Mesopotamian journal of Quran studies
Vol.2024, **pp**. 16–23
DOI: https://doi.org/10.58496/MJQS/2024/003 ; ISSN: 3005–9933
https://mesopotamian.press/journals/index.php/MJQS

An AI model for Parsing the Text of Holy Quran Sentences

نموذج الذكاء الاصطناعي لاعراب جمل من نصوص القرآن الكريم

Haval H. Ameen [1, *], AbdulSattar M. Khidhir [2]

[1] *Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.*

[2] *Northern Technical University, Mosul Technical Institute, Mosul, Iraq.*

هفال حجي امين [1، *]، عبدالستار محمد خضر [2]

[1]قسم علوم الحاسوب ، كلية علوم الحاسوب والرياضيات ، جامعة الموصل ، الموصل ، العراق.

[2]جامعة التقنية الشمالية معهد التقني في الموصل , الموصل , العراق.

ABSTRACT

The Holy Quran is of immense importance to many societies due to the position that this book holds for Muslims around the world and the religious teachings it contains. The Holy Quran employs a high-standard Arabic language, which requires analysis and simplification of expressions to enhance comprehension and application of its teachings. The digitization and combining of the Holy Quran with computing operations have made it easier to discover the vast amounts of information contained within its verses. Using these applications, academics and erudite researchers have successfully developed norms that govern the study of Qur'anic knowledge, thus expanding the illustration of the Quran. NLP is a well-known branch of study that has been the subject of research for many years. The intrinsic intricacy of this field has resulted in slower progress when compared to other areas of inquiry. Furthermore, despite having the most sophisticated syntax, structure, and verb conjugation of all-natural languages, Arabic has received comparatively little attention. As a result, there is an urgent need for study in this area to aid in the discovery of inclusive lexical knowledge. This research is concerned with the parsing of holy Quranic sentences. A neural network was used Consisting of two layers to training the token word attribute depending on the neural network input words characteristics. This research is based on a knowledge base that consists of 160 features was used, and 26 features related to the grammatical case were selected. The first group comprised an entered dataset with 128,221 rows, whereas the second group had 99.540 rows. We ran seven different experiments within each group and achieved a 96% accuracy rate. This accuracy was reached by using a training set size of 90,000 rows and a testing set size of 9,000 rows. Furthermore, we achieved 94% accuracy within the broader dataset of 128,221 rows by employing a training set size of 90,000 rows and a testing set size of 38,000 rows.

الخلاصة

يحظى القرآن الكريم بأهمية كبيرة في العديد من المجتمعات بسبب المكانة التي يحملها هذا الكتاب للمسلمين في جميع أنحاء العالم وما يحتويه من تعاليم دينية. يستخدم القرآن الكريم لغة عربية رفيعة المستوى، مما يتطلب تحليل وتبسيط العبارات لتعزيز الفهم وتطبيق تعاليمه. إن رقمنة القرآن الكريم ودمجه مع العمليات الحاسوبية سهّلت اكتشاف الكم الهائل من المعلومات الواردة في آياته. وباستخدام هذه التطبيقات، نجح الأكاديميون والباحثون المثقفون في تطوير القواعد التي تحكم دراسة المعرفة القرآنية، وبالتالي التوسع في توضيح القرآن. البرمجة اللغوية العصبية هي فرع من فروع الدراسة المعروفة التي كانت موضوع البحث لسنوات عديدة. وقد أدى التعقيد الجوهري لهذا المجال إلى تقدم أبطأ بالمقارنة مع مجالات البحث الأخرى. علاوة على ذلك، على الرغم من امتلاك اللغة العربية لتركيب الجملة والبنية وتصريف الأفعال الأكثر تطورًا بين اللغات الطبيعية، إلا أنها لم تحظ باهتمام كبير نسبيًا. ونتيجة لذلك، هناك حاجة ملحة للدراسة في هذا المجال للمساعدة في اكتشاف المعرفة المعجمية الشاملة. يهتم هذا البحث بإعراب الجمل القرآنية الكريمة. تم استخدام شبكة عصبية مكونة من طبقتين لتدريب سمة الكلمة المميزة بالاعتماد على خصائص كلمات إدخال الشبكة العصبية. يعتمد هذا البحث على قاعدة معرفية تتكون من 160 ميزة تم استخدامها، وتم اختيار 26 ميزة تتعلق بالحالة النحوية. ضمت المجموعة الأولى مجموعة بيانات تم إدخالها تحتوي على 128,221 صفًا، بينما ضمت المجموعة الثانية 99,540 صفًا. أجرينا سبع تجارب مختلفة داخل كل مجموعة وحققنا معدل دقة يصل إلى 96%. تم الوصول إلى هذه الدقة باستخدام حجم مجموعة تدريب يبلغ 90.000 صف وحجم مجموعة اختبار يبلغ 9000 صف. علاوة على ذلك، حققنا دقة بنسبة 94% ضمن مجموعة البيانات الأوسع المكونة من 128,221 صفًا من خلال استخدام حجم مجموعة تدريب يبلغ 90,000 صف وحجم مجموعة اختبار يبلغ 38,000 صف.

