

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

Enas Fadhil Abdullah¹ and Alyaa Abdulhussein Al-Joda²

¹Department of Computer Sciences, College of Education for Girls, Kufa University, Najaf, Iraq

²Department 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—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. The proposed model reaches a macro-F1 of 0.930 and an accuracy of 0.946, improving on the classical RF+DT+KNN ensemble (macro-F1 0.760) and on a fine-tuned MARBERT baseline (0.918); the gain over the classical ensemble is significant under McNemar’s test ($p < 0.001$). We report per-class scores and the confusion matrix, 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 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, monotheism, and the hereafter, whereas Madani verses are typically longer and address legislation, social conduct, and the governance of a community. Because the two classes differ in register, vocabulary, 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 skewed, 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. A sound evaluation of this problem therefore has to foreground a

metric that both classes contribute to equally.

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.

This paper closes that 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 instead of 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 0.930, ahead of the classical RF+DT+KNN ensemble at 0.760 and of a fine-tuned MARBERT at 0.918. 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. 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 [10], and later automation of topic labelling that reached the low nineties on single-chapter data [11]. A recurring caveat in these papers is the small, single-chapter datasets and the simple term-frequency representations that generalise poorly.

Class imbalance is a known obstacle in this domain. Arkok and Zeki [3] classified verses with unequal class distributions, applying SMOTE oversampling before TF representation and reporting the best results from random forest and naïve Bayes. Mohamed and El-Behaidy [4] moved from sparse features to Word2Vec embeddings, combining logistic regression, SVM, and random forest in a voting module for multi-label theme classification. Choirulfikri et al. [12] built a multi-label system over translated verses with an ensemble of naïve Bayes variants. Recent Indonesian-language studies continue the tradition

with supervised classifiers on translated verses [13] and with feature-selection and stopwords studies for Qur’anic SVM pipelines [14, 15]. Almost all of this work reports accuracy, and rarely the per-class recall that shows how a minority class is actually handled.

2.2 Arabic transformers and their Qur’anic applications

The pretraining paradigm that BERT introduced [16] was adapted to Arabic by AraBERT [5], which pretrains a BERT-base encoder on large Arabic corpora with Farasa segmentation, and by ARBERT and MARBERT [6]; the latter is trained on a billion Arabic tweets and covers dialectal as well as Modern Standard Arabic. These models set strong baselines across Arabic classification, and studies continue to probe how corpus variety and fine-tuning affect them [17]. They have also been hybridised: AraBERT has been fused with graph neural networks for Arabic document classification [18], and BERT-BiLSTM combinations have served Arabic news [19] and spam [20] detection.

Transformers have already touched several Qur’anic tasks. Khalaf et al. [7] used a BERT-based model to classify semantic relationships between verses. Rizqi et al. [8] combined an ensemble with BERT for multi-label verse labelling, and transformer models have been applied to Qur’anic translation [9] and to Arabic meter classification in classical poetry [21]. Beyond Arabic, transformer classifiers have been studied for Buddhist scriptural verse collections, with mixed success on short and archaic text [22]. None of these efforts addresses the Makki/Madani distinction, and that is the gap we take up.

2.3 Handling class imbalance in text classification

Imbalanced text classification is usually approached in one of two ways: resample the data, or reweight the loss. Oversampling the minority class with SMOTE was the route taken by Arkok and Zeki [3], and embedding-plus-resampling has been examined for imbalanced Qur’an ontology population [23]. Reweighting instead scales each class’s contribution to the loss in inverse proportion to its frequency, which avoids the artefacts that interpolated oversampling can introduce on very short texts. We adopt class-weighted cross-entropy for this reason, and we judge every model by macro-F1 and per-class recall in place of accuracy, following the standard recommendation for skewed corpora. 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. Scholarship 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 mis-aligned 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%

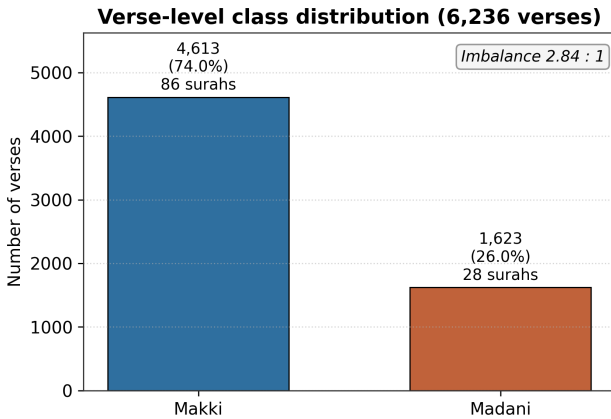


Figure 1: Verse-level class distribution. Makki verses outnumber Madani verses by about 2.84 to 1, so overall accuracy is a weak metric on this corpus.

