Journal Title
XX(X):1–7

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Machine Learning Based Classification of Qur'anic Verses into Makki and Madani

Enas Fadhil Abdullah[1] and Alyaa Abdulhussein Al-Joda[2]

Abstract

The Holy Quran has about six thousand verses from various historical and cultural contexts, making classification difficult in computational linguistics and Quranic studies. This research aims to classify Quranic verses into two types: Meccan and Medinan. Meccan verses typically focus on matters of faith, while Medinan verses address social organization and community governance. The research utilizes a dataset of 6,236 Quranic verses. A hybrid text processing methodology was applied, beginning with data cleaning and the use of the Arabic linguistic root (ISRI Stemmer) to reduce semantic dimensions. This was followed by feature extraction using the TF-IDF statistical weighting system, combined with structural features such as verse length, to ensure the highest classification efficiency. The SelectKBest feature selection algorithm was used to select the top 500 linguistic features, and standard scaling was applied. An ensemble voting classifier was constructed, integrating three different logic algorithms: Random Forest (RF), Decision Trees (DT), and k-Nearest Neighbors (KNN). The system assessment employed the GroupKFold cross-validation test to assure results reliability and prevent data leaking. The results showed that integrating linguistic and structural variables with ensemble models can automatically analyze and classify Qur'anic text with 82.33% accuracy.

Keywords

The Holy Quran, Meccan and Medinan, Machine Learning (ML), Group Classification, Text Classification, Natural Language Processing (NLP).

Introduction

Apart from being a book of religion, Muhammad's Qur'an is divine revelation. In formal Arabic, it is a linguistic and intellectual masterpiece Pitchay and Ridzuan (2016). Each of its 114 chapters and almost 6,000 poems is precise, eloquent, and profound. Islamic scholarship divides Qur'an verses into Makki and Madani based on revelation time and place Safeena and Kammani (2013). In addition, the Makki verses, presented before the Prophet left for Madinah, emphasize monotheism and faith.

Afterward, Madani poetry addressed legislation and social conduct Kurniawan and Adebisi (2022). Considering certain chapters have Makki and Madani passages with stylistic overlaps, categorizing verses is difficult. Performing manual classification is difficult Hasan and Mamun (2022). Despite the Qur'an's unique orthography and morphology, root-based derivations, lack of capitalization, complex diacritic system, and poetic rhythm, automated and data-driven classification methods can help Al-Shidi *et al.* (2025). Further, by applying diverse methods, computational linguistics has made major contributions to Qur'anic studies in recent years, including NLP for grammatical analysis, semantic search, ontology development, and text classification Ibrahim and Abdullah (2023).

However, Arabic NLP remains about a decade behind English NLP in terms of related resources and tools, due to the complexity of Arabic syntax and the scarcity of Qur'anic corpora Abdullah *et al.* (2021). Lately, Machine learning (ML) methods and algorithms have been adopted in many fields, and automatic Qur'anic verse classification is one of the most significant, due to its demonstration of high efficiency in text categorization tasks because they require moderate data volumes, offer interpretability, and maintain strong performance on small, imbalanced datasets Khan *et al.* (2010). In this paper, the Qur'anic verses are classified into Makki and Madani categories using three classical ML algorithms, including KNN, DT, and RF, which are implemented on the features extracted from the Arabic text using TF-IDF representation after performing a series of preprocessing operations.

This paper sections are organized as follow: in Section 2 the related literature s about classifying Qur'anverses are presented, in Section 3 the utilized dataset, and the proposed methodology steps are described, in Section 4, all the evaluation results from the three ML algorithms for classifying the Qur'an verses are shown along with a full discussion, finally Section 5 show the reached conclusions and future work.

