Vol.4, No.5, Page 2064- 2074, 2026 https://doi.org/10.59628/jast.v4i5.2677 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

Multi-Label Classification of Qur'anic Similes: A Computational Approach to Arabic Rhetorical Theory

Wadee A. Nashir[1], A S. Al-Hegami[2], B. Al-Fuhaidi[1], Wedad AL-Sorori[1] and Naseebah Maqtary[1] *

[1]**Department of Computer Science, Science and Technology University, Sana'a, Yemen,**
[2]**Faculty of Computer and Information Technology, Sana'a University, Sana'a, Yemen.**

*Corresponding author: n.maqtary@ust.edu.ye

Abstract

Similes (*tashbīh*) serve as a cornerstone of Qur'anic eloquence, functioning as a vital cognitive tool for conveying complex theological concepts through accessible imagery. Despite the sophisticated taxonomies established by classical Arabic rhetoricians, contemporary computational approaches often rely on single-label classification paradigms. This methodological reductionism fails to capture the inherent "rhetorical overlap" where a single verse embodies multiple, non-mutually exclusive categories, such as being simultaneously explicit (*mursal*) and representational (*tamthīl*). This study addresses this gap by formalizing Qur'anic simile classification as a multi-label learning task, bridging the divide between classical linguistic theory and modern Natural Language Processing. Utilizing an expert-annotated dataset of 364 verses grounded in authoritative classical exegeses, we evaluated the performance of several Arabic-specific Transformer models, including AraBERT, CamelBERT, and MARBERT. Quantitative results demonstrate that MARBERT achieved superior performance, reaching a Micro F1-score of 0.7685 and a Macro F1-score of 0.6003, significantly outperforming traditional statistical baselines. The findings revealed a high label density across the corpus, providing empirical validation for the synergistic and multidimensional nature of Qur'anic rhetorical figures. Beyond technical metrics, this study contributes to "computational hermeneutics" by demonstrating that classical rhetorical categories function as structured, learnable knowledge. By successfully modeling overlapping categories, this study offers a novel methodology for Digital Humanities and provides a scalable framework for the automated analysis of highly sophisticated Classical Arabic texts.

Article Info

Keywords:
Arabic NLP, Qur'anic Rhetoric, Simile Classification, Multi-Label Learning, Computational Rhetoric, Transformer Models, Classical Arabic.

Article History:
Received: 23-December-2025,
Revised: 11-February-2026,
Accepted: 05-March-2026,
Published: 28 May 2026.

1. INTRODUCTION

In the landscape of Islamic civilization, the Qur'an stands as the foundational miracle of linguistic eloquence, where rhetorical devices (*balāghah*) serve as the primary vehicle for conveying complex theological truths. Among these devices, the simile (*tashbīh*) occupies a position of paramount importance, functioning not merely as an ornamental literary figure but as a vital cognitive tool for bridging the gap between abstract divine concepts and human perception [1, 2]. Classical scholars, most notably [c]Abd al-Qāhir al-Jurjānī in his seminal work *Asrār al-Balāghah*, conceptualized *tashbīh* as a dynamic process of semantic transfer that facilitates the comprehension of divine attributes and eschatological realities through systematic likenesses [3]. Al-Zamakhshar = ifurther elaborated on this device, arguing that mastering Arabic rhetoric is indispensable for uncovering divine intention (*murād Allāh*), thereby positioning rhetorical analysis as a cornerstone of Qur'anic hermeneutics [4].

The classical theoretical framework produced a meticulous taxonomy of similes based on the struc-

tural presence of comparison markers (*adāt*) and the degree of elaboration in terms of resemblance (*wajh al-shabah*). This classification includes categories such as *Mursal* (explicit), *Mu'akkad* (implicit), *Balīgh* (heightened/condensed), and *Tamthīlī* (representational/narrative). However, the sophistication of Qur'anic eloquence lies in the fact that these categories are rarely static; they represent a system where meanings interact synergistically to create what scholars term "semantic expansiveness." Despite this theoretical richness, contemporary computational approaches in Arabic Natural Language Processing (NLP) have largely treated simile detection and classification through the lens of methodological reductionism. Most existing studies rely on binary classification or single-label paradigms, which assume that a textual unit can belong to only one rhetorical category [5, 6]. This "categorical mutual exclusivity" assumption fundamentally misrepresents the multidimensional nature of Qur'anic discourse and leads to a significant loss of rhetorical depth in computational modeling [7].

The insufficiency of single-label classification is most evident when examining verses that exhibit systematic "rhetorical overlap." For instance, in Surat al-Kahf (Q 18:45), the life of this world is likened to water sent from the sky that causes vegetation to grow, withers, and scatters. This verse constitutes a *tashbīh mursal* due to the explicit comparison marker (*ka-mā'*), yet it simultaneously functions as a *tashbīh tamthīlī* by constructing a complex, multi-stage narrative scenario [8]. Forcing a model to choose a single label for such a verse ignores the structural co-occurrence documented by classical rhetoricians and diminishes the ecological validity of the resulting NLP models. Consequently, there is a critical need to shift the computational paradigm toward Multi-Label Learning (MLL) architectures that can model the label correlations and interpretive polyvalence inherent in Qur'anic text.

In response to these challenges, the primary goal of this research is to bridge the gap between classical Arabic rhetoric and modern computational linguistics by formalizing Qur'anic simile classification as a multi-label learning task. The objectives of this study are fourfold: first, to formalize the task of simile classification within a multi-label framework; second, to construct a high-quality supervised dataset grounded in the classical taxonomies of Al-Jurjānī and Al-Zamakhsharī; third, to evaluate the performance of state-of-the-art Arabic Transformer models—such as AraBERT, MARBERT, and CamelBERT—in capturing the nuanced features of rhetorical text; and finally, to analyze whether these modern NLP architectures can capture the systematic rhetorical overlaps theorized by classical scholars.

This research represents a methodologically novel synthesis that contributes to Arabic NLP and Digital Humanities through several key interventions. We introduce the first multi-label simile classification task for Qur'anic text, moving beyond the limitations of traditional single-label models. Furthermore, we provide a rhetorically grounded annotation schema that ensures that the computational labels maintain scholarly depth and authenticity. By conducting an empirical evaluation using Arabic transformers, we offer new insights into how these models handle the morphological and semantic complexities of Classical Arabic. Ultimately, this approach exemplifies a form of "computational hermeneutics," where neural network architectures are used to formalize and test interpretive frameworks derived from centuries of humanistic scholarship [9].

