: 12 and the total number of rows: 6237) This is a sample of the CSV file of Core Quran Text Meaning and Aya Explanation for every Chapter and every Aya - CSV3 The first Row is columns as below:

(id,jozz,sura_no,sura_name_en,sura_name_ar,page,line_start,line_end,aya_no,aya_text,aya_text_emlaey,aya_tafseer)
Row Sample as below

(1,1,1,Al-Fātiḥah," الفَاتِحةبِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ ١","بسم الله الرحمن الرحيم",1,2,2,1,"[1] أبتدئ قراءة القرآن باسم الله مستعيناً به، سمجٱللَّهِعلم على الرب -تبارك وتعالى- المعبود بحق دون سواه، وهو أخص أسماء الله تعالى، ولا يسمى به غيره سبحانه وتعالى.سمجٱلرَّحۡمَٰنِذي الرحمة العامة الذي وسعت رحمته جميع الخلق، سمجٱلرَّحِيمِسمج
بالمؤمنين، وهما اسمان من أسمائه تعالى، يتضمنان إثبات صفة الرحمة لله تعالى، كما يليق بجلاله.")

Methodology

The following fig 1 describes the research solution proposed model including data acquisition, Modeling, and data analysis and visualization. We got a dataset for the holy Quran CSV files from the public official developers' portal for authorized King Fahd Glorious Qur'an Printing Complex [13].

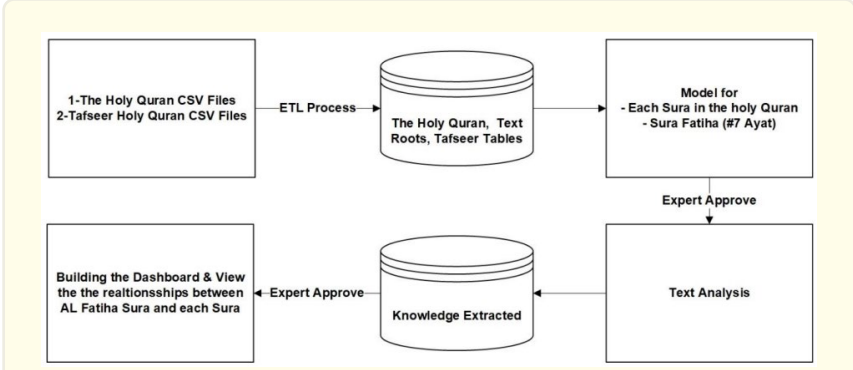

Figure 1: Research Solution Model.

The following three phases show the breakdown structure of the proposed methodology:

Data acquisition phase (Preprocessing Text and Database)

We built the ETL for the whole Quran Database by extracting CSV Files so it can be converted into SQL relational database using a migration tool to convert CSV format to table design (SrNo, Juz, JuzNameArabic, JuzNameEnglish, SurahNo, SurahNameArabic, SurahNameEnglish, SurahMeaning, AyahNo, EnglishTranslation, OrignalArabicText, ArabicText, ArabicWordCount, ArabicLetterCount) then applying the normalization techniques with other CVS files, this is can be done by the following steps:

1) *Core Quran Text for every chapter - CSV 1.*
2) *Core Quran text Roots for every chapter - CSV 2.*
3) *Core Quran Text Meaning for every Chapter - CSV3.*

After the Migration steps, we implement relationships between tables and each other to start the second phase of modeling and discover relationships between the word root of Fatiha Sura and every sura in the whole holy Quran.

Modeling and Extraction Phase

We had built two models for the whole Quran by implementing the following procedures:

1) *Building Object Model including Quran Text Aya, word, root, Tafseer properties.*
2) *Discover the Distinct Roots and relationships between roots and Fatiha Sura.*

We had used a formula to calculate their term frequency percentage for each Sura to the whole Quran by using the following formula and Vlookup function [14]:

$$STR\ \% = DC\ (WR)/(TDRS)$$

STR: *Sura term Root Percentage,* DC: *DISTINCT COUNT,* WR: *Word Root,* TDRS: *Total Distinct Root of Sura*.

