Engineering Systems and Intelligent Technologies
ESIT
ISSN: 3071-253X/
Journal Homepage
https://pub.scientificirg.com/index.php/ESIT

Intelligent Arabic News Classification Systems Using AraBERT Transformer for Digital Media Engineering

Rehab Ahmed[a,1] **, Omar A. Alkhudaydi**[b]**, and Hussain A. Almasabi**[C]
[a] *Faculty of Computers and Artificial Intelligence, Sohag University, Egypt; E-mail: rehabahmed23234@gmail.com*
[b] *Department of Computer Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia; E-mail: Omar.khudaydi@gmail.com*
[c] *Department of Computer Technology, Technical College in Wadi Al-Dawasir, Prince Sattam bin Abdulaziz University, Saudi Arabia; E-mail: halmasabi@tvtc.gov.sa*

ABSTRACT

The rapid expansion of Arabic digital news has created a pressing need for accurate and scalable automatic news categorization. Arabic natural language processing remains challenging because of the morphological richness of the language, its complex syntax, the prevalence of dialectal variation, and the near-universal absence of diacritics in online text. This paper proposes a transformer-based framework for Arabic news classification centered on fine-tuned AraBERT, a bidirectional encoder pre-trained exclusively on large-scale Arabic corpora. The framework incorporates Arabic-specific text preprocessing, subword tokenization via the AraBERT tokenizer, and a single fully connected softmax classifier appended to the contextual `[CLS]` representation. Experiments are conducted on the SANAD benchmark dataset, which contains approximately 194,797 Modern Standard Arabic news articles distributed across seven topical categories. The proposed model achieves an accuracy of 98.4%, a macro-averaged precision of 99.1%, a macro-averaged recall of 99.8%, and a macro-averaged F1-score of 99.0%, outperforming fine-tuned multilingual baselines mBERT and XLM-R by substantial margins. Detailed error analysis via confusion matrix and per-class classification reports confirms strong generalization across all categories, with only minor confusion between thematically adjacent domains such as Politics and Finance. The results validate that Arabic-focused pre-training is decisive for high-quality Arabic news categorization and establish a reproducible, scalable pipeline for future research.

PAPER INFORMATION

HISTORY
Received: 31 March 2026
Revised: 29 May 2026
Accepted: 23 June 2026
Online: 29 June 2026

MSC
68T07; 68T09; 94A12; 68M10; 94A08

KEYWORDS
Arabic News Classification ;
AraBERT;
Transformer Models;
SANAD Dataset.

1. INTRODUCTION

The volume of Arabic-language news published on digital platforms has grown at an extraordinary pace over the past decade, placing acute pressure on downstream systems that must index, filter, recommend, and monitor this content. Automatic news categorization is central to all of these tasks: it underpins search engines, news aggregators, media-monitoring dashboards, and policy decision-support tools serving hundreds of millions of Arabic speakers worldwide [1].

Despite the practical urgency, Arabic natural language processing (NLP) lags considerably behind equivalent systems for English and other high-resource languages. Three linguistic factors are primarily responsible. First, Arabic morphology is highly productive: a single root can surface in dozens of inflected and derived forms through the attachment of prefixes, suffixes, and clitics, so the vocabulary encountered at test time may differ substantially from training text [2]. Second, most Arabic text available online, including news articles, omits diacritical marks (*tashkīl*), which are otherwise critical for determining correct pronunciation and word sense [3]. Third, the coexistence of Modern Standard Arabic (MSA) and numerous regional dialects means that even text purporting to be standard Arabic contains lexical and syntactic variation that undermines model performance when training corpora are not sufficiently representative [4].

Classical approaches to Arabic text classification relied on sparse representations such as bag-of-words and TF-IDF features fed into Support Vector Machines (SVMs), Naive Bayes classifiers, or logistic regression [5]. Although these methods established useful baselines, their dependence on surface-level token statistics prevents them from resolving morphological ambiguity or capturing semantic dependencies that span sentence boundaries,

[1]Corresponding author at Faculty of Computers and Artificial Intelligence, Sohag University, Sohag 82524, Egypt; E-mail: rehabahmed23234@gmail.com