The remainder of this paper is organized as follows: Section 2 reviews related work on computational metaphor detection and the current state of Arabic rhetorical analysis. Section 3 establishes the theoretical rhetorical framework, detailing the classical typology of similes and justifying the overlapping categories. Section 4 describes the construction of the expert-annotated dataset, and Section 5 delineates the methodology, including task formalization and model architecture. Section 6 presents the experimental results and the error analysis. Section 7 provides a multidimensional discussion of the findings from both linguistic and technical perspectives. Finally, Section 8 concludes the paper and suggests potential avenues for future research.

2. RELATED WORK

The computational study of figurative language has evolved significantly over the past decade, shifting from rule-based approaches to sophisticated deep learning architectures. Foundational work in metaphor processing, pioneered by Shutova [10, 11], established that metaphorical expressions can be detected through distributional anomalies and violations of the selectional preferences. This evolution was institutionalized through the Workshop on Metaphor in NLP series, which established benchmark datasets such as the VU Amsterdam Metaphor Corpus (VUA) [12]. However, a critical examination of these foundational efforts reveals a persistent methodological bottleneck: the overwhelming majority of studies conceptualize figurative language identification as a binary classification task—metaphor versus non-metaphor or simile versus non-simile. Recent critiques suggest that this homogenization obscures the rich typological diversity of literary texts. For instance, Li et al. [13] demonstrated that binary paradigms fail to challenge modern language models, while Kuo and Carpuat showed that improvements on binary benchmarks, such as VUA, do not effectively transfer to more complex, domain-specific corpora.

Recent research specifically targeting simile detection has begun to move beyond simple identification toward more nuanced typological frameworks. Mpouli [14] pio-

neered the automatic annotation of similes by distinguishing structural features, while Chen et al. [15] revealed that although language models capture relational semantics, they struggle with the triadic structure of tenors, vehicles, and grounds. Further advancements by Wang [16] in creating datasets for novel metaphors and similes demonstrate that typological classification is computationally feasible. Despite these strides, existing research remains constrained by single-label frameworks. This gap is particularly acute in the context of Arabic rhetoric (*balāghah*), where classical traditions articulate sophisticated taxonomies that resist such binary reduction. No existing work has yet addressed a multi-dimensional typological classification where a single rhetorical expression may simultaneously belong to multiple categories—a phenomenon central to the Qur'anic discourse.

The application of Natural Language Processing (NLP) to Arabic introduces distinct challenges rooted in morphological complexity, diacritical ambiguity, and the significant linguistic distance between Modern Standard Arabic (MSA) and Classical Arabic [17, 18]. While the introduction of AraBERT [19] and its successors marked a watershed for Arabic NLP, these models are primarily trained on contemporary MSA sources, leaving the stylistic subtleties of classical *balāghah* largely underexplored. Within the subfield of Qur'anic computation, research has predominantly focused on morphological analysis and syntactic parsing [20]. This has culminated in the development of the Extended Quranic Treebank (EQTB) and the 'Noor' framework, which introduced automated ellipsis resolution to achieve 100% syntactic and morphological coverage of the Quran in a machine-readable format [21, 22]. However, as Atwell [23] observed, rhetorical analysis still remains a nascent frontier compared to these foundational structural layers. Furthermore, while large-scale resources like the 'Diwan' corpus have recently emerged to provide multi-dimensional annotations for Arabic poetry [24], pioneering studies by Abouhagar [5] and Zahrawi et al. [8] on Qur'anic figurative language continue to employ binary or independent classification approaches that do not account for the co-occurrence of multiple rhetorical devices within a single textual unit.

The limitations of these single-label approaches can be addressed by adopting the Multi-Label Text Classification (MLTC) paradigm, which is designed for domains characterized by overlapping, non-mutually-exclusive categories [25]. In the context of Arabic literary analysis, recent frameworks like 'Maqasid' [26] have demonstrated the necessity of MLTC for capturing thematic interdependence in poetry through hybrid CNN-BiLSTM architectures. Such theoretical advances emphasize the importance of label correlation modeling, where dependencies between categories are explicitly captured using graph neural networks or attention mechanisms [27, 28]. The integration of transformer architectures with multi-label learning has yielded substantial gains in other complex semantic domains, such as emotion detection and toxicity classification. In emotion detection, studies have shown that texts frequently express simultaneous, overlapping emotions that single-label models fail to capture [29, 30]. Similarly, toxicity classification benefits from multi-label formulations as a single comment may simultaneously exhibit various types of hate speech [31].

By bridging the success of multi-label classification in these domains with the study of Arabic rhetoric, a significant research gap can be closed. Just as a text can express multiple simultaneous emotions, a Qur'anic simile often embodies multiple rhetorical categories, such as being simultaneously *mufaṣṣal* (detailed) and *tamthīlī* (representational). The current study addresses the converging gaps in figurative language research, Arabic NLP, and multi-label applications. By moving beyond binary detection to capture the typological complexity of classical *balāghah*, this research provides the first systematic application of multi-label learning to Qur'anic rhetorical analysis, recognizing that the text's sophistication demands computational methods commensurate with its literary complexity.

3. RHETORICAL FRAMEWORK

To accurately model the complexity of Qur'anic similes, it is essential to establish a robust rhetorical framework that bridges classical Arabic linguistic theory with modern computational paradigms. This section provides the theoretical grounding for the study, transforming abstract rhetorical concepts into a structured format that is suitable for machine learning. First, we delineate the classical typology of *tashbīh* (simile) as articulated by masters of Arabic eloquence. Second, we provide a theoretical justification for categorical overlap, demonstrating how the non-mutually exclusive nature of these categories necessitates a multi-label approach. Finally, we operationalized these rhetorical insights into a formal annotation schema, creating the necessary bridge between classical *Balāgha* and automated text classification.

3.1. Classical Typology of Simile in Arabic Rhetoric

Within the tripartite architecture of Arabic *Balāgha* (rhetoric), *tashbīh* (simile) serves as a foundational pillar of *ᶜIlm al-Bayān* (science of clarity). ᶜAbd al-Qāhir al-Jurjānī, in his seminal *Asrār al-Balāgha*, conceptualizes *tashbīh* not as a mere ornamental comparison but as a cognitive operation that likens one entity to another in form and appearance to achieve a specific communicative end [1]. This linguistic mapping involves four constitutive elements known as *arkān al-tashbīh*: the tenor (*al-mushabbah*), the vehicle (*al-mushabbah bihi*), the tool of comparison (*adāt al-tashbīh*), and the point of resemblance (*wajh al-shabah*). Al-Zamakhsharī further

elevates the status of the simile in his exegetical work *al-Kashshāf*, demonstrating that Qur'anic similes function as hermeneutical instruments that translate abstract theological truths into accessible, experientially grounded cognitive frames [32].

