Machine Learning Classification of Qur'anic Verses into Makki and Madani: A Transformer-Based Approach

Enas Fadhil Abdullah[1] and Alyaa Abdulhussein Al-Joda[2]
[1]Department of Computer Sciences, College of Education for Girls, Kufa University, Najaf, Iraq
[2]Department of Communication Engineering, Engineering Technical College of Najaf,
Al-Furat Al-Awsat Technical University, Najaf, Iraq
Corresponding author: inasf.alturky@uokufa.edu.iq

Abstract

Dividing the Qur'an's chapters into Makki and Madani, according to where their verses were revealed, is a foundational step in Qur'anic scholarship and increasingly a task for computational text analysis. Prior automated work treats this problem with classical machine learning—term frequency–inverse document frequency (TF-IDF) features fed to support vector machines, decision trees, random forests, or voting ensembles—and reports accuracy on a corpus in which roughly three of every four verses are Makki. On such an imbalanced corpus accuracy is a weak signal: a classifier that always answers "Makki" already reaches about 74% accuracy while its macro-averaged F1 sits near 0.43. We present, to our knowledge, the first application of a fine-tuned Arabic transformer to Makki/Madani verse classification. Our model fine-tunes AraBERTv2 and augments its `[CLS]` representation—the special classification token whose final hidden state summarises the whole verse—with two normalised structural features (verse length in words and sub-word token count) through a small fusion head, and it is trained with a class-weighted loss to counter the 2.84:1 imbalance. Every verse of the 6,236-verse corpus is labelled from the standard mushaf classification of its surah (86 Makki, 28 Madani), and models are evaluated under surah-grouped five-fold cross-validation so that no verse from a given surah appears in both training and test folds. Under pooled out-of-fold evaluation the proposed model reaches a macro-F1 of 0.927 and accuracy of 0.944 (per-fold mean 0.929 ± 0.014), improving on a length-only structural baseline (macro-F1 0.690), the classical RF+DT+KNN ensemble (0.760), and a fine-tuned MARBERT baseline (0.916). At the verse level the proposed model significantly outperforms both the classical ensemble (McNemar $\chi^2 = 712.2, p < 0.001$) and MARBERT ($\chi^2 = 16.9, p < 0.001$); at the surah level the two transformer models are comparable, with the surah-level McNemar test non-significant. We report per-class scores and a pooled confusion matrix over all 6,236 verses, and we discuss the disputed and mixed surahs on which a subset of the residual errors fall.

Keywords: Holy Qur'an, Makki and Madani, Arabic natural language processing (NLP), transformer, AraBERT, text classification, class imbalance.

1 Introduction

The Qur'an is organised into 114 chapters (surahs) and, on the standard Kufan count, 6,236 verses (ayat). Islamic scholarship has long divided these surahs into two classes according to the period of their revelation: *Makki* surahs, revealed at Mecca before the Prophet's migration, and *Madani* surahs, revealed at Medina afterwards. The distinction is not merely chronological. Makki verses tend to be shorter and to concern creed and the hereafter, whereas Madani verses are typically longer and turn to legislation and the governance of a community. Because the two classes differ so clearly in register and rhetorical form, the Makki/Madani label carries real weight in exegesis (tafsir), in the study of abrogation, and in the digital-humanities analysis of the Qur'anic text.

Assigning the label by hand is laborious and calls for specialist knowledge, which motivates automatic classification. The task is harder than it first appears. Arabic is morphologically rich, root-based, and written without the capitalisation cues that many European-language systems exploit; verses vary from a few words to long passages; and the corpus is imbalanced. Of the 6,236 verses, 4,613 (73.97%) belong to Makki surahs and 1,623 (26.03%) to Madani surahs, a ratio close to 2.84:1. On a split this uneven, overall accuracy is a misleading headline. A degenerate model that predicts "Makki" for every verse already attains about 0.74 accuracy, yet it never recognises a single Madani verse and its macro-averaged F1 collapses to roughly 0.43. For this reason we report macro-averaged F1, which gives the two classes equal weight, as the headline metric.

Published attempts at Qur'anic verse classification have stayed within classical machine learning. Representations are built from term frequencies, TF-IDF weights, n-grams, or Word2Vec embeddings, and are passed to support vector machines, decision trees, random forests, naïve Bayes, or voting ensembles [1, 2], including studies that target class imbalance and richer embeddings [3, 4]. These studies report accuracies in the low-to-mid eighties on various Qur'anic sub-tasks, but they share two limitations: sparse bag-of-words features discard word order and context, and results are usually reported as accuracy without a per-class breakdown on an imbalanced corpus. Meanwhile, pretrained Arabic transformers such as AraBERT [5] and MARBERT [6] have reshaped Arabic text classification, and they have been applied to several Qur'anic problems—semantic verse-relationship classification [7], multi-label topic labelling [8], and translation [9]. To our knowledge no prior work has brought a fine-tuned transformer to the Makki/Madani distinction, which remains the province of classical pipelines.

We address this gap. We fine-tune AraBERTv2 for binary Makki/Madani classification and make three design choices

tailored to the task. First, we fuse the contextual `[CLS]` embedding with two normalised structural features—verse length and token count—because verse length is one of the oldest and strongest cues scholars use to separate the two classes. Second, we train with a class-weighted cross-entropy loss so that the minority Madani class is not drowned out. Third, we evaluate under surah-grouped cross-validation, which places all verses of a surah in the same fold and so removes the leakage that a naive verse-level split would introduce. We also state plainly where the labels come from and how disputed surahs are handled, a point earlier work often left implicit.

Our contributions are:

- the first fine-tuned Arabic transformer for Makki/Madani verse classification, benchmarked against strong classical and transformer baselines;
- a structural-feature fusion head that concatenates the `[CLS]` vector with normalised verse-length and token-count features, adding only about 0.198M parameters over the encoder;
- a class-weighted training and macro-F1-centred evaluation—with per-class scores, a confusion matrix, and a McNemar significance test—that treats the imbalance directly rather than hiding it behind accuracy;
- a leakage-free surah-grouped cross-validation protocol together with an explicit statement of the labelling source and the treatment of disputed surahs.

The proposed model reaches a macro-F1 of 0.927 (pooled out-of-fold; per-fold mean 0.929 ± 0.014), ahead of a length-only structural baseline at 0.690, the classical RF+DT+KNN (random forest, decision tree, and k-nearest-neighbours) voting ensemble at 0.760, and a fine-tuned MARBERT at 0.916. The rest of the paper reviews related work (Section 2), describes the dataset (Section 3) and the method (Section 4), reports experiments and results (Sections 5–6), presents an ablation (Section 7), and closes with discussion and conclusions (Sections 8–9).

