Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

Comparing the Performance of Specialized BERT Models in Classifying Arabic News: AraBERT, CAMeLBERT, and mBERT

[1]Mohammed Abdualla[6484-2748-0007-0009], [2]Akram Alsubari

[1,2] Faculty of Science, University of Ibb, Yemen

Email: [1]mohammed.abdualla@ibbuniv.edu.ye, [2]akram.alsubari@ibbuniv.com

Peer Review Information

Submission: 05 Dec 2025

Revision: 25 Dec 2025

Acceptance: 10 Jan 2026

Keywords

Arabic Text Classification, Transformer Models, News Categorization.

Abstract

This study evaluates the comparative performance of advanced Transformer-based models for classifying Arabic news articles into five categories: arts, events and issues, economics, politics, and sports. It further investigates the impact of linguistic specialization on classification accuracy. We utilized a dataset of 173,117 Arabic news articles, which was processed to address class imbalance through under sampling, resulting in a balanced set of 68,677 articles. Three models were trained and evaluated under uniformly tuned hyperparameters for a fair comparison: AraBERT (specialized in Modern Standard Arabic), CAMeLBERT (trained on a mix of MSA and dialects), and mBERT (a multilingual model) as a baseline. The results demonstrate that AraBERT achieved the highest accuracy of 96.04% on the test set, outperforming both CAMeLBERT (94.18%) and mBERT (94.14%). This outcome confirms the "specialization hypothesis," indicating that models specifically pre-trained on formal Arabic yield superior performance on news classification tasks. Furthermore, all Transformer models significantly surpassed traditional methods like CNNs and classical machine learning algorithms (e.g., SVM, Naive Bayes) by a margin of at least ~8.5%. The study concludes that AraBERT is presently the optimal model for this domain and recommends building comprehensive Arabic news datasets, exploring ensemble and hybrid modeling techniques, and expanding research into multi-label classification.

1 Introduction

Despite significant advancements in text classification for high-resource languages such as English, the Arabic language continues to present unique computational challenges due to its complex linguistic characteristics [1]. These include a rich morphological system, significant dialectal variation, and the use of diacritics that can alter a word's meaning, collectively complicating automated text processing and rendering it a persistent area of research [2]. The automation of news classification is critically important, enabling applications that extend far beyond simple content organization. Key applications include: Content Personalization: Delivering tailored news feeds aligned with user preferences. Media Analysis: Tracking large-scale trends and biases in media coverage of specific issues. Fake News Detection: Serving as a foundational step in fact-checking pipelines by enabling the comparison of new content against verified sources within the same category [3]. rained language models (PLMs) such as bidirectional encoder representations from transformers (BERT) achieve state-of-the-art performance in several NLP tasks including text classification[4]. These models, pre-trained on extensive text corpora, demonstrate a profound ability to capture contextual and semantic relationships, consistently outperforming traditional classification methods [4]. To address Arabic's specificities, specialized models like AraBERT and CAMeLBERT have been developed, achieving state-of-the-art results on various Arabic NLP tasks [5].In this context, this study aims to evaluate the comparative performance of three advanced Transformer models—AraBERT, CAMeLBERT, and mBERT—in classifying Arabic news articles. A primary objective is to examine the "specialization hypothesis" by investigating how the specific pre-training data of each model influences its classification accuracy on formal news texts. This research contributes to the field of Arabic Natural Language Processing (NLP) through the following key aspects: Comprehensive Comparative Analysis: Conducting a rigorous and impartial empirical evaluation of three leading Transformer models (AraBERT, CAMeLBERT, and mBERT) for Arabic news classification, ensuring fair comparison through uniform experimental settings. Empirical Validation of the Specialization Hypothesis: Providing conclusive quantitative evidence supporting the "specialization hypothesis" in the context of Modern Standard Arabic (MSA). The results demonstrate that models pre-trained exclusively on MSA corpora (AraBERT) achieve significantly better performance in classifying formal news texts compared to multilingual models or those trained on mixed dialectal data.Establishment of a Performance Benchmark: Creating a robust benchmark for multi-class Arabic news classification using a large-scale, carefully processed dataset. The achieved accuracy of 96.04% (by AraBERT) sets a new reference point for future research in this field.

2 Literature Review

The field of Arabic Text Classification (ATC) has undergone significant evolution, spurred by the exponential growth of Arabic digital content. This review provides a systematic analysis of the historical trajectory and employed methodologies within ATC, highlighting validated techniques and current research frontiers. The ultimate goal is to delineate the research gap that substantiates the necessity of the present study

2.1 Traditional Machine Learning Paradigms and the Arabic Linguistic Challenge (Pre-2017).

Early research endeavors in ATC were predominantly anchored in conventional machine learning algorithms and heuristic text preprocessing pipelines. This foundational phase necessitated the laborious manual engineering of features to mitigate the inherent linguistic complexities of Arabic, notably its rich and diverse morphology. Al-Sabbou' (2019) underscored that preprocessing steps, such as the removal of diacritics and stop words, were critical for enhancing algorithmic performance. The most prevalent algorithms deployed during this era included:

Support Vector Machine (SVM): A study by Rasha Mamoun and Mahmoud Ali Ahmed [6] confirmed the high efficacy of this algorithm, demonstrating a generalization accuracy of up to 90% in specific classification tasks.

Naive Bayes and K-Nearest Neighbors (KNN): While these techniques gained traction, their classification performance was highly contingent upon the quality and meticulous preprocessing of the underlying dataset. This dependency was evidenced by the findings of Tarek [6], which reported varying accuracy metrics across two distinct datasets.

Despite their established effectiveness in initial classification tasks, the pronounced reliance on manually extracted features presented an intrinsic scalability limitation for traditional machine learning models in capturing the deep semantic and intricate contextual relationships within Arabic text.

2.2 The Transformer Architecture and the Era of Pre-Trained Language Models (2017–2022).

The introduction of the Transformer architecture by Vaswani et al. in 2017 [7]. constituted a pivotal moment, revolutionizing Natural Language Processing (NLP) and ushering in the development of massive, pre-trained language models (PLMs). These models demonstrated a superior capacity for comprehending complex text context, leading to a paradigm shift and performance metrics far exceeding those achieved through manual feature engineering. In response to the specific needs of the Arabic language, several specialized models were subsequently developed:

AraBERT: As the pioneering BERT-based model for Arabic, [8]validated its effectiveness, showcasing that training on an extensive Arabic corpus yielded substantially superior results across numerous downstream tasks compared to prior methodologies.