Classical taxonomy organizes similes into distinct categories based on the structural presence or absence of these elements. One primary axis distinguishes between *tashbīh mursal* (loose), where the comparison particle—such as *ka-* (like) or *mithl*—is explicitly mentioned, and *tashbīh mu'akkad* (confirmed), where the particle is suppressed to intensify the identification between tenor and vehicle [33]. The secondary axis focuses on the point of resemblance: *tashbīh mufaṣṣal* (detailed) provides an explicit articulation of the shared attribute, whereas *tashbīh mujmal* (concise) leaves the point of resemblance implicit, thereby demanding higher cognitive engagement from the reader to infer the intended meaning.

At the apex of this structural hierarchy lies *tashbīh al-balīgh* (the eloquent simile), which represents a synthesis of the emphatic and concise. In this form, both the comparison tool and the point of resemblance are elided, creating maximal semantic density and interpretive openness [34]. Parallel to these structural divisions is the category of *tashbīh al-tamthīlī* (representational simile). Unlike discrete similes that rely on isolated attributes, the *tamthīlī* form derives its point of resemblance from a holistic configuration of multiple interacting elements that form a composite scenario or "gestalt" image. Al-Jurjānī characterizes this as a "compound similarity" where the meaning emerges from the relational scenario rather than individual components, a view echoed by Ibn ᶜĀshūr, who underscores its role as a cognitive-pedagogical device that resonates with the audience's lifeworld [1, 35].

3.2. Overlapping Rhetorical Categories

A critical theoretical bridge between classical rhetoric and modern computational modeling is the recognition that these taxonomic categories are mutually exclusive. Instead, they operate as orthogonal dimensions of the analysis. For example, the axes of "particle presence" (*mursal/mu'akkad*) and "attribute explicitness" (*mujmal/mufaṣṣal*) are independent; a simile can simultaneously possess a particle and remain implicit in its point of resemblance. This non-exclusivity is not a theoretical flaw but a reflection of the multidimensional nature of rhetorical eloquence, where a single linguistic unit can activate multiple classificatory labels concurrently.

The Qur'anic verse Q18:45, which likens worldly life to the lifecycle of vegetation influenced by rainwater, serves as a paradigmatic case of this overlap. From a structural perspective, the verse is *mursal* because of the explicit particle *ka-* (as). Simultaneously, it is *mujmal* because the central point of resemblance—the transience and ephemerality of life—is not lexically stated but must be reconstructed by the reader. Furthermore, it is classified as *tamthīlī* because the resemblance emerges from the entire temporal sequence of growth, withering, and dispersal, rather than a single element. This "rhetorical polysemy" proves that traditional multi-class classification, which enforces a single label per instance, is theoretically incompatible with the sophisticated reality of *balāgha*. Consequently, this theoretical overlap provides the necessary justification for adopting multi-label learning (MLL) architectures, which permit a verse to activate several labels simultaneously, based on different rhetorical dimensions.

3.3. Annotation Schema

To transform this classical framework into a computationally tractable format, we operationalized the aforementioned categories into an annotation schema that comprised six primary labels. These labels (L1–L6) correspond to *Mursal*, *Mu'akkad*, *Mujmal*, *Mufaṣṣal*, *Balīgh*, and *Tamthīlī*. Each label was governed by a rigorous decision protocol designed to ensure inter-annotator reliability. For instance, the *tamthīlī* label is assigned only if the point of resemblance is identified as emerging from a composite scenario rather than a discrete attribute. This systematic approach allows the translation of abstract literary concepts into a binary vector format suitable for modern NLP models.

The schema also incorporates logical constraints derived from rhetorical theories to maintain internal consistency. While categories like *tamthīlī* can co-occur with any other label, certain labels remain mutually exclusive by definition, such as *mursal* and *mu'akkad*. Furthermore, the label for *tashbīh al-balīgh* is treated as a composite label that is activated only when both the *mu'akkad* and *mujmal* conditions are satisfied. By formalizing these relationships, the schema preserves the analytical depth of the classical tradition while providing the structured knowledge required to train transformer-based models to capture the nuanced, overlapping features of Qur'anic similes.

4. DATASET CONSTRUCTION

The empirical foundation of this study rests on a specialized corpus of 364 Qur'anic verses systematically extracted from the complete text of the Qur'an based on authoritative rhetorical and exegetical references. The selection process prioritized verses identified as containing simile structures (*tashbīh*) by classical masterworks, most notably Al-Zamakhsharī's *al-Kashshāf* [32] and Ibn ᶜĀshūr's *al-Taḥrīr wa-al-Tanwīr* [35], which serve as gold standards for identifying the nuanced boundaries of Qur'anic figurative language. This rigorous sourcing strat-

egy ensured that the dataset captured not only explicit similes marked by comparative particles but also more complex, implicit, and representational forms that are often overlooked by automated keyword-based extraction methods. By anchoring the data selection in these classical sources, the study maintains a high degree of "rhetorical authenticity," ensuring that the computational models are trained on instances recognized by the tradition as the pinnacle of Arabic eloquence.

To ensure the highest level of theoretical fidelity and expert validation, the annotation process followed rigorous protocols similar to those employed in recent landmark Quranic linguistic projects [21]. The task was conducted by a panel of three specialized linguists, each holding a doctorate in Classical Arabic Rhetoric. These experts employed a deductive labeling approach, applying the classical typology outlined in the rhetorical framework (Section 3) to all 364 instances. Crucially, the annotation was designed as a multi-label task from its inception; annotators were instructed to identify every applicable rhetorical category for each verse rather than forcing a singular and reductive classification. To maintain consistency and minimize subjectivity, a Delphi-style consensus protocol was utilized, whereby discrepancies in labeling were resolved through a joint review of classical exegeses until final agreement was reached. This expert-led methodology produced a robust "ground truth" that accounts for the interpretive plurality and structural complexity characteristic of the Qur'anic text.