[1]Department of Computer Sciences, College of Education for Girls, Kufa University, Najaf, Iraq, inasf.alturky@uokufa.edu.iq, safaad.alkhafaji@uokufa.edu.iq
[2]Department of Communication Engineering, Engineering Technical College of Najaf, Al-Furat Al-Awsat Technical University, Najaf, Iraq, dr.alyaa@atu.edu.iq

Corresponding author:
Department of Computer Sciences, College of Education for Girls, Kufa University, Najaf, Iraq.
Email: inasf.alturky@uokufa.edu.iq

Related Works

Many recent research have used machine learning and text mining to classify Quranic verses. These works aimed to bridge Islamic experts' manual annotation and current intelligence approaches to improve Quran retrieval and comprehension. In this section of the thesis, some of the most relevant works regarding classifying Qur'an verses are presented. In Nassourou (2012) presented a study to categorize the Quranic chapters according to the major phases of the messenger ship of Prophet Mohammad's using ML algorithms using a dataset of 114 chapters of Quran separated into two sets: Meccan and Medinan after applying a preprocessing variant steps, two classifiers were applied including Support Vector Machine (SVM) a custom Naïve Bayes, showing a precision equal to 90% and 69% in sequence. A limitation of this work is its reliance on a small, labeled dataset.

In Adeleke *et al.* (2017), the automation of Quranic verse labeling was performed using text classification algorithms to distinguish verses related to "Shahadah" (faith) from those related to "Prayers". Two feature selection algorithms were implemented for filtering, and classification was then performed using KNN, SVM, and NB, achieving accuracy rates of 77.5%, 71.25%, and 86.3%, respectively. The study demonstrated the potential of feature selection in improving Quranic text classification. However, it was limited by a small dataset (only two topics and two hundred verses) and a simple feature representation (term frequency).

In Adeleke *et al.* (2018a), another selection of features, the Group-Based approach, is implemented to improve the classification of Quranic verses from multiple sources, in which the employed datasets are composed of 451 verses from Surah Al-Baqarah and Surah Al-Anaam, categorized into three topics and gathered from two sources. The text was preprocessed, and the features were extracted before classification. Five feature selection algorithms were evaluated, including Information Gain, Pearson Correlation Coefficient, Chi-square, ReliefF, and Correlation-based Feature Selection (CFS), and the results were classified using NB, LibSVM, KNN, and J48. The highest accuracy obtained with LibSVM and CFS, 94.4%, demonstrated the high discrimination capability. Despite achieving high accuracy, it has high computational demands and generalizes poorly on small, expanded datasets.

In Adeleke *et al.* (2019a), the goal was to label Quranic verses from Surah Al-Baqarah (286 verses) into three topics, using three datasets derived from the English translation. Also, the features are selected in this work using two techniques: Information Gain (IG) and Chi-square (CH), after preprocessing and feature extraction. Finally, classification was conducted using SVM, NB, J48, and KNN, and NB achieved the highest result at 93.9%, but the main limitation of this study is its focus on only one chapter.

In Adeleke and Samsudin (2018), a hybrid feature selection technique (IG-CFS) has been implemented to reduce computational cost by combining the strengths of IG and CFS. The experiments were conducted on 451 verses from Surah Al-Baqarah and Surah Al-Anaam. Each verse was represented using TF-IDF weighting, and the data were prepared in three versions: translation only (QTrans), tafsir only (QTaf), and combined translation + tafsir (QTrans+Taf). For classification, NB, LibSVM, KNN, and J48 are used, and the highest accuracy was achieved with LibSVM. The restriction in this work is the use of a small dataset (two chapters only).

In Adeleke *et al.* (2019b), which implements the chi-square instead of IG along with CFS and uses it as a hybrid feature selection, three ML algorithms were implemented for classification: NB, SVM, and J48. The SVM achieved the best accuracy of 93.6%.

In Arkok and Zeki (2020), the focus was on classifying Quranic verses with unequal class distributions, where the SMOTE technique was applied before classification. Two datasets were used: one for Tawheed (962 verses) and Shirk (118 verses), and another for Meccan (86 chapters) and Medinan (28 chapters) which are classified after representing the features by TF using NB, Bayes Net, LibSVM, KNN, J48, and RF and the best results achieved from RF and NB that was equal to 96%.

