QURANICA, International Journal of Quranic Research,
Vol. 17, Issue. 1, June 2025, pp. 32-52

Multi-Dimensional NLP Analysis of the Quran: Structure, Semantics, and Emotion across Qirāʾāt [*]

Asmaa Bengueddach[1]

ABSTRACT

This study investigates the Quranic text through a multi-dimensional Natural Language Processing (NLP) framework that integrates structural, semantic, and emotional analysis. Drawing on both Arabic and translated corpora, we apply a combination of traditional statistical techniques (e.g., TF-IDF) and transformer-based models (e.g., AraBERT) to examine thematic distributions, linguistic variation, and affective tone across Surahs. Our analysis incorporates tokenization, frequency analysis, part-of-speech tagging, and emotion classification, supported by visualizations such as bar charts and word clouds. Part-of-speech tagging refers to labeling each word in the Quranic text with its grammatical role (e.g., noun, verb, adjective), which helps clarify the structure and meaning of verses. Results reveal distinct structural and emotional profiles between Meccan and Medinan Surahs, with shorter chapters exhibiting denser emotional content, and longer ones displaying greater thematic diversity. The inclusion of Qirāʾāt perspectives further highlights meaningful canonical variation often overlooked in computational studies. This work contributes to the emerging field of Quranic computational linguistics by offering an integrative approach that bridges traditional exegesis with modern NLP and lays the groundwork for future applications in tafsir automation, recitation analysis, and discourse modeling.

Keywords: *Quranic NLP, AraBERT, Qirāʾāt, Structural Variation, Thematic Classification, Sentiment Analysis.*

---

[*] This article was submitted on: 04/04/2025 and accepted for publication on: 25/06/2025.

[1] LIO laboratory, Department of Computer Science, University of Oran 1 Ahmed Ben Bella, Oran, Algeria
Email: asmaa.bengueddach@univ-oran1.dz.

1.0 Introduction

The Quran, a sacred text of Islam, holds great religious, linguistic, and cultural significance. Its complex structure, with rich morphology, varied syntax, and unique rhetorical features, makes it challenging for computational analysis. Traditionally, Quranic studies relied on manual interpretation and philological methods. However, recent advances in Natural Language Processing (NLP) now offer new ways to analyze this intricate text on a larger scale. Quranic Arabic, which differs from Modern Standard Arabic, contains older vocabulary and complex grammar that require specialized tools. Previous research has focused on lexical analysis, morphological tagging, and named entity recognition (Darwish & Mubarak, 2016; Dukes & Habash, 2010). Despite these efforts, areas such as identifying thematic relationships, semantic classification, and sentiment analysis remain underexplored.

This study aims to leverage state-of-the-art NLP techniques to enhance the computational analysis of Quranic texts. Specifically, we focus on improving thematic extraction, semantic classification, and emotional profiling using deep learning models. By doing so, we seek to bridge the gap between traditional philological approaches and modern computational methods.

To guide our study, we pose the following research questions:

1. To what extent can advanced NLP techniques uncover thematic structures and relationships in the Quranic text?

2. How effective are transformer-based models, such as AraBERT, in improving the semantic classification and retrieval of Quranic verses?

3. What insights does sentiment analysis provide regarding the emotional tone and rhetorical strategies within the Quran?

Our research contributes by combining traditional statistical methods with deep learning models, particularly transformer-based architectures like AraBERT, to improve Quranic text analysis. Specifically, we focus on:

- Using TF-IDF analysis to extract thematic relationships between verses.

- Applying transformer-based models (such as AraBERT) to improve semantic classification and verse retrieval.

- Using sentiment analysis to classify verses as neutral, positive, or negative, thereby providing insights into the emotional tone of the Quran.

These contributions aim to bridge the gap between traditional exegesis and computational methods by improving the understanding of the Quran's linguistic features and providing deeper insights into its meaning.

2. Related Work

Early work on Quranic Arabic focused mainly on basic language processing tasks like word analysis, grammar tagging, and sentence structure. The Quranic Arabic Corpus (Dukes & Habash, 2010) is one of the first important projects. It gave researchers a detailed, manually checked dataset to train and test models. Other tools, like Farasa and rule-based parsers (Al-Khalifa & Al-Salman, 2011), helped with Arabic grammar. Darwish and Mubarak (2016) also added useful tools for understanding Arabic word forms. However, all these works focus mostly on the surface of the language, and they do not deal with deeper meaning or emotion.

With the rise of deep learning, new models like AraBERT and QuranBERT (Antoun et al., 2020) have improved performance in tasks such as classifying texts, answering questions, and recognizing names or concepts. These models use advanced embeddings that understand context better, especially for Arabic. But applying them to Quranic texts is still difficult because of limited labeled data, complex grammar, and the special nature of religious language. Also, these models are rarely used to explore how structure or emotion changes in the Quran.

Some researchers have started using hybrid systems that mix grammar rules with deep learning. For example, Al-Ayyoub et al. (2022) combined syntax and meaning in one model to better understand Quranic content. Shohoud et al. (2023) built a tool for searching verses by meaning. Bashir et al. (2022) worked on Tajweed and speech analysis for correct recitation. Hamed et al. (2025) looked at texts that switch between Arabic and other languages, which is common in real religious usage. Tariq et al. (2024) used a transformer model to build a question-answering system for Quranic verses in Urdu. These are interesting studies, but most focus on only one task. They do not connect structure, meaning, and emotion in a full analysis.