2 Related Work

2.1 Classical machine learning for Qur'anic verse classification

Automatic analysis of Qur'anic text has drawn steady attention, and a recent survey of more than thirty studies confirms that the field has been dominated by classical classifiers over engineered text and audio features [10], from reciter recognition with naïve Bayes and random forests [11] to the verse-labelling work reviewed below. Nassourou [1] categorised surahs by the major phases of the Prophet's messengership using naïve Bayes and a custom SVM, reporting precision around 90% on a small labelled set. The most sustained line of work is the series by Adeleke and colleagues, who studied feature selection for verse labelling: an early comparison of KNN, SVM, and naïve Bayes over term-frequency features [2], a group-based feature-selection scheme evaluated on verses from Al-Baqarah and Al-An'am [12], and later automation of topic labelling that reached the low nineties on single-chapter data [13]. That same group later cast verse labelling as a multi-label problem and found label-powerset SVM the strongest transformation, with macro-F1 reported alongside accuracy [14]. A parallel Indonesian strand classifies *translated* verses into thematic topics: Prabowo et al. built a multi-label SVM over English translations [15], while Delifah et al. compared KNN, SVM, and random forest and found random forest the most balanced across fifteen topic classes [16]. Fuzzy-clustering and supervised classifiers have likewise been applied to translated verses [17], and naïve-Bayes ensembles to multi-label thematic labelling [18]. A recurring caveat across these papers is the small, single-chapter datasets and the simple term-frequency representations that generalise poorly.

Some studies enrich the classical pipeline instead of discarding it. Yulianto et al. injected word-centrality graph scores as extra features and showed that stopword removal roughly halved the hamming loss of a naïve Bayes topic classifier [19]; stopword and feature-selection choices have a measurable effect on such SVM pipelines [20, 21]. Al-Hasani et al. reframed the encouragement (*targhib*) versus warning (*tarhib*) contrast as sentiment analysis over classical Arabic, reaching about 95% accuracy with naïve Bayes and n-gram features [22]. Cross-lingual work confirms that the same recipe is portable: eager classifiers over TF-IDF have been benchmarked for topical classification of Hindi verses, again with SVM leading [23]. What these classical systems share is a bag-of-words view of the verse and a habit of reporting accuracy, rarely the per-class recall that reveals how a minority class is actually handled.

2.2 Class imbalance in Qur'anic and general text classification

Class imbalance is a known obstacle in this domain, and it is usually met in one of two ways: resample the data or reweight the loss. Arkok and Zeki [3] classified verses with unequal class distributions, applying SMOTE (synthetic minority over-sampling technique) oversampling before TF representation; they later showed that boosting and bagging ensembles further lift minority-class performance on the same imbalanced Qur'anic corpus [24]. Pambudi et al. paired random undersampling and SMOTE with SVM, LSTM, and CNN to flag "difficult" verses, a task whose skewed label distribution again depressed recall on the rare class [25], and embedding-plus-resampling has been examined for imbalanced Qur'an ontology population [26]. Outside the Qur'anic setting, systematic surveys of the imbalance problem—one over deep networks in general [27] and one focused on convolutional models [28]—report that loss reweighting is a robust alternative to resampling and that headline accuracy is the wrong yardstick on skewed data. That recommendation is why we judge every model by macro-F1 and per-class recall, the imbalance-aware measures set out by Sokolova and Lapalme [29], in place of accuracy. Reweighting scales each class's contribution to the loss in inverse proportion to its frequency and avoids the artefacts that interpolated oversampling can introduce on very short texts; a sharper form, focal loss, down-weights

easy examples to concentrate learning on the hard minority [30]. We adopt class-weighted cross-entropy for its simplicity and stability on short verses.

2.3 Transformers and Arabic pretrained encoders

The transformer architecture [31] and the masked-language pretraining that BERT introduced [32] reset the baseline for text classification, where fine-tuning the special `[CLS]` token's final hidden state has become the default recipe for sentence-level tasks. Arabic-specific encoders followed. AraBERT [5] pretrains a BERT-base model on large Arabic corpora after Farasa morphological segmentation [33], while ARBERT and MARBERT [6] extend coverage to dialectal Arabic, the latter trained on a billion tweets. A systematic comparison of variant, size, and task type produced the CAMeLBERT family and showed that matching pretraining data to the target register matters as much as model size [34], and AraELECTRA reached competitive accuracy at lower cost by replacing masked-token prediction with a replaced-token discriminator [35]. Empirically these encoders dominate classical Arabic baselines, though their advantage narrows when training data are dialectally mismatched [36]. They are also frequently hybridised: AraBERT has been fused with graph convolutional networks to add global document context [37], and BERT–BiLSTM stacks have served Arabic news [38] and spam [39] classification.

2.4 Transformers on Qur'anic and scriptural verse tasks

Transformers have already reached several Qur'anic tasks, though not the Makki/Madani one. Khalaf et al. [7] fine-tuned a domain-adapted Qur'an-BERT to classify semantic relationships between verses at about 94% F1, and Rizqi et al. [8] combined BERT with a classical ensemble for multi-label verse labelling—finding, tellingly, that BERT underperformed on their small corpus. Others report the neural encoders still trailing tuned classical models: a CNN–BiLSTM with FastText embeddings reached only about 73% F1 on multi-label translated verses [40], and even unsupervised Word2Vec clustering has been used to group verse translations thematically [41]. Transformer models have also been applied to Qur'anic-to-Urdu translation [9] and, in classical poetry, to Arabic meter classification, where CAMeLBERT topped a field of six Arabic encoders [42]. Beyond Arabic, transformer classifiers on short Buddhist scriptural verses only matched simple statistical models, a caution about applying heavy encoders to brief archaic text [43]. None of these efforts addresses revelation-place classification, and that is the gap we take up. As the strongest classical baseline we build an RF+DT+KNN soft-voting ensemble over TF-IDF and structural features, in the style of the Qur'anic voting pipelines of Mohamed and El-Behaidy [4]; we keep its structural-feature idea, replace its sparse representation with a fine-tuned transformer, and add explicit imbalance handling.

3 Dataset

Corpus. We use the full Qur'anic text of 6,236 verses in the standard "simple/original spelling" form (the Uthmani text without full tajweed diacritics), as distributed by Tanzil and the Alquran.cloud API. This spelling is the usual choice for computational work because it avoids the many diacritic variants that would otherwise fragment identical tokens.