Below Table 1 displays a matrix report showing Sura word root matching compared to others in the holy Quran, in the x column dimension is sorting all of the Holy Quran Sura and the y dimension is sorting every Sure and percentage of root matching with AL Fatiha Sura and others in the Holy Quran.

Sura Name	*AlFatiha*	*AlBaqara*	*Alomran*	*ALNessa*	*AlMaada*	*AlAnam*	*AlAraf*	*AlAnfal*	*AlTawba*	*Younes*
الفاتحة	100%	3%	4%	4%	4%	4%	4%	5%	4%	6%
البقرة	100%	100%	74%	75%	76%	71%	74%	82%	75%	83%
آل عمران	95%	58%	100%	63%	66%	60%	60%	78%	66%	75%
النساء	90%	60%	65%	100%	67%	63%	58%	76%	69%	73%
المائدة	95%	55%	61%	60%	100%	59%	58%	70%	66%	74%
الأنعام	90%	52%	57%	57%	59%	100%	62%	66%	59%	79%
الأعراف	100%	60%	63%	59%	65%	70%	100%	72%	66%	82%
الأنفال	71%	37%	45%	43%	44%	41%	40%	100%	46%	55%

التوبة	81%	47%	53%	54%	57%	51%	51%	64%	100%	66%
يونس	86%	41%	48%	45%	51%	54%	50%	61%	52%	100%
هود	81%	44%	50%	48%	51%	55%	53%	64%	57%	70%
يوسف	81%	41%	46%	44%	49%	52%	49%	59%	50%	61%
الرعد	71%	35%	40%	38%	41%	41%	40%	49%	43%	53%
إبراهيم	76%	31%	38%	36%	38%	39%	38%	48%	38%	51%
الحجر	71%	26%	30%	30%	32%	34%	34%	41%	34%	43%
النحل	95%	46%	51%	50%	51%	57%	53%	62%	53%	66%
الإسراء	81%	41%	48%	47%	48%	51%	48%	62%	51%	65%
الكهف	81%	43%	48%	47%	51%	52%	49%	58%	53%	62%
مريم	76%	33%	40%	36%	37%	38%	37%	45%	38%	47%
طه	90%	41%	44%	42%	46%	47%	48%	54%	46%	58%
الأنبياء	67%	36%	41%	39%	42%	47%	43%	51%	42%	59%
الحج	90%	42%	47%	46%	47%	48%	46%	60%	51%	60%
المؤمنون	81%	36%	41%	40%	43%	47%	43%	52%	42%	59%
النور	76%	38%	43%	43%	45%	43%	41%	53%	48%	54%
الفرقان	71%	33%	37%	37%	38%	42%	40%	50%	42%	55%
الشعراء	71%	35%	39%	37%	42%	43%	44%	51%	42%	55%
النمل	86%	35%	42%	39%	42%	45%	43%	52%	43%	60%
القصص	76%	40%	45%	44%	47%	46%	45%	55%	47%	57%
العنكبوت	76%	32%	39%	36%	39%	43%	40%	54%	43%	55%

Table 1: Sample Matching and Term Frequency Matrix Report.

Text Analysis and Visualization

Similarity Percentage Relationship: getting the similarity based on the relation of the word root and meaning between every Sura and Fatiha Sura using Text Categorization techniques and visualizing the information.