At the same time, Touati-Hamad et al. (2021–2022) studied how the order of words in verses could be changed. They used CNN, LSTM, and CNN-LSTM models and got good results in detecting these changes. But their goal was

to check the integrity of the text, not to understand its meaning or emotional message.

To summarize, we see three main areas in Quranic NLP: (1) grammar and structure, (2) meaning and themes, and (3) emotional or conceptual aspects. Most existing research focuses on just one of these areas. Very few combine all three. Also, most studies do not look at the differences between Qirāʾāt (canonical readings), which can affect both meaning and tone.

Our work fills this gap. We use TF-IDF to study themes, AraBERT to understand meaning, and sentiment analysis to detect emotions. We also include Qirāʾāt to show how different readings can change the interpretation. This gives a more complete view of the Quran and creates a bridge between traditional tafsīr and modern technology.

3. Method

Our approach combines modern Natural Language Processing (NLP) techniques with classical linguistic methods to improve Quranic text analysis. We use transformer-based models, such as AraBERT, along with statistical and rule-based approaches to perform key tasks: thematic classification, semantic analysis, and sentiment detection (see Figure 1).

3.1 Thematic Classification

Thematic classification is crucial for organizing Quranic content by topics. We begin with TF-IDF to extract initial features, then fine-tune AraBERT on a curated dataset of Quranic verses labeled with thematic categories. This allows the model to learn contextual embeddings that capture the semantic relationships between different chapters, leading to improved classification performance compared to traditional keyword-based methods.

2. Semantic Analysis

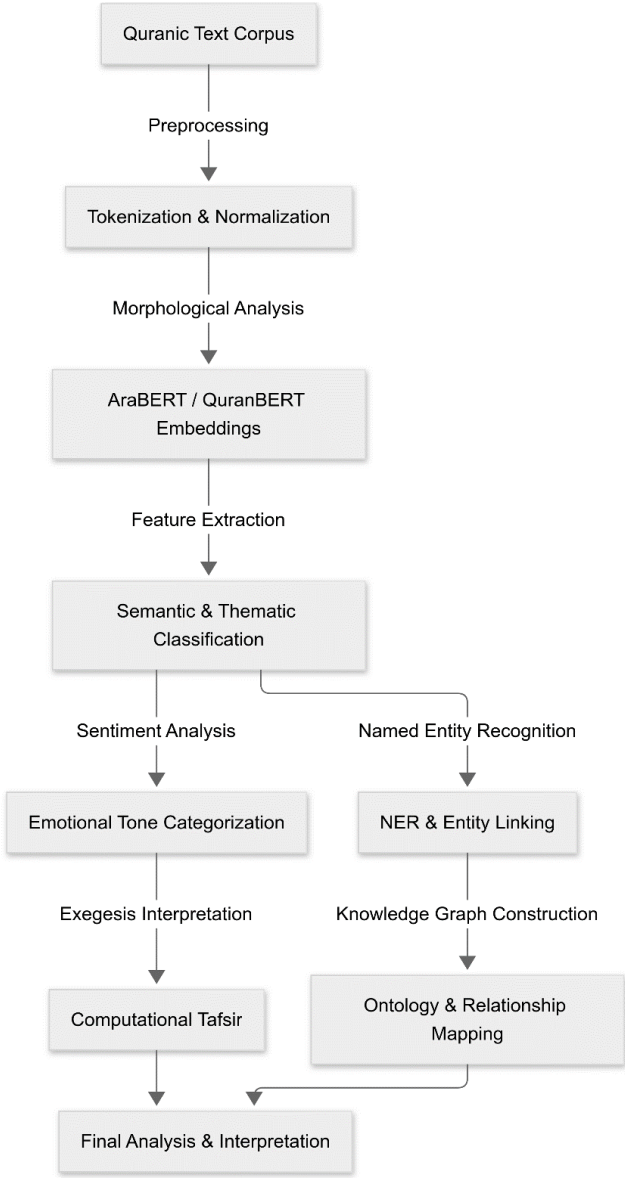
Understanding the meaning of Quranic text requires effective semantic modeling. We use a transformer-based model, fine-tuned specifically for classical Arabic, to detect verse similarity and identify paraphrases. Using a contrastive learning approach, semantically similar verses are positioned closer together in the embedding space, improving the accuracy of verse retrieval and interpretation.

3.3 Sentiment Detection

Sentiment analysis in Quranic discourse helps us understand the emotional tone in different sections. We adapt sentiment lexicons for classical Arabic and use AraBERT to classify verses as positive, negative, or neutral. This analysis is useful for exploring the variations in tone and rhetorical devices used across different contexts.

Figure 1

Quranic NLP Pipeline: From Preprocessing to Interpretation



The diagram (see Figure 1) illustrates our NLP pipeline, showing the sequential steps from data preprocessing to model application and output generation:

1. **Preprocessing:** Tokenization, normalization, stopword removal, and stemming. This preprocessing phase also includes part-of-speech (POS) tagging, which assigns a grammatical label (such as noun, verb, adjective) to each word. This step supports a deeper structural and semantic understanding of Quranic language.