*Corresponding author haval.21csp106@student.uomosul.edu.iq

Keywords

الكلمات المفتاحية

NLP ,Word Embedding , AI for Quran , Arabic parsing , Deep Learning

Received	Accepted	Published online
استلام البحث	قبول النشر	النشر الالكتروني
14/4/2024	18/6/2024	10/7/2024

1. INTRODUCTION

The Holy Quran is one of the most important books in the world, especially in the Islamic world, and contains a wealth of knowledge from many other fields. For Muslims, an all-encompassing understanding of the Qur'an is crucial since, combined with the Sunnah, it forms the foundation of the faith. Particular scientific efforts have been made in the field of Qur'anic studies with the goal of interpreting its scientific components and reaping their benefits. The digitization and combining of the Holy Quran with computing operations have made it easier to discover the vast amounts of information contained within its verses. Using these applications, academics and erudite researchers have successfully developed norms that govern the study of Qur'anic knowledge, thus expanding the illustration of the Qur'an [6]. NLP is a well-known branch of study that has been the subject of research for many years. The intrinsic intricacy of this field has result-ed in slower progress when compared to other areas of inquiry. Furthermore, despite having the most sophisticated syntax, structure, and verb conjugation of all-natural languages, Arabic has received comparatively little attention. As a result, there is an urgent need for study in this area to aid in the discovery of inclusive lexical knowledge [10]. The morphological aspects of Arabic sentences and words defy the syntax and semantics of Arabic text, and the Arabic language is rich and considered the most sophisticated of all languages. The sentence's intricacy is inherent in both its syntax and semantics. It also includes a large number of vocabularies, such as synonyms, antonyms, and word roots used as nouns or verbs [8]. This level of complexity has an impact on the ability to learn, analyze, and automate this language. Word roots combine with various vowel forms to generate basic verbs and nouns. It can begin with sophisticated techniques in the grammatical derivation. The Arabic Language Statement includes a variety of verb forms. Each verb's conjugation is determined by a variety of criteria [9].

The remainder of the document is organized as follows: An illustration from the Related Works in Section 2. In Section 3, the suggested model's theory is explained. Furthermore, the approach of the suggested model is explained in Section 4. The experimental and findings discussion is highlighted in Part 5, the conclusions and future work are illustrated in Section 6, and a list of references is produced in Section 7.