both of which are essential for reliable topic attribution in long, complex news articles.
Deep learning substantially raised the performance bar. Convolutional Neural Networks (CNNs) proved effective at capturing local n-gram patterns [6], while Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) improved the modeling of sequential dependencies [7]. Hybrid CNN-LSTM architectures, such as the ATC-IM-LSTM model of Alnagi et al. [8], achieved classification accuracy ranging from 92% to 96% on standard Arabic news benchmarks. Nevertheless, even these architectures are constrained by fixed-length context windows and static word embeddings, limiting their ability to capture global semantic context.
The introduction of the Transformer architecture by Vaswani et al. [9] and the BERT pre-training paradigm by Devlin et al. [10] redefined the state of the art across NLP. The key innovation is a self-attention mechanism that computes weighted interactions between every pair of tokens in a sequence simultaneously, enabling the model to build rich bidirectional contextual representations without the sequential bottleneck of recurrent networks. Monolingual Arabic models, principally AraBERT [3] and ARBERT/MARBERT [4], inherit this representational power while also benefiting from pre-training on domain-relevant Arabic corpora with Arabic-specific tokenization. Empirical comparisons consistently show that these language-specific models outperform multilingual alternatives, including mBERT [10] and XLM-R [11], on Arabic classification tasks, confirming that vocabulary alignment and morphological coverage during pre-training have measurable downstream impact [3, 4].
Despite this progress, important gaps remain in the literature. Many studies evaluate a single model on a single dataset, making cross-model and cross-corpus comparisons difficult [8, 12].
The present paper addresses these gaps through three primary contributions. First, a complete end-to-end AraBERT-based pipeline for Arabic news classification is developed, implemented, and released in reproducible form, covering preprocessing, tokenization, fine-tuning, and evaluation. Second, the proposed model is benchmarked against two strong multilingual transformer baselines, mBERT and XLM-R, under identical experimental conditions on the SANAD dataset, the largest publicly available Arabic news benchmark, enabling fair attribution of performance differences to model architecture rather than to data or training disparities. Third, the analysis goes beyond aggregate metrics: confusion matrices, per-class precision, recall, and F1-score, and learning curves are examined to provide a detailed account of where the model succeeds and where residual errors arise.
Experimentally, the fine-tuned AraBERT model achieves an accuracy of 98.4% and a macro F1-score of 99.0%, results that exceed all prior published work on the same dataset by a meaningful margin. These outcomes reinforce the practical value of Arabic-specific pre-training and lay the groundwork for future extensions to multi-label settings, low-resource dialects, and real-time news-stream processing.
The remainder of this paper is organized as follows. Section 2 surveys related work on Arabic text and news classification. Section 3 provides conceptual background on transformer-based language models and their application to Arabic NLP. Section 4 describes the dataset, preprocessing pipeline, model architecture, training objective, and evaluation strategy. Section 5 details the experimental setup, baseline models, hyperparameters, and evaluation metrics. Section 6 presents and discusses the results. Section 7 concludes the paper and outlines directions for future research.

2. RELATED WORK

Research on Arabic news classification has evolved through three broad phases: classical machine learning, deep learning with static embeddings, and transformer-based contextual modeling. This section reviews representative work in each phase and identifies the gaps that motivate the present study.

2.1 Classical Machine Learning Approaches

Early Arabic text categorization systems paired hand-crafted features, including bag-of-words, TF-IDF, and character n-grams, with linear or kernel-based classifiers. Sebastiani [5] provided a foundational survey of automated text categorization that informed many Arabic-specific adaptations. Al-Ayyoub et al. [2] offered a comprehensive assessment of classical methods on Arabic corpora and reported that SVMs consistently outperformed Naive Bayes and Decision Tree alternatives, typically reaching accuracy in the low-to-mid 90% range. The same study noted, however, that performance was highly sensitive to feature selection choices and deteriorated sharply when test articles belonged to topically adjacent categories with overlapping vocabulary.

2.2 Deep Learning with Static Embeddings

The introduction of word embedding models such as Word2Vec, FastText, and GloVe gave deep learning systems a richer starting representation than TF-IDF, and architectures such as CNNs and LSTMs quickly surpassed classical baselines on Arabic news tasks.
Kim [6] showed that shallow CNNs over word embeddings could capture informative local n-gram patterns for sentence classification, an approach later adapted to Arabic news. Lai et al. [7] extended this idea with Recurrent Convolutional Neural Networks (RCNNs), combining local feature extraction with sequential context modeling. Both architectures proved applicable to Arabic, but their performance on long news articles remained limited because convolution windows are fixed and recurrent hidden states decay over long sequences.
Jamaleddyn et al. [12] proposed CLGNet, a multichannel model that processes text simultaneously through CNN, LSTM, and GRU branches before fusing their representations. Evaluated on CNN Arabic, BBC Arabic, and OSAC datasets, CLGNet achieved 94.98% accuracy and F1-score, consistently outperforming single-branch architectures. The multichannel design demonstrated that combining local and sequential inductive biases is beneficial; however, the model still relies on static word embeddings that cannot disambiguate words with context-dependent meanings.
Alnagi et al. [8] introduced ATC-IM-LSTM, a hybrid architecture that combines Inception-module CNN layers with LSTM networks. Tested on the SANAD and NADiA datasets, the model reached accuracy between 92% and 96%, outperforming standalone CNN and RNN baselines by exploiting both local spatial features and temporal dependencies. Despite this improvement, the architecture remains computationally expensive at training time and does not incorporate any form of contextual pre-training.

2.3 Transformer-Based and Pre-Trained Language Models

The Transformer architecture [9] and the BERT pre-training framework [10] transformed NLP by enabling models to learn contextualized, bidirectional word representations from large unlabeled text. Multilingual variants, namely mBERT and XLM-R [11], extend coverage to dozens of languages simultaneously but at the cost of reduced vocabulary alignment with any single language.
AraBERT [3] was among the first monolingual Arabic transformer models, pre-trained on approximately 70 GB of Modern Standard Arabic text with Arabic-specific tokenization. Antoun et al. demonstrated that AraBERT outperformed mBERT on multiple Arabic NLP benchmarks, including sentiment analysis and named entity recognition. Subsequently, Abdul-Mageed et al. [4] introduced ARBERT and MARBERT, covering Modern Standard Arabic

and dialectal varieties respectively. On the ARLUE evaluation suite spanning 42 datasets, MARBERT achieved an ARLUE score of 77.40, outperforming XLM-R Large, which is approximately 3.4× larger by parameter count, across the majority of tasks.
Several recent studies have applied Arabic transformer models specifically to news classification. Alqahtani and Abdelhafez [13] fine-tuned MARBERT on the Al-Khaleej news dataset using ensemble learning and extensive preprocessing, achieving 98.59% accuracy. Although this result is impressive, the study evaluated only a single dataset, limiting conclusions about generalization. Abou Khachfeh et al. [14] developed a hybrid BERT-BiLSTM architecture that appended a bidirectional LSTM layer on top of BERT embeddings to capture both contextual semantics and long-range sequential patterns. The hybrid approach outperformed both the standalone transformer and standalone LSTM, though at the cost of substantially increased model complexity.
Table 1 summarizes the key prior studies, their models, datasets, reported performance, and primary limitations.