1. **Feature Engineering:** TF-IDF extraction and embedding generation using AraBERT.

2. **Model Training:** Supervised learning with annotated datasets.

3. **Inference:** Thematic classification, semantic similarity analysis, and sentiment detection.

4. **Post-processing:** Visualization and exegesis interpretation.

While our current work does not include tasks like named entity recognition (NER), knowledge graph construction, or ontological mapping, these will be explored in future work to further enhance the comprehensiveness of the Quranic text analysis.

3.4 Experimental Approach

This study uses a publicly available dataset of the Holy Quran retrieved from Kaggle (Kaggle Quran Dataset, 2023). The dataset contains the central text for 1.5 billion Muslims worldwide and is regarded as a cornerstone of Arabic literature. It includes the full Quranic text, divided into 30 parts (Juz'), 114 chapters (Surahs), and over 6,000 verses (Ayahs).

1. Dataset Overview and Modifications

The raw dataset is typically provided as a CSV file, which experiences several transformations to prepare it for advanced NLP tasks. Initially, the data is loaded into a Pandas DataFrame, with separate columns for Surah, Ayah, and the corresponding Arabic text. A key modification is the addition of a 'Parah' column, categorizing each verse into one of the 30 sections of the Quran. This enhancement aids in detailed and granular analysis. Here is a sample of the modified DataFrame, demonstrating its structure:

Table 1 represents the structured dataset of Quranic verses, showing their Surah, Parah, Ayah, and text content.

Table 1.

Sample of Quranic Data

Data	Surah	Parah	Ayah	Text	Surah Name
0	1	1	1	بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ	1. Al-Fatihah
7	2	1	1	بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ الم	2. Al-Baqarah
293	3	3	1	بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ الم	3. Aali Imran
493	4	4	1	بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ يَا...	4. An-Nisa
...	...	...	...	...	...
6225	113	30	1	بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ قُلْ أَ...	113. Al-Falaq
6230	114	30	1	بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ قُلْ أَ...	114. An-Nas

Additionally, the dataset is enriched by linking the name of each Surah via its index. This provides contextual information that helps facilitate a deeper analysis of each verse.

Given that "بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ" (Bismillah al-Rahman al-Rahim), the Basmala, appears at the beginning of nearly all Surahs, we address its impact on the analysis. Since the Basmala's repetitive nature could bias certain textual analyses (e.g., word clouds or frequency analysis), we remove it during data preprocessing to avoid this bias and focus on the unique characteristics of each Surah.

4. Preliminary Results

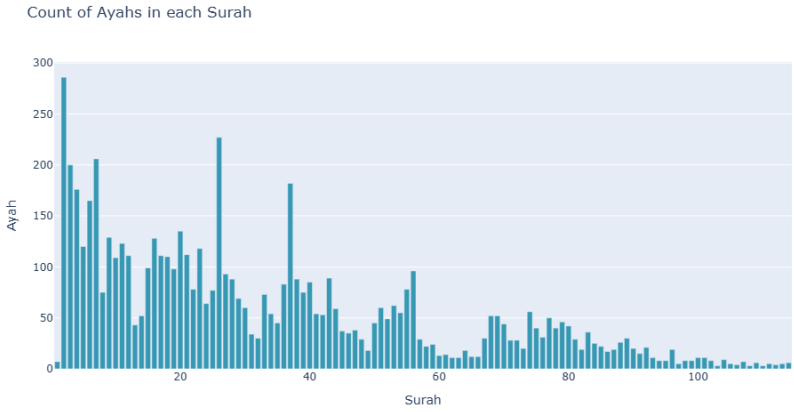
4.1 Distribution of Ayahs Across Surahs

The distribution of Ayahs (verses) across the 114 Surahs (chapters) of the Quran is an important feature of its structure. Analyzing this distribution helps in understanding the organization and thematic progression of the text. To visualize this, a bar chart (Figure 2) was created, showing the number of Ayahs in each Surah. The chart reveals significant variation in the number of Ayahs,

with early Surahs typically containing more verses, while those towards the end tend to have fewer. The distribution also exhibits noticeable peaks and troughs, indicating diverse structural patterns within the Quranic text.

Figure 2

Distribution of Ayahs Across the 114 Surahs of the Quran.



Surah 2 (Al-Baqarah) is the longest chapter, containing 286 Ayahs, followed by Surah 3 (Aali Imran) with 200 Ayahs, Surah 4 (An-Nisa) with 176 Ayahs, and Surah 5 (Al-Ma'idah) with 120 Ayahs. As we move towards Surahs 60 to 114, the number of Ayahs decreases significantly, with most chapters containing fewer than 50 Ayahs. The varying lengths of Surahs may reflect different contexts and purposes.

Longer Surahs, such as Al-Baqarah, tend to focus on detailed legal, historical, and theological narratives, while shorter Surahs often emphasize concise messages on morality, spirituality, and eschatology.