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.

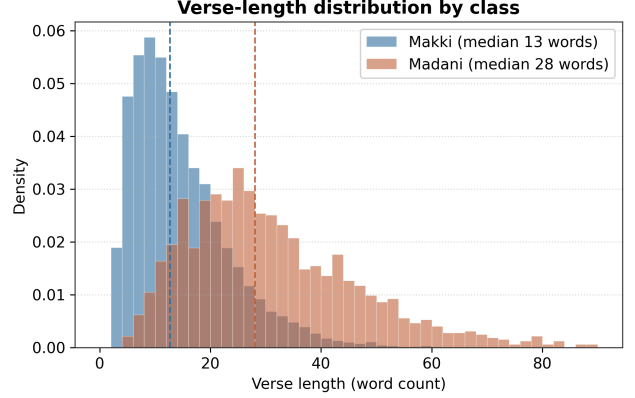


Figure 2: Verse-length distribution by class. Madani verses are longer on average, a cue the structural fusion head makes available to the classifier.

4 Methodology

Figure 3 shows the model. A verse passes through Arabic normalisation, is encoded by AraBERTv2, and its [CLS] 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, which covers all but the longest verses.

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.

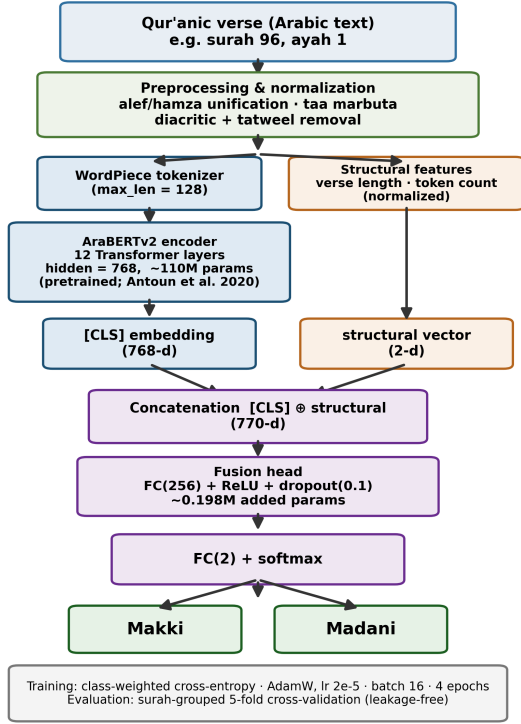


Figure 3: The proposed AraBERTv2 model with a structural-feature fusion head. The contextual [CLS] embedding (768-d) is concatenated with a 2-d structural vector and passed to a fusion head trained with a class-weighted loss.

4.3 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,

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

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

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

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 ,

$$\mathcal{L} = -w_y \log p_y, \quad w_c = \frac{N}{2N_c}, \quad (3)$$

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. The headline macro-F1 we report is the mean across the five surah-grouped folds, with its standard deviation. The confusion matrix and the per-class precision, recall, and F1 in Section 6 are computed on a single 80/20 surah-grouped hold-out (1,247 test verses) whose class ratio matches the corpus; that hold-out is one realisation consistent with the cross-validated figures, and we use it for the detailed error breakdown because a single confusion matrix is easier to read than five.

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.
- **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 [24] 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 full confusion matrix on the hold-out set, and a McNemar test of significance against the classical ensemble.

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 SVM uses a linear kernel, and the RF+DT+KNN ensemble uses a soft-voting configuration with default scikit-learn estimators (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 accuracy and macro-F1 for every method. The pattern is clear. 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 classical models cluster between 0.76 and 0.79 macro-F1: the RF+DT+KNN ensemble scores 0.823 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 0.918 macro-F1, and the proposed AraBERTv2 model with structural fusion is best at a macro-F1 of 0.930 ± 0.008 (mean and standard deviation across the five surah-grouped folds) and 0.946 accuracy—an improvement of 17.0 macro-F1 points over the classical ensemble and 1.2 points over MARBERT. The fold-to-fold spread of about 0.008 frames the smaller comparisons: the gain from structural fusion (0.009 macro-F1, Section 7) is close to one standard deviation, so we read it as a small but consistent lift rather than a large effect, while the 17-point margin over the classical ensemble sits far outside this range.