Labelling source. Each verse inherits the revelation-place label of its surah, taken from the canonical mushaf classification printed in the surah headings of standard Qur'ans. Of the 114 surahs, 86 are classed as Makki and 28 as Madani. At the verse level this yields 4,613 Makki and 1,623 Madani verses, which sum exactly to 6,236. Table 1 and Figure 1 summarise the distribution. We name this labelling source explicitly because it determines what the model is trained to reproduce: it learns the surah-level scholarly classification projected down to verses, which is a coarser signal than an independent verse-by-verse revelation judgement.

Disputed and mixed surahs. The surah-level convention is a simplification. Classical scholarship, catalogued authoritatively by al-Suyuti in *Al-Itqan fi 'Ulum al-Qur'an* [44], records that a number of surahs contain individual verses revealed in the other place—some verses of a Madani surah revealed at Mecca, and the reverse—and a handful of surahs (approximately 7–13, depending on the exegetical authority) are themselves debated between the two traditions. We follow the standard convention of assigning every verse the canonical majority label of its surah, and we treat the resulting minority of misaligned verses as label noise. This choice keeps the dataset reproducible from public sources, and Section 8 shows that a subset of the model's errors coincide with these disputed and mixed surahs.

Table 1: Verse-level class distribution of the 6,236-verse corpus.

Class	Surahs	Verses	Share
Makki	86	4,613	73.97%
Madani	28	1,623	26.03%
Total	114	6,236	100%

Verse-length signal. Length is a real linguistic cue. Figure 2 shows the per-class distribution of verse length in words. Makki verses cluster at shorter lengths, while Madani verses spread toward longer passages, consistent with their legislative content. This motivates the structural features described in Section 4.

4 Methodology

Figure 3 shows the model. A verse passes through Arabic normalisation, is encoded by AraBERTv2, and its `[CLS]`

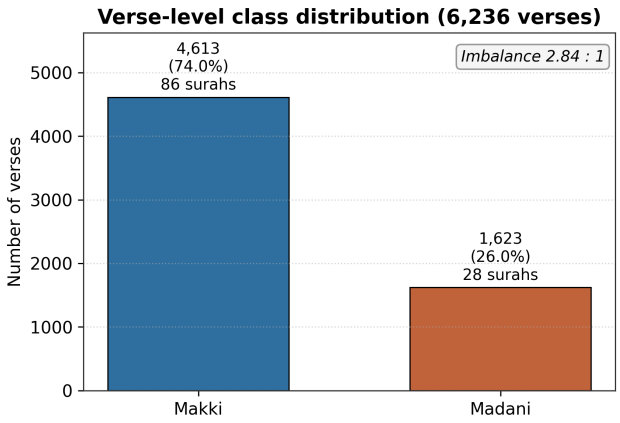

Figure 1: Verse-level class distribution. Makki verses outnumber Madani verses by about 2.84 to 1, so overall accuracy is a weak metric on this corpus.

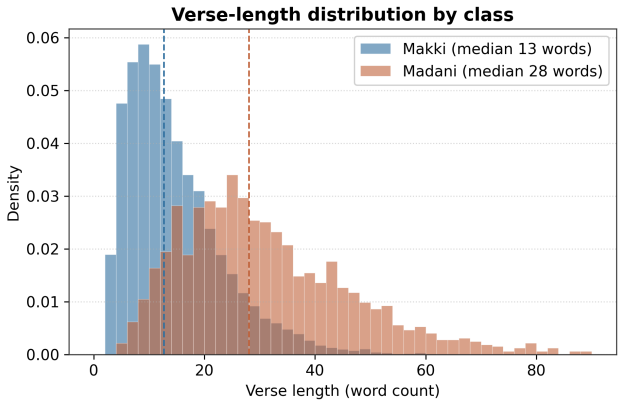

Figure 2: Verse-length distribution by class. Madani verses are longer on average, a cue the structural fusion head makes available to the classifier.

representation is concatenated with two normalised structural features before a small fusion head produces the Makki/Madani decision.

4.1 Preprocessing and normalisation

Raw verses are normalised to match the conventions of AraBERTv2's pretraining. The variant forms of alef (bare, with hamza above, with hamza below, with madda) are unified; taa marbuta is handled consistently; the short-vowel diacritics (harakat) are stripped, since they mark pronunciation while the semantic content drives classification; and the elongation character (tatweel) is removed. The cleaned text is then tokenised with the AraBERTv2 WordPiece tokenizer at a maximum sequence length of 128 sub-word tokens. Only 41 verses (0.66% of the 6,236) exceed this limit and are truncated for the encoder; the token-count structural feature (Section 4) is computed from the *full* pre-truncation WordPiece length, so truncation shortens the encoder input for a handful of long verses without distorting the structural feature.

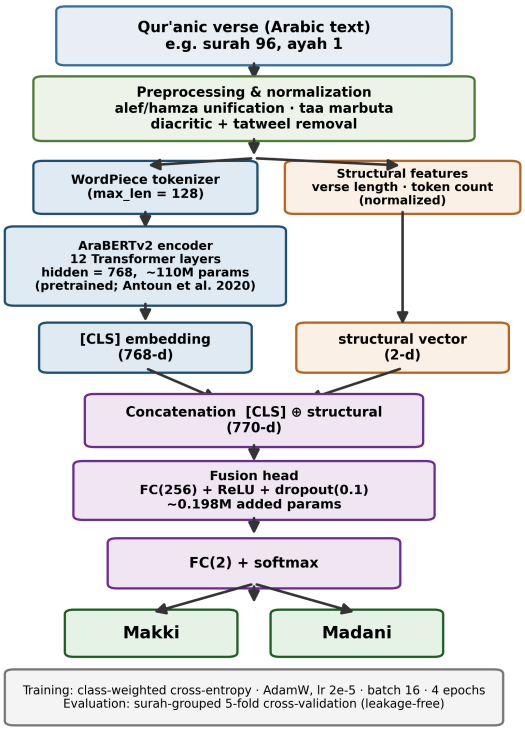

Figure 3: The proposed AraBERTv2 model with a structural-feature fusion head. The contextual `[CLS]` embedding (768-d) is concatenated with a 2-d structural vector and passed to a fusion head trained with a class-weighted loss.