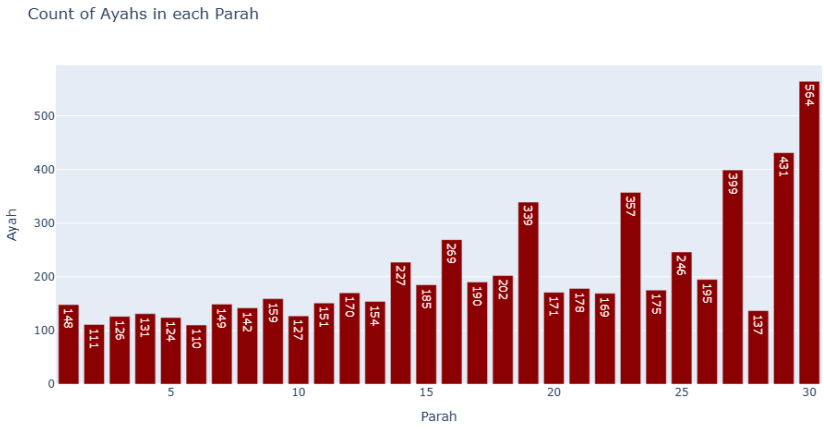
The longer Surahs, typically revealed in Medina, address issues like community organization, laws, and social matters. The shorter Surahs, generally from the Meccan period, emphasize core theological principles and moral teachings. This variation in Ayah length highlights the different thematic focuses and the span of revelation periods, offering insight into the historical and contextual development of the Quranic text.

4.2 Distribution of Ayahs Across Parahs

The distribution of Ayahs (verses) across the 30 Parahs (sections) of the Quran provides valuable insight into the text's structure and thematic development. To visualize this, a bar chart (Figure 3) was created, showing the number of Ayahs in each Parah.

Figure 3

Distribution of Ayahs Across the 30 Parahs of the Quran.



As seen in Figure 3, the number of Ayahs per Parah varies across the Quran. Early and middle Parahs tend to have fewer Ayahs, while the final Parahs contain significantly more. The variable lengths of the Parahs may reflect different contexts and thematic arrangements. The chart shows variable distribution in the number of Ayahs and emphasizes the changes in context over the book.

4.3 Comparative Analysis of Word Count Distribution: Longest vs. Shortest Surahs

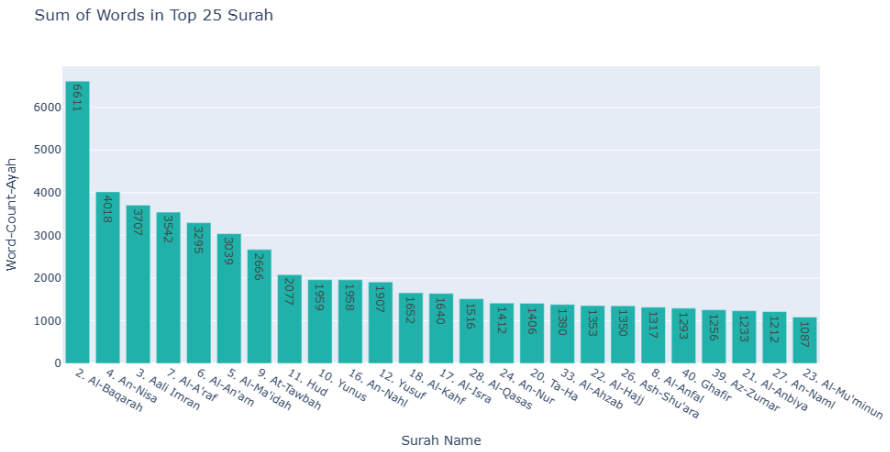
Analyzing the word counts in the Quran provides insights into the variable length and thematic emphasis of different Surahs (chapters). To explore this, we compare the distribution of word counts in the top 25 longest Surahs with that of the bottom 25 shortest Surahs. This comparative analysis helps highlight the diversity in narrative and thematic structures present in the Quranic text. As shown in Figure 4-a, the word counts in the top 25 Surahs vary significantly, ranging from 6611 words in Al-Baqarah (Surah 2) to 1087 words in Al-

Mu'minun (Surah 23). There's a steep decline from Al-Baqarah to the subsequent Surahs, with An-Nisa (Surah 4) having 4018 words and Aali Imran (Surah 3) having 3707 words. Figure 4.b illustrates that the word counts in the bottom 25 Surahs are substantially lower and range from approximately 10 words in Al-Kawthar (Surah 108) to 73 words in Al-A'la (Surah 87). The most striking difference is the scale of word counts. The longest Surah (Al-Baqarah) contains over 600 times more words than the shortest Surah (Al-Kawthar). The top 25 Surahs are generally much longer than the bottom 25. The longer Surahs often contain detailed narratives, legal discussions, and historical context, whereas the shorter Surahs tend to convey essential theological or moral messages concisely.