In Adeleke *et al.* (2021), the authors presented a multi-label classification (MLC) for labeling the Quranic verses into multiple overlapping categories, such as Faith (Iman) and Worship (Ibadah). The features are extracted first using TF-IDF, and three ML models are used, including SVM, NB, KNN, and J48, evaluated on a dataset of 1,098 English-translated verses collected from Abdullah Yusuf Ali's translation and Ibn Kathir's tafsir, achieving the highest accuracy of 86% with SVM.

In Mohamed and El-Behaidy (2021), semantic word embeddings (Word2Vec) were employed to achieve MLC of verses using a system with four phases: preprocessing, feature extraction, classification, and voting. For classification, three ML classifiers are used: Logistic Regression (LR), SVM, and RF, and an ensemble voting module is subsequently implemented to select the most probable topic for each verse. The interest was on Surah Al-Baqarah, including 286 verses, with a precision rate of 75% using Word2Vec embeddings.

In Choirulfikri *et al.* (2022), an MLC system for Quranic verses was proposed using a dataset composed of over six thousand English-translated Quranic verses, labeled under 15 topics, and an ensemble of NBs to assign each verse to multiple thematic topics simultaneously. The classification was performed using four NB variants, including Gaussian (GNB), Multinomial (MNB), Complement (CNB), and Bernoulli (BNB), along with a majority voting ensemble, showing the best performance from the Bernoulli NB model, with a Hamming Loss of 0.1175.

Taking into account the previously developed related literature and its limitations, the main contribution of the proposed system is the following:

- Develop a text classification system for Qur'anic verses to distinguish between Makki and Madani verses using machine learning.
- Design an effective preprocessing framework that includes normalization, stop word removal, and stemming to improve the quality and discriminative power of Qur'anic text features.
- Contrast and assess RF, DT, and KNN ML classifiers to identify the best Qur'anic verse classification method.

Proposed Methodology

The classification of verses from the Holy Quran into Meccan and Medinan is accomplished through the employment of a complex hybrid methodology that applies semantic analysis of linguistic origins and the structure of the Quranic text. The categorization process is comprised of several interconnected steps, which are designed to maximize accuracy and scientific reliability. In order to clean up the raw text, the linguistic rooting process comes after the removal of diacritical marks and the standardization of Arabic script. In this phase, terms are traced back to their triliteral roots, which helps to reduce the size of the data set and standardize the semantics of the Quranic vocabulary.

A statistically significant numerical representation of the text was also created by using TF-IDF to capture word and binary expression weights (N-grams). This was done in order to create a numerical representation of the text. It is important to note that the methodology did not merely focus on the linguistic aspect; rather, it also included an additional component known as "verse length" (Word Count...).

As a result of its close connection to the stylistic patterns of both Meccan and Medinan revelations, as well as in order to improve the effectiveness of the models and circumvent the issue of "multidimensionality," the SelectKBest test, which is based on the Chi-Square distribution, was utilized in order to select the top 500 linguistic features that have a strong statistical correlation to the target group. In addition, standard scaling was utilized in order to guarantee that the impact of numerical and linguistic characteristics was equally distributed. This methodology adopted a Voting Ensemble system, which integrates three algorithms with different working philosophies: RF, DT, and KNN.

This is in contrast to traditional studies, which were based on a single model. On the basis of the average confidence probabilities that were generated by the three models, the ultimate decision was rendered. An alternative cross-validation mechanism that is based on GroupKFold was substituted for the standard partitioning in order to guarantee the generalizability of the system. Eighty percent of the data was used for training, and twenty percent was used for testing. This ensured that there was no interference between the two groups, which contributed to the results having a high level of scientific credibility. Figure 1 provides a clear illustration of the order in which these processes occur.