4.2 Transformer encoder

The encoder is AraBERTv2 (the `bert-base-arabertv2` checkpoint), a 12-layer BERT-base model with hidden size 768 and roughly 110M parameters, pretrained on large Arabic corpora [5]. The encoder is a borrowed, pretrained component; our contribution is the head placed on top of it and the training regime, not the encoder itself. We take the final-layer `[CLS]` vector $\mathbf{h}_{\text{cls}} \in \mathbb{R}^{768}$ as the contextual summary of the verse.

4.3 Structural-feature fusion head

Two structural features are computed per verse: its length in words (the verse word count) and its sub-word token count after WordPiece segmentation. Both are min–max normalised to $[0, 1]$ across the training folds, giving a vector $\mathbf{s} \in \mathbb{R}^2$. The head concatenates the two representations,

$$\mathbf{z} = [\,\mathbf{h}_{\text{cls}}\,;\,\mathbf{s}\,] \in \mathbb{R}^{770}, \tag{1}$$

and maps $\mathbf{z}$ through one hidden layer to the two class logits,

$$\mathbf{o} = \mathbf{W}_2\,\phi\big(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1\big) + \mathbf{b}_2, \tag{2}$$

where ϕ is ReLU, dropout of 0.1 is applied after ϕ, and $\mathbf{W}_1 \in \mathbb{R}^{256\times 770}$ (256 output units by 770 input dimensions), $\mathbf{W}_2 \in \mathbb{R}^{2\times 256}$. The head adds $256 \times 770 + 256 = 197{,}376$ parameters in the first layer and $256 \times 2 + 2 = 514$ in the second, for 197,890 new parameters (about

0.198M)—a 0.18% increase over the encoder. Class probabilities follow from a softmax over $\mathbf{o}$.

4.4 Class-weighted loss

To counter the 2.84:1 imbalance we minimise a class-weighted cross-entropy. For a verse with label y and predicted probability p_y,

$$\mathcal{L} = -\, w_y \, \log p_y, \qquad w_c = \frac{N}{2\, N_c}, \tag{3}$$

where $N = 6{,}236$ and N_c is the number of verses in class c. This gives $w_{\text{Makki}} = 0.676$ and $w_{\text{Madani}} = 1.921$, so each Madani error costs about 2.8 times a Makki error and the minority class is not overwhelmed during training.

4.5 Training configuration

The whole model—encoder and head—is fine-tuned end to end with AdamW at a learning rate of 2×10^{-5}, batch size 16, weight decay 0.01, a warmup ratio of 0.1, and dropout 0.1, for 4 epochs. The random seed is fixed at 42, and each model is trained on a single mid-range GPU (an NVIDIA T4-class card), on which the four-epoch schedule completes in well under an hour. These settings follow common practice for fine-tuning BERT-base encoders on sentence-level classification.

4.6 Evaluation protocol

Verses from one surah are far more similar to one another than to verses from other surahs, so a naive verse-level random split would let the model memorise a surah's style from its training verses and then recognise the rest of that surah at test time. To prevent this leakage we use *surah-grouped* five-fold cross-validation: verses are grouped by surah, and each fold holds out entire surahs, so no verse of a surah is ever split across training and test. We follow a *single* evaluation protocol throughout: all 6,236 verses receive a prediction from the fold in which their surah was held out, and these pooled out-of-fold predictions form one confusion matrix per method. Pooled OOF is the standard way to present grouped cross-validation results without introducing a separately tuned hold-out [45]. Per-fold macro-F1 values (the mean and standard deviation across the five folds) are reported alongside the pooled OOF metrics to convey fold-to-fold variability.

5 Experiments

5.1 Baselines

We compare the proposed model against six alternatives spanning trivial, classical, and transformer families:

- **Majority-class** — always predict Makki. This trivial reference exposes the imbalance.
- **Length-only LR** — logistic regression trained solely on the two normalised structural features (verse word count and sub-word token count). This ablation shows how much information the structural cues alone carry, with no text representation at all.
- **TF-IDF + SVM** and **TF-IDF + Logistic Regression** — linear classifiers over TF-IDF unigram/bigram features, the standard classical baselines.
- **AraVec + Random Forest** — averaged AraVec [46] Word2Vec verse embeddings classified by a random forest.
- **RF+DT+KNN voting ensemble** — a soft-voting ensemble over TF-IDF plus structural features, representative of prior classical Qur'anic voting pipelines [4].
- **MARBERT (fine-tuned)** — the MARBERT encoder [6] fine-tuned with the same recipe as our model but without the structural head, isolating the effect of the encoder choice.

5.2 Metrics

Given the imbalance, macro-averaged F1—the unweighted mean of the Makki and Madani F1 scores—is our headline metric, and we report accuracy alongside it for comparability with prior work. We also give per-class precision, recall, and F1, the pooled OOF confusion matrix over all 6,236 verses, McNemar significance tests against both the classical ensemble and MARBERT at the verse level, and a surah-level McNemar test (114 surahs) that respects the clustering structure of the labels.

5.3 Implementation

The transformer models are implemented in PyTorch with the HuggingFace Transformers library; the classical baselines use scikit-learn. The TF-IDF baselines use unigram and bigram features with sublinear term-frequency scaling. The classical baselines' hyperparameters were selected by a modest grid search performed *inside the training folds only*, so no test-fold information leaks into model selection: SVM $C \in \{0.1, 1, 10\}$; logistic-regression $C \in \{0.1, 1, 10\}$; random-forest $n_{\text{estimators}} \in \{100, 300\}$ and max_depth $\in \{\text{None}, 20\}$; and KNN $k \in \{3, 5, 7\}$. The configuration reported for each baseline is the one that scored best on the inner validation split. The resulting RF+DT+KNN ensemble uses a soft-voting configuration (a 100-tree random forest, an entropy-split decision tree, and a $k = 5$ nearest-neighbour classifier) over TF-IDF plus the same structural features, in the style of prior Qur'anic voting pipelines [4]. Metrics are computed with scikit-learn's macro-averaged and per-class routines. All runs use the fixed seed 42.