Quantitatively, the final dataset comprised 364 unique verses, which collectively yielded 712 label assignments across six defined rhetorical categories. This distribution results in a high label density, with an average of 1.96 labels per verse, a figure that empirically confirms the non-mutually exclusive nature of Arabic rhetorical devices. While certain simpler structures were assigned a single label, a significant majority of the corpus (approximately 72%) manifested two or more overlapping categories, with the most complex instances activating up to four distinct labels simultaneously. These statistics provide clear evidence of the "rhetorical overlap" that necessitates a multi-label learning approach, as a standard multi-class paradigm would have failed to capture nearly 65% of the total rhetorical information in the dataset. This high-density corpus, hereafter referred to as **Sima**, provides a rigorous benchmark for evaluating the capacity of modern transformer models to decode the multilayered eloquence of the Qur'an. To facilitate reproducibility and future research in Qur'anic NLP, the Sima dataset and the implementation framework are made publicly available at https://github.com/NoorBayan/Sima.

5. METHODOLOGY

This section delineates the computational framework and experimental design employed to classify the Qur'anic similes. The methodology is structured to ensure technical rigor and replicability by integrating modern transformer-based architectures with specialized multi-label learning techniques that align with classical rhetorical theory. Our methodological choices—specifically, the adoption of pre-trained language models and a deterministic evaluation protocol—are strategically driven by the necessity to capture the high semantic density of Qur'anic verses while mitigating the statistical limitations of a finite corpus. We begin by formally defining the classification task and its mathematical formulations. Subsequently, we describe the model architecture, focusing on the adaptation of Arabic-specific encoders to overlapping rhetorical categories. This is followed by a detailed account of the training configuration and the experimental setup. Finally, we provide an overview of the evaluation metrics used to assess the model performance across global and label-specific dimensions.

5.1. Task Definition

The classification of Qur'anic similes is formally defined as a supervised multi-label learning task in which each input verse is associated with a set of non-mutually exclusive rhetorical categories. Unlike traditional single-label classification, this task accounts for the overlapping nature of Arabic rhetorical devices, where a single textual unit may simultaneously instantiate multiple categories. Formally, given an input verse x, the model aims to learn a mapping function $f(x) \rightarrow y$, where $y \in \{0,1\}^k$ is a binary vector of length k that represents the presence or absence of each rhetorical label. This formulation allows for $\sum y \geq 0$, enabling the model to capture scenarios in which multiple rhetorical devices co-occur, which is essential for preserving the semantic and stylistic depth of the Qur'anic text.

5.2. Model Architecture

The proposed research strategically utilizes a transformer-based architecture to generate high-dimensional contextualized representations of the input text, leveraging the self-attention mechanism to capture long-range dependencies within the Qur'anic verse [36]. This architectural choice is predicated on the Transformer's superior ability to resolve the complex syntactic structures and implied meanings inherent in Classical Arabic, which traditional sequential models (such as RNNs) often fail to capture. The complete end-to-end computational pipeline, which facilitates the transition from raw linguistic input to multi-label rhetorical predictions, is illustrated in Figure 1. This architectural design is specifically engineered to preserve nuanced semantic and syntactic layers essential for sophisticated rhetorical analysis.

As depicted in the architectural diagram, the core

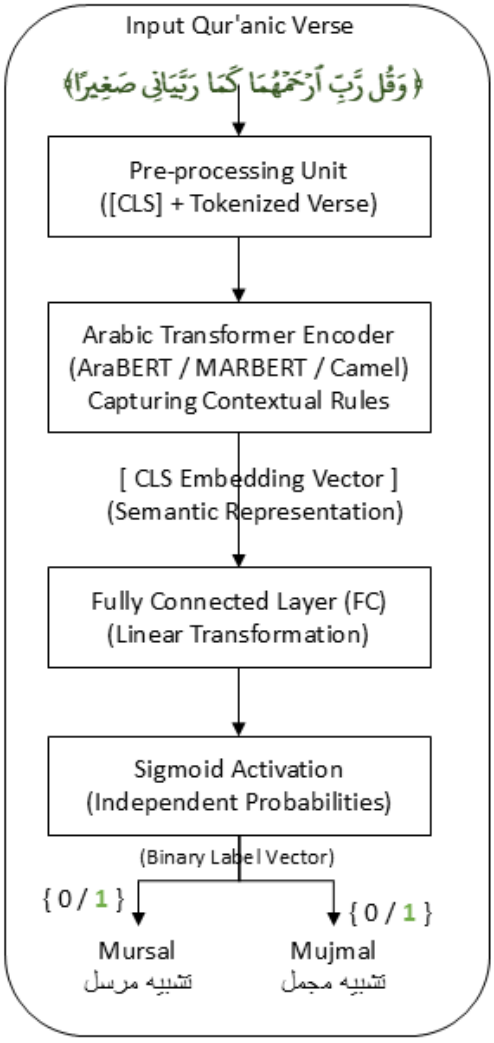


Figure 1. The proposed multi-label classification architecture for Qur'anic similes, illustrating the flow from the input verse through the Transformer encoder to the independent sigmoid output layer.

of the system consists of a transformer encoderutilizing state-of-the-art Arabic-optimized variants such as AraBERT [19], MARBERT [37], or CamelBERT [38]. The model processes the tokenized input to produce an aggregated semantic vector derived from the special classification token (`[CLS]`), which serves as a holistic latent representation of the rhetorical structure of the verse.

To adapt this framework for the multi-label task, the encoder is coupled with a task-specific classification head comprising a linear transformation layer that maps the hidden state to the label space dimension. Crucially, a sigmoid activation function is applied to the output of this layer, diverging from the standard softmax function, which is typical of multi-class paradigms. The adoption of sigmoid activation is not arbitrary; it is theoretically motivated by the requirement to model each rhetorical category as an independent, Bernoulli trial. This mathematical formulation enables the model to estimate independent probabilities for each label, thereby facilitating the identification of overlapping rhetorical features and capturing the inherent non-mutual exclusivity of classical *balāghah* categories.

5.3. Training Setup

The training process is optimized using Binary Cross-Entropy (BCE) loss, which treats each label as an independent binary classification problem, penalizing incorrect predictions for each rhetorical category separately [39]. Given the specialized nature of the dataset, which represents an exhaustive census of Qur'anic similes rather than a random sample, strict measures were taken to ensure experimental reproducibility and a fair comparison between architectures. Accordingly, the dataset was partitioned into training, validation, and test sets (70/15/15) using a fixed random seed (42). This deterministic splitting strategy ensures that the validation and test sets remain consistent benchmarks for all models, mitigating the stochastic variance often associated with smaller corpora [40].