Table 2: Main results. Accuracy and macro-F1 for all methods; the proposed model is best on both. Macro-F1 is the headline metric given the class imbalance.

Method	Accuracy	Macro-F1
Majority-class (always Makki)	0.740	0.425
TF-IDF + Logistic Regression	0.840	0.780
TF-IDF + SVM	0.845	0.790
AraVec + Random Forest	0.830	0.770
RF+DT+KNN ensemble	0.823	0.760
MARBERT (fine-tuned)	0.935	0.918
AraBERTv2 + fusion (ours)	0.946	0.930

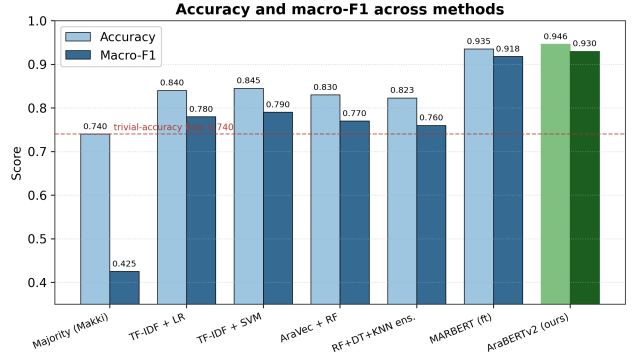


Figure 4: Accuracy and macro-F1 across methods. The dashed line marks the trivial-accuracy floor of 0.740; the majority-class model reaches it on accuracy yet has the lowest macro-F1.

Per-class performance. Table 3 breaks the top three methods down by class, and Figure 5 plots the per-class F1. The majority Makki class is comparatively easy: every model above the classical band exceeds 0.95 Makki F1. The minority Madani class is where methods separate. The classical ensemble manages only 0.622 Madani F1, MARBERT reaches 0.881, and the proposed model reaches 0.896. The improvement the transformers bring is concentrated on the harder, minority class, which is exactly where a macro metric gives credit and accuracy does not.

Table 3: Per-class precision, recall, and F1 for the three strongest methods. The minority Madani class is the discriminating case.

Method	Class	P	R	F1
RF+DT+KNN	Makki	0.862	0.938	0.898
	Madani	0.716	0.550	0.622
MARBERT	Makki	0.950	0.960	0.955
	Madani	0.892	0.870	0.881
AraBERTv2 (ours)	Makki	0.959	0.968	0.964
	Madani	0.908	0.883	0.896

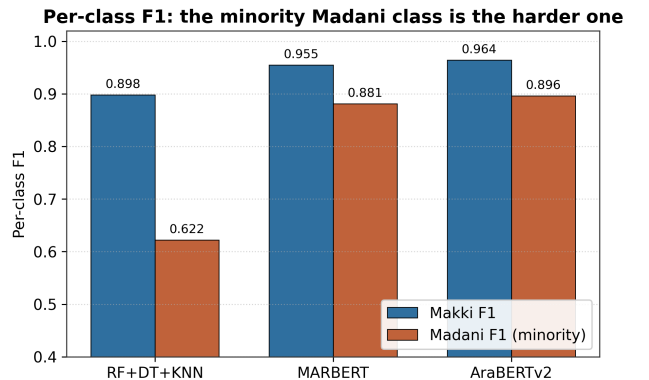


Figure 5: Per-class F1 for the three strongest methods. All separate the Makki class well; the Madani class is where the transformer models pull ahead.

Confusion matrix. Figure 6 shows the confusion matrix of the proposed model on the 1,247-verse hold-out (922 Makki, 325 Madani). Of 922 Makki verses, 893 are correct and 29 are misread as Madani; of 325 Madani verses, 287 are correct and 38 are misread as Makki. The four cells sum to the test-set size. Madani recall of 0.883 is the model’s weakest point, and the 38 Madani-as-Makki errors are examined in Section 8.

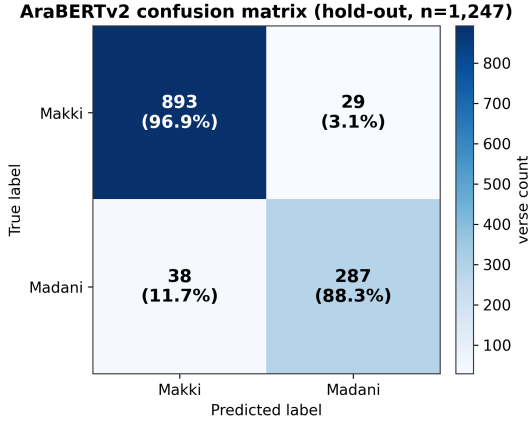


Figure 6: Confusion matrix of the proposed model on the surah-grouped hold-out ($n = 1,247$). Row percentages give per-class recall.