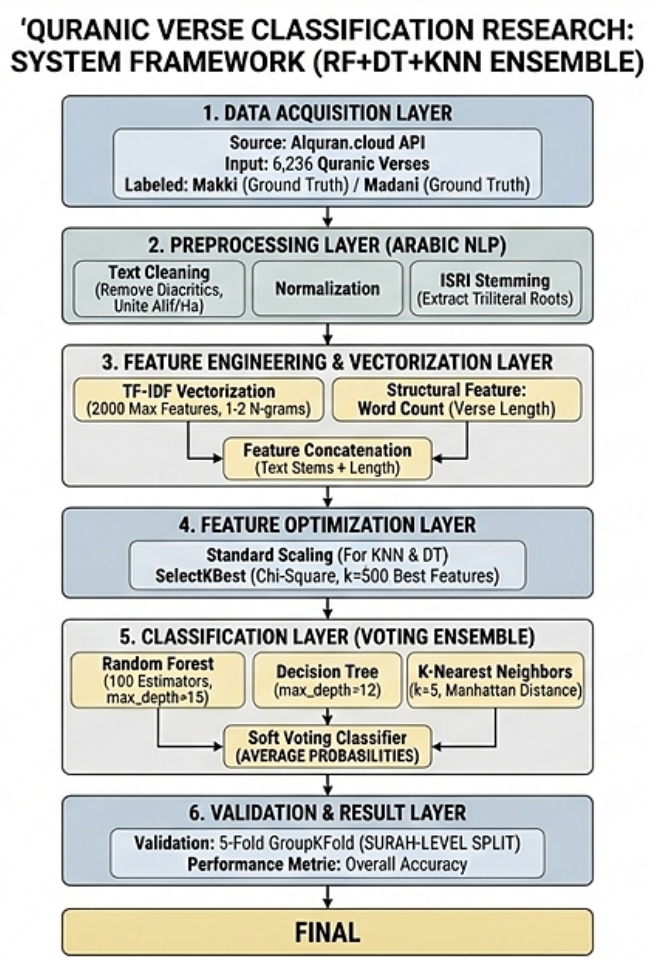

Figure 1. the proposed classification system for the Quran verses.

Dataset Description

A single dataset, Alquran.cloud API, was used, one of the most reliable digital sources for the Holy Quran. The text used is simple "original spelling" (without complex Tajweed diacritics) to facilitate computer processing and contains 6,236 verses, which is the actual number of verses in the Holy Quran.

Preprocessing

It is considered a fundamental procedure when the goal is to classify the Quranic text, as it prepares the raw textual data for effective feature extraction and final classification Alhaidar and Aliwy (2022); Abdullah *et al.* (2021). Before computational study, the Quran's traditional Arabic morphology and linguistic complexity must be standardized. Three steps are performed to prepare the Quranic textual data in this paper: Normalization, stop-word removal, and Stemming.

Normalization It is the process of standardizing the textual data in the Quran to eliminate inconsistencies arising from different writing styles, diacritics, and special characters Kadhim *et al.* (2025a). In the Arabic language, the same letter can be shown using different orthographic forms; for instance, the letters *Alef with Hamza Above*, *Alef with Hamza Below*, and *Alef with Madda* are unified into the standard *Alef*, and *Taa Marbuta* is converted to *Haa* in some cases Saiegh-Haddad and Henkin-Roitfarb (2014). Additionally, diacritics including "fatha", "kasra", and "damma" are removed, since they represent vowel sounds not necessary in the semantic classification. By normalizing the data, all ambiguity is removed and the data are represented in a unified format, which is very crucial for Quranic datasets because morphological and orthographic variations may cause the same word to be treated as different tokens, thereby reducing classification accuracy. Normalization can be obtained using Equation (1) for the raw data T_r. Aliero *et al.* (2023).

$$T_n = f(T_r) \tag{1}$$

Stop-word Removal The term "Stop words" refers to the words that frequently appear and do not carry any semantic meaning, thus increasing the noise in the textual data Kadhim *et al.* (2025b). The Arabic language contains a set of these words, such as the prepositions "from", "in", "on", and "to", which are employed to link words together and do not contribute to identifying the thematic content of a verse in the Quran Namly *et al.* (2019). Consequently, removing these stop words enhances the dataset and includes only words of contextual meaning, thereby improving both the efficiency and the accuracy of feature extraction. Considering $W = \{w_1, w_2, \ldots, w_n\}$ is the set of all words in a verse, and S is the predefined list of stop words, then the cleaned set W' is defined as shown in Equation (2) El-Khair (2017).