Model optimization was conducted using the AdamW optimizer with decoupled weight decay to mitigate overfitting [41]. Based on preliminary hyperparameter tuning, the final training configuration was standardized with a learning rate of 2×10^{-5} (3×10^{-5} for AraBERT), per-device batch size of 8, and training duration of 15 epochs. To convert the continuous probability outputs into final binary predictions, a classification threshold of $\tau = 0.5$ was applied to the sigmoid outputs. This rigorous setup ensured a robust and fully replicable training environment capable of capturing the subtle nuances of Classical Arabic rhetoric.

5.4. Evaluation Metrics

To comprehensively assess the model performance in a multi-label context, three specialized metrics were employed. The Micro-averaged F1-score was used to provide a global assessment of the effectiveness of the system by aggregating the true and false positives across all labels, effectively weighting the results by label frequency. Conversely, the Macro-averaged F1-score was used to evaluate the model's performance on a per-label basis, ensuring that less frequent but rhetorically significant categories were given equal weight in the final assessment [42]. Finally, the Hamming Loss is included to measure the fraction of incorrectly predicted labels averaged over the entire dataset [43]. Together, these metrics account for partial correctness and provide a balanced view of the model's ability to navigate the complex, overlapping taxonomies of Qur'anic simile classification.

6. EXPERIMENTS AND RESULTS

This section presents an empirical evaluation of the proposed multi-label classification framework, providing a

comprehensive assessment of its performance in identifying overlapping Qur'anic rhetorical categories. The primary objective was to measure the efficacy of deep contextualized representations compared to traditional statistical benchmarks. We began by establishing a performance baseline using frequency-based features, followed by an in-depth analysis of state-of-the-art Arabic Transformer models. To provide a transparent view of the learning process, we examined the training dynamics and convergence patterns of the optimal model. Finally, the section concludes with a qualitative error analysis that categorizes predictive limitations, offering insights into the structural and lexical challenges inherent in automated rhetorical analyses.

6.1. Baseline Models

To establish a comparative performance benchmark, a traditional machine learning pipeline was implemented using Term Frequency-Inverse Document Frequency (TF-IDF) as the feature extraction method and Logistic Regression as the classifier. TF-IDF was configured to transform the raw Qur'anic text into a numerical representation based on word frequency and inverse document distribution, capturing the lexical importance of tokens in the corpus. This baseline was included to evaluate the efficacy of frequency-based statistical methods in handling multi-label rhetorical classification compared with contextualized deep learning architectures. Quantitatively, the TF-IDF baseline achieved a Micro F1-score of 0.6161 and a Macro F1-score of 0.3077, with a Hamming Loss of 0.2560. These results indicate a limited capacity for capturing label correlations, particularly for low-frequency categories, thereby serving as the lower bound for performance evaluation.

6.2. Transformer-Based Results

The performance of three Arabic-specific Transformer models—AraBERT [19], CamelBERT-CA [38], and MARBERT [37]—was evaluated over 15 training epochs. Table 1 summarizes the peak performance metrics achieved by each model in the validation phase.

A quantitative comparison demonstrates that all transformer-based models substantially outperformed the TF-IDF baseline across all metrics, as shown in Figure 2. MARBERT emerged as the superior model, achieving the highest Micro F1-score (0.7685) and Macro F1-score (0.6003), while maintaining the lowest Hamming Loss (0.1577).

The steady progression of these metrics for the top-performing model is illustrated in Figure 3(a). CamelBERT-CA showed a competitive Micro F1-score of 0.7433, whereas AraBERT peaked at 0.7304. The observed increase in the Macro F1-scores across the transformer models, particularly in MARBERT, suggests

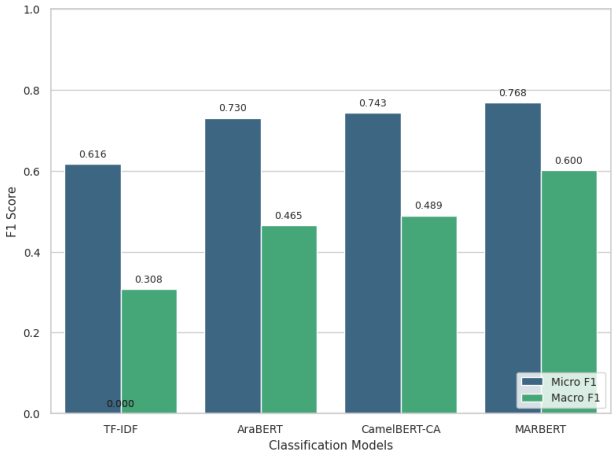


Figure 2. Visual comparison of Micro and Macro F1-scores across all evaluated models, highlighting the significant performance gain of MARBERT over statistical baselines.

an improved capability to classify infrequent labels compared with the statistical baseline. Furthermore, the convergence patterns indicate that most models reached their optimal validation performance between epochs 7 and 9, as evidenced by the training and validation loss trends depicted in Figure 3(b). Subsequently, the validation loss exhibited fluctuations, suggesting the onset of potential overfitting on the specialized dataset.

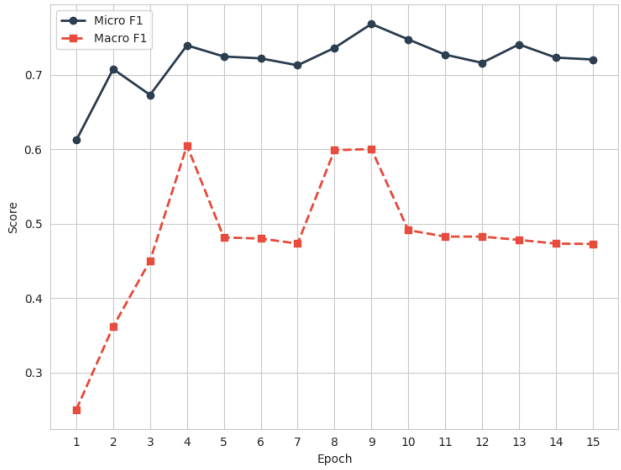
6.3. Error Analysis

A systematic review of the models' predictions revealed specific patterns of error related to the data distribution and structural complexity. The most common error types observed were "Partial Match" and "Label Omission." In cases of label omission, the models frequently failed to identify labels with low representation in the dataset, such as *Mu'akkad* (7.4%) and *Mufaṣṣal* (6.0%). This suggests a sensitivity to data sparsity, where the limited number of training instances for these categories impeded the model's ability to learn distinct discriminative features.