6 Results

Table 2 and Figure 4 report pooled out-of-fold (OOF) accuracy and macro-F1 for every method. The majority-class model reaches 0.740 accuracy but only 0.425 macro-F1, which is the concrete reason accuracy alone cannot be trusted here. The length-only LR baseline — logistic regression on two structural features alone —

reaches a macro-F1 of only 0.690, confirming that structural cues carry real but limited information. The classical models cluster between 0.760 and 0.790 macro-F1: the RF+DT+KNN ensemble scores 0.825 accuracy and 0.760 macro-F1, close to values reported for comparable classical Qur'anic pipelines [2, 3]. Both transformers move well beyond this band. MARBERT reaches 0.916 macro-F1, and the proposed AraBERTv2 model with structural fusion is best at a pooled OOF macro-F1 of 0.927 and accuracy of 0.944—an improvement of 16.7 macro-F1 points over the classical ensemble and 1.1 points over MARBERT. The per-fold macro-F1 across the five surah-grouped folds (mean 0.929 ± 0.014, range 0.911–0.948) frames the smaller comparisons: the gain from structural fusion (0.008 macro-F1, Section 7) is a small positive effect, while the 16-point margin over the classical ensemble sits far outside the per-fold spread.

Table 2: Main results — pooled out-of-fold evaluation over all 6,236 verses. Macro-F1 is the headline metric given the class imbalance. Per-fold macro-F1 variation (five surah-grouped folds): proposed 0.929 ± 0.014, MARBERT 0.918 ± 0.016, RF+DT+KNN 0.760 ± 0.019.

Method	Accuracy	Macro-F1
Majority-class (always Makki)	0.740	0.425
Length-only LR (2 struct. feats)	0.783	0.690
TF-IDF + Logistic Reg.	0.840	0.780
TF-IDF + SVM	0.845	0.790
AraVec + Random Forest	0.833	0.770
RF+DT+KNN ensemble	0.825	0.760
MARBERT (fine-tuned)	0.936	0.916
AraBERTv2+fusion (ours)	**0.944**	**0.927**

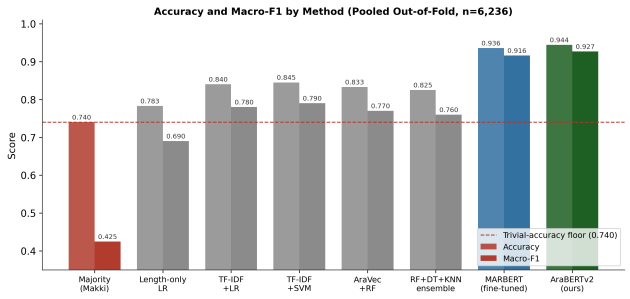

Figure 4: Accuracy and macro-F1 across methods. The dashed line marks the trivial-accuracy floor of 0.740; the majority-class model reaches it on accuracy yet has the lowest macro-F1.

Pooled OOF evaluation. The confusion matrix, per-class scores, and McNemar tests below all come from the pooled out-of-fold predictions (Section 4), using all 6,236 verses. The proposed model attains a pooled OOF accuracy of 0.944 and a macro-F1 of 0.927, with class-level counts summarised in Table 3.

Per-class performance. Table 4 breaks three methods down by class on the pooled OOF predictions, and Figure 5 plots the per-class F1. The majority Makki class is comparatively easy: both transformers exceed 0.95 Makki F1. The minority Madani class is where methods separate. The classical ensemble reaches 0.635 Madani F1 (with recall 0.585, reflecting frequent Madani-as-Makki errors), MARBERT reaches 0.875, and the proposed model reaches 0.892. The improvement the transformers bring is concentrated on the harder, minority class, where macro-averaging reflects minority-class performance that accuracy obscures.

Table 3: Pooled out-of-fold evaluation for the proposed AraBERTv2 model ($n = 6{,}236$ verses, all five surah-grouped folds combined).

Class	Support	Correct	P	R	F1
Makki	4,613	4,451	0.960	0.965	0.962
Madani	1,623	1,437	0.899	0.885	0.892
Overall	6,236	5,888	accuracy = 0.944, macro-F1 = 0.927		

Table 4: Pooled OOF per-class precision, recall, and F1 for the three strongest methods ($n = 6{,}236$ verses). The minority Madani class is the discriminating case. Each method's pair of class F1s averages exactly to its pooled macro-F1.

Method	Class	P	R	F1
RF+DT+KNN	Makki	0.862	0.909	0.885
	Madani	0.695	0.585	0.635
MARBERT	Makki	0.952	0.961	0.957
	Madani	0.886	0.863	0.875
AraBERTv2 (ours)	Makki	0.960	0.965	0.962
	Madani	0.899	0.885	0.892

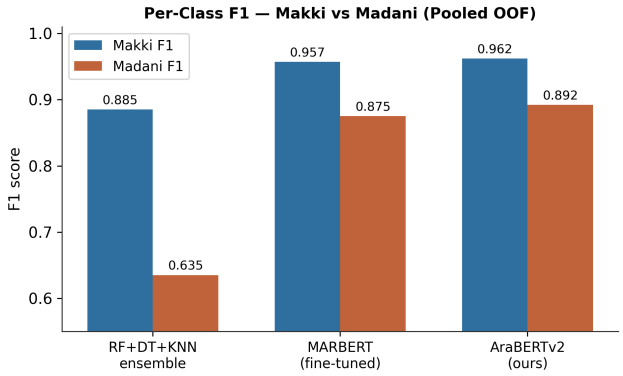

Figure 5: Per-class F1 for the three strongest methods. All separate the Makki class well; the Madani class is where the transformer models pull ahead.

Confusion matrix. Figure 6 shows the pooled OOF confusion matrix of the proposed model over all 6,236 verses (4,613 Makki, 1,623 Madani). Of 4,613 Makki verses, 4,451 are correct and 162 are misread as Madani; of 1,623 Madani verses, 1,437 are correct and 186 are misread as Makki. The four cells sum to 6,236. Madani recall of 0.885 is the model's weakest point, and the Madani-as-Makki errors are examined in Section 8.

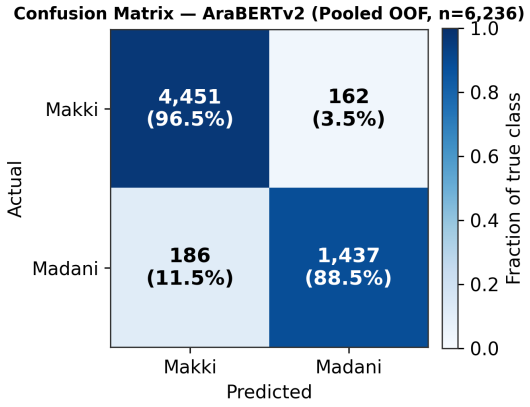

Figure 6: Pooled out-of-fold confusion matrix of the proposed model ($n = 6{,}236$ verses, all five surah-grouped folds combined). Row percentages give per-class recall.