$$W' = W - S = \{w_i \in W \mid w_i \notin S\} \tag{2}$$

Stemming It refers to the process of reducing inflected or derived words in Arabic to their roots, thereby illustrating the base meaning of a term Ameen *et al.* (2025). For example, the Arabic verb meaning "they write" is reduced to the root meaning "write", from which many derivations are formed, such as "I wrote" and "written". Likewise, the Arabic words meaning "believers" and "faith" are derived from the same root meaning "secure" Singh and Gupta (2016). The light stemming method is the most widely used and efficient approach for Arabic language word stemming, eliminating common prefixes and suffixes (e.g., the definite article, plural suffixes, and possessive suffixes) while preserving the core word meaning. The general representation for stemming is shown in Equation (3), for the word w, and both P and Suf are any removable prefixes and suffixes, respectively Abdullah *et al.* (2018).

$$S(w) = w - (P + Suf) \tag{3}$$

Feature Extraction TF-IDF

It is a weighting scheme employed to measure and assess the importance of a specific word related to a document in a collection of words, in which it is illustrated and depends on the number of appearances of a specific word and increases proportionally Wendland *et al.* (2021). When classifying textual data, TF-IDF is employed as a filtering scheme and has a high significance in reducing dimensionality by removing redundant and irrelevant words. The TF-IDF technique is a product of two statistical weighting methods, which are Term Frequency (TF) and Inverse Document Frequency (IDF), and can be obtained using Equation (4) for the document d_i, where the TF_{ij} is the number of words, and for the number of documents D, the df_i is the count of the words in D Das and Alphonse (2023).

$$w_{ij} = TF_{ij} \times \log\left(\frac{D}{df_i}\right) \tag{4}$$

Machine Learning-based Classification

The final stage of the proposed automatic classification system involves applying an ensemble learning model to differentiate between Meccan and Medinan verses. In this step, TF-IDF metrics and structural parameters like verse length (Word Count) form the system's key inputs. A soft voting classifier was created using three algorithms: RF (to manage complex linguistic structures and feature interactions), DT (to develop separating rules with the most influential keywords), and KNN (to categorize verses based on their geometric similarity in vector space) to maximize reliability and overcome algorithm weaknesses.

This research used 5-fold GroupKFold cross-validation instead of typical classification methods. Sectioning the data into fences treats them as independent blocks, allowing the model to be tested on fences it never encountered during training. Data leakage and automated saving are eliminated by this mechanism's accurate assessment of the system's generalization potential.

Decision Tree (DT) This classification method creates a hierarchical tree structure with internal nodes representing feature-based decision rules. In contrast, the leaf nodes refer to the class labels (Makki or Madani). The dataset is recursively partitioned to increase class purity, using the Gini Index or Information Gain (Entropy) as a splitting criterion. For the probability of a class in a subset, the entropy of a dataset is computed as shown in Equation (5) **?**.

$$E(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{5}$$

Random Forest (RF) It is a classification and regression method that aggregates a large number of trees for decision-making, where each tree is trained on a random subset of the dataset and features, and each tree votes for a class, with the majority vote determining the final classification Salman *et al.* (2024). It can be represented mathematically using Equation (6) for the class prediction of the tree for an input x, where T is the number of trees Chen *et al.* (2022).

$$H(x) = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \tag{6}$$

K-Nearest Neighbor (KNN) It is one of the popular instance-based learning algorithms widely employed in pattern recognition due to its simplicity, in which it classifies a verse depending on the majority class among its k closest samples in the feature space, and the object sample will be classified in a group with the largest number of same label neighbors, in which after computing distances, the algorithm selects the KNN and assigns the class most common among them Kadhim *et al.* (2025c). The similarity between two verses is typically computed using the Euclidean distance as shown in Equation (7) Ukey *et al.* (2023).