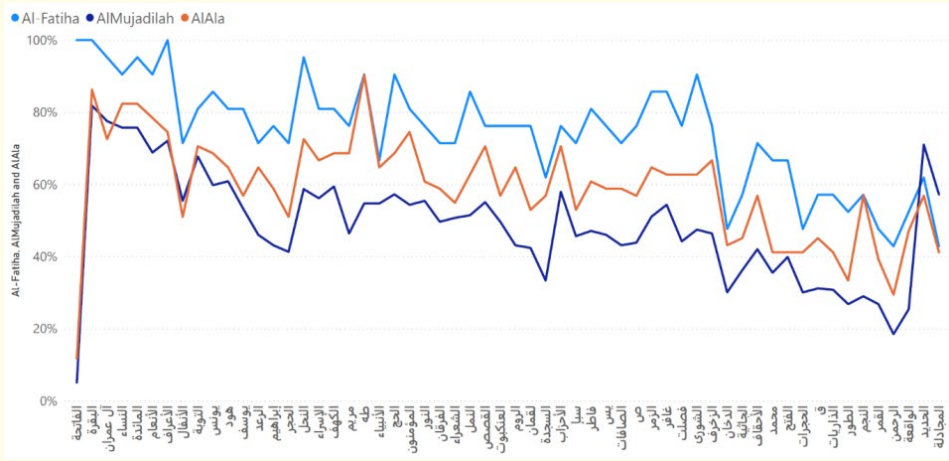

Figure 2: Fathia, Al Ala and Almujadilah Sura and most Semilarty Relation using croping and semilarty text.

Above Fig.2 displays the most Quran Text Word distinct Roots frequency and Shows that the roots of Al Fatiha, Al Alaa, and al mujadilah are the most common use in the whole Quran using cropping and term frequency and every Sura.

Result

The proposed methodology shows that the proposed idea is efficient in implementing a dataset for the whole Holly Quran and Tafseer in a relational database as well as normalizing the tables using 1NForm, 2NForm, and third normal form then implementing a model and measures calculations for each Sura. Below are some sample reports that will be helpful for Arabic language research and enhance search engines throw text meaning as well as text roots. Report Describes every Sura, Total roots distinct of Al Omran, Al Bakra, and AL Fatiha, and Total roots of the Quran Karim. Furthermore, it displays a Column Chart for every sura and total distinct root, Fatiha Roots, Whole Quran Word Redundancy and Root Redundancy, Bakra, and Al Omran Roots. We show the results of the output statistics and reports to the Judgment expert (Islamic and Arabic researchers) to collect the Feedback and the results are good for them and reports are helpful for the metrics, visualization, clustering, and text mining measurement using Text Categorization techniques. Here table 2 represents samples of the Fatiha text roots, word redundancy, and root redundancy.

ID	*Word*	*Root*	*Word Redundancy*	*Root Redundancy*
1	بِسْمِ	سمى	3	70
2	اللَّهِ	ءله	2153	2851
3	الرَّحْمَـٰنِ	رحم	45	338
4	الرَّحِيمِ	رحم	34	338
5	الْحَمْدُ	حمد	26	68
6	لِلَّهِ	ءله	116	2851
7	رَبِّ	ربب	128	979

8	الْعَـٰلَمِينَ	علم	61	853
9	الرَّحْمَـٰنِ	رحم	45	338
10	الرَّحِيمِ	رحم	34	338
11	مَـٰلِكِ	ملك	33	206
12	يَوْمِ	يوم	217	475
13	الدِّينِ	دين	47	124
14	إِيَّاكَ	ءيو	1	242
15	نَعْبُدُ	عبد	6	275
16	وَإِيَّاكَ	ءيو	1	242
17	نَسْتَعِينُ	عون	1	11
18	اهْدِنَا	هدى	1	325
19	الصِّرَٰطَ	صرط	6	45
20	الْمُسْتَقِيمَ	قوم	5	660
21	صِرَٰطَ	صرط	32	45
22	الَّذِينَ	ـ'	811	8480
23	أَنْعَمْتَ	نعم	7	144
24	عَلَيْهِمْ	على	215	1441
25	غَيْرِ	غير	69	153
26	الْمَغْضُوبِ	غضب	1	24
27	عَلَيْهِمْ	على	215	1441
28	وَلَا	ـ'	605	8480
29	الضَّآلِّينَ	ضلل	6	182

Table 2: Sura AL Fatiha words and roots redundancy.