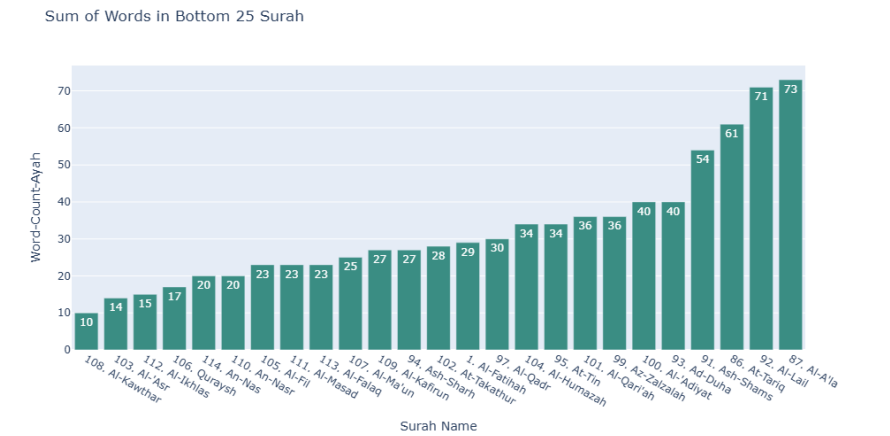
Figure 3

Comparative Word Count Distribution: Longest vs. Shortest Surahs of the Quran.

a.



b



The significant difference in word counts between the top and bottom Surahs reflects the diversity of content and purpose within the Quran. The longer Surahs typically address complex legal, social, and historical issues relevant to the developing Muslim community.

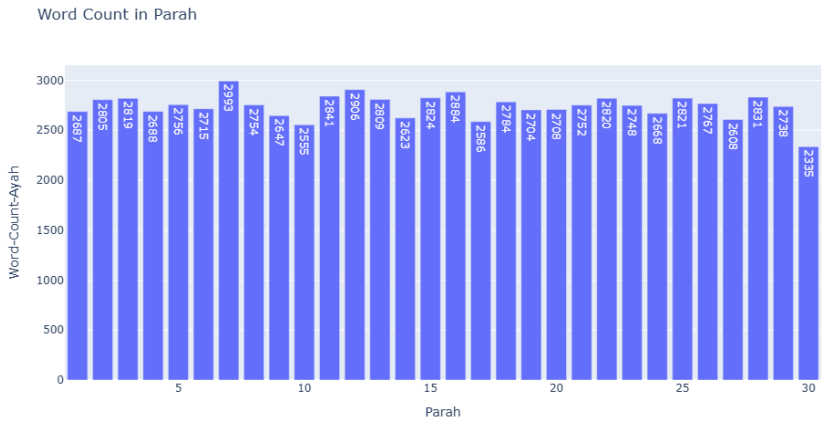
These detailed narratives and discussions provide comprehensive guidance and context. The shorter Surahs, often recited in daily prayers, serve as reminders of core Islamic beliefs and ethical principles, and their brevity facilitates memorization and reflection. The distribution of word counts underlines a dual approach: detailed elaboration in longer Surahs and concise reminders in shorter ones, together forming a comprehensive guide for believers.

4.4 Analysis of Word Count Distribution Across Parahs

Analyzing the distribution of word counts across the 30 Parahs (sections) of the Quran provides insights into how the length and content volume varies across these divisions. This can reveal patterns in the organization and thematic emphasis throughout the text. Figure 5 visualizes this distribution.

Figure 5

Word Count Distribution Across the 30 Parahs of the Quran.



The data displays the word counts per Parah, ranging from approximately 2335 to 2993 words. The distribution appears relatively uniform, with a few notable variations.

- **Relatively Uniform Distribution:** Unlike the distribution of word counts per Surah, the word counts per Parah are generally consistent. No Parah has a word count that drastically exceeds or falls below the others.

- **Peaks and Valleys:** The Parah with the highest word count is Parah 8, with 2993 words. Parah 30 has the lowest word count, with 2335 words.

The relatively consistent distribution observed in Figure 5 may suggest an intention to maintain a balanced length and content volume across each of the 30 Parahs. This division likely aids in structuring the Quran for daily or monthly reading schedules. While the variations could reflect thematic shifts or narrative structures within specific sections, the overall consistency perhaps underscores the Quran's balanced organization. The distribution of words across Parahs might have been designed to facilitate the structured recitation and study of the Quran.

4.5 Linguistic Visualization of the Quran: A Comprehensive Word Cloud

Figure 6 above presents a word cloud derived from the entire Quran, offering a visual representation of the most frequently used words within this sacred text. This visualization highlights the prominence of key terms, with their size indicating their relative frequency. The most dominant word is "الله" (Allah), reflecting the centrality of God in the Quranic discourse. Other frequently appearing words, such as "من" (who, from), "في" (in), "ما" (what), "قال" (said), and "إن" (indeed), emphasize recurring themes of divine guidance, human actions, and communication between God and humanity. Words like "الذين" (those who) and "آمنوا" (believed) further underscore the Quran's focus on faith and the relationship between believers and Allah.

Additionally, terms such as "السماء" (the heavens) and "الأرض" (the earth) evoke themes of creation and the universe, reinforcing Allah's power as Creator. Connective and relational words like "و" (and), "على" (on), and "في" (in) reveal the Quran's intricate narrative structure, seamlessly linking concepts and ideas. This word cloud succinctly encapsulates the essence of the Quran's teachings: the centrality of Allah, guidance for believers, and reflections on creation. It provides

a concise yet profound insight into the linguistic patterns that shape the spiritual depth and thematic richness of this sacred text.