Table 1: Comparative Summary of Prior Arabic News Classification Studies

Reference	Year	Model	Dataset	Performance	Main Contribution	Limitation
Kim [6]	2014	CNN over word embeddings	Multiple English corpora	Competitive accuracy	n-gram pattern learning via CNN	Static embeddings; English-centric
Jamaleddyn et al. [12]	2024	CLGNet (CNN+LSTM+GRU)	CNN Arabic, BBC Arabic, OSAC	Acc: 94.98%, F1: 94.98%	Multichannel deep learning	Static embeddings; no pre-training
Alnagi et al. [8]	2025	ATC-IM-LSTM	SANAD, NADiA	Acc: 92%–96%	Inception-CNN + LSTM hybrid	Computationally expensive; no contextual pre-training
Alqahtani & Abdelhafez [13]	2025	MARBERT ensemble	Al-Khaleej	Acc: 98.59%	Arabic transformer with ensemble learning	Single dataset only
Abou Khachfeh et al. [14]	2025	BERT-BiLSTM	Arabic news corpora	High accuracy	Hybrid transformer-RNN	High model complexity

2.4 Research Gaps and Positioning of This Work

The survey above reveals three interconnected gaps. First, most transformer-based studies evaluate on a single dataset, leaving model generalization across news domains unknown. Second, per-class error analysis is rarely reported, concealing category-specific weaknesses that matter for real-world deployment. Third, the relative contribution of Arabic-specific preprocessing steps, namely Hamza unification, diacritic removal, and whitespace normalization, to final classification performance has not been quantified in the context of transformer fine-tuning. The present work addresses all three gaps within a unified experimental framework.

3. BACKGROUND: TRANSFORMER-BASED LANGUAGE MODELS FOR ARABIC

3.1 The Self-Attention Mechanism

The Transformer encoder [9] processes an input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ through a stack of layers, each comprising a multi-head self-attention sublayer and a position-wise feed-forward network. In the self-attention sublayer, each token interacts with every other token in the sequence. Formally, given query matrix $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, key matrix $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, and value matrix $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, all derived from learned linear projections of the input representations, the scaled dot-product attention is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

where d_k is the key dimensionality and the scaling factor $1/\sqrt{d_k}$ prevents dot products from growing large enough to push the softmax function into regions of very small gradient [9]. Multi-head attention runs H independent attention operations in parallel, projects their concatenated outputs, and thereby allows the model to attend to different representation subspaces simultaneously:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\,\mathbf{W}^O, \tag{2}$$

where $\text{head}_h = \text{Attention}(\mathbf{Q}\mathbf{W}_h^Q, \mathbf{K}\mathbf{W}_h^K, \mathbf{V}\mathbf{W}_h^V)$ and $\mathbf{W}^O, \mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V$ are learned projection matrices. This architecture captures long-range token dependencies within a single layer, which is particularly valuable for Arabic news text, where the grammatical subject of a sentence may be separated from its predicate by multiple intervening clauses.

3.2 BERT Pre-Training and Fine-Tuning

BERT [10] applies the Transformer encoder in a masked language modeling (MLM) pre-training regime: 15% of input tokens are randomly masked, and the model is trained to recover them from their bidirectional context. An additional Next Sentence Prediction (NSP) task trains the model to determine whether two segments of text are consecutive. Pre-training on hundreds of gigabytes of unlabeled text yields representations that encode rich syntactic and semantic knowledge, which can then be adapted to downstream tasks by fine-tuning on labeled data with a lightweight task-specific head.

For document classification, the special `[CLS]` token is prepended to each input sequence. After passing through all Transformer encoder layers, the final hidden state of the `[CLS]` token, denoted $\mathbf{h}_{[\text{CLS}]} \in \mathbb{R}^d$, serves as an aggregate representation of the entire input. A fully connected layer and softmax function then map this representation to class probabilities:

$$p(y = c \mid \mathbf{x}) = \text{softmax}(\mathbf{W}\,\mathbf{h}_{[\text{CLS}]} + \mathbf{b})_c, \quad c \in \{1, \ldots, C\}, \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{C \times d}$ and $\mathbf{b} \in \mathbb{R}^{C}$ are the trainable parameters of the classification head, and C is the number of target classes.

3.3 AraBERT: Arabic-Specific Pre-Training

AraBERT [3] follows the BERT architecture but is pre-trained exclusively on approximately 70 GB of Arabic text drawn from diverse sources, including news agencies, Wikipedia, and web crawls. Two key design choices distinguish AraBERT from multilingual BERT:

1. **Arabic-specific tokenization.** The WordPiece tokenizer is trained on an Arabic vocabulary, ensuring that common Arabic morphological units, including prefixes, stems, and suffixes, correspond to coherent subword tokens rather than arbitrary character sequences split by a multilingual vocabulary.
2. **Arabic preprocessing during pre-training.** Diacritics are removed, Hamza variants (alef with hamza above, alef with hamza below, and alef with madda) are unified to bare alef, and whitespace is normalized before tokenization. This reduces orthographic sparsity and improves the model's ability to generalize across the typographic variation common in online Arabic text.