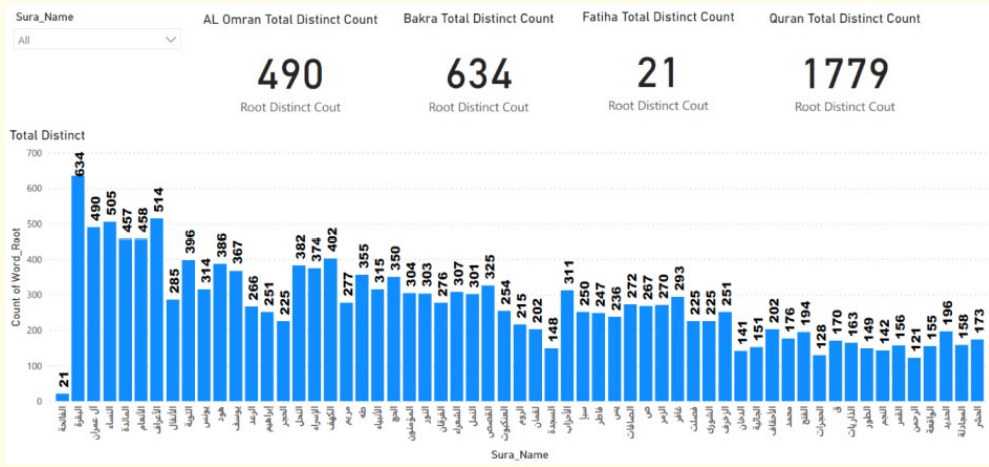

Figure 3: Fathia, Al Omran distinct Roots Redundancy and other Statistics.

Below Fig.3 displays samples of measurement report showing Al Omran, AL Bakra, Al Fatiha, and whole Quran Text Word distinct Roots Redundancy and Showing the report in Column type Chart.

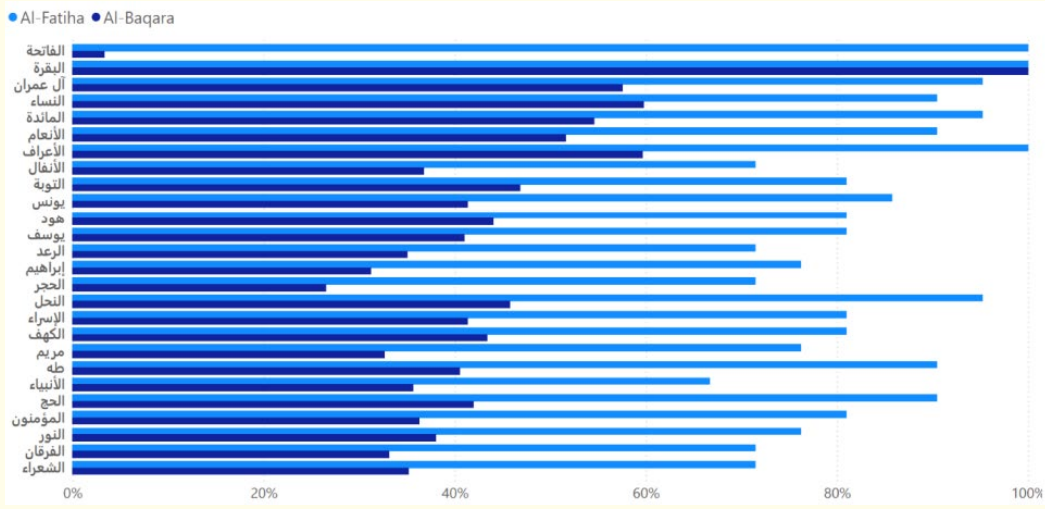

Figure 4: Fathia, Al Bakra Roots distinct Similarity Comparison.

Above Fig.4 displays compare between Al Fatiha, AL Bakra Quran Word distinct Roots Redundancy and every Sura in the Holy Quran and shows the report in a stacked bar chart.

Moreover, Search enhancement of the search and here the following table 3 represents samples of the search enhancement after adding the roots and meaning and shows the percent difference.

Keyword Sample	*Total Number of words*	*Total rows retrieved without roots and meaning*	*Total rows retrieved without roots and meaning in %*	*Total rows retrieved with roots and meaning*	*Total rows retrieved with roots and meaning in %*	*Differences in %*
الصابرين (the patient)	77779	20	0.026%	103	0.13%	0.11%
كريم (generous)	77779	27	0.035%	47	0.06%	0.03%
الصادقين (the truthful)	77779	19	0.024%	155	0.20%	0.17%
المحسنين (benefactors)	77779	29	0.037%	149	0.19%	0.15%

Table 3: Performance Enhancement Evaluation.