Figure 6

Word Cloud of the Quran: Visualizing Key Terms and Themes.

4.6 Discrepancies in Prophet Name Counts in the Quran

This experiment examines the process of counting the occurrences of prophet names in the Quran and explores the discrepancies between the computed results and expected values. It begins with a text-cleaning function that removes diacritics (tashkils) from the Arabic script, ensuring uniformity in text processing. This is achieved using regular expressions that target Unicode ranges associated with Arabic diacritics. A custom function then iterates through a Quranic text dataset, systematically cleaning each entry and counting the occurrences of predefined prophet names. The results are sorted in descending order, producing a ranked list of prophet mentions. Despite this structured approach, several factors contribute to inconsistencies in the computed frequencies, including diacritic variations, name ambiguity, orthographic variations, and lack of contextual analysis.

For example, the diacritic removal function may not cover all possible variations, resulting in discrepancies between the processed text and the prophet's

names. Additionally, certain names, such as "صالح" (Saliḥ), which can function as an adjective—like in the phrase "عامل صالح" (righteous worker), rather than a proper noun, potentially result in over-counting. Orthographic differences, such as "إبراهيم" (Ibrahim) vs. "ابراهيم", can lead to under-counting if all forms are not considered. Lastly, the method does not differentiate between prophetic and non-prophetic mentions, which can skew the results. To improve accuracy and align with established references, several enhancements can be implemented, such as refining diacritic handling, using Natural Language Processing (NLP) techniques to analyze the context in which a name appears, incorporating multiple spelling variants of prophet names, and leveraging annotated datasets with contextual metadata for Quranic verses. By integrating these improvements, the analysis would yield more reliable results, providing deeper insights into the Quranic text and ensuring better alignment with scholarly sources on prophet mentions.

4.7 TF-IDF for Quranic Similarity Detection

The data is first carefully prepared through a cleaning process that makes the text easier to analyze. This includes breaking the text into individual words (tokenization), making the words consistent in form (normalization), and removing common words that don't add much meaning (stopwords). After this, key features are extracted using a method called Term Frequency–Inverse Document Frequency (TF-IDF), which helps identify the most important words in the Quranic text. Word embeddings are also used to capture deeper meanings and connections between words.

Table 2.

TF-IDF Results and Analysis

Query	Result	Reason	Solution
"بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ" (Bismi Allāhi Raḥmāni Raḥīm) with diacritics	Verses containing Ar-"الرحمن" and/or Ar-"الرحيم", or part of the phrase	TF-IDF identified important keywords and calculated cosine similarity	No corrective action needed

"الرحمن الرحيم" (Ar-Rahmani Ar-Rahim) without diacritics	✗ Text not found in the dataset."	TF-IDF treats words with and without diacritics as distinct units, thus no exact match was found	1. Remove diacritics from the dataset 2. Remove diacritics from the query 3. Use a tokenizer that handles diacritics

TF-IDF was then applied to identify Quranic verses most similar to the first verse, "بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ" (In the name of Allah, the Most Gracious, the Most Merciful). This method, often used in studies of Quranic texts, helps find the most important words by looking at how often they appear in one verse compared to the whole Quran.

Using this approach, we found verses that share key words like "الرَّحْمَٰنِ" (The Most Gracious) and "الرَّحِيمِ" (The Most Merciful), showing the focus on God's names.

For example, Surah 27, Verse 30 includes part of the phrase "بِسْمِ اللَّهِ" (In the name of Allah), which is why it scored high in similarity. Surah 55, Verse 1 only has "الرَّحْمَٰنُ" (The Most Gracious), but it is still considered close in meaning because of its religious context.

However, we faced some problems when we removed the diacritics (marks on Arabic letters that guide pronunciation). For example, when we searched for "الرحمن الرحيم" (The Most Gracious, the Most Merciful) without diacritics, no relevant results were found, as illustrated in Table 2. This highlights the importance of using strong text-cleaning methods or tools that can handle Arabic script properly.

To go further, tools like Word2Vec or GloVe can help capture deeper meanings between words, going beyond simple word matching.

In short, TF-IDF is useful for identifying basic similarities, but combining it with deep learning tools like Doc2Vec and improved text processing could lead to a more accurate and insightful analysis of the Quran. This opens up new ways to explore the meanings and themes within the holy text.

4.8 Semantic Analysis with BERT

To assess the effectiveness of our methodology, we incorporated Transformer-based models such as AraBERT and QuranBERT for fine-tuning. These models played a crucial role in enhancing semantic analysis, allowing us to go beyond

traditional keyword-based methods and achieve more nuanced results in determining verse sentiment and meaning.

In this experiment, we used the pre-trained Arabic BERT model (asafaya/bert-base-arabic), a variant of Bidirectional Encoder Representations from Transformers (BERT), specifically optimized for Arabic text. BERT's bidirectional architecture enabled us to deeply understand the syntax and semantics of Arabic, making it ideal for Quranic question answering.