2. RELATED WORK

Bashir, Muhammad Huzaifa, et al. (2023) [1]. This paper surveys the different efforts in the field of Qur'anic NLP, serving as a synthesized compendium of works (tools, data sets, approaches) covering the gamut from automated morpho-logical analysis to correction of Qur'anic recitation via speech recognition. Mul-tiple approaches are discussed for several tasks, where appropriate. Finally, we outline future research directions in this field. Alargrami, A. M.; Eljazzar, M. M. (2020) [2]. This research might be considered one of the first to provide an effi-cient distributed word representation model for several NLP tasks in the Islamic domain. The Word Embedding Model and the algorithm built on top of it are implemented in the Imam application, which allows users to ask the program to search for any data connected to the Islamic domain and provide an answer. The information is taken from several sources (Maliks Muwataa, Musnad Ahmad Ibn-hanbal, Sahih Muslim Hadith, Sahih Al-Bukhari, Sunan Al-darimi, and others). More than 90,000 documents (text blocks) from ten different books were ob-tained. After numerous successive pipeline operations of data cleaning, preprocessing, and normalization, the word2vec model was created using the skip-gram technique. Finally, the model was tested in various ways, first using K-means clustering and then a nonlinear dimensionality reduction technique to represent the data in 2D dimensions, and then using word similarity to test its capacity to understand the Quranic language. The tests clearly show that the model can per-form well in a variety of NLP Arabic and Islamic tasks. Touati-Hamad Z et al. (2021) [3]. aim to use deep learning algorithms to automatically authenticate the Quranic material layout. The Long Short-Term Memory (LSTM) method was used in this study, and the findings obtained a test accuracy of 99.98% on a dataset constructed using Tanzil website data. Khadhim, Sahira Alawi, and Hashim Wasi Chiad. (2022) [4]. These letters could have appeared at the

beginning of a Quranic verse or anywhere else. The study solely looks at letters with the three diacritical markings (Fathah, Dammah, and Kasrah) and includes the recipient for each reading. The study does not cover alternate readings of letters that do not have the three diacritical markings. Biltawi, M., Awajan, A., and Tedmori, S. (2017) [11]. The paper provides a complete introduction to lexical semantics using examples from the Arabic language, and then surveys the efforts of schol-ars attempting to develop ontologies for the Arabic language. Naili, Marwa, Anja Habacha Chaibi, and Henda Hajjami Benghezala. (2017) [12]. Word vector repre-sentations are extremely effective in a variety of natural language processing tasks because they capture the semantic meaning of words. In this context, there are three recognized methods: LSA, Word2Vec, and GloVe. In this research, these strategies will be examined in the context of topic segmentation for both Arabic and English. Furthermore, Word2Vec is extensively investigated, utilizing various models and approximation approaches. As a result, we discovered that LSA, Word2Vec, and GloVe are all language-dependent. Chaimae, A. Z. R. O. U. M. A. H. L. I., José F. Aldana Montes, and Maciej Rybinski. Word2Vec provides the best word vector representation, although it is dependent on the mod-el used. (2020) [13]. This paper describes the cleaning and pre-processing processes used to create three different training datasets. We provide a full overview of the pro-cedures we took to generate 180 different word embedding models with Word2Vec and CBOW. We use a combination of extrinsic and intrinsic evaluation approaches to assess the quality of word embeddings. The preliminary re-sults suggest that these models can generate meaningful word embeddings despite the higher logical complexity of the Arabic language. We concluded that the source of the training dataset has a considerable impact on the type of infor-mation acquired by the model. Furthermore, the hyperparameters of the training architecture and the nature of an NLP task influence its accuracy Sabbeh, S.F.; Fasihuddin, H.A. (2023) [14]. We compare classic and contextualized word embeddings in sentiment analysis.The four most prevalent word embedding algorithms were used in both trained and pre-trained versions. The chosen embedding combines traditional and contextualized techniques.Classical word embedding algorithms include GloVe, Word2vec, and Fast Text. In contrast, ARBERT functions as a contextualized embedding model. Because word embedding is more typically used as an input layer in deep networks, we selected the deep learning architectures BiLSTM and CNN for sentiment categorization. To achieve these goals, the tests were run on several benchmark datasets, including HARD, Khooli, AJGT, ArSAS, and ASTD.Finally, a comparative analysis was performed on the experimental model results.Our results.Our findings reveal that the produced embedding using one technique outperforms the pretrained version of the same technique by 0.28 to 1.8% accuracy, 0.33 to 2.17% precision, and 0.44 to 2% recall. Furthermore, the contextualized transformer based embedding model BERT outperformed both pretrained and trained versions. Furthermore, the results suggest that BiLSTM outperforms CNN by around 2% in three datasets: HARD, Khooli, and ArSAS, whereas CNN performed 2% better in two smaller datasets: AJGT and ASTD. Yagi S., Elnagar A.,and Fareh S[15]. To validate the suggested benchmark, we created a set of embedding models from various textual sources. Then we assessed them both intrinsically using the specified benchmark and extrinsically with two natural language processing tasks: Arabic Named Entity Recognition and Text Classification. The study concludes that the suggested benchmark accurately reflects this morphologically diverse language and dis-criminates against word embeddings.