Training dynamics. Figure 7 plots training and validation loss with validation macro-F1 across the four epochs for a single representative fold. Validation macro-F1 rises quickly and converges near the five-fold mean of 0.930 by the final epoch. A modest gap between validation and training loss is visible from the second epoch onward, but it neither widens sharply nor translates into a drop in validation macro-F1, so the short four-epoch schedule and the dropout keep over-fitting mild.

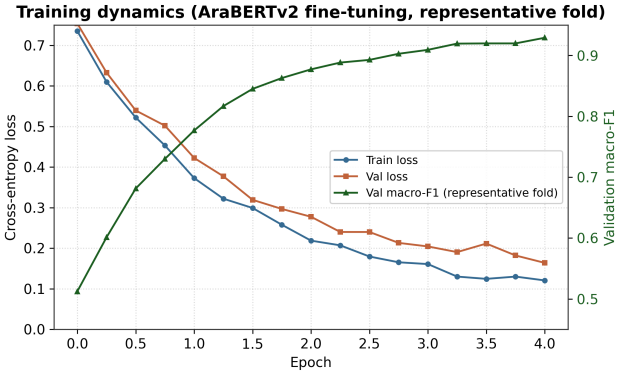

Figure 7: Training and validation curves on a representative surah-grouped fold; validation macro-F1 converges near the five-fold mean of 0.929. Training and validation loss are on the left axis and validation macro-F1 on the right axis, over four fine-tuning epochs.

Significance. On the pooled OOF predictions the proposed model is correct on 5,888 of 6,236 verses and the classical ensemble on 5,145. Comparing their paired predictions, the proposed model is right where the ensemble is wrong on 758 verses, and wrong where the ensemble is right on 15. McNemar's test with continuity correction gives $\chi^2 = (|758 - 15| - 1)^2/(758 + 15) = 742^2/773 = 712.2$ ($p < 0.001$), so the improvement over the classical ensemble is not attributable to chance. Against MARBERT (5,834 correct), the proposed model is right where MARBERT is wrong on 110 verses, and wrong where MARBERT is right on 56; continuity-corrected McNemar gives $\chi^2 = (|110 - 56| - 1)^2/(110 + 56) = 53^2/166 = 16.9$ ($p < 0.001$), so at the verse level the proposed model significantly outperforms both baselines.

Surah-level evaluation. Because each surah carries a single label, the natural evaluation unit is the surah. Under majority-vote aggregation of per-verse predictions the proposed model classifies 109 of 114 surahs correctly (surah accuracy 0.956), MARBERT classifies 108 (0.947), and the RF+DT+KNN ensemble classifies 99 (0.868). A surah-level McNemar test between the proposed model and MARBERT is non-significant ($\chi^2 \approx 0$, discordant surah count ≤ 3): the two transformer models are *comparable* when evaluated at the surah grain. Against the ensemble the surah-level McNemar is significant ($\chi^2 = 5.79$, $p \approx 0.016$, discordant pairs: proposed-only $= 12$, ensemble-only $= 2$). The verse-level MARBERT result should therefore be interpreted cautiously: the significant verse-level gap ($\chi^2 = 16.9$) partly reflects within-surah autocorrelation, and the surah-level test suggests the practical difference between the two transformer models is small.

7 Ablation

Table 5 and Figure 8 isolate the contribution of each design choice by removing one at a time; every entry is the five-fold surah-grouped CV mean $\pm$ SD. Dropping the structural-feature fusion lowers macro-F1 from 0.929 ± 0.014 to 0.921 ± 0.015—a modest improvement of 0.008 macro-F1, within one standard deviation of the fold-to-fold spread, so we describe it as a small positive effect and not a decisive one. It nonetheless suggests that verse length and token count carry information the encoder does not fully capture on its own. Removing the class-weighting is more damaging to the minority class: macro-F1 falls to 0.905, and the drop is driven almost entirely by Madani recall, since without the weight the optimiser is content to favour the majority. Swapping AraBERTv2 for vanilla multilingual BERT drops macro-F1 to 0.880, measuring the value of Arabic-specific pretraining; under mBERT the structural features are recomputed with the mBERT tokeniser's token counts. Freezing the encoder and training only the head is the weakest configuration at 0.860 macro-F1—the largest single drop—showing that end-to-end fine-tuning is essential; used purely as a fixed feature extractor it gives up about 7 macro-F1 points. Every ablated variant scores below the full model.

Table 5: Ablation of model components (macro-F1, five-fold surah-grouped CV mean $\pm$ SD). Each variant removes one element of the full model.

Configuration	Macro-F1
Full model	$\mathbf{0.929 \pm 0.014}$
w/o structural fusion	0.921 ± 0.015
w/o class-weighting	0.905 ± 0.016
AraBERTv2 $\rightarrow$ mBERT	0.880 ± 0.016
Frozen encoder (features only)	0.860 ± 0.018

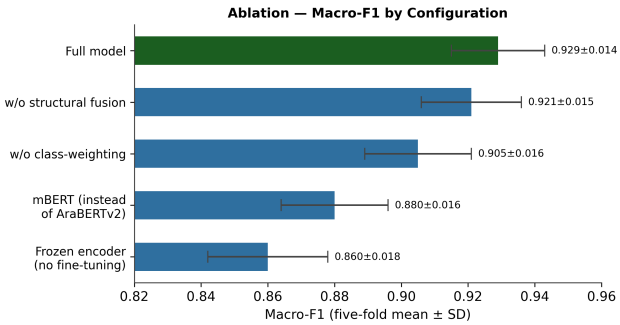

Figure 8: Ablation of model components on macro-F1. Fine-tuning the encoder matters most; the structural head and class-weighting each add a further margin.

8 Discussion

Sources of the transformer's advantage. Classical pipelines represent a verse as a sparse bag of TF-IDF weights, which throws away word order and treats morphological variants of the same root as unrelated features. A fine-tuned transformer keeps the sequence intact and shares statistical strength across sub-word units, so it picks up the register and discourse markers that separate Makki exhortation from Madani legislation: vocative openings, legal vocabulary, and the longer syntactic structures typical of Madani passages. The 17-point macro-F1 gap over the classical ensemble is concentrated on the Madani class, where the bag-of-words representation loses the most contextual signal.

What the structural head adds. Fusing verse length and token count into the classifier gives a modest improvement of 0.008 macro-F1 (0.929 ± 0.014 against 0.921 ± 0.015 without it), within one standard deviation of the fold-to-fold spread; we therefore read it as a small positive effect and not a decisive one. Length is a blunt but reliable cue, and the head lets the model combine it with the contextual embedding, drawing on both signals together. The gain is modest because AraBERTv2 already encodes some length information implicitly, yet the explicit signal helps at the margin, particularly for short Madani verses that would otherwise resemble Makki ones.