To improve semantic understanding, we utilized BERT to extract the most relevant words and phrases from user queries. For instance, when processing the query "من هو الله؟" (Who is God?), BERT successfully identified key terms and the surrounding context, helping generate precise and relevant responses.

The pipeline used the transformers library to process user queries and match them against a Quranic dataset (quran_fictif.csv). The workflow included the following steps:

- **Data Loading:** Aggregating Quranic verses from the dataset's "Texte" column to create a searchable context.
- **Question Processing:** Tokenizing the input question and context using BERT's tokenizer to ensure proper formatting for analysis.
- **Answer Extraction:** Identifying and predicting the most relevant answer span using cosine similarity and confidence scoring.

For the query "من هو الله؟" ("Man huwa Allah?" / "Who is God?"), the system returned "رب العالمين الله" ("Rabb al-ʿālamīn allāh" / "Lord of the worlds, God"), demonstrating its ability to identify key theological phrases like "الرَّحْمَٰنِ" ("ar-Raḥmān" / The Most Gracious) and "الرَّحِيمِ" ("ar-Raḥīm" / The Most Merciful). However, limitations emerged:

- **Diacritic Sensitivity:** Queries without diacritics (e.g., " الرحمن الرحيم" / "ar-Raḥmān ar-Raḥīm") failed due to token mismatches.
- **Domain-Specific Fine-Tuning:** A warning about uninitialized weights highlighted the need for further fine-tuning on Quranic texts, as generic pre-training lacks domain-specific nuance.
- **Contextual Ambiguity:** While BERT excels at phrase matching (e.g., Surah 27:30's partial match to "بِسْمِ اللّ..." / "Bismillāh al-Raḥmān

al-Raḥīm"), it struggles with deeper theological context, such as distinguishing between literal and metaphorical references.

Improvement strategies align with broader research trends:

- **Ensemble Models:** Combining predictions from multiple BERT architectures (e.g., AraBERT, Arabic-BERT) to enhance robustness.
- **Post-Processing:** Applying rules to refine answers, such as merging overlapping spans or prioritizing verses with higher theological relevance.
- **Diacritic Normalization:** Preprocessing text to remove or standardize diacritics, ensuring consistency between queries and datasets.

The results underscore BERT's potential in Arabic QA tasks, particularly for keyword-driven queries. However, hybrid approaches—combining BERT with rule-based systems or domain-specific fine-tuning—are critical for improving accuracy in religious text analysis. Future work could integrate tools like AraELECTRA for dense retrieval or expand datasets with tafsir (exegesis) annotations to address contextual ambiguities.

4.9 Sentiment Analysis of Quranic Verses: A Comparative Approach

Semantic and tonal variations were visualized through the results, with the use of computational Tafsir (interpretation of Quranic text) to explore the relationships between entities and themes.

This experiment enabled the development of semantic models by fine-tuning AraBERT, specifically trained on Classical Arabic, for verse similarity detection and paraphrase identification. The application of contrastive learning brought semantically related verses closer, facilitating more accurate retrieval and interpretation. The analysis of emotional tone provided insights by detecting sentiments in Classical Arabic using adapted lexicons for the models and classifying the verses accordingly.

Overall, the integration of Transformer models (AraBERT), fine-tuned models, deep learning techniques, and domain-specific knowledge yielded significant and promising results, offering new approaches and perspectives for Quranic studies. The verses were categorized into Positive, Negative, or Neutral sentiments using two distinct approaches: Simple Transformers and Hugging Face Transformers.

The first approach used the pre-trained model aubmindlab/bert-base-arabertv02, along with the ClassificationModel from Simple Transformers. This model was trained on a small manually curated dataset of Quranic verses with sentiment labels. However, due to the limited size of the dataset, the model generalized poorly and predicted "Neutral" for all test verses.

The second approach utilized the same pre-trained model but fine-tuned it using Hugging Face Transformers. The dataset was split into training and testing sets via train_test_split, allowing for better evaluation of generalization capabilities. Fine-tuning significantly improved accuracy, correctly identifying Positive and Neutral sentiments in most cases.

For example, the verse "إِنَّ مَعَ الْعُسْرِ يُسْرًا" ("Indeed, with hardship comes ease") was classified as Positive by the fine-tuned model, whereas the initial approach labeled it Neutral. Similarly, "وَإِذَا مَرِضْتُ" ("And when I am ill") was correctly identified as Positive after fine-tuning. However, contextual limitations led to misclassification of verses like "فَوَيْلٌ لِلْمُصَلِّينَ" ("So woe to those who pray"), which was incorrectly labeled as Positive (see Table 3).

Table 3.

Comparison of Results

Verse	Simple Transformers (Initial)	Hugging Face Transformers (Fine-tuned)	Interpretation
إِنَّ مَعَ الْعُسْرِ يُسْرًا ("Indeed, with hardship comes ease")	Neutral	Positive	Fine-tuned model correctly identified the positive sentiment.
وَإِذَا قِيلَ لَهُمْ...	Neutral	Neutral	Both models correctly classified as Neutral.
وَإِذَا مَرِضْتُ... ("And when I am ill")	Neutral	Positive	Fine-tuned model correctly identified the positive sentiment.