3. THEORY

One of the contributing elements to this diversity is a disagreement in grammatical laws, which results in discrepancies in Quran parsing. For example, some people interpret specific words based on linguistic reasoning rather than the underlying narrative. In the Quran, certain letters are treated differently, such as the disjointed letters (Muqat-taʿat) at the beginning of chapters (Surah) or those impacted by diacritical markings (Fathah, Dammah, Kasrah). As a result, these dis-parities in parsing and grammatical orientations emerge [5].

3.1 Embedding

In machine learning, the idea of embedding refers to the conversion of categorical variables or text input into numerical representations known as embeddings. These embeddings successfully capture the semantic meaning or relationships that exist among the data pieces. When working with categorical variables like gender or job titles, embeddings help to represent each category as a vector. These vectors are obtained by training on data, and it has been discovered that categories with similar or related characteristics have similar vector representations. Text data, such as sentences or documents, uses embeddings to encode the meaning of words or phrases. Each word in the text is represented by a fixed-length vector. Words with similar semantic meanings tend to have equivalent vector representations. Embeddings for text data can be pre-trained on large corpora using approaches such as Word2Vec, GloVe, or BERT, or they can be trained for a specific job or domain. Embeddings are useful in machine learning because they may turn categorical or text input into a format that machine learning algorithms can easily understand and process. They help algorithms under-stand patterns and associations between data pieces and are widely used in activities such as natural language processing, recommender systems, and sentiment analysis.

4. METHODOLOGY

4.1 Dataset Collection

The data was taken from the website https://corpus.quran.com/ [16] and was modified and rearranged according to each Arabic syllable in the word, and the number of rows became 128,221 rows for the first group and 99540 rows for the second group.

4.2 Preprocessing

The data was taken and modified, and 26 different grammatical cases were taken, as shown in Table 1. especially the parsing of verbs and nouns in terms of nominative, accusative, jussive, and genitive. And dividing the Holy Qur'an into words according to the grammatical cases—26 different cases. Converting parsing cases into 160-bit codes to represent each case. The database is divided into input and output, where the number of entries was 7 cases, and the output is the next case (the eighth case), as shown in Fig. 1 shifting the states by one, meaning that the seven new states start from the eighth state and the previous output state is added. It is an input to the new state and continues until the end of the data set that showing in Fig 2.

TABLE I. CODING 26 PARSING CASES

LOCATION	FORM	TAG			PERF	ROOT:	3MP	SUFFIX	PRON:3 MP	POS:N	M	INDEF	NOM	P	مرفوع		
(1:1:1:1)	bi	P	PREFIX|bi+	بِ	0	0	0	0	0	0	0	0	0				
(1:1:1:2)	somi	N	STEM|POS:N|LEM:{som|ROOT:smw|M|GEN	سْمِ	0	1	0	0	0	1	1	0	0				
(1:1:2:1)	{ll^ahi	PN	STEM|POS:PN|LEM:{ll^ah|ROOT:Alh|GEN	ٱللَّهِ	0	1	0	0	0	0	0	0	0	1	0	0	
(1:1:3:1)	{l	DET	PREFIX|Al+	ال	0	0	0	0	0	0	0	0	0	0	1	0	
(1:1:3:2)	r^aHoma`n	ADJ	STEM|POS:ADJ|LEM:r^aHoma`n|ROOT:rHm|MS|GEN	رَّحْمَٰنِ	0	1	0	0	0	0	0	0	0	0	0	0	
(1:1:4:1)	{l	DET	PREFIX|Al+	ال	0	0	0	0	0	0	0	0	0	0	1	0	
(1:1:4:2)	r^aHiymi	ADJ	STEM|POS:ADJ|LEM:r^aHiym|ROOT:rHm|MS|GEN	رَّحِيمِ	0	1	0	0	0	0	0	0	0	0	0	0	
(1:2:1:1)	{lo	DET	PREFIX|Al+	ال	0	0	0	0	0	0	0	0	0	0	1	0	
(1:2:1:2)	Hamodu	N	STEM|POS:N|LEM:Hamod|ROOT:Hmd|M|NOM	حَمْدُ	0	1	0	0	0	1	1	0	1	0	0	0	
(1:2:2:1)	li	P	PREFIX|l:P+	لِ	0	0	0	0	0	0	0	0	0	0	1	0	