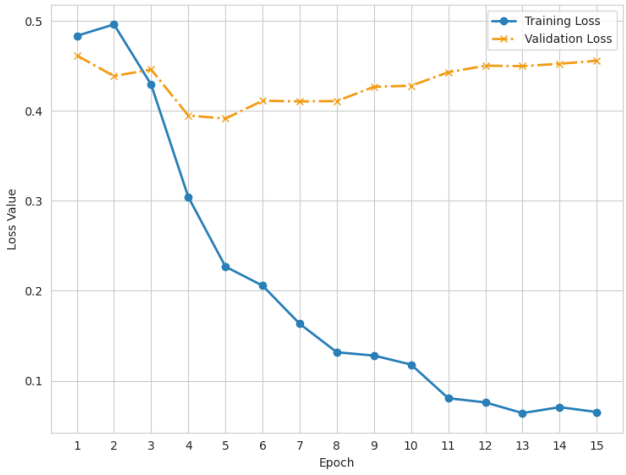
Additionally, the models exhibited lower accuracy on verses with high label density (3+ labels), which constitute 22.3% of the dataset. For these complex instances, the models often predicted the dominant labels (*Mursal* or *Mujmal*), while failing to activate concurrent secondary labels. Structurally, errors were more frequent in verses where comparative markers were absent, leading to misclassifications between the *Balīgh* and *Mu'akkad* categories. These errors are attributable to the high degree of lexical overlap between these classes at the input level, where the model's classification head struggled to differentiate between subtle syntactic variations. Finally, a small subset of errors resulted from "Over-prediction," where the model assigned frequent labels to verses that did not contain them, likely as a result of the high prior probability of these labels within the training distribution.

Table 1. Comparative performance metrics of the baseline and Transformer-based models for multi-label Qur'anic simile classification.

Model	Best Epoch	Micro F1	Macro F1	Hamming Loss
TF-IDF (Baseline)	-	0.6161	0.3077	0.2560
AraBERT	7	0.7304	0.4648	0.1845
CamelBERT-CA	7	0.7433	0.4888	0.1726
MARBERT	**9**	**0.7685**	**0.6003**	**0.1577**



(a) Metric progression for MARBERT.



(b) Training vs. Validation Loss.

Figure 3. Training dynamics of the MARBERT model, showing (a) the evolution of F1 metrics and (b) the convergence of loss functions over 15 epochs.

7. DISCUSSION

The empirical findings presented in the previous section demonstrate a significant shift in the computational capacity to decode the rhetorical layers of the Qur'an, moving from simple statistical detection to nuanced, multi-dimensional classification. This section provides a qualitative and theoretical synthesis of these results, moving beyond quantitative metrics to interpret the underlying linguistic and cognitive factors that drive the model performance. We begin by offering a linguistic interpretation of the results, specifically analyzing how transformer-based architectures capture—or struggle with—the overlapping nature of *Balāgha* categories. Subsequently, we discuss the broader implications of these findings for classical Arabic rhetorical theory, positioning the results as computational evidence of the structured and synergistic nature of Qur'anic eloquence. Finally, we provide a critical reflection on the study's limitations, addressing the challenges of data sparsity and the inherent subjectivity involved in annotating divine discourse.

7.1. Linguistic Interpretation of Results

The experimental findings offer profound computational validation of the structural complexity inherent in Qur'anic similes. The high label density (1.96) and the fact that over 72% of the corpus carries multiple labels suggest that rhetorical figures in the Qur'an do not function in isolation but through a synergistic overlap of categories. The superior performance of MARBERT over the statistical baseline indicates that identifying these categories requires more than keyword matching; it necessitates attention to contextual and semantic nuances that classical rhetoricians termed *naẓm* (the delicate arrangement of words).

The disparity between the Micro and Macro F1 scores (0.76 vs. 0.60) reveals a significant linguistic insight: while the models excel at identifying dominant rhetorical modes such as *Mursal* and *Tamthīlī*, they struggle with the "rhetorical exceptions"—the rare *Mu'akkad* and *Mufaṣṣal* forms. This suggests that the models have effectively learned the "prototypical" Qur'anic simile (explicit and scenario-based) but find the "heightened" or "condensed" eloquence of *Balīgh* similes more challenging.

For instance, in a verse such as Q14:18 (see Figure 4), where deeds are compared to ashes scattered by the wind, the model's success in identifying the *Tamthīlī* (representational) aspect while potentially omitting the *Mufaṣṣal* (detailed point) reflects a bias toward holistic imagery over discrete syntactic specifications.

مَّثَلُ الَّذِينَ كَفَرُوا بِرَبِّهِمْ أَعْمَالُهُمْ كَرَمَادٍ اشْتَدَّتْ بِهِ الرِّيحُ فِي يَوْمٍ عَاصِفٍ

Figure 4. Qur'anic verse Q14:18, representing a complex simile structure where the model must navigate the overlap between representational (*Tamthīlī*) and detailed (*Mufaṣṣal*) categories.

Furthermore, the "Partial Match" errors identified in the analysis underscore the difficulty of "rhetorical polysemy." When a verse instantiates three or more categories, the model often captures the most structurally evident label (e.g., the presence of a comparative particle) while overlooking deeper semantic layers (e.g., the implicit nature of the point of resemblance). This behavior suggests that while modern NLP can capture the surface-level markers of *Balāgha*, the deeper interpretive layers—where multiple rhetorical functions converge—still require more sophisticated semantic grounding.

7.2. Implications for Arabic Rhetorical Theory

This study provides a significant computational bridge to classical Arabic rhetorical theory, specifically supporting the long-held scholarly view of the non-exclusivity of rhetorical devices. By formalizing *tashbīh* classification as a multi-label task, we provide empirical evidence for the synergistic nature of *Balāgha* that classical scholars such as al-Jurjānī and al-Zamakhsharī described [1, 32]. The high rate of co-occurrence between the *Mursal* and *Tamthīlī* labels suggests that Qur'anic discourse intentionally uses explicit comparative tools to anchor complex, multi-element narrative scenarios, rendering abstract truths more accessible.

The results further position Arabic rhetoric as a form of "structured knowledge." The ability of transformer models to achieve substantial F1 scores indicates that the categories defined by medieval scholars are not merely subjective aesthetic judgments but are rooted in consistent linguistic and semantic patterns that are computationally learnable. This validates classical taxonomy as a rigorous analytical system. Rather than viewing rhetoric as a fluid, unpredictable art, this study demonstrates that it functions as a sophisticated, rule-governed system in which different levels of meaning—syntactic, semantic, and representational—interleave within a single textual unit.