Discussion

Some big samples of a good command Arabic language are the holy Quran especially on natural language processing and the relation between Arabic documents and other specific documents also Books introduction, Topics abstraction, or content summary. The holy Quran is different from other religious texts in that it is believed to be the literal words of God in the Arabic language without any modification by humans.

The proposed research implements data modeling, statistics reports, and enhancement of the search in text meaning related to Arabic language and the holy Quran plus but still, we need more research and discovery for data classification inside the holy Quran like events, reasons for going down, the Paradise, pronouns, stories, Morality, Knowledge of Mecca or Medina, Urban and travel knowledge.

The sequence of going down, terrestrial and celestial, etc. so this model solution will be helpful for both specialized researchers in Islamic and Arabic studies.

Conclusion and Future Work

The holy Quran contains information that covers many domains, Fatiha is the first chapter of the Holy Quran which contains a small brief about the whole Quran Karim thus we need to apply a deep analysis of the text of the Fatiha Chapter to Prove the inheritance/Abstraction/Introduction relationship between it and other Quran chapters in more scientific research paper.

We had successed to build a new proposed model and reports and this idea shows a good solution for both specialized people in Islamic studies as well as non-specialized people. After Discovery the correlation and similarity relationships between Fatiha and other Quranic Chapters we can prove that the Fatiha Sura is a parent Chapter of the whole Quran.

A further step in the proposed system is to fully automate the process as well as cover more comparison points in the Holly Quran using more techniques and methods in deep learning and ETL processing.

References

1. Otter DW, Medina JR and Kalita JK. “A survey of the usages of deep learning for natural language processing”. IEEE transactions on neural networks and learning systems 32.2 (2020): 604-624.
2. Yousif J. “Neural computing-based part of speech tagger for the Arabic language: a review study”. International Journal of Computation and Applied Sciences IJOCAAS 5.1 (2018).
3. Kumar L and Bhatia PK.” Text mining: concepts, process and applications”. Journal of Global Research in Computer Science 4.3 (2013): 36-39.

4. Wolf T., et al. "Huggingface's transformers: State-of-the-art natural language processing". arXiv preprint (2019).
5. Dhanda P and Sharma N. "Extract Transform Load Data with ETL Tools". International Journal of Advanced Research in Computer Science 7.3 (2016).
6. Mohamed EH and El-Behaidy WH. "An ensemble multi-label themes-based classification for holy Qur'an verses using Word2Vec embedding". Arabian Journal for Science and Engineering 46.4 (2021): 3519-3529.
7. H Sayoud. "Automatic authorship classification of two ancient books: Quran and Hadith". 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA) (2014): 666-671.
8. 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 (2017).
9. M Alhawarat, M Hegazi and A Hilal. "Processing the Text of the Holy Quran: a Text Mining Study". International Journal of Advanced Computer Science and Applications (IJACSA) 6.2 (2015).
10. ABMS Said., et al. "Applying ontological modeling on quranic 'nature' domain". 2016 7th International Conference on Information and Communication Systems (ICICS) (2016): 151-155.
11. Sds Alromima W., et al. "Ontology-based model for Arabic lexicons: An application of the Place Nouns in the Holy Quran". In 2015 11th International Computer Engineering Conference (ICENCO) IEEE (2015): 137-143.
12. Foda KM., et al. "A Qur'anic Code for Representing the Holy Qur'an (Rasm Al-'Uthmani)". In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, IEEE (2013): 304-309.
13. Official King Fahd Glorious Quran Printing Complex. "Holy Quran Software Developers official Portal" (2022).
14. Seal KC. "A generalized PERT/CPM implementation in a spreadsheet". INFORMS Transactions on Education 2.1 (2001): 16-26.
15. MA Sherif and Axel-Cyrille Ngonga Ngomo. "Semantic Quran". Semantic Web 6.4 (2015).
16. Aktas ME and Akbas E. "Text classification via network topology: A case study on the holy quran". In 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE (2019): 1557-1562.
17. Adeleke AO., et al. "Comparative analysis of text classification algorithms for automated labelling of Quranic verses". Int. J. Adv. Sci. Eng. Inf. Technol 7.4 (2017): 1419.