TABLE II. DISTRIBUTING THE DATA INTO SEVEN INPUTS AND ONE OUTPUT

	input							output
	1	2	3	4	5	6	7	8
بِ	2	92	46	1	42	1	42	1
سْمِ	92	46	1	42	1	42	1	95
اللَّهِ	46	1	42	1	42	1	95	1
ال	1	42	1	42	1	95	1	46
رَّحْمَٰنِ	42	1	42	1	95	1	46	92
ال	1	42	1	95	1	46	92	1
رَّحِيمِ	42	1	95	1	46	92	1	138
ال	1	95	1	46	92	1	138	1
حَمْدُ	95	1	46	92	1	138	1	42
لِ	1	46	92	1	138	1	42	1
لَّهِ	46	92	1	138	1	42	1	42
رَبِّ	92	1	138	1	42	1	42	92
الْ	1	138	1	42	1	42	92	92
عَٰلَمِينَ	138	1	42	1	42	92	92	1
ال	1	42	1	42	92	92	1	92
رَّحْمَٰنِ	42	1	42	92	92	1	92	15
ال	1	42	92	92	1	92	15	104
رَّحِيمِ	42	92	92	1	92	15	104	4

4.3 Dataset Splitting

The dataset was entered into two groups: the first group has 128211 rows and is divided into training data and testing data, with seven different divisions for the purpose of training and evaluating the model. The divisions were as in Fig. 3. The second group has 99540 rows and is divided into training data and testing data, with seven different divisions for the purpose of training and evaluating the model. These divisions were as in Fig. 4.

TABLE III. DIVISION FIRST GROUP 128211 ROWS

Experiment Code	Training Rows	Testing Rows
A	3000	1000
B	3000	125000
C	10000	3000
D	30000	10000
E	30000	98000
F	90000	38000

TABLE IV. DIVISION FIRST GROUP 99540 ROWS

Experiment code	Training Rows	Testing Rows
A	3000	1000
B	3000	96000
C	10000	3000
D	30000	10000
E	30000	69000
F	90000	9000

The data set is evaluated, and a different percentage is chosen for each training case of the model after selecting the cases randomly. Training The model has two layers, where the first layer consists of a different set of nodes and the second layer consists of 160 nodes for each case. Two types of training were used. The Holy Quran in its entirety.

4.4 Model Used

There were two sets of data used: the full Holy Qur'an (128,221 rows) and the entire Holy Qur'an without prepositions and definiteness (99,540 rows). The model utilized for the two groups consisted of two layers of models. The accurate calculation of the time spent in execution and the influence of these letters on the process of establishing accuracy in the model requires taking into account both the time required for implementation and the impact of these letters on the model. Specifically, we use the sequential model, which has two distinct levels. The first layer contains 16000 units and 1120 inputs and employs the ReLU activation algorithm. The second layer, on the other hand, contains 160 units and uses the SoftMax activation algorithm. To compile the model, use the model Compile' function. Furthermore, we use the fit model method for additional optimization and the model evaluate function to assess its performance. Finally, we save the model as the last step.

5. EXPERIMENTS AND RESULTS

The experiments took several shapes, depending on the data entered and the number of units entered in the functions, as well as the partition of the input data into training and testing the data; the results were obtained as shown in Fig. 5, Fig. 6, Fig. 7 and Fig.8 show the results as well.