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{7}$$

Classification criteria

The methodology used in this research for classifying verses relies on integrating traditional criteria of Qur'anic sciences with automatically extracted statistical and grammatical characteristics. Four fundamental criteria were identified upon which the ensemble model was based to distinguish between Meccan and Medinan revelation. First, the structural criterion: Verse length is considered one of the strongest statistical criteria in this research. Programmatically, the

word count feature was used to represent this criterion. The quantitative analysis showed a consistent pattern associating the shortness of verse with a fast tempo in the Meccan phase, in contrast to the reduced length and detailed elaboration of the Medinan verses on legal matters. Second, the criterion based on semantics and stems: The research was based on triliteral roots as a main criterion.

Keywords related to each environment were extracted using ISRI Stemming technology. While the Meccan criteria included the issues of creed, warning and resurrection (such as the roots: punished, oath, paradise), the Medinan criteria included the vocabulary of rulings, obligations and transactions (like the roots: tunnel, wrote, usury). Third: The Discourse & N-grams Criterion. The system employed the N-grams technique to capture complex forms of discourse that are considered as an essential standard in Qur'anic sciences. This consisted of tracking vocative forms like "O you who believe" for Medinan and "O people" or "No" for Meccan, offering algorithms strong contextual discrimination. Fourth: The Statistical Weight Criterion; the particular weight of each word and expression was defined according to the TF-IDF criterion. This criterion is based on marginalizing general words (stop words) and focusing on words that possess "discriminative power." That is, the words that frequently appear in the Meccan category and rarely appear in the Medinan category, which transformed the Qur'anic text from descriptive words into precise geometric values in the vector space.

Results and Discussion

Systems for classifying the Quran verses are presented in this section. The implementation environment was run on a HP Pavilion laptop running Microsoft Windows 10 (64-bit) and an Intel Core™ i7-8750H processor with 16 GB of RAM. Initially, Initially, the following four tables list the top coefficients used in each of the three algorithms, in addition to the parameters of the Logistic Regression model, which we will use as a reference model for comparison.

Table 1. RF Hyperparameters

Parameter	Value
n_estimators	100
max_depth	15
min_samples_split	5
random_state	42
class_weight	Balanced

Table 2. DT Hyperparameters

Parameter	Value
Criterion	Entropy
max_depth	12
min_samples_leaf	4
Splitter	Best
random_state	42

Now we will test each algorithm individually, in addition to the Logistic Regression test, which serves as a comparative model. Then we will test our proposed model on the previously mentioned dataset. The following table shows the accuracy results for each model.

Table 3. KNN Hyperparameters

Parameter	Value
n_neighbors	5
weights	Distance
metric	Manhattan
algorithm	Auto
p	1

Table 4. Logistic Regression Hyperparameters

Parameter	Value
solver	Liblinear
penalty	l2
C	1.0
max_iter	100
random_state	42

Table 5. Comparing the results of machine learning algorithms

ML model	Accuracy
RF	79.90%
DT	80.02%
KNN	77.72%
LR	82.01%
Proposed	82.33%

The table above shows the superiority of the Logistic Regression model over the three algorithms when operating independently. It also shows the slight superiority of our proposed model over Logistic Regression, which is a positive finding and highlights the importance of the proposed model that integrates the three algorithms using a voting approach to leverage the advantages of each.

It should be noted that the slightly lower accuracy value we obtained, compared to the values in the reference studies we cited, is due to our use of the Group K Fold technique to prevent data leakage, while the higher accuracy values in some of the references we mentioned are due to the use of the traditional data partitioning method. This demonstrates the reliability and accuracy of our results.

Conclusion

This research developed an advanced intelligent system for classifying the verses of the Holy Quran into Medinan and Meccan verses employing ensemble ML techniques. The system has effectively overcome the limitations of traditional classifications, combining linguistic features obtained from roots (Stems) with structural characteristics like the length of a verse.