7.3. Limitations

Despite the promising performance of these models, several limitations must be acknowledged. First, regarding the dataset size, the corpus comprised 364 verses. This constraint is intrinsic to the domain, representing an exhaustive census of similes identified in authoritative classical exegeses, rather than a scalable random sample. While this ensures rhetorical authenticity, the resulting sparsity creates a "Macro F1 bottleneck," where rare categories such as *Mufaṣṣal* and *Mu'akkad* are underpredicted. Furthermore, regarding statistical robustness, the finite nature of the data limited the feasibility of extensive protocols, such as k-fold cross-validation, which might fragment these rare label co-occurrence patterns. Instead, a fixed-seed split was employed to ensure experimental reproducibility and validate the model's capacity within this specialized, expert-verified domain.

Second, the inherent subjectivity of rhetorical annotations remains a challenge. Although the dataset was validated by experts, the boundaries between certain categories and such as *Mujmal* (implicit) versus *Mufaṣṣal* (explicit)—can occasionally be "fuzzy" depending on the exegetical school followed. This subjectivity introduces a level of label noise that may affect the model's certainty, as reflected in the fluctuations in the validation loss. Finally, the study is limited to the text of the Qur'an; therefore, the learned rhetorical patterns may not be directly transferable to other Classical Arabic genres (such as pre-Islamic poetry) without further domain-specific adaptation, as the Qur'anic style possesses a unique rhetorical signature.

8. CONCLUSION AND FUTURE WORK

This study formalized Qur'anic simile classification as a multi-label learning task, transcending the binary and single-label constraints that have historically limited computational rhetoric. By integrating classical taxonomies with state-of-the-art Arabic Transformers, we provided empirical validation for the "rhetorical overlap" theorized by classical scholars, proving that nuanced, interleaved rhetorical functions are computationally modelable. The superior performance of the MARBERT model confirms that modern NLP can capture high-level aesthetic features when grounded in a rigorous and expert-validated framework. Ultimately, this work contributes to Digital Humanities by establishing a scalable "computational hermeneutics" bridge, demonstrating that classical rhetorical wisdom can be effectively operationalized to deepen our analytical understanding of sophisticated Classical Arabic discourse.

Several promising avenues for future research have emerged from this framework. A primary direction involves expanding the scope to a broader spectrum of *balāghah* devices, such as metaphor (*istiᶜāra*) and irony (*tahakkum*), to a comprehensive computational model of Qur'anic stylistics. Furthermore, future efforts should transition from classification to generative paradigms capable of producing natural language explanations and detailed exegetical interpretations of rhetorical structures. Finally, investigating cross-surah generalization will be critical for evaluating model robustness across diverse revelation periods and thematic contexts. Such advancements will refine this methodology into a holistic tool for both scholars and practitioners, facilitating the automated stylistic analysis of the vast and intricate heritage of Classical Arabic literature.

REFERENCES

[1] A. Q. i. A. R. al-Jurjānī, *Asrār al-Balāgha [The Secrets of Eloquence]*, 1st, M. M. Shākir, Ed. Cairo, Egypt: Dār al-Madanī, 1991, p. 548, Originally composed ca. 1078 CE; critically edited by Maḥmūd Muḥammad Shākir, ISBN: 978-...

[2] H. Ahmad and N. A. Ghafar, "AI as interpretive aid in qur'anic stylistics: Ethical foundations for digital hermeneutics," *Int. J. Arts Soc. Sci.*, vol. 8, no. 9, 2025, Published online Sept 2025, ISSN: 2581-7922.

[3] K. M. Shuqair, "An ornamentalist view of metaphor in arabic literary theory," *J. Crit. Stud. Lang. Lit.*, vol. 2, no. 2, pp. 33–52, 2021.

[4] K. Malik, N. Habibi, and I. Patrah, "Rhetoric and balaghah: The significance of Zamakhshari's contributions to linguistic studies," *Ihya Al-Arabiyah: J. Pendidikan Bahasa dan Sastra Arab.*, vol. 11, no. 1, pp. 87–104, 2025.

[5] L. A. Abouhagar, "Automatic identification of arabic figurative language," Master's Thesis, King Fahd University of Petroleum and Minerals, 2023. [Online]. Available: https://eprints.kfupm.edu.sa/id/eprint/142742/.

[6] H. El Rifai, L. Al-Qadi, and A. Elnagar, "Arabic text classification: The need for multi-labeling systems," *Neural Comput. Appl.*, vol. 34, no. 3, pp. 1495–1515, 2022. DOI: 10.1007/s00521-021-06390-z. [Online]. Available: https://link.springer.com/article/10.1007/s00521-021-06390-z.

[7] M. Alqahtani and E. Atwell, "A review of semantic search methods to retrieve information from the Qur'an corpus," in *Corpus Linguistics 2015*, Leeds University, 2015.

[8] R. T. A. Zahrawi, S. N. S. Abdullah, and A. Sarirete, "Advancing literary analysis with Python: A comprehensive study of simile detection and classification in the translation of Al-Abrat," *SAGE Open*, vol. 15, no. 1, 2025. DOI: 10.1177/21582440251378859.

[9] J. W. Mohr, R. Wagner-Pacifici, and R. L. Breiger, "Toward a computational hermeneutics," *Big Data & Soc.*, vol. 2, no. 2, pp. 1–8, 2015. DOI: 10.1177/2053951715613809.

[10] E. Shutova, "Computational approaches to figurative language," *Lang. Linguist. Compass*, vol. 5, no. 6, pp. 299–319, 2011.

[11] E. Shutova, "Design and evaluation of metaphor processing systems," *Comput. Linguist.*, vol. 41, no. 4, pp. 579–623, 2015.

[12] G. J. Steen, A. G. Dorst, J. B. Herrmann, T. Kaal A. Krennmayr, and T. Pasma, *A method for linguistic metaphor identification*. John Benjamins Publishing, 2010.

[13] E. Sanchez-Bayona et al., "Metaphor and large language models: When surface features matter more than deep understanding," *arXiv preprint arXiv:2507.15357*, 2025. [Online]. Available: https://arxiv.org/abs/2507.15357.

[14] S. Mpouli, "Annotating similes in literary texts," in *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*, Association for Computational Linguistics, 2017. [Online]. Available: https://aclanthology.org/W17-7403/.

[15] W. Chen et al., "Probing simile knowledge from pre-trained language models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, 2022, pp. 5875–5887. DOI: 10.18653/v1/2022.acl-long.404. [Online]. Available: https://aclanthology.org/2022.acl-long.404/.

