



RESEARCH ARTICLE

OPEN ACCESS

HYBRID LEXICAL-CONTEXTUAL LEARNING FOR ARABIC NEWS CLASSIFICATION: INTEGRATING TF-IDF WITH ARABERT EMBEDDINGS

Yasir Hadi Farhan¹, Mustafa Tareq², Boumedyen Shannaq³, Said Almaqbali⁴, Oualid Ali⁵

¹Department of Artificial Intelligence, College of Information Technology, University of Fallujah, Iraq.

²College of Computer Science, University of Technology, Baghdad, Iraq.

³College of Business, Management Information System Department, University of Buraimi, Al Buraimi, Oman.

⁴Multimedia University Melaka, Malaysia.

⁵College of Arts & Science, Applied Science University Manama, Kingdom of Bahrain.

¹<http://orcid.org/0000-0002-2378-9177>, ²<http://orcid.org/0000-0002-3499-8257>, ³<http://orcid.org/0000-0001-5867-3986>.

⁴<http://orcid.org/0009-0002-2159-5288>, ⁵<http://orcid.org/0009-0001-1433-017X>.

E-mail: yasir.hadi@uoanbar.edu.iq, mustafa.t.abd@uotechnology.edu.iq, [*boumedyen@uob.edu.om](mailto:boumedyen@uob.edu.om),
Said.matr.said@student.mmu.edu.my, oualid.ali@asu.edu.bh.

ARTICLE INFO

Article History

Received on: March 02, 2026

Reviewed on: April 03, 2026

Accepted on: May 01, 2026

Published on: July 31, 2026

Keywords:

Arabic Text Classification,
 Hybrid Feature Fusion,
 TF-IDF,
 AraBERT,
 Multi-Class News Categorization.

ABSTRACT

Arabic text classification is still a daunting undertaking because of the rich morphology, derivational complexity as well as lexical variability that is inherent in the language. Though transformer-based pretrained