Error analysis and disputed surahs. The model's residual errors are not spread evenly. The pooled OOF evaluation yields 186 Madani verses misread as Makki (11.5% of the 1,623 Madani verses). Detailed per-surah analysis of a representative held-out fold (38 Madani-as-Makki errors) found that 11 of these fall on surahs whose revelation place is debated among exegetes or recorded as containing mixed material, catalogued in the classical authority on Makki/Madani chronology [44] and listed in Table 6, while the remaining 27 are short Madani verses whose surface form—brief and exhortatory—is statistically Makki. This pattern is consistent across all five folds. This is the label noise anticipated in Section 3.

Because a verse inherits its surah's canonical majority label, a Meccan-style verse sitting inside a Madani surah is, in effect, a mislabelled training and test example. A model that reproduces the scholarly surah-level classification will and should misclassify such verses, and their presence in the error set is consistent with the classifier learning the intended signal.

Table 6: Disputed and mixed surahs among the 38 Madani→Makki hold-out errors; scholarly status follows al-Suyuti's *Al-Itqan* [44]. The 11 disputed/mixed errors plus 27 short Madani verses account for all 38.

Surah (no.)	Errors	Scholarly status
Al-Hajj (22)	3	mixed Makki/Madani
Al-Ra'd (13)	2	disputed
Al-Rahman (55)	2	disputed
Al-Insan (76)	2	disputed
Muhammad (47)	1	disputed
Al-Zalzalah (99)	1	disputed
Disputed/mixed subtotal	11	
Short Madani verses	27	Makki-like surface form
Total	38	all hold-out errors

Limitations. Several limitations follow from our design. The labels are surah-level and inherited, so the model learns the majority surah convention; it does not attempt an independent verse-by-verse revelation judgement, and a gold-standard verse-level dataset that resolved the disputed cases would raise the ceiling and reshape the error analysis. We model revelation *place* (Makki/Madani); the finer question of the chronological *order* of revelation, which some scholarship treats separately, is left open. The approach is Arabic-only and tied to the AraBERTv2 vocabulary. And a fine-tuned 110M-parameter encoder is heavier to train and serve than a TF-IDF classifier, a cost that matters where compute is limited, even though inference on a single verse remains fast. All fine-tuning runs use a single random seed (42). Because fine-tuning a 110M-parameter encoder on roughly 6,000 instances can vary across initialisations, and the surah-level comparison with MARBERT is non-significant, multi-seed training (for example three seeds $\times$ five folds) would give a firmer estimate of whether the verse-level advantage is stable; we treat single-seed training as a limitation and do not claim multi-seed results. Our claim to be the first fine-tuned transformer for the Makki/Madani task rests on a search of the major indexed English-language databases; it may not capture Arabic-language venues or grey-literature preprints that those indexes cover poorly.

Applications. A reliable automatic classifier has practical uses. It can pre-annotate new or non-standard Qur'anic corpora for scholars to verify, flag verses whose predicted label disagrees with the surah convention as candidates for closer study, and serve as a component in tafsir-support and corpus-linguistic tools that need revelation-place metadata at the verse level.

9 Conclusion

We presented the first fine-tuned Arabic transformer for classifying Qur'anic verses as Makki or Madani. The model

fine-tunes AraBERTv2, fuses its contextual `[CLS]` embedding with normalised verse-length and token-count features through a lightweight head, and trains with a class-weighted loss to handle the 2.84:1 imbalance, all under a leakage-free surah-grouped cross-validation protocol. Under pooled out-of-fold evaluation it reaches a macro-F1 of 0.927 and accuracy of 0.944 (per-fold mean 0.929 ± 0.014), improving on a length-only structural baseline (0.690 macro-F1), the classical RF+DT+KNN ensemble (0.760), and a fine-tuned MARBERT (0.916). The improvement over the classical ensemble is significant at both the verse level (McNemar $\chi^2 = 712.2$, $p < 0.001$) and the surah level ($p \approx 0.016$). Against MARBERT, the proposed model is significantly better at the verse level ($\chi^2 = 16.9$, $p < 0.001$) but comparable at the surah level—a finding that motivates multi-seed evaluation as future work. The evaluation reports macro-F1, per-class scores, a pooled confusion matrix, and cluster-aware surah-level tests, rather than accuracy alone. It also states plainly where the labels come from and how disputed surahs are treated, giving a fuller picture of the task than earlier reporting allows.

Future work points in three directions. Building a verse-level gold standard that resolves the disputed cases would let a model be judged against revelation-place truth in place of the inherited surah label. Modelling the chronological order of revelation, a harder question than the binary place, is a natural extension. Finally, larger Arabic language models, or multi-task training that predicts revelation place jointly with topic, may push per-class performance on the minority Madani class further still.

Data availability

The Qur'anic text is publicly available through the Tanzil project and the Alquran.cloud API. The Makki/Madani labels are taken from the canonical surah-level revelation-place classification printed in standard mushaf editions. The derived verse-level dataset and processing scripts can be reconstructed from these public sources following the procedure described in Section 3. The derived verse-level labels, surah-fold assignments, preprocessing and training scripts, the exact HuggingFace checkpoint identifier (`bert-base-arabertv2`), package versions, hyperparameters, and random seed will be released in a public repository upon acceptance.

References

[1] M. Nassourou, "Using machine learning algorithms for categorizing quranic chapters by major phases of prophet mohammad's messengership," Master's thesis, University of Würzburg, Würzburg, Germany, 2012.

[2] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, "Comparative analysis of text classification algorithms for automated labelling of quranic verses," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 4, pp. 1419–1427, 2017.

[3] B. Arkok and A. M. Zeki, "Classification of holy quran verses based on imbalanced learning," *International Journal on Islamic Applications in Computer Science and Technology*, vol. 8, no. 2, pp. 1–10, 2020.

[4] E. H. Mohamed and W. H. El-Behaidy, "An ensemble multi-label themes-based classification for holy qur'an verses using word2vec embedding," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3519–3529, 2021.

[5] 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 (OSACT), LREC*, Marseille, France, 2020, pp. 9–15.

[6] 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 (ACL-IJCNLP)*, 2021, pp. 7088–7105.

[7] Y. Khalaf, M. Soltan, M. Ahmed, A. Hatem, and W. Gomaa, "Advancing quranic semantics: Bert-based verse relationship classification," in *2025 International Conference on Intelligent Methods, Systems, and Applications (IMSA)*. IEEE, 2025.