TABLE V. RESULT WITH 128221 FOR 16000 NODES

Experiment code	Accuracy	Time
A	0.6903	80.4 S
B	0.6953	568 S
C	0.8663	268.9 S
D	0.9467	843.8 S
E	0.9504	1176.3 S
F	0.9484	2648.8 S

TABLE VI. RESULT WITH 99540 ROWS FOR 16000 NODES

Experiment code	Accuracy	Time
A	0.6817	87.5 S
B	0.6907	454.5 S
C	0.8539	333.2 S
D	0.9551	913.3 S
E	0.9542	1087 S
F	0.9624	2431.8 S

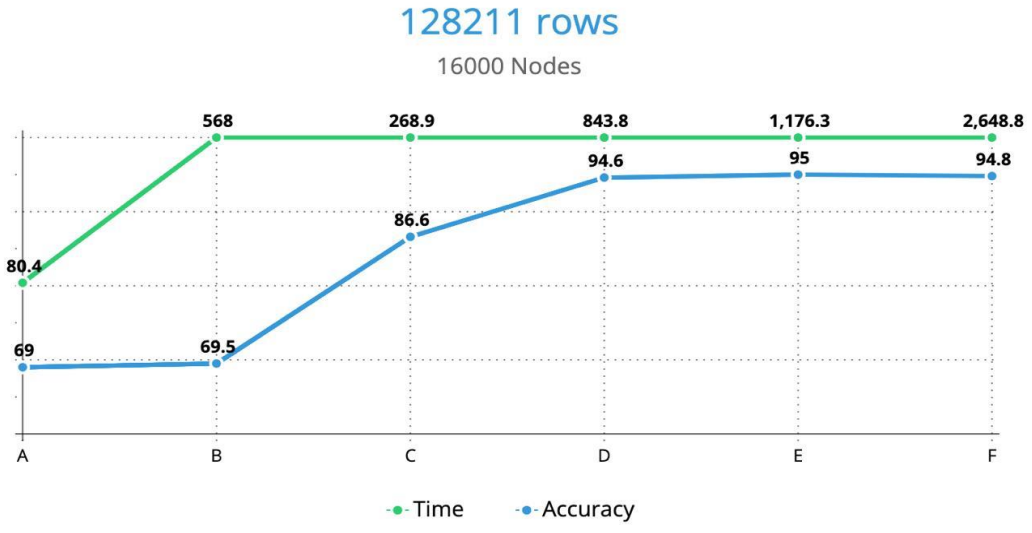

Fig. 1. Result chart for 128221 Rows

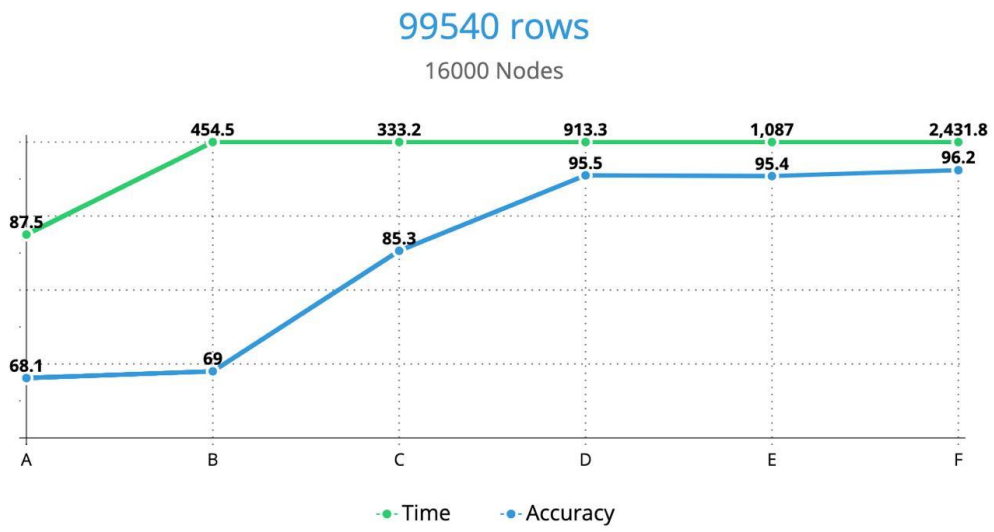

Fig .2. Result chart for 97957 Rows

6. CONCLUSION AND FUTURE WORK

We used two unique sets of datasets. The first group comprised an entered dataset with 128,221 rows, whereas the second group had 99.540 rows. We ran seven different experiments within each group and achieved a 96% accuracy rate. This accuracy was reached by using a training set size of 90,000 rows and a testing set size of 9,000 rows. Furthermore, we achieved 94% accuracy within the broader dataset of 128,221 rows by employing a training set size of 90,000 rows and a testing set size of 38,000 rows.