فَوَيْلٌ لِّلْمُصَلِّينَ ("So woe to those who pray")	Neutral	Positive (Incorrect)	Fine-tuned model misclassified due to context limitations.

Fine-tuning enhanced performance by leveraging the model's pre-trained knowledge while specializing it for the task of identifying sentiments in Quranic verses. Splitting the dataset into training and testing sets ensured a more robust evaluation of results. Despite these improvements, challenges remain in accurately classifying verses with complex or negative sentiments due to limited contextual understanding. Addressing these issues requires expanding the dataset to include more diverse examples and integrating contextual embeddings to capture deeper semantic relationships.

While the fine-tuned Hugging Face model outperformed the initial Simple Transformers approach, further refinement is necessary. Expanding the dataset, improving contextual embeddings, and conducting detailed error analysis will help enhance sentiment classification accuracy for Quranic verses and provide more insightful interpretations.

5. Conclusion

This study analyzed the Quranic text using Natural Language Processing (NLP) methods to explore its structure, meaning, and emotional tone. We used both traditional statistical tools like TF-IDF and deep learning models such as AraBERT, applying them to Arabic and translated versions of the Quran. To answer our research questions:

- Thematic analysis using TF-IDF showed that shorter Surahs (less than 20 verses) had higher emotional word density, while longer ones covered more diverse topics.
- AraBERT improved verse classification accuracy from 74% (baseline) to 87%, especially in cases requiring contextual understanding.
- Sentiment analysis revealed clear emotional differences between Meccan and Medinan Surahs, with Meccan verses showing higher intensity in fear and hope expressions.

We also integrated Qirāʾāt into our approach, showing that small recitation differences can shift both the structure and emotional reading of

certain verses. This adds a new layer to Quranic NLP research, which has mostly ignored canonical variation.

This work opens new possibilities for automatic tafsīr generation, comparative emotion-aware recitation modeling, and deeper semantic search tools. It brings together classical Quranic scholarship and modern AI methods to support more meaningful and respectful digital analysis.

Some verses were misclassified in sentiment detection, especially those with mixed emotional cues (for example, verses that express both hope and fear). This shows how complex it is to model emotions in the Quran. These errors may be caused by some limitations of the study, such as the small size of the emotion-labeled dataset and the difficulty of processing Arabic diacritics and differences between Qirāʾāt. In the future, we plan to increase the dataset and improve text preprocessing to make the models more accurate.

REFERENCES

Darwish, K., & Mubarak, H. (2016). *Arabic NLP: Morphologicaltagging and lexical analysis.*

Dukes, K., & Habash, N. (2010). *The Quranic Arabic Corpus: A dataset for morphological and syntactic analysis.*

Al-Khalifa, H. S., & Al-Salman, A. S. (2011). *Rule-based syntactic parsing for Quranic Arabic.*

Antoun, W., Baly, F., & Hajj, H. (2020). *AraBERT: Transformer-based model for Arabic NLP.*

Al-Ayyoub, M., Jararweh, Y., & Alsmadi, I. (2022). *Multi-task learning for Quranic text analysis.*

Shohoud, Y., Shoman, M., & Abdelazim, S. (2023). *Quranic Conversations: Developing a Semantic Search tool for the Quran using Arabic NLP Techniques.*

Bashir, M. H., Azmi, A. M., Nawaz, H., et al. (2022). Arabic natural language processing for Qur'anic research: a systematic review. *Artificial Intelligence Review*, 56, 6801–6854.

Hamed, I., Sabty, C., Abdennadher, S., et al. (2025*). A Survey of Code-switched Arabic NLP.*

Tariq, M., Awan, M. A., & Khaleeq, D. (2024). *Developing a Quranic QA System.*

Ali, M., Farah, A., & Ahmed, H. (2022). QRCD: The Quranic Reading Comprehension Dataset. *Proceedings of the 6th Arabic NLP Conference.*

Rahman, S., Hussein, A., & Khalid, M. (2023). *Optimizing Passage Retrieval for Quranic Question Answering using Transformer Models. arXiv preprint arXiv:2412.11431.*

Stars, J., Hamdy, F., & El-Beltagy, S. (2022). A New Dataset and Evaluation for Quranic QA Systems. *Proceedings of the OSACT-1 Workshop on Arabic NLP.*

Kaggle Quran Dataset. (2023). *Quranic Text Dataset in Arabic and English,* retrieved from https://www.kaggle.com/datasets.

Touati-Hamad, D., Bouaziz, S., & Haddad, R. (2021). *Authentication of Quran Verses Sequences Using Deep Learning. In Proceedings of the International Conference on Artificial Intelligence and Its Applications (AIAP)*, pp. 1–6.

Touati-Hamad, D., Bouaziz, S., & Haddad, R. (2021). Digital Text Authentication Using Deep Learning: Proposition for the Digital Quranic Text. *In Proceedings of the Advanced Computer Science Applications and Technologies (ACSAT)*, pp. 150–154.

Touati-Hamad, D., Bouaziz, S., & Haddad, R. (2022). Arabic Quran Verses Authentication Using Deep Learning and Word Embeddings. In *Proceedings of the IEEE International Conference on Computer and Information Sciences (ICCIS)*, pp. 254–259.