The results of the experiments demonstrated that the utilization of "model voting" (which involves RF, DT, and KNN) results in greater stability and accuracy than the utilization of individual models. The accuracy rate of the system is eighty-two-point three percent. This paper provides evidence that the digital analysis of the style of the Qur'anic discourse has the potential to be a trustworthy simulation of human deduction in the field of Qur'anic sciences. Furthermore, it demonstrates that this technique has the potential to open up new avenues for the digital service of the Qur'anic text.

This study uses Group K Fold technology in training to ensure outcomes reliability and prevent data leakage. Also, despite the good results, this area has huge growth potential. RNNs or text transformer models, like the BERT model for Arabic, are one of the most popular future research directions to capture deep semantic contexts that traditional algorithms can miss. In addition, AI was able to distinguish among Meccan (powerful and focused) and Medinan (legislative and detailed) discourses (82.33%).

References

Pitchay SA and Ridzuan F (2016) A systematic review analysis for Quran verses retrieval. *Journal of Engineering and Applied Sciences* 100(3): 629–634.

Safeena R and Kammani A (2013) Quranic computation: A review of research and application. In: *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, pp. 203–208.

Kurniawan MA and Adebisi A (2022) Ulumul Qur'an: Classification of Makkiyah Madaniah verses in the Qur'an. *Jurnal Ilmiah Az-Ziqri Kajian Keislaman dan Kependidikan* 1(3): 1–10.

Hasan R and Mamun FE (2022) Numerical statements of the Holy Quran based on verses, words, and letters in Makki and Madani Sura. *International Journal of Culture and Religious Studies* 2(2).

Al-Shidi A, Nordin FHB, Mohamed II, *et al.* (2025) A literature review on natural language processing techniques for Qur'anic studies: Challenges and insights. *Frontiers in Signal Processing* 5: 1535166.

Ibrahim R and Abdullah EF (2023) Leveraging sentiment analysis of drugs review-based drugs recommender system. In: *International Conference on Innovations in Data Analytics*, pp. 229–238. Singapore: Springer Nature Singapore.

Abdullah EF, Alasadi SA and Al-Joda AA (2021) Text mining based sentiment analysis using a novel deep learning approach. *International Journal of Nonlinear Analysis and Applications* 12(Special Issue): 595–604.

Khan A, Baharudin B, Lee LH, *et al.* (2010) A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology* 1(1): 4–20.

Nassourou M (2012) *Using Machine Learning Algorithms for Categorizing Quranic Chapters by Major Phases of Prophet Mohammad's Messengership*. Master's Thesis, Universität Würzburg, Germany.

Adeleke AO, Samsudin NA, Mustapha A, *et al.* (2017) Comparative analysis of text classification algorithms for automated labelling of Quranic verses. *International Journal of Advanced Science Engineering and Information Technology* 7(4): 1419.

Adeleke AO, Samsudin NA, Mustapha A, *et al.* (2018a) A group-based feature selection approach to improve classification of Holy Quran verses. In: *International Conference on Soft Computing and Data Mining*, pp. 282–297. Cham: Springer International Publishing.

Adeleke A, Samsudin N, Mustapha A, *et al.* (2019a) Automating Quranic verses labeling using machine learning approach. *Indonesian Journal of Electrical Engineering and Computer Science* 16(2): 925–931.

Adeleke A and Samsudin N (2018) A hybrid feature selection technique for classification of group-based Holy Quran verses. *International Journal of Engineering and Technology* 7(4.31): 228–233.

Adeleke A, Samsudin NA, Othman ZA, *et al.* (2019b) A two-step feature selection method for Quranic text classification. *Indonesian Journal of Electrical Engineering and Computer Science* 16(2): 730–736.

Arkok B and Zeki AM (2020) Classification of Holy Quran verses based on imbalanced learning. *International Journal of Islamic Applications in Computer Science and Technology* 8(2).