Conflicts Of Interest

The author declares no conflict of interest in relation to the research presented in the paper.

Funding

No grant or sponsorship is mentioned in the paper, suggesting that the author received no financial assistance.

Acknowledgment

The author extends gratitude to the institution for fostering a collaborative atmosphere that enhanced the quality of this research.

References

[1] M. H. Bashir, A. M. Azmi, H. Nawaz, W. Zaghouani, M. Diab, A. Al-Fuqaha, and J. Qadir, "Arabic natural language processing for Qur'anic research: A systematic review," Artificial Intelligence Review, vol. 56, no. 7, pp. 6801-6854, 2023.

[2] A. M. Alargrami and M. M. Eljazzar, "Imam: Word Embedding Model for Islamic Arabic NLP," in 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Oct. 2020, pp. 520-524.

[3] Z. Touati-Hamad, M. Laouar, and I. Bendib, "Authentication of Quran Verses Sequences Using Deep Learning," in Proceedings of the International Conference on Recent Advances in Mathematics and Informatics, Tebessa, 2021, pp. 1-4.

[4] S. Pudaruth, S. Soyjaudah, and R. Gunputh, "Classification of Legislations using Deep Learning," The International Arab Journal of Information Technology, vol. 18, no. 5, pp. 651-662, 2021.

[5] S. A. Khadhim and H. W. Chiad, "Parsing Facets of Letters Inflicted by the Three Diacritics in the Holy Quran," resmilitaris, vol. 12, no. 2, pp. 4092-4101, 2022.

[6] F. Beirade, A. Hamid, and D. E. Zegour, "Semantic query for Quranic ontology," Journal of King Saud University-Computer and Information Sciences, vol. 33, no. 6, pp. 753-760, 2021.

[7] M. Muhtadi Khazani, M. Mohamad, et al., "A Framework for Semantic Knowledge Representation of Al-Quran Based on Word Dependencies," in 2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP), 2021
[8] M. Beseiso, A. R. Ahmad, and R. Ismail, "An Arabic language framework for semantic web," in 2011 International Conference on Semantic Technology and Information Retrieval, June 2011, pp. 7-11.
[9] M. Poprat and D. Jena, "Building a biowordnet by using wordnet's data formats and wordnet's software infrastructure—A failure story," Softw. Eng. Test. Qual. Assurance Natural Lang. Process., vol. 4, pp. 31-39, 2008.
[10] M. T. B. Othman, M. A. Al-Hagery, and Y. M. E. Hashemi, "Arabic Text Processing Model: Verbs Roots and Conjugation Automation," IEEE Access, vol. 8, pp. 103913-103923, 2020. doi: 10.1109/ACCESS.2020.2999259
[11] M. Biltawi, A. Awajan, and S. Tedmori, "Towards building a frame-based ontology for the Arabic language," learning, 2017.
[12] M. Naili, A. Habacha Chaibi, and H. Hajjami Ben Ghezala, "Comparative study of word embedding methods in topic segmentation," Procedia Computer Science, vol. 112, pp. 340-349, 2017.
[13] C. A. Z. Roumahli, J. F. Aldana Montes, and M. Rybinski, "Comparative study of Arabic word embeddings: evaluation and application," International Journal of Computer Information Systems and Industrial Management Applications, vol. 12, pp. 14-14, 2020.
[14] S. F. Sabbeh and H. A. Fasihuddin, "A Comparative Analysis of Word Embedding and Deep Learning for Arabic Sentiment Classification," Electronics, vol. 12, no. 6, 1425, 2023.
[15] S. E. Yagi, A. Elnagar, and S. Fareh, "A benchmark for evaluating Arabic word embedding models," Natural Language Engineering, vol. 29, no. 4, pp. 978-1003, 2023.
[16] K. Dukes and N. Habash, "Morphological Annotation of Quranic Arabic," in Lrec, May 2010, pp. 2530-2536.