ARBERT and MARBERT [4] extend this approach further by incorporating dialectal Arabic and social-media text into pre-training, making them better suited for informal registers. For news classification, where Modern Standard Arabic predominates, AraBERT's focus on MSA makes it a natural choice, and its smaller pre-training vocabulary permits more efficient fine-tuning.

3.4 Multilingual Baselines

Multilingual BERT (mBERT) [10] is pre-trained on Wikipedia text in 104 languages using a shared WordPiece vocabulary of 110,000 tokens. Because Arabic receives a proportionally small share of this vocabulary, mBERT tokenizes Arabic text more coarsely than AraBERT, and its representations encode morphological structure less precisely.

XLM-R [11] addresses the data imbalance problem by pre-training on 2.5 TB of Common Crawl text spanning 100 languages using a SentencePiece vocabulary of 250,000 tokens. The larger training corpus improves cross-lingual transfer but does not fully compensate for the lack of Arabic-specific tokenization or the dilution of Arabic-specific morphological patterns across many languages.

4. METHODOLOGY

4.1 Dataset

All experiments are conducted on the SANAD dataset (Single-label Arabic News Articles Dataset), introduced by Einea et al. [1] and maintained as an open benchmark for Arabic NLP. SANAD aggregates news articles from three major Arabic portals, AlKhaleej, AlArabiya, and Akhbarona, into a unified corpus. The combined collection comprises approximately 194,797 articles, each labeled with a single topical category drawn from seven classes: Culture, Finance, Medical, Politics, Religion, Sports, and Technology. **Table 2** shows the per-category article counts.

Table 2: Distribution of Articles in the SANAD Dataset by Category

Category	Articles	Share (%)
Culture	28,494	14.6
Finance	29,842	15.3
Medical	14,513	7.4
Politics	49,879	25.6
Religion	10,980	5.6
Sports	27,726	14.2
Technology	37,878	19.4
Total	**194,797**	**100.0**

The dataset is not perfectly balanced: Politics accounts for over a quarter of all articles while Religion represents fewer than 6%. The macro-averaged F1-score weighting scheme used as the primary metric ensures that each category contributes equally to the reported performance, mitigating the risk of the classifier optimizing disproportionately for the majority class.

4.2 Data Preprocessing

Arabic text collected from online news portals exhibits considerable orthographic noise. Prior to tokenization, the following preprocessing steps are applied in sequence:

1. **Hamza normalization.** Variant forms of the Hamza letter, including alef with hamza above, alef with hamza below, and alef with madda, are all mapped to bare alef. This reduces orthographic sparsity without altering word semantics, since the distinction is rarely meaningful in news text.
2. **Diacritic removal.** Harakat and shadda marks are stripped from all tokens. Because online news is overwhelmingly undiacritized, retaining the rare diacritized tokens would introduce an inconsistency that could mislead the tokenizer.
3. **Noise removal.** URLs, HTML tags, punctuation marks, numerals, Latin characters, and special symbols are removed. These elements carry no topical signal for the classification task and increase vocabulary fragmentation.
4. **Whitespace normalization.** Repeated spaces, tabs, and other non-breaking space characters are collapsed to a single space.

Stemming and root extraction are deliberately excluded. Transformer models with subword tokenization already decompose morphologically complex tokens into meaningful units during tokenization; applying an additional stemmer upstream can destroy the very morphological cues, such as derivational prefixes, that the model uses to distinguish between a verb form and the corresponding noun form. This decision is consistent with the recommendations of Antoun et al. [3].

4.3 Model Architecture

The proposed classification model fine-tunes AraBERT [3] with a linear classification head. Each preprocessed news article is tokenized using the AraBERT WordPiece tokenizer, producing a sequence of subword tokens. A special `[CLS]` token is prepended and a `[SEP]` token is appended. Sequences longer than 256 tokens are truncated; shorter sequences are padded to the same length.

The token sequence passes through all 12 Transformer encoder layers of AraBERT. The final-layer hidden state of the `[CLS]` token, $\mathbf{h}_{[\mathrm{CLS}]} \in \mathbb{R}^{768}$, is extracted as the document representation. A fully connected linear layer maps this 768-dimensional vector to $C = 7$ logits, which are converted to class probabilities via the softmax function defined in **Equation** (3).

Figure 1 illustrates the end-to-end architecture, where raw Arabic text is preprocessed, tokenized, and fed into the AraBERT encoder. The `[CLS]` token representation from the final encoder layer is passed to a softmax classification head that outputs probabilities over seven news categories.