Adeleke A, Samsudin NA, Rahim MHA, *et al.* (2021) Multi-label classification approach for Quranic verses labeling. *Indonesian Journal of Electrical Engineering and Computer Science* 24(1): 484–490.

Mohamed EH and El-Behaidy WH (2021) An ensemble multi-label themes-based classification for Holy Qur'an verses using Word2Vec embedding. *Arabian Journal for Science and Engineering* 46(4): 3519–3529.

Choirulfikri MR, Lhaksamana KM and Al Faraby S (2022) A multi-label classification of Al-Quran verses using ensemble method and naïve Bayes. *Building of Informatics, Technology and Science* 3(4): 473–479.

Alhaidar HA and Aliwy AH (2022) Automatic extraction of semantic relation among verses of Quran's. *NeuroQuantology* 20(8): 7703–7714. DOI: 10.14704/nq.2022.20.8.NQ44796.

Abdullah EF, Lafta AA and Alasadi SA (2021) Information gain-based enhanced classification techniques. In: *Next Generation of Internet of Things: Proceedings of ICNGIoT 2021*, pp. 499–511. Singapore: Springer Singapore.

Kadhim SM, Paw JKS, Tak YC, *et al.* (2025a) Robust security system: A novel facial recognition optimization using coronavirus-inspired algorithm and machine learning. *Iraqi Journal of Computer Science and Mathematics* 6(2): 19.

Saiegh-Haddad E and Henkin-Roitfarb R (2014) The structure of Arabic language and orthography. In: *Handbook of Arabic Literacy: Insights and Perspectives*, pp. 3–28. Dordrecht: Springer Netherlands.

Aliero AA, Bashir SA, Aliyu HO, *et al.* (2023) Systematic review on text normalization techniques and its approach to non-standard words.

Kadhim S, Paw JKS, Tak YC, *et al.* (2025b) Deep learning for robust iris recognition: Introducing synchronized spatiotemporal linear discriminant model-iris.

Namly D, Bouzoubaa K, Tajmout R, *et al.* (2019) On Arabic stop-words: A comprehensive list and a dedicated morphological analyzer. In: *International Conference on Arabic Language Processing*, pp. 149–163. Cham: Springer International Publishing.

El-Khair IA (2017) Effects of stop words elimination for Arabic information retrieval: A comparative study. *arXiv preprint* arXiv:1702.01925.

Ameen ST, Yussof S, Ahmad A, *et al.* (2025) MB-ConvLSTM: A novel hybrid deep learning model for accurate sign language recognition. *International Journal of Web Information Systems*.

Singh J and Gupta V (2016) Text stemming: Approaches, applications, and challenges. *ACM Computing Surveys* 49(3): 1–46.

Abdullah EF, Al-Sultany GA and Nawaf HN (2018) Rough set-based context suggestions. *Journal of Theoretical and Applied Information Technology* 96(21).

Wendland A, Zenere M and Niemann J (2021) Introduction to text classification: Impact of stemming and comparing TF-IDF and count vectorization as feature extraction technique. In: *European Conference on Software Process Improvement*, pp. 289–300. Cham: Springer International Publishing.

Das M and Alphonse PJA (2023) A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset. *arXiv preprint* arXiv:2308.04037.

Costa VG and Pedreira CE (2023) Recent advances in decision trees: An updated survey. *Artificial Intelligence Review* 56(5): 4765–4800.

Salman HA, Kalakech A and Steiti A (2024) Random forest algorithm overview. *Babylonian Journal of Machine Learning* 2024: 69–79.

Chen H, Wu L, Chen J, *et al.* (2022) A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management* 59(2): 102798.

Kadhim S, Paw JKS, Tak YC, *et al.* (2025c) An optimized machine learning models by metaheuristic corona virus optimization algorithm for precise iris recognition. *Advances in Artificial Intelligence and Machine Learning* 5(1): 3389–3408.

Ukey N, Yang Z, Li B, *et al.* (2023) Survey on exact kNN queries over high-dimensional data space. *Sensors* 23(2): 629.