[16] X. Wang, "Normed dataset for novel metaphors, novel similes, literal and anomalous sentences in chinese," *Front. Psychol.*, vol. 13, p. 922 722, 2022. DOI: 10.3389/fpsyg.2022.922722. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.922722/full.

[17] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. on Asian Lang. Inf. Process.*, vol. 8, no. 4, Article 16, 2009. DOI: 10.1145/1644879.1644881. [Online]. Available: https://dl.acm.org/doi/10.1145/1644879.1644881.

[18] Y. Al Moaiad, M. Alobed, M. Alsakhnini, and A. M. Momani, "Challenges in natural arabic language processing," *Int. J. Appl. Sci. Eng.*, vol. 21, no. 3, pp. 1–9, 2024. DOI: 10.6703/IJASE.202409_21(3).004.

[19] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 9–15. [Online]. Available: https://aclanthology.org/2020.osact-1.2/.

[20] M. H. Bashir et al., "Arabic natural language processing for qur'anic research: A systematic review," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 6801–6854, 2023. DOI: 10.1007/s10462-022-10313-2. [Online]. Available: https://link.springer.com/article/10.1007/s10462-022-10313-2.

[21] W. A. Nashir, A. M. Mohsen, A. A. Al-Shargabi, M. K. Nour, and B. B. Al-Onazi, "A complete, multi-layered quranic treebank dataset with hybrid syntactic annotations for classical arabic processing," *Data Brief*, vol. 62, p. 111 940, 2025. DOI: 10.1016/j.dib.2025.111940.

[22] W. A. Nashir, A. M. Mohsen, A. A. Al-Shargabi, M. K. Nour, and B. B. Al-Onazi, "Noor: A pipeline framework for classical arabic parsing," *IEEE Access*, vol. 13, pp. 167 538–167 559, 2025. DOI: 10.1109/ACCESS.2025.3607915.

[23] E. S. Atwell, "Using the web to model modern and quranic arabic," in *Arabic Corpus Linguistics*, T. McEnery, A. Hardie, and N. Younis, Eds., Edinburgh, UK: Edinburgh University Press, 2019, pp. 100–119, ISBN: 9780748677375.

[24] B. B. Al-Onazi, W. A. Nashir, and A. A. Al-Shargabi, "Diwan: Constructing the largest annotated corpus for Arabic poetry," *IEEE Access*, vol. 13, pp. 58 927–58 941, 2025. DOI: 10.1109/ACCESS.2025.3551161.

[25] J. Gong et al., "Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification," *IEEE Access*, vol. 8, pp. 30 885–30 896, 2020. DOI: 10.1109/ACCESS.2020.2972751. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8988213.

[26] R. G. Al-Anazi, W. A. Nashir, B. B. Al-Onazi, A. A. Alhamad, and A. A. Al-Shargabi, "Maqasid: A hybrid CNN-BiLSTM framework for nuanced thematic classification of Arabic poetry," *IEEE Access*, vol. 13, pp. 189 421–189 443, 2025. DOI: 10.1109/ACCESS.2025.3621112.

[27] S. Zhu, J. Qu, and X. Huang, "Multi-label text classification with global and local label correlation," *arXiv preprint arXiv:1711.11475*, 2017.

[28] Y. Ma, G. Sun, L. Sun, and Z. Zhao, "Label-specific dual graph neural network for multi-label text classification," *arXiv preprint arXiv:2106.01574*, 2021.

[29] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. Gelbukh, "Multi-label emotion classification in texts using transfer learning," *Expert Syst. with Appl.*, vol. 213, p. 118 534, 2023. DOI: 10.1016/j.eswa.2022.118534.

[30] Q. Liu et al., "Mlgn: A multi-label guided network for improving text classification," *IEEE Access*, vol. 11, pp. 76 453–76 464, 2023. DOI: 10.1109/ACCESS.2023.3297821.

[31] J. Bäumler, L. Blöcher, L. J. Frey, X. Chen, M. Bayer, and H. Stuckenschmidt, "A survey of machine learning models and datasets for the multi-label classification of textual hate speech in english," *arXiv preprint arXiv:2504.08609*, 2025. [Online]. Available: https://arxiv.org/abs/2504.08609.

[32] M. i. Ù. Al-Zamakhshari, *Al-Kashshaf àn Haqa'iq Ghawamid al-Tanzil [The Revealer of the Truths of the Obscurities of the Revelation]* (Silsilat al-Turath al-Islami), 1st, À. A. À. al-Mawjud and À. M. Muàwwad, Eds. Riyadh, Saudi Arabia: Maktabat al-Ùbaykan, 2009, vol. 1, Edited by Àdil Ahmad Àbd al-Mawjud and Àli Muhammad Muàwwad.

[33] J. D. Al-Qazw=ın=ı, *Al-=Iḍ=aḥ f=ı ᶜUl=um al-Bal=agha [Clarification in the Sciences of Rhetoric]*, 4th, M. c. M. Khaf=aj=ı, Ed. Beirut: D=ar al-J=ıl, 2010.

[34] A. Y. Y. i. A. B. i. M. i. A. al-Sakkākī, *Miftāḥ al-ᶜUlūm [The Key to the Sciences]*, 2nd, N. Zarzūr, Ed. Beirut, Lebanon: Dār al-Kutub al-ᶜIlmiyya, 1987, p. 621, Critical edition; originally compiled in 13th century CE (626 AH).

[35] M. Ibn ᶜĀshūr, *Al-Taḥrīr wa-al-Tanwīr [The Verification and Illumination]*. Tunis, Tunisia: Al-Dār al-Tūnisiyya li-al-Nashr, 1984, Classical tafsīr work in Arabic literature.

[36] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[37] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for arabic," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, 2021, pp. 7088–7105. DOI: 10.18653/v1/2021.acl-long.551.

[38] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in arabic pre-trained language models," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP)*, Kyiv, Ukraine (Virtual): Association for Computational Linguistics, 2021, pp. 92–104. DOI: 10.18653/v1/2021.wanlp-1.8. [Online]. Available: https://aclanthology.org/2021.wanlp-1.10.pdf.

[39] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. on Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014. DOI: 10.1109/TKDE.2013.39.

[40] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Machine Learning and Knowledge Discovery in Databases*, 2011, pp. 145–158. DOI: 10.1007/978-3-642-23808-6_10.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7.

[42] M. S. Sorower, "A literature survey on algorithms for multi-label learning," Oregon State University, Tech. Rep., 2010.

[43] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, no. 2, pp. 135–168, 2000. DOI: 10.1023/A:1007649029923.