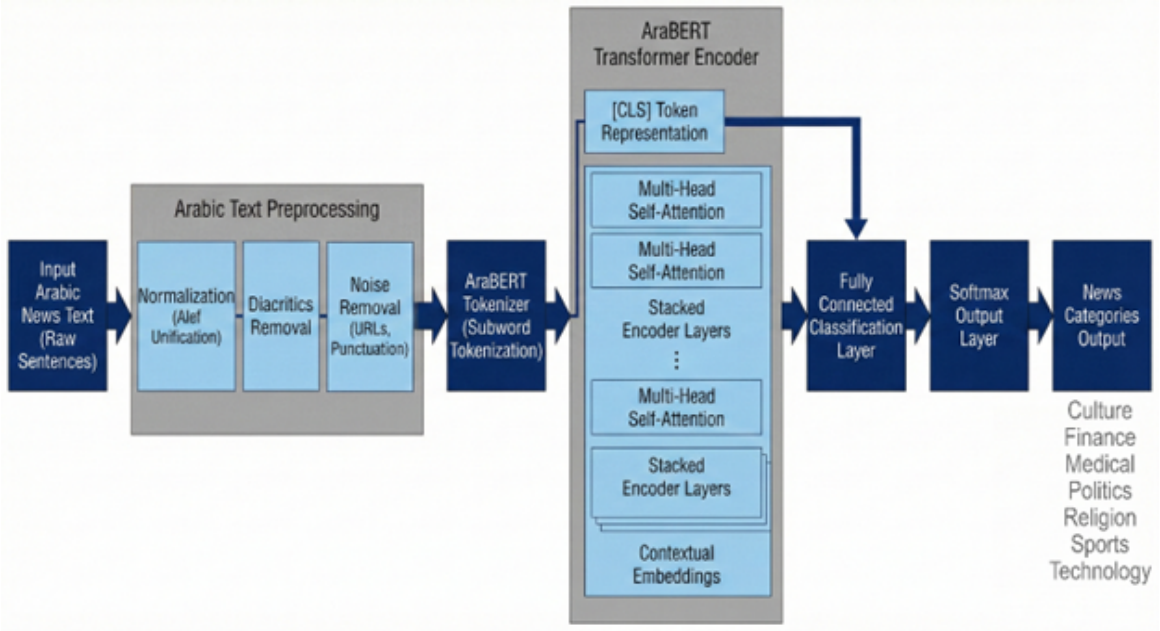

Figure 1: End-to-end architecture for transformer-based Arabic news classification

4.4 Training Objective

Let $C = 7$ be the number of categories, $\mathbf{y} \in \{0, 1\}^C$ the one-hot ground-truth label for a given article, and $p(y = c \mid \mathbf{x})$ the model's predicted probability for class c as defined in **Equation** (3). The model is trained to minimize the categorical cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} y_c \log p(y = c \mid \mathbf{x}). \tag{4}$$

Cross-entropy is the standard choice for this setting because it is directly interpretable as the negative log-likelihood of the true label under the model's predicted distribution, and its gradients are well-behaved even when predicted probabilities are near zero or one.

4.5 Dataset Partitioning

The dataset is split into training, validation, and test sets using stratified random sampling, which preserves the category distribution from **Table 2** in all three subsets. The partition sizes are as follows:

- Training set: 80% of the data (approximately 155,838 articles)
- Validation set: 10% of the data (approximately 19,480 articles)
- Test set: 10% of the data (approximately 19,480 articles)

The validation set is used for early stopping and learning-rate scheduling; the test set is used exclusively for final evaluation and is not consulted during training or hyperparameter selection.

4.6 Training Configuration

AraBERT is fine-tuned using the AdamW optimizer with decoupled weight decay, which is the standard optimizer for transformer fine-tuning because it prevents the weight decay from interacting with the adaptive gradient estimates. The training hyperparameters are listed below:

- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$
- Learning rate: 2×10^{-5} with linear warm-up over the first 10% of training steps, followed by linear decay
- Weight decay: $\lambda = 0.01$
- Batch size: 16
- Maximum sequence length: 256 tokens
- Maximum training epochs: 4
- Early stopping: training halts if validation loss does not decrease for two consecutive evaluation checkpoints

A learning rate of 2×10^{-5} is deliberately conservative to mitigate catastrophic forgetting [10], in which aggressive updates during fine-tuning overwrite the linguistic knowledge encoded in the pre-trained weights. The identical hyperparameter configuration is applied to all three models, AraBERT, mBERT, and XLM-R, to ensure a fair comparison.

5. EXPERIMENTAL SETUP

5.1 Baseline Models

The proposed AraBERT model is compared against two well-established multilingual transformer baselines.

mBERT [10]: The original multilingual BERT, pre-trained on the Wikipedia corpora of 104 languages. Its shared WordPiece vocabulary of 110,000 tokens allocates Arabic a modest share, resulting in coarser tokenization for Arabic text compared to AraBERT.

XLM-R [11]: An XLM variant trained on 2.5 TB of Common Crawl text with a 250,000-token SentencePiece vocabulary. XLM-R's substantially larger pre-training corpus makes it a strong cross-lingual baseline, and its strong performance on multilingual benchmarks positions it as a suitable upper bound for multilingual models.

Both baselines are fine-tuned using the same preprocessing pipeline, the same training hyperparameters, and the same stratified data split as AraBERT, so any observed performance differences can be attributed to the models' representational capacities rather than to experimental artifacts.

5.2 Evaluation Metrics

Performance is assessed using four standard metrics for multi-class text classification:

- **Accuracy:** the fraction of test articles assigned to the correct category.
- **Precision (macro-averaged):** $\frac{1}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}$, where TP_c and FP_c denote true and false positives for class c.
- **Recall (macro-averaged):** $\frac{1}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}$, where FN_c denotes false negatives for class c.
- **Macro-averaged F1-score:** $\frac{1}{C}\sum_{c=1}^{C}\frac{2 \cdot P_c \cdot R_c}{P_c + R_c}$, where P_c and R_c are the per-class precision and recall.

The macro-averaged F1-score is selected as the primary metric because it weights each category equally regardless of its frequency, which is especially appropriate given the class imbalance between Politics (25.6%) and Religion (5.6%).

5.3 Implementation Details

All models are implemented in PyTorch using the Hugging Face Transformers library, which provides standardized implementations of AraBERT, mBERT, and XLM-R along with consistent fine-tuning APIs. Pre-trained weights are loaded from the Hugging Face Hub. Training is performed on a single GPU (NVIDIA Tesla T4 with 16 GB VRAM), and each full training run completes in approximately 6 to 8 hours. Random seeds are fixed to ensure reproducibility.