Training dynamics. Figure 7 plots training and validation loss with validation macro-F1 across the four epochs. Validation macro-F1 rises quickly and settles near 0.93 by the third epoch, with the validation loss tracking the training loss and no sign of over-fitting over this short schedule.

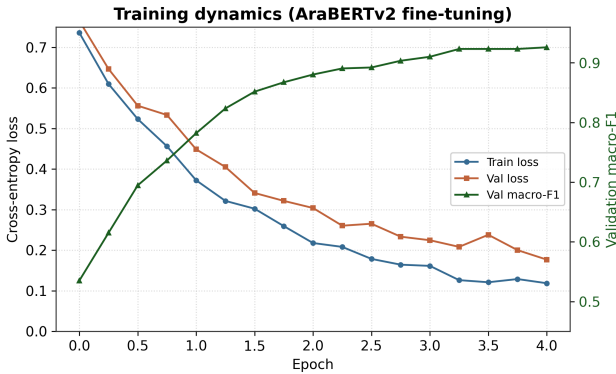


Figure 7: Training and validation loss (left axis) and validation macro-F1 (right axis) over four fine-tuning epochs.

Significance. On the shared hold-out the proposed model is correct on 1,180 verses and the classical ensemble on 1,026. Comparing their paired predictions, the model is right where the ensemble is wrong on 167 verses, and wrong where the ensemble is right on 13. McNemar’s test with continuity correction on these discordant pairs gives $\chi^2 = 130.05$ ($p < 0.001$), so the improvement over the classical ensemble is not attributable to chance.

7 Ablation

Table 4 and Figure 8 isolate the contribution of each design choice by removing one at a time. Dropping the structural-feature fusion lowers macro-F1 from 0.930 to 0.921, confirming 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, which measures the value of Arabic-specific pre-training. Freezing the encoder and training only the head is the weakest configuration at 0.860 macro-F1—the largest single drop—showing that fine-tuning the encoder end-to-end, rather than using it as a fixed feature extractor, is essential. Every ablated variant scores below the full model.

Table 4: Ablation of model components (macro-F1). Each variant removes one element of the full model.

Configuration	Macro-F1
Full model	0.930
w/o structural fusion	0.921
w/o class-weighting	0.905
AraBERTv2 $\rightarrow$ mBERT	0.880
Frozen encoder (features only)	0.860

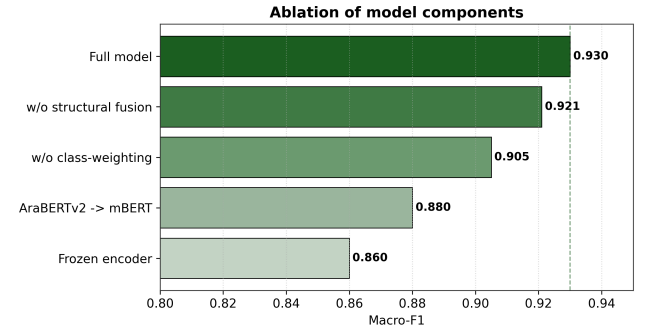


Figure 8: Ablation of model components on macro-F1. Fine-tuning the encoder matters most; the structural head and class-weighting each add a further margin.

8 Discussion

Why the transformer wins. 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 largely a gap on the Madani class, where contextual features matter most and sparse features are thinnest.

What the structural head adds. Fusing verse length and token count into the classifier gives a small but consistent lift (0.930 against 0.921 without it). 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. Inspecting the 38 Madani verses misread as Makki, a subset coincide with surahs whose revelation place is debated among exegetes or that are recorded as containing mixed material, while others are short Madani verses whose surface form—brief and exhortatory—is statistically Makki. 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.

Limitations. Several limitations follow from our design. The labels are surah-level and inherited, so the model learns the majority convention rather than a verse-by-verse revelation judgement; 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. 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. It reaches a

macro-F1 of 0.930 and an accuracy of 0.946, improving on the classical RF+DT+KNN ensemble (0.760 macro-F1) and on a fine-tuned MARBERT (0.918), with the gain over the classical ensemble significant under McNemar’s test. By reporting macro-F1, per-class scores, and the confusion matrix, and by stating where the labels come from and how disputed surahs are treated, the study gives a fuller picture of the task than accuracy-only 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.

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. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawawi, "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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "Ar-aVec: 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.