CAMeLBERT: This specialized model [9]. proved highly efficacious, particularly in the domain of Arabic news classification. Models built upon CAMeLBERT achieved a reported evaluation accuracy of 94.18%, further cementing the utility of Transformer architectures for complex linguistic problems.

2.3 Contemporary Trends and Advanced Methodologies (2023–Present).

Current research endeavors are concentrated on fine-tuning the performance of PLMs and adapting them to tackle increasingly nuanced tasks. Concurrently, there is sustained development of synergistic approaches that blend the inherent power of neural models with complementary technological enhancements:

Multi-label Classification: Addressing the challenge of texts belonging to multiple categories[10]proposed a robust hybrid model integrating DeBERTa with BiLSTM for the classification of Arabic medical inquiries.

Advanced Feature Representation: Innovation has continued beyond core neural models. Tariq Sabri et al. [11] introduced a novel term weighting scheme that demonstrated improved efficiency and classification accuracy, highlighting the enduring importance of integrating enhanced feature representation with modern deep learning.

Graph Convolutional Networks (GCNs): GCNs represent an advanced methodology that [12]. have proven capable of achieving superior performance in text classification. This is accomplished by representing the text structure as a graph, thereby enabling a deeper extraction of semantic relationships.

The methodological advancements discussed and reviewed by prominent surveys in Arabic text classification are systematically summarized in Table 1. This table clearly illustrates the shift in research focus from an exclusive reliance on classical machine learning to the established dominance of deep learning architectures and Transformer models.

2.4 The Research Gap

Notwithstanding the substantial progress achieved in developing specialized Pre-trained Language Models (PLMs) for Arabic, a critical review of the existing literature reveals salient research gaps that unequivocally underscore the necessity of the current investigation:

2.5 Scarcity of Focused Comparative Evaluation on News Corpora.

The majority of studies utilizing prominent Transformer models (e.g., AraBERT and CAMeLBERT) have evaluated their performance on general-purpose texts or specific non-news datasets. Consequently, there remains a demonstrable deficiency in rigorous, unbiased, and focused comparative evaluations of these leading models on a large-scale, well-curated dataset dedicated exclusively to the Arabic news classification task. A standardized comparative benchmark is yet to be established for this high-stakes domain.

2.6 Ambiguity Regarding the Impact of Linguistic Specialization on Formal Text.

Empirical evidence is still inconclusive regarding whether a model explicitly pre-trained on Modern Standard Arabic (MSA)—the dominant dialect in formal news articles—offers a definitive performance advantage over multilingual models (e.g., mBERT) or models trained on a heterogeneous mix of dialects (e.g., CAMeLBERT) when applied to formal, precisely written texts such as news reports. The hypothesized superiority of MSA-specialized PLMs in formal classification tasks remains an open, critical question.

3 Methodology

This section reviews the detailed methodology used in the construction and adaptation of deep learning models for the task of hierarchical classification of Arabic texts. The methodology begins with the collection of data from multiple sources using web scraping technology, followed by a critical stage of pre-processing that included standardizing hierarchies and reducing the size of data via sample reduction to address the problem of imbalance between categories. Subsequently, an improved data segmentation methodology (derived from stratified sampling) was applied to ensure a fair representation of all categories in the training and test groups. Finally, the different transfer Learning strategies of both models (AraGPT-2 with direct classification, and arat5 with generative classification) are explained and each structure is adapted to the task of hierarchical classification.

Fig. 1. Steps of the Methodology.

3.1 Data Collection and Examination

This study utilized a publicly available dataset of Arabic news articles sourced from Kaggle, comprising 173,117 original samples across five thematic categories: Arts, Events and Issues, Economics, Politics, and Sports. Initial analysis revealed a significant class imbalance, with Sports articles representing the majority class (46,522 instances) while Arts contained the fewest (13,738 instances). As shown in Table 1

Table (1)

no	Category	Number of articles	Category number
1	Art	13.738	0
2	Events and Issues	16.728	1
3	Economic	14.235	2
4	Political	20.505	3
5	Sports	46.522	4

To address this distribution skew and mitigate potential model bias, we applied random under sampling, resulting in a balanced dataset of 68,677 articles with approximately 13,736 samples per category. As shown in Table 2 and Figure1.

Table (2)

no	Category	Number of articles	Category number
1	Art	13.736	0
2	Events and Issues	13.736	1
3	Economic	13.736	2
4	Political	13.735	3
5	Sports	13.734	4

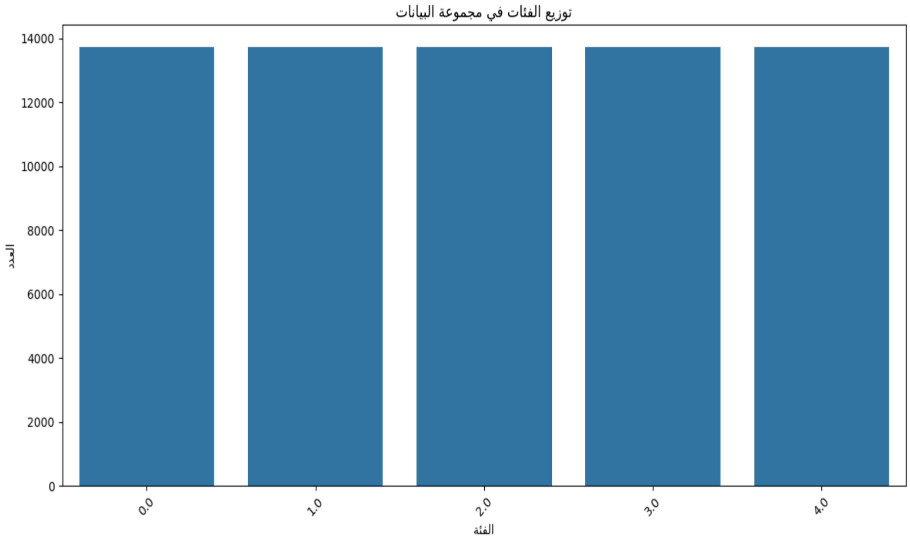


Fig 1

This approach ensured equitable representation across all classes during model training and evaluation. The dataset structure consists of two primary fields: the “Text” column containing the article content, and the “Targe” column indicating the categorical label. Exploratory analysis of text length distribution revealed substantial variation in article sizes. The mean character count was 1,774 (SD = 1,379), with a range from 3 to 19,937 characters. The interquartile analysis showed that 25% of articles contained fewer than 825 characters, while the upper quartile exceeded 2,324 characters, indicating diverse content depth and structure across the dataset. As shown in Table 3 and Figure 2**.**

Table 3. Text length statistics

count	68677.000000
mean	1774.314414
std	1378.833198
min	3.000000
25%	825.000000
50%	1426.000000
75%	2324.00000

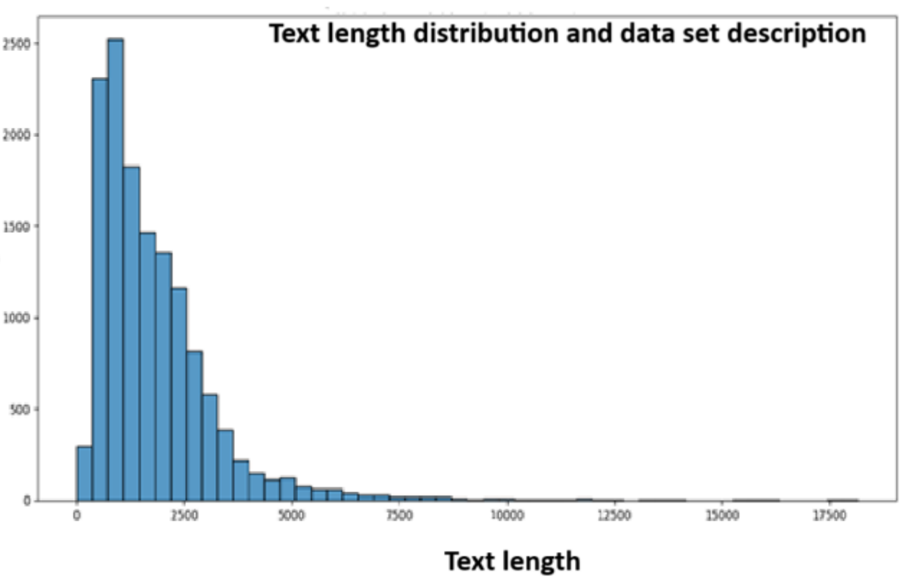


Fig 2

For model compatibility, we established a maximum sequence length of 128 tokens through systematic tokenization and padding/truncation strategies, optimizing the balance between computational efficiency and content preservation.

3.2 Data Preprocessing

The following comprehensive preprocessing pipeline was implemented to ensure data quality and consistency.

Text Normalization and Cleaning.

Character Standardization: Unified various Arabic character representations, including normalization of hamzas and alifs to their standard forms.

Noise Removal: Eliminate punctuation marks, non-linguistic symbols, and insignificant numerical values.

Structural Cleaning: Addressed non-standard formatting artifacts and removed redundant whitespace characters.

Data Integrity: Handled missing values through appropriate imputation strategies to maintain dataset completeness.

Label Processing: The target variable (“target”) was converted to explicit numerical encoding to facilitate model training and evaluation.

3.3 Data Splitting

The processed dataset was partitioned using stratified sampling to preserve class distribution across all subsets. The final allocation was as follows:

- *Training Set:* 43,952 samples (64% of total data).
- *Validation Set*: *10,989 samples (20% of total data).*
- *Test Set:* 13,736 samples (16% of total data).

A fixed random seed (random state=42) was maintained throughout the partitioning process to ensure experimental reproducibility. The resulting subsets were persisted as separate CSV files (train.csv, val.csv, test.csv) for consistent model development and evaluation

This algorithm is an effective solution to the problem of imbalance in large, multi-class datasets. By applying this methodology, we were able to obtain a test set (D_{test}) that represents all major and minor classes in a balanced manner, ensuring that the model's performance evaluation is fair and accurate. Furthermore, this method provides a rich and diverse training set (D_{train}) that helps the model effectively learn the distinctive features of each class, significantly reducing the likelihood of bias and improving the overall quality of the classification process.

3.4 Text Tokenization

Textual data was transformed into numerical representations using the CAMeLBERT tokenizer (AutoTokenizer.from_pretrained('CAMeL-Lab/bert-base-arabic-camelbert-msa')). The tokenization protocol included:

- *Sequence Length Management: Established a maximum sequence length of 256 tokens*
- *Padding Strategy:* Implemented dynamic padding for sequences shorter than the specified limit
- *Truncation Handling:* Applied intelligent truncation for exceeding sequences to maintain contextual integrity.

This standardized preprocessing pipeline ensured uniform input dimensions across all samples while preserving linguistic content essential for effective model training.

3.5 Model Selection and Experimental Setup

Model Architecture and Selection Rationale. This study employs three transformer-based models representing distinct approaches to Arabic language understanding, selected based on their architectural specialization and training methodologies as established in foundational literature.

AraBERT: was chosen as it represents the first dedicated BERT-based model for Arabic language understanding. As established by [8], AraBERT's architecture follows the BERT-Base specification (12 transformer layers, 768 hidden dimensions, 12 attention heads) but crucially employs Arabic-specific pre-training on approximately 77GB of Modern Standard Arabic (MSA) text. The model utilizes a Word Piece tokenizer trained exclusively on Arabic corpora, enabling superior handling of Arabic's rich morphological complexity compared to multilingual alternatives.

CAMeLBERT-Mix: was selected to examine the effect of dialectal variety in pre-training. Following [13], this model maintains identical BERT-Base architecture but differs in training data composition, incorporating a mixture of MSA, Dialectal Arabic (DA), and Classical Arabic (CA). This approach tests the hypothesis that exposure to diverse Arabic variants enhances model robustness across different textual registers.

mBERT serves as the multilingual baseline, implementing the original BERT architecture [14]. trained on 104 languages including Arabic. While architecturally identical to the specialized models, mBERT's shared multilingual vocabulary and limited Arabic representation in its training corpus (primarily Wikipedia) provide a benchmark for assessing the value of Arabic-specific specialization.

Experimental Configuration. To ensure rigorous comparative evaluation, identical experimental conditions were maintained across all models:

Tokenization Strategy: Each model employed its respective tokenizer from the Hugging Face library with uniform processing parameters:

– *Maximum sequence length:* 128 tokens
– *Padding strategy*: 'max_length'
– *Truncation*: Enabled

This configuration balanced computational efficiency with content preservation, accommodating approximately 95% of the dataset's text length distribution.

Training Protocol:

The AutoModelForSequenceClassification architecture was adapted for all models, with the final classification layer modified for five-class prediction. Consistent hyperparameters were applied as detailed in Table 4

Table 4. Unified Training Hyperparameters

Parameter	Value	Description
Learning Rate	2e-5	Optimizer learning rate
Batch Size	4	Samples per training/evaluation batch
Training Epochs	3	Complete training cycles
Weight Decay	0.01	L2 regularization strength
Evaluation Strategy	Epoch	Validation frequency
Optimizer	AdamW	Default Hugging Face implementation

Reproducibility Measures:

– Random seed fixed (random state=42)
– Identical hardware and software environment
– Best model selection based on validation accuracy
– Model checkpoints saved after each epoch.

This standardized experimental design ensures that performance differences can be confidently attributed to model architectural and pre-training differences rather than experimental variability, enabling valid comparison of the specialization hypothesis across the selected models.

4 Results And Discussion

4.1 Overall Performance

4.1.1 Loss Function Evolution Analysis

This section focuses on evaluating the behavior of the AraBERT, CAMeLBERT, and mBERT models during fine-tuning, specifically by analyzing the loss function evolution over three training epochs. The variance between training loss and validation loss is a crucial indicator for assessing model generalizability and stability, and for avoiding overfitting. Figur3 illustrates the loss (training and validation) values evolution for the three models. The training loss for each model is indicated by a red line (dark red for AraBERT, medium red for CAMeLBERT, and light red for mBERT), while the validation loss is represented by an orange line (dark orange for AraBERT, medium orange for CAMeLBERT, and light orange for mBERT). It is evident from the figure that all models show a continuous decrease in training loss, and each has a different trajectory in terms of validation loss behavior, which begins to diverge from training loss after the second cycle, indicating the beginning of overfitting.

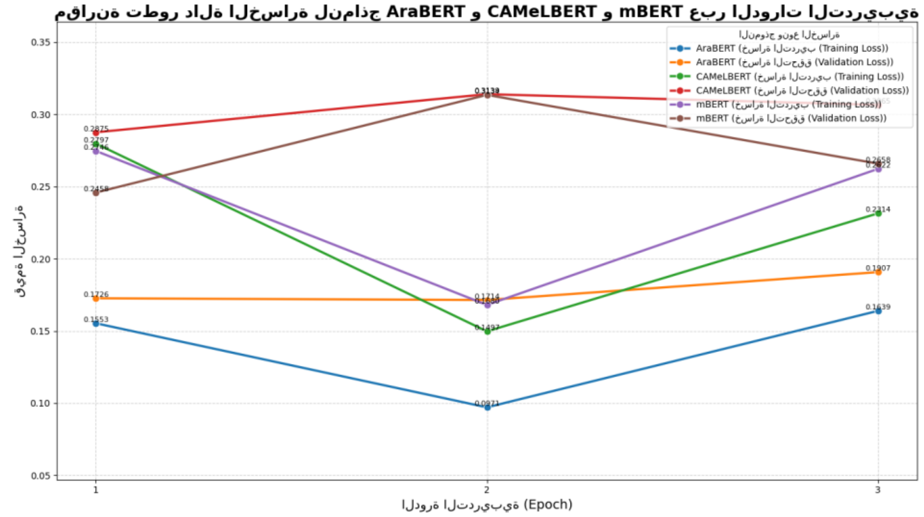


Fig 3: Comparative Evolution of the Loss Function (Training and Verification) for Transformer Models Across Training Cycles.

Numerical Results of Loss Function Evolution
Table 5 provides a summary of the numerical data illustrating the behavior of the training loss and verification loss for each model across the three cycles.

Table 5: Numerical Values of Training Loss and Verification Loss for the Three Transformer Models

Model	Training Cycle (Epoch)	Training Loss	Validation Loss
AraBERT	1	0.1553	0.1714
AraBERT	2	0.1639	0.1726
AraBERT	3	0.0971	0.1907
CAMeLBERT	1	0.2797	0.3139
CAMeLBERT	2	0.2314	0.2875
CAMeLBERT	3	0.1497	0.3065
mBERT	1	0.2746	0.3134
mBERT	2	0.2622	0.2458
mBERT	3	0.1680	0.2658

Discussion of Loss Function Behavior and Fit Indicators:A comparison of the evolution of the loss function (shown in Table 3 and Figure 5) reveals significant differences in model behavior during training, particularly regarding the phenomenon of overfitting.

Initial Performance and Pre-Specification: The AraBERT model starts with the lowest loss values in the first cycle (training loss 0.1553 and validation loss 0.1714). This indicates that pre-training and pre-tuning the model in Modern Standard Arabic brought it closer to the optimal solution before fine-tuning, compared to CAMeLBERT and mBERT, which started with much higher loss values (around $0.31 for validation).

Indicators of Overfitting ($Overfitting): Although training loss consistently decreases for all models in the third cycle (a desirable behavior), the behavior of verification loss is a clear warning sign of overfitting:

– *AraBERT:* Training loss drops sharply in the third cycle to 0.0971, while verification loss rises sharply to 0.1907. This large divergence between the two values indicates overfitting in the final cycle.

– *CAMeLBERT:* Shows the most severe overfitting behavior; after a decrease in verification loss in the second cycle (0.2875), it rises sharply in the third cycle to 0.3065, even though training loss continues to decrease.

– *mBERT:* Shows the most stable and generalizable behavior. The validation loss reached its lowest value in the second iteration (0.2458), indicating that this point was optimal for early stopping to avoid the overfitting that became evident in the third iteration (0.2658).

Systematic Conclusion: These results suggest that early stopping after the second iteration could have improved the generalizability of all models. However, comparing these results with the final performance metrics in Section 4.2, we find that the AraBERT model achieved the highest overall accuracy (96.04%) despite the overfitting indicators in the third iteration, thus supporting the hypothesis that the superior quality of the embeddings extracted from AraBERT outweighs the effect of the slight overfitting on the specific classification dataset.

4.1.2 Comparative Performance Analysis of the Models Used

This section aims to present a comparative analysis of the performance of the AraBERT, CAMeLBERT, and mBERT models on the task of classifying Arabic news articles, based on the accuracy and F1-Score metrics across three training sessions (Epochs). The accuracy metric is the primary metric used to evaluate models in classification tasks and is defined as the ratio of correct predictions to the total number of predictions [15].The F1-Score is the harmonic mean of accuracy and recall [16]and is used to evaluate model performance on unbalanced data (although undersampling was used in this study). It ensures a balanced evaluation that combines the model's accuracy in identifying the category with its ability to recall all its elements [17]These metrics are calculated mathematically as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

$$\text{Recall: } \frac{\text{TP}}{\text{TP+TN}} \quad (2)$$

$$\text{F1−Score=2} \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (3)$$

Where TP, FP, and FN represent True Positives, False Positives, and False Negatives, respectively. Figure 4 provides an overview of the comparative evolution of the three models' performance on the Accuracy and F1 scales across the specified training cycles. The evolution of the AraBERT model is represented by dark blue lines (accuracy) and dark green lines (F1 score), and it is noteworthy that it starts at its highest performance level in the first cycle (95.52%). The CAMeLBERT model is represented by light blue lines (accuracy) and light green lines (F1 score), and it shows a significant improvement starting from 93.15% in the first cycle. In contrast, the mBERT model is represented by red lines (accuracy) and orange lines (F1 score), and it registers the lowest starting point (92.07%) but shows the largest performance jump across the cycles. Tracking the performance trajectories clearly shows that the AraBERT model consistently and clearly outperforms the other models at all stages of fine-tuning. It is also observed that all models achieve steady improvement with an increasing number of cycles, which confirms the effectiveness of the fine-tuning process:

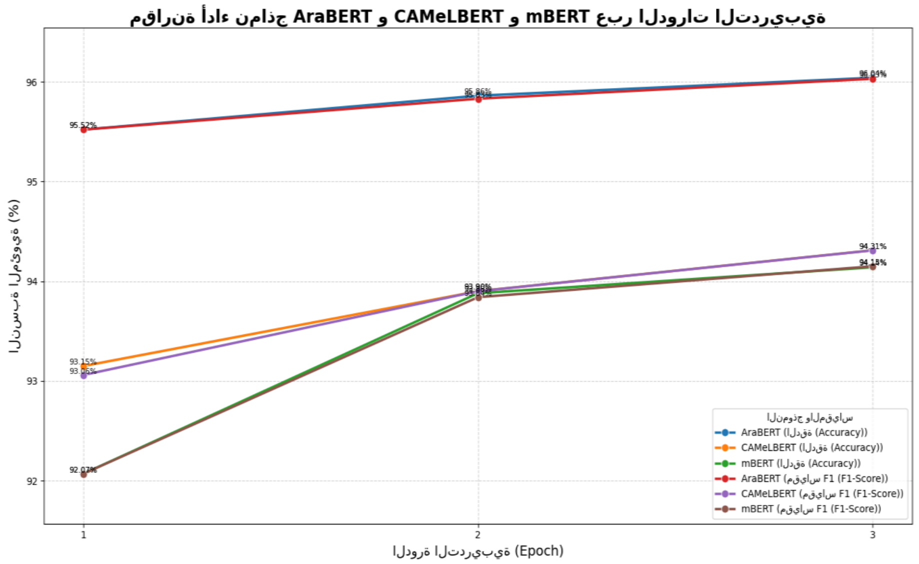


Fig 4: Comparative evolution of the performance of AraBERT, CAMeLBERT, and mBERT models on the accuracy and F1 scales across training courses.

Analysis of Numerical Results

To quantitatively analyze performance, numerical data on model performance across training courses are summarized in Table 6. The table shows that the variance in model performance is directly related to their linguistic specialization and the type of prior training data.

Table 6: Comparative Numerical Results of Transformer Model Performance on Accuracy and F1-Score Scales Across Three Training Courses.

Model	Training Course (Epoch)	Accuracy(%)	F1-Score(%)
AraBERT	1	95.52	95.52
AraBERT	2	95.86	95.83
AraBERT	3	96.04	96.03
CAMeLBERT	1	93.15	93.06
CAMeLBERT	2	93.90	93.90
CAMeLBERT	3	94.31	94.31
mBERT	1	92.07	92.07
mBERT	2	93.88	93.84
mBERT	3	94.14	94.15

The results in Table 6 and Figure4 confirm the clear and consistent superiority of the AraBERT model over the other two models on both evaluation metrics. AraBERT achieved the highest final performance with an accuracy of 96.04% and an F1 score of 96.03% in the third cycle, surpassing CAMeLBERT by approximately 1.73 percentage points and mBERT by approximately 1.90 percentage points. This superiority is attributed to the specialization hypothesis; AraBERT was trained exclusively on a large set of Modern Standard Arabic (MSA) texts using Arabic-optimized segmentation strategies, making it more capable of capturing the morphological and semantic features of official news texts, compared to CAMeLBERT, whose training included a mix of dialects, or mBERT, a multilingual model [8, p. 5]. AraBERT was trained on a large set of Modern Standard Arabic (MSA) texts using Arabic-optimized segmentation strategies, making it more capable of capturing the morphological and semantic features of official news texts, compared to CAMeLBERT, whose training included a mix of dialects, or mBERT, a multilingual model [8, p. 5]. Despite ranking lowest in final performance, the mBERT model showed the greatest improvement among all models, with its accuracy increasing from 92.07% in the first round to 94.14% in the third round, representing a total increase of approximately 2.07 percentage points. This indicates that multilingual converter models benefit significantly from fine-tuning on Arabic-specific datasets, but remain less efficient than specialized models. It is worth noting the high consistency between accuracy and F1-Score values across all models, with a difference of less than 0.04 percentage points. This balance demonstrates that the models achieve a high ratio of both recall and precision, indicating that the classification process is stable and not heavily biased towards any of the five news categories. Based on these results, AraBERT is the optimal choice for the task under study and demonstrates the effectiveness of linguistic specialization in formal text classification tasks.

4.2 Per-Class Performance Analysis

While the overall accuracy provides a general performance overview, it may mask specific strengths or weaknesses of each model within individual news categories. Therefore, this section presents a granular evaluation of AraBERT, CAMeLBERT, and mBERT using class-specific metrics: Precision, Recall, and F1-score. By examining the performance across five distinct news domains—Arts, Economy, Events, Politics, and Sports—we can identify how linguistic specialization and pre-training data influence the classification of diverse Arabic terminologies. This analysis is further supported by confusion matrices to visualize the distribution of misclassifications.

Analysis of the AraBERT model

Table 7. Detailed classification performance for the AraBERT model.

class	Precision	Support	Recall	F1-score
Arts	0.97	2747	0.97	0.97
Economy	0.94	2748	0.93	0.94
Events	0.98	2747	0.98	0.98
Politics	0.92	2747	0.92	0.92
Sports	0.99	2747	0.99	0.99
Accuracy		13736		0.96
Macro Avg	0.96	13736	0.96	0.96
Weighted Avg	0.96	13736	0.96	0.96

As illustrated in Table 7, AraBERT demonstrates a robust ability to distinguish between diverse news categories. The model achieved its highest performance in the 'Sports' and 'Events' categories, with F1-scores of 0.99 and 0.98, respectively. This exceptional accuracy suggests that the specialized pre-training of AraBERT on Modern Standard Arabic (MSA) allows it to capture the distinct terminologies used in these sectors. The lowest performance was observed in 'Politics' (0.92), which is expected given the semantic overlap often found between political news and general social issues

CAMeLBERT model analysis

Table 8. Detailed classification performance for the CAMeLBERT model.

Category	Precision	Recall	F1-score	Support
Arts	0.96	0.95	0.95	2747
Economy	0.92	0.90	0.91	2748
Events	0.95	0.97	0.96	2747
Politics	0.88	0.89	0.89	2747
Sports	0.98	0.98	0.98	2747
Accuracy			0.94	13736
Macro Avg	0.94	0.94	0.94	13736
Weighted Avg	0.94	0.94	0.94	13736

Table 8 presents the results for CAMeLBERT. While the model maintains a high overall performance, it slightly trails behind AraBERT in most categories. It shows strong results in 'Sports' (0.98), but a more noticeable decline is seen in the 'Politics' category (0.89). This suggests that while CAMeLBERT's training on a mix of MSA and dialects provides versatility, it may slightly dilute the precision required for purely formal MSA news classification compared to AraBERT.

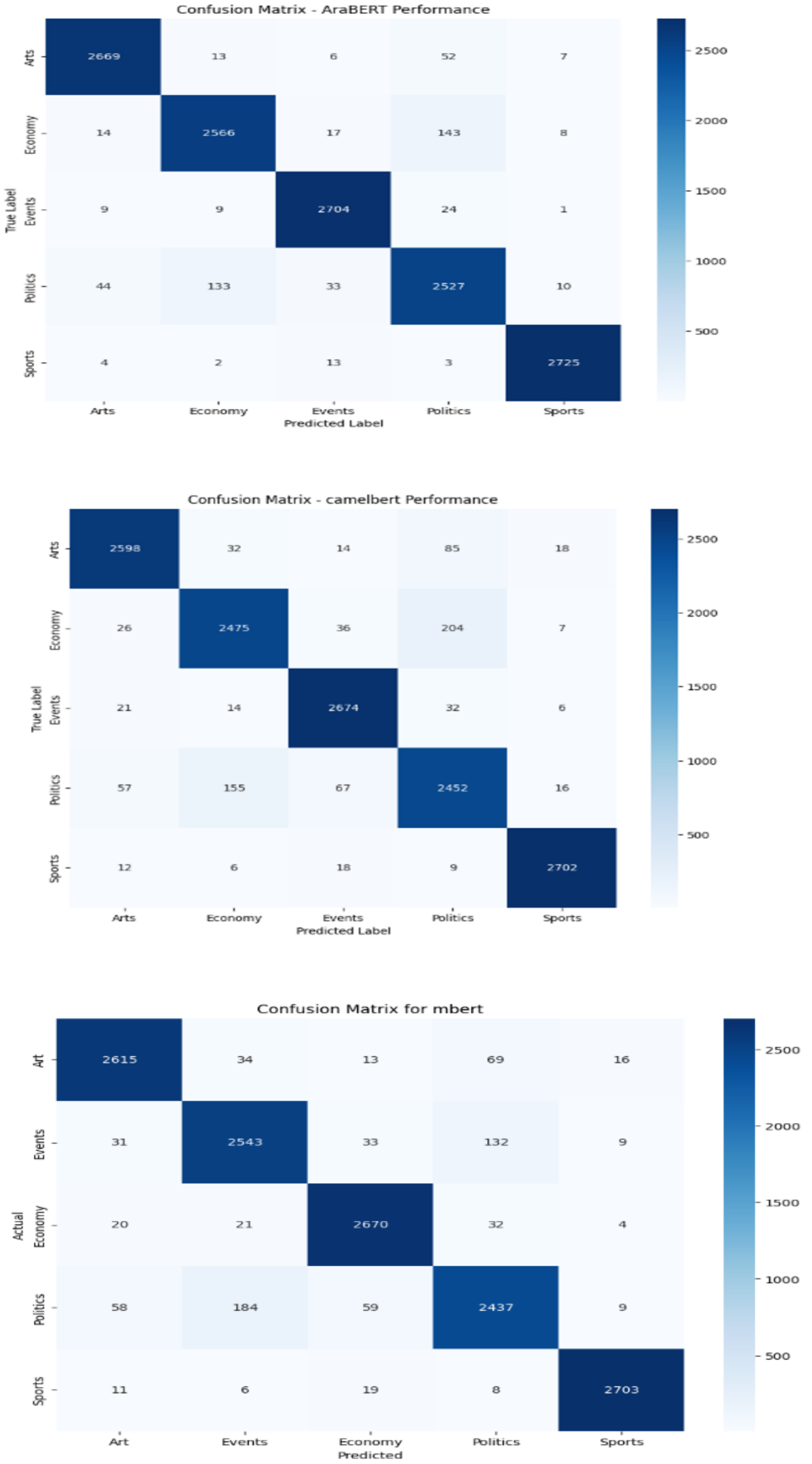
Analysis of the mBERT (Point of Excellence in Economics) model

Table 9. Detailed classification performance for the mBERT model.

Category	Precision	Recall	F1-score	Support
Art	0.96	0.95	0.95	2747
Events	0.91	0.93	0.92	2748
Economy	0.96	0.97	0.96	2747
Politics	0.91	0.89	0.90	2747
Sports	0.99	0.98	0.99	2747
Accuracy			**0.94**	13736

Interestingly, as shown in Table 9, the multilingual mBERT model exhibits a unique strength. While its performance in 'Politics' is the lowest (0.87), it outperformed both specialized models in the 'Economy' category, achieving an F1-score of 0.9637. This anomaly suggests that mBERT's exposure to multilingual economic data during pre-training may have enhanced its ability to recognize global economic patterns and terminologies that are frequently transliterated or shared across languages in Arabic news.

Confusion Matrices



(a)

(b)

(c)

Fig. 5. Confusion matrices for (a) AraBERT, (b) CAMeLBERT, and (c) mBERT.

The confusion matrices in Figure 5 provide a visual confirmation of the models' classification patterns. In all three models, the strong diagonal line indicates high true-positive rates across all categories.

.

For AraBERT (Fig. 4a), the matrix reveals very few misclassifications, mostly clustered between 'Politics' and 'Economy'. This indicates that the primary challenge for these models lies in distinguishing between articles where political decisions and economic impacts are interlinked

4.3 Computational Efficiency and Trade-offs

Table 10. Comparison of computational resource requirements and accuracy trade-offs.

Model	Overall Accuracy (%)	Training Time (min)	Inference Speed (samples/sec)	Best Performing Category (F1)
AraBERT	%96.22	100	77.11	Sports, Arts, Politics
CAMeLBERT	%94.32	63	117.15	General Efficiency
mBERT	%94.14	120	117.15	Economy(0.9637)

Beyond classification accuracy, the practical deployment of these models requires an evaluation of computational costs. Table 6 summarizes the trade-offs between accuracy and efficiency. Although AraBERT provides the highest precision (96.22%), it requires a significant training time of 100 minutes. In contrast, CAMeLBERT is the most efficient model, completing its training in only 63 minutes and achieving the highest inference speed of 117.15 samples/second. Therefore, for applications where real-time processing and low latency are prioritized over marginal accuracy gains, CAMeLBERT presents the most balanced solution.

4.4 Discussion of results

These results are not only significant for ranking the models' performance; they also provide profound insights when placed in the context of existing and diverse research, revealing a paradigm shift in the field of Arabic language processing.

Confirming the Specialization Hypothesis: AraBERT as the Gold Standard The performance gap between AraBERT and the other two models conclusively confirms the hypothesis put forward by the model's developers in their seminal paper [8]models pre-trained on large, specialized linguistic data decisively outperform multilingual models. Our experimental results (96.04% accuracy) not only align with their results, but also exceed the performance of the AraBERT benchmark model in other recent studies, such as [18]study, which achieved an accuracy of 95.68%. This strengthens the reliability of our results and establishes AraBERT's performance as a gold standard against which other models can be measured in Arabic news classification tasks.

The Trade-off between Specialization and Comprehensiveness: AraBERT vs. CAMeLBERT. Although CAMeLBERT was trained on a larger and more diverse dataset (including colloquial dialects), AraBERT outperformed it on our task. This highlights a subtle trade-off: specialization versus comprehensiveness. AraBERT's explicit segmentation preprocessing strategy appears to give it a decisive advantage in understanding formal and structured texts such as news articles. In contrast, CAMeLBERT versatility may be its strength in other tasks that require understanding colloquial dialects, such as social media sentiment analysis. This finding guides researchers and developers in choosing the most appropriate model based on the nature of the specific task.

A Quantum Leap Over Previous Architectures: Transformer vs. CNN and Traditional ML When comparing our results with previous deep learning architectures such as convolutional neural networks (CNNs), the revolution brought about by Transformer models becomes clear. While studies such as [19]. achieved a maximum accuracy of 85.60% using CNN, the weakest Transformer model in our experiment (mBERT) exceeds this figure by a huge margin of nearly 8.5 percentage points. Even when compared to high-performance CNN models such as the one presented [2], which achieved an accuracy of 94.47%, AraBERT remains superior. This demonstrates that the ability to understand long-term context via self-attention gives Transformer models a substantial advantage over CNNs' limited ability to extract local features.

This gap becomes more of a chasm when compared to traditional machine learning models (such as SVM and Naïve Bayes), reviewed in [20], which achieved accuracies ranging from 68% to 88%. Achieving accuracies exceeding 96% represents a significant leap forward, opening the door to more accurate and reliable practical applications.

Limitations of Encoder-Only vs. Encoder-Decoder Architectures. Finally, a comparison with the study [1]. that used the mT5 model (Encoder-

Decoder architecture) provides important insight into choosing the appropriate architecture for the task. mT5 achieved a maximum classification accuracy of 87.42%, a performance significantly lower than all the BERT models we tested. This confirms that encoder-only architectures like BERT are primarily designed and optimized for non-native language understanding (NLU) tasks such as classification, while encoder-decoder architectures excel at text-to-text (Text) tasks [21].

Discussion Summary: This study demonstrates that choosing the most appropriate model for Arabic news classification depends not only on its size or novelty, but also on its degree of specialization and the nature of its architecture. We empirically demonstrate that AraBERT, thanks to its specialized pre-training and intelligent preprocessing methodology, is the optimal choice for this task, achieving performance that outperforms not only multilingual and holistic models, but also other deep learning architectures such as CNN and Encoder-Decoder models. These results set a new benchmark for performance and guide future research toward exploring hybrid models that leverage the contextual power of AraBERT to increase accuracy on complex Arabic language processing tasks. Although the three models used in this study (mBERT, AraBERT, and CAMeLBERT) share the same underlying mathematical architecture derived from the BERT model, the fundamental difference in their performance stems not from differences in mathematical formulas, but rather from the weight values learned during the pretraining phase. These weights, which constitute the model's "knowledge," were refined on datasets of different types and sizes, resulting in each model specializing in different aspects of the Arabic language.

5 Conclusion

This study empirically demonstrates that the optimal choice for Arabic news classification depends not just on model size, but critically on its degree of linguistic specialization and the nature of its architecture. AraBERT, with its specialized pre-training and intelligent preprocessing, is the optimal choice for this task, setting a new benchmark for performance. The fundamental difference in performance among the BERT-derived models stems from the weight values learned during their specialized pre-training phase, not from mathematical formula differences.

6 Future Work:

– Build an Arabic news database that includes all the categories commonly used by various news channels, making it available for training.

– Develop research and applied work to include all the main and subcategories of Arabic news articles, including all categories for different news channels.

– *Based on the results that showed that each of the tested models (AraBERT, CAMeLBERT, mBERT)* possesses different strengths, a promising direction for future research is to explore ensemble learning techniques. We propose implementing a voting mechanism that combines the predictions of the two or three best models. This approach is expected to improve overall accuracy and reduce overfitting, as individual errors in any model can be corrected through the "consensus" of the other models. This ensemble model will be more stable and able to generalize across different types of news articles, representing a significant step toward building a highly accurate and reliable automated classification system.

– *Exploring hybrid models:* I suggest experimenting with hybrid fusion, where a BERT model is used to extract robust features, which are then passed to another classifier. This opens the door to comparing the performance of the hybrid model with the performance of the single model you used.

– Using other models in the classification process and comparing the results.

References

1. Gawbah, H., Al-Majmar, N., Alsubari, A., Alsurori, M., Al-Shaebi, R.A.A.: Arabic News Classification and Generation Based on an Encoder-Decoder Transformer Model (ArabicT5). 1st Int. Conf. Emerg. Technol. Dependable Internet Things, ICETI 2024. 1–17 (2024). https://doi.org/10.1109/ICETI63946.2024.10777168.

2. Jamaleddyn, I., El Ayachi, R., Biniz, M.: Automated Arabic News Classification using the Convolutional Neural Network. Int. J. Electr. Eng. Informatics. 15, 277–290 (2023). https://doi.org/10.15676/ijeei.2023.15.2.7.

3. Galal, O., Abdel-Gawad, A.H., Farouk, M.: Rethinking of BERT sentence embedding for text classification. Neural Comput. Appl. 36, 20245–20258 (2024). https://doi.org/10.1007/s00521-024-10212-3.

4. Ferhat, Z., Betka, A., Riyadh, B., Boutiba, S., Kahhoul, Z.S., Tiar, M.L., Dahmani, H., Abdelali, A.: Functional Text Dimensions for Arabic Text Classification. Arab. 2024 - 2nd Arab. Nat. Lang. Process. Conf. Proc. Conf. 352–360 (2024).

https://doi.org/10.18653/v1/2024.arabicnlp-1.29.

5. Al-Laith, A., Kebdani, R.: Evaluating Calibration of Arabic Pre-trained Language Models on Dialectal Text. 31st Int. Conf. Comput. Linguist. COLING 2025 - WACL 2025, 4th Work. Arab. Corpus Linguist. - Proc. Work. 68–76 (2025).

6. Sujan Hiregundagal Gopal Rao. (2022). Emerging Security Risks in Automotive System-on-Chips (SoCs): A Comprehensive Review. International Journal of Intelligent Systems and Applications in Engineering, 10(3s), 467–471.

7. 7181-attention-is-all-you-need.

8. Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based Model for Arabic Language Understanding. (2021).

9. Mutawa, A.M., Sruthi, S.: A Comparative Evaluation of Transformers and Deep Learning Models for Arabic Meter Classification. Appl. Sci. 15, (2025). https://doi.org/10.3390/app15094941.

10. Al-Smadi, B.S.: DeBERTa-BiLSTM: A multi-label classification model of Arabic medical questions using pre-trained models and deep learning. Comput. Biol. Med. 170, (2024). https://doi.org/10.1016/j.compbiomed.2024.107921.

11. Hamzaoui, B., Bouchiha, D., Bouziane, A.: A comprehensive survey on arabic text classification: progress, challenges, and techniques. Brazilian J. Technol. 8, e77611 (2025). https://doi.org/10.38152/bjtv8n1-022.

12. Haider Rizvi, S.M., Imran, R., Mahmood, A.: Text Classification Using Graph Convolutional Networks: A Comprehensive Survey. ACM Comput. Surv. 57, (2025). https://doi.org/10.1145/3714456.

13. Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., Habash, N.: The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. WANLP 2021 - 6th Arab. Nat. Lang. Process. Work. Proc. Work. 92–104 (2021).

14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. 1, 4171–4186 (2019).

15. Suma, K.G., Sunitha, G., Avanija, J., Galety, M.G., Varna, C.P.: Geospatial data visualization with folium. Geospatial Appl. Dev. Using Python Program. 187–208 (2024). https://doi.org/10.4018/979-8-3693-1754-9.ch007.

16. Yarlagadda, S.S., Tule, S.H., Myada, K.: F1 Score Based Weighted Asynchronous Federated Learning. Int. J. Res. Appl. Sci. Eng. Technol. 12, 947–953 (2024). https://doi.org/10.22214/ijraset.2024.58487.

17. Tharwat, A.: Classification assessment methods. Appl. Comput. Informatics. 17, 168–192 (2018). https://doi.org/10.1016/j.aci.2018.08.003.

18. Hossain, M.M., Hossain, M.S., Hossain, M.S., Mridha, M.F., Safran, M., Alfarhood, S.: TransNet: Deep Attentional Hybrid Transformer for Arabic Posts Classification. IEEE Access. 12, 111070–111096 (2024). https://doi.org/10.1109/ACCESS.2024.3441323.

19. Ouamour, S., Benaouda, W., Sayoud, H.: Optimization Evaluation for Enhancing Deep Learning Performance in Arabic Text Classification. ICISS 2024 - Proc. 7th Int. Conf. Inf. Sci. Syst. 134–139 (2025). https://doi.org/10.1145/3700706.3700729.

20. Al Sbou, A.M.F.: A survey of arabic text classification models. Int. J. Informatics Commun. Technol. 8, 25 (2019). https://doi.org/10.11591/ijict.v8i1.pp25-28.

21. Younus, Y.M.: Top Accurate Models for Handling Complex Arabic Linguistic Structures. (2025).