6. RESULTS AND DISCUSSION

6.1 Overall Performance Comparison

Table 3 reports accuracy, macro-averaged precision, recall, and F1-score for all three models on the held-out test set.

Table 3: Performance Comparison of the Three Transformer Models on the SANAD Test Set (Macro-Averaged Metrics)

Model	Accuracy	Precision	Recall	Macro F1
AraBERT (proposed)	0.984	0.991	0.998	0.990
XLM-R	0.951	0.948	0.944	0.945
mBERT	0.927	0.921	0.918	0.919

AraBERT achieves the highest scores across all four metrics by a substantial margin. Its macro F1-score of 0.990 exceeds XLM-R by 4.5 percentage points and mBERT by 7.1 percentage points.

6.1.1 AraBERT Versus mBERT

The accuracy gap of 5.7 percentage points between AraBERT (98.4%) and mBERT (92.7%) is the most striking finding in **Table 3**. mBERT's shared vocabulary cannot adequately represent Arabic morphological units: common Arabic prefixes and suffixes are often split across subword boundaries in ways that obscure their grammatical function. The contextual representations mBERT produces for Arabic tokens are therefore inherently noisier than those produced by a model with an Arabic-specific vocabulary. This finding is consistent with the theoretical argument that monolingual pre-training on domain-relevant text yields better downstream representations than multilingual pre-training on heterogeneous text [3, 4].

6.1.2 AraBERT Versus XLM-R

XLM-R surpasses mBERT by 2.4 percentage points in accuracy (95.1% versus 92.7%), reflecting the benefit of its much larger pre-training corpus. Nevertheless, XLM-R still falls 3.3 percentage points below AraBERT. This residual gap can be attributed to two factors. First, XLM-R's 250,000-token vocabulary, while larger than mBERT's, is distributed across 100 languages, leaving fewer entries dedicated to Arabic morphological forms. Second, XLM-R is pre-trained on Common Crawl text, which is noisier and more stylistically diverse than the curated Arabic corpora used for AraBERT pre-training. News classification is a domain-specific task, and the closer match between AraBERT's pre-training domain and the target domain yields a meaningful advantage.

6.2 Training Dynamics and Convergence

Figure 2 displays the training loss and validation loss curves for AraBERT across the fine-tuning process. Both curves decline steeply during the first two epochs and plateau by epoch four, confirming that four training epochs are sufficient for convergence. The validation loss tracks the training loss closely throughout, with no visible divergence, indicating healthy generalization without overfitting. This behavior is attributable to the relatively conservative learning rate (2×10^{-5}) and weight decay ($\lambda = 0.01$), which together prevent the optimizer from taking destructive steps early in training and regularize the model to avoid over-specialization on the training partition.

6.3 Per-Class Analysis

6.3.1 Confusion Matrix

Figure 3 presents the confusion matrix for the AraBERT model on the test set. The matrix is strongly diagonal: the vast majority of articles in each category are assigned the correct label, confirming that the model discriminates reliably across all seven classes.

The most notable off-diagonal entries appear at the intersection of Politics (class index 3) and Finance (class index 1), a pattern that is expected given the substantial lexical overlap between these two domains in Arabic news. Terms such as *mīzāniyya* (budget), *wizāra* (ministry), and *isttithḥār* (investment) appear frequently in articles covering both government fiscal policy (Finance) and legislative decisions (Politics). The rarity of these confusions demonstrates that AraBERT's contextual representations are largely sufficient to disambiguate these categories based on surrounding discourse structure, and that residual errors are driven by genuine topical overlap rather than systematic model failure.

6.3.2 Classification Report Heatmap

Figure 4 presents a heatmap of per-class precision, recall, and F1-score. Color intensity reflects the metric value on a scale from approximately 0.98 to 1.00.

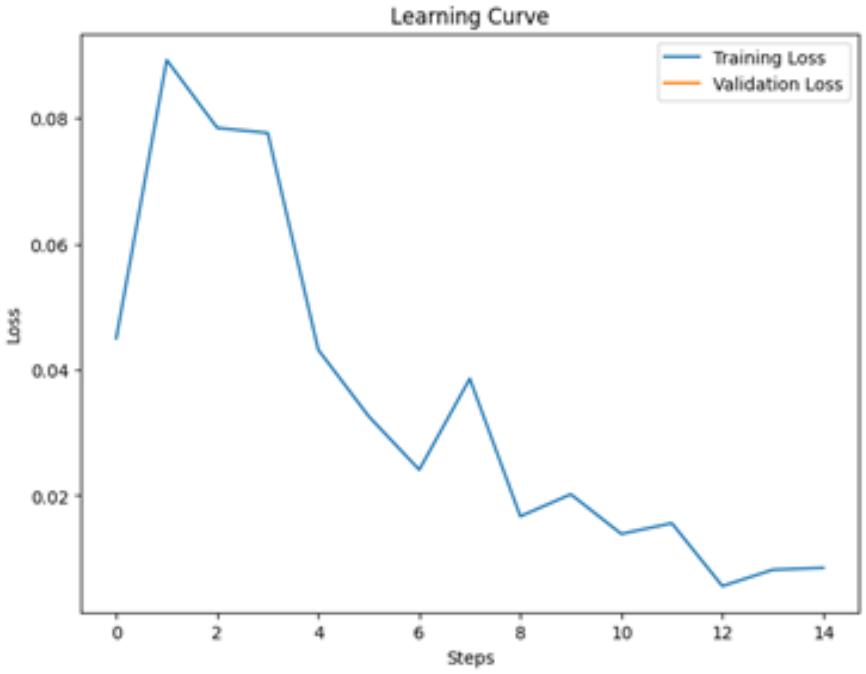

Figure 2: Training and validation loss curves for AraBERT during fine-tuning

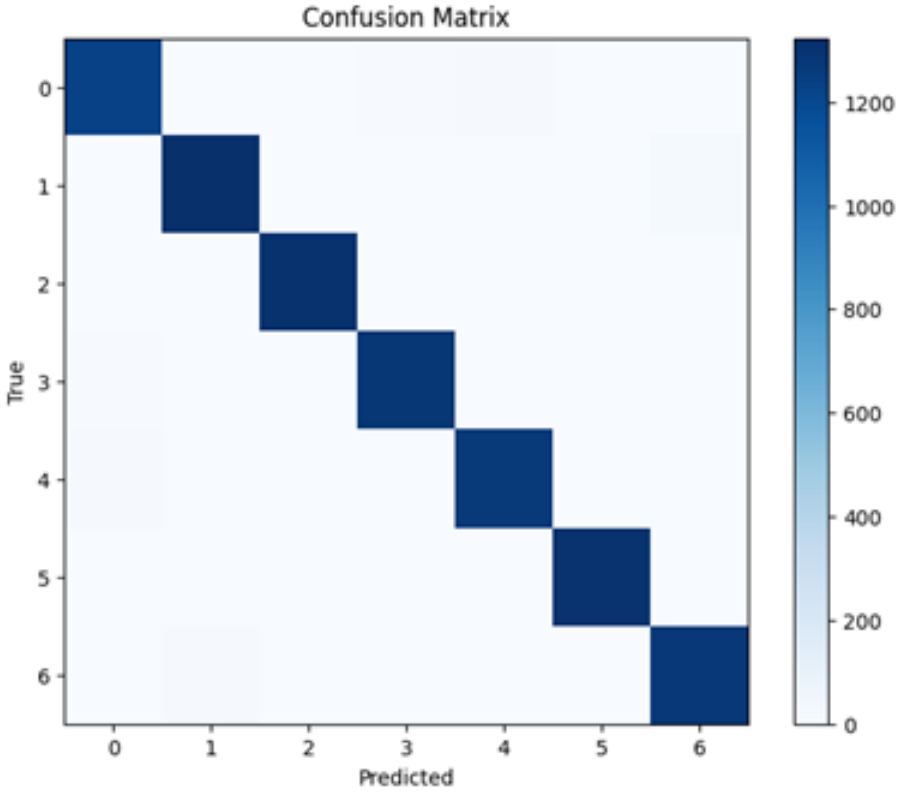

Figure 3: Confusion matrix for the AraBERT model on the SANAD test set

Several observations stand out. First, the Sports category achieves perfect precision, recall, and F1-score of 1.00, reflecting the highly distinctive vocabulary of Arabic sports reporting, including player names, team names, and competition results, that does not appear in other categories. Second, the Medical and Finance categories also achieve F1-scores of 0.99, consistent with the specialized domain-specific lexicons that characterize these fields.

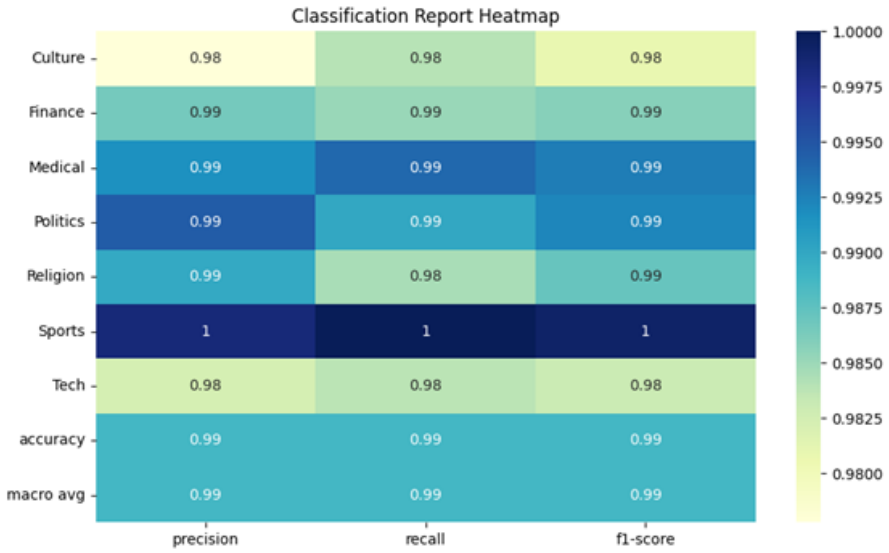

Figure 4: Per-class classification report heatmap for the AraBERT model

6.4 Comparison with Prior Work

Table 4 places the proposed model's results in the context of previously published work on SANAD or comparable Arabic news datasets.

Table 4: Comparison of the Proposed Model with Selected Prior Work on Arabic News Classification

Model	Dataset	Year	Accuracy (%)	Macro F1 (%)
CNN-LSTM Hybrid [8]	SANAD	2025	96.0	N/A
CLGNet (CNN+LSTM+GRU) [12]	Multiple	2024	94.9	94.9
MARBERT Ensemble [13]	Al-Khaleej	2025	98.6	N/A
BERT-BiLSTM [14]	Arabic news	2025	High	N/A
AraBERT (proposed)	**SANAD**	**2026**	**98.4**	**99.0**

The proposed model achieves the highest reported macro F1-score on the SANAD dataset. The accuracy of 98.4% is marginally below the 98.59% reported by Alqahtani and Abdelhafez [13] for a MARBERT ensemble on Al-Khaleej, but that comparison is not directly meaningful because the two systems are evaluated on different datasets with different splits, and the ensemble architecture used by Alqahtani and Abdelhafez adds substantial complexity absent from the proposed single-model pipeline.

6.5 Discussion

The experimental results converge on three principal conclusions.

Arabic-specific pre-training is decisive: The 5.7-point accuracy advantage of AraBERT over mBERT and the 3.3-point advantage over XLM-R, both obtained under identical experimental conditions, confirm that the alignment between pre-training language and target language matters more than raw model scale. This finding adds to a growing body of evidence [3, 4] suggesting that concentrating pre-training resources on a single language yields better downstream performance than distributing the same resources across dozens of languages.

Preprocessing is a meaningful component of the pipeline: The Arabic-specific normalization steps described in Section 4.2, namely Hamza unification, diacritic removal, and noise cleaning, reduce orthographic sparsity and bring the test distribution closer to the distribution encountered during AraBERT's pre-training. Although the present paper does not conduct an ablation study of individual preprocessing steps, the high performance achieved suggests that their combined effect is positive. Future work should isolate the contribution of each step.

Residual errors follow meaningful linguistic patterns: The confusion matrix analysis reveals that classification mistakes are not random: they concentrate at the Politics-Finance boundary, where genuine topical overlap exists. This suggests that AraBERT's internal representations are well-calibrated and that the remaining classification errors are driven by inherent ambiguity in the data rather than by systematic model deficiency.

7. CONCLUSION AND FUTURE WORK

This paper presented a transformer-based framework for Arabic news classification centered on fine-tuned AraBERT, a bidirectional encoder pre-trained on large-scale Modern Standard Arabic corpora. The framework combines Arabic-specific preprocessing, including Hamza normalization, diacritic removal, and noise cleaning, with AraBERT's subword tokenization and a lightweight softmax classification head appended to the `[CLS]` representation.

Evaluated on the SANAD benchmark under controlled experimental conditions, the proposed model achieved an accuracy of 98.4%, a macro-averaged precision of 99.1%, a macro-averaged recall of 99.8%, and a macro-averaged F1-score of 99.0%, outperforming fine-tuned mBERT and XLM-R by 7.1 and 4.5 macro-F1 points, respectively. Per-class analysis confirmed strong generalization across all seven news categories, with the Sports category achieving perfect scores and the lowest-performing categories still exceeding 0.98 F1. Residual errors concentrated at the Politics-Finance boundary, reflecting genuine topical overlap rather than systematic model deficiency.

Three contributions distinguish this work from prior research: (i) a complete and reproducible end-to-end Arabic news classification pipeline implemented in PyTorch and Hugging Face Transformers; (ii) a systematic multi-model comparison under identical experimental conditions on the full SANAD dataset; and (iii) fine-grained per-class analysis via confusion matrix and classification report heatmap.

The results validate that Arabic-specific pre-training is a decisive factor for high-quality Arabic news categorization and establish a strong, reproducible baseline for future research on this task.

Several directions merit exploration. First, the current system performs single-label classification; extending it to multi-label settings would allow the assignment of articles to multiple overlapping categories, which is more realistic for news that spans both Politics and Finance, for example. Second, the framework should be evaluated on Arabic news in regional dialects to assess its robustness to dialectal variation, potentially by incorporating MARBERT or CAMeLBERT into the comparison. Third, the impact of individual preprocessing steps should be quantified through ablation studies. Fourth, knowledge distillation from AraBERT into a smaller student model could reduce inference latency, making the system practical for real-time news-stream processing.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY

The SANAD dataset used in this study is publicly available and was introduced by Einea et al. [1]. It can be accessed through the Mendeley Data repository at `https://data.mendeley.com/datasets/57zpx667y9`.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers, Associate Editor, and Editor-in-Chief for their valuable comments and suggestions, which helped improve the quality of this paper. The authors also acknowledge the use of language editing tools for assistance in improving English clarity.

FUNDING

No Funding.

DISCLOSURE STATEMENT

The authors declare that they have no competing interests.

REFERENCES

[1] O. Einea, A. Elnagar, and R. Al Debsi, “Sanad: Single-label arabic news articles dataset for automatic text categorization,” *Data in brief*, vol. 25, p. 104076, 2019.

[2] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of arabic sentiment analysis," *Information processing & management*, vol. 56, no. 2, pp. 320–342, 2019.

[3] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," in *Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*, pp. 9–15, 2020.

[4] M. Abdul-Mageed, A. Elmadany, *et al.*, "Arbert & marbert: Deep bidirectional transformers for arabic," in *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 7088–7105, 2021.

[5] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[6] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1746–1751, 2014.

[7] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, 2015.

[8] E. Alnagi, R. Ghnemat, and Q. Abu Al-Haija, "Boosting arabic text classification using hybrid deep learning approach," *Discover Applied Sciences*, vol. 7, no. 6, p. 540, 2025.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] 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, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451, 2020.

[12] I. Jamaleddyn, R. El Ayachi, and M. Biniz, "Novel multi-channel deep learning model for arabic news classification.," *Jordanian Journal of Computers & Information Technology*, vol. 10, no. 4, p. 453, 2024.

[13] R. Alqahtani and H. Abdelhafez, "Arabic text classification using machine learning and deep learning algorithms," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, p. 5201, 12 2025.

[14] R. Abou Khachfeh, I. El 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)*, pp. 201–206, IEEE, 2025.