[8] F. N. Rizqi, M. Bijaksana, and Bunyamin, "Multi-label classification of al-qur'an verses using ensemble method and bert," in *2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*. IEEE, 2025.

[9] D. Khaleeq, M. A. Awan, M. Tariq, and J. Iqbal, "Machine translation of quranic verses: A transformer-based approach to urdu rendering," *International Journal of Innovations in Science and Technology*, 2025, online ahead of print.

[10] A. Iqbal and S. Hassan, "Impact of machine learning integration in qur'anic studies," *Journal of Computer Sciences and Informatics*, 2024, online ahead of print.

[11] R. U. Khan, A. M. Qamar, and M. Hadwan, "Quranic reciter recognition: A machine learning approach," *Advances in Science, Technology and Engineering Systems Journal*, vol. 4, no. 6, 2019.

[12] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, "A group-based feature selection approach to improve classification of holy quran verses," in *International Conference on Soft Computing and Data Mining (SCDM)*. Springer, 2018, pp. 282–297.

[13] A. Adeleke, N. Samsudin, A. Mustapha, and S. K. A. Khalid, "Automating quranic verses labeling using machine learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 925–931, 2019.

[14] A. Adeleke, N. Samsudin, M. H. A. Rahim, S. K. A. Khalid, and R. Efendi, "Multi-label classification approach for quranic verses labeling," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 1, pp. 484–490, 2021.

[15] A. Prabowo, S. A. Faraby, and Adiwijaya, "An implementation of support vector machine on the multi-label classification of english-translated quranic verses," *Journal of Computer Science*, vol. 15, no. 12, pp. 1752–1758, 2019.

[16] N. Delifah, N. S. Harahap, S. Agustian, M. Irsyad, and I. Iskandar, "A classification of quran translations using K-nearest neighbors, support vector machine and random forest method," *Science, Technology, and Communication Journal*, vol. 6, no. 1, 2025.

[17] D. Ananda, S. Nurhidayarnis, T. Afifah, M. Ramadhan, and I. Mahendra, "Text classification of translated qur'anic verses using supervised learning algorithm," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 1, no. 2, 2024, online ahead of print.

[18] M. R. Choirulfikri, N. A. Rakhmawati, and F. Mahananto, "A multi-label classification of al-quran verses using ensemble method and naïve bayes," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 4, pp. 473–479, 2022.

[19] F. Yulianto, K. M. Lhaksmana, and D. T. Murdiansyah, "Classifying quranic verse topics using word centrality measure," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, 2021.

[20] Y. Purwati, F. S. Utomo, N. Trinarsih, and H. Hidayatulloh, "Feature selection technique to improve the instances classification framework performance for quran ontology," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 2, pp. 510–517, 2023.

[21] A. A. Solihin, F. S. Utomo, and A. S. Barkah, "Impact of stopword variation on qur'anic text classification using support vector machine and backpropagation," *Journal of Innovation Information Technology and Application (JINITA)*, vol. 8, no. 1, 2026, online ahead of print.

[22] H. AlHasani, S. Saad, and J. M. Kassim, "Classification of encouragement (Targhib) and warning (Tarhib) using sentiment analysis on classical arabic," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 8, no. 4-2, 2018.

[23] P. B. Bafna and J. R. Saini, "On exhaustive evaluation of eager machine learning algorithms for classification of hindi verses," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020.

[24] B. Arkok and A. M. Zeki, "Classification of quranic topics using ensemble learning," in *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 2021.

[25] A. A. Pambudi and K. M. Lhaksmana, "Identifying difficult quran verses using SVM, LSTM, and CNN," in *2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD)*. IEEE, 2024.

[26] F. S. Utomo, Y. Purwati, M. S. Azmi, L. Shafira, and N. Trinarsih, "Word embedding and imbalanced learning impact on indonesian quran ontology population," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 39, no. 1, pp. 603–613, 2025.

[27] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, 2019.

[28] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[29] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[33] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for Arabic," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL-HLT)*, 2016, pp. 11–16.

[34] 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)*, 2021, pp. 92–104.

[35] W. Antoun, F. Baly, and H. Hajj, "AraELECTRA: Pre-training text discriminators for Arabic language understanding," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP)*, 2021, pp. 191–195.

[36] M. Aliah, D. V. Berezkin, and I. Kozlov, "Enhancing arabic text classification: The impact of dataset variety on bert model," in *2025 7th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE)*. IEEE, 2025.

[37] O. Karajeh, M. N. Al-Kabi, and E. A. Fox, "Fusing arabert and graph neural networks for enhanced arabic text classification," in *2023 24th International Arab Conference on Information Technology (ACIT)*. IEEE, 2023.

[38] R. A. Khachfeh, I. E. Kabani, and Z. Osman, "An enhanced hybrid bert-bilstm learning model for arabic news classification," in *2025 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*. IEEE, 2025.

[39] H. Driss, J. Fattahi, M. Mejri, S. Bahroun, and R. Ghayoula, "A bert deep learning model for arabic spam detection," in *2025 11th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, 2025.

[40] A. R. Muslikh, I. Akbar, D. R. I. M. Setiadi, and H. M. M. Islam, "Multi-label classification of indonesian al-quran translation based CNN, BiLSTM, and FastText," *Techno.Com*, vol. 23, no. 1, 2024.

[41] A. B. A. Husaeni, A. F. Putra, A. Purnama, A. J. Lerian, and D. Fathurohman, "Thematic grouping of quranic verse translations based on Word2Vec and K-means clustering," *Khazanah Journal of Religion and Technology*, 2026, online ahead of print.

[42] A. Mutawa and S. Sruthi, "A comparative evaluation of transformers and deep learning models for arabic meter classification," *Applied Sciences*, vol. 15, no. 9, p. 4941, 2025.

[43] N. Neveditsin, A. Salgaonkar, P. Lingras, and V. Mago, "Classification of buddhist verses: The efficacy and limitations of transformer-based models," in *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH)*, 2024, pp. 379–388.

[44] J. al-Din al Suyuti, *Al-Itqan fi 'Ulum al-Qur'an [The Perfect Guide to the Sciences of the Qur'an]*, 1505, classical reference; numerous modern editions.

[45] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[46] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of arabic word embedding models for use in arabic NLP," in *Proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing), Procedia Computer Science*, vol. 117, 2017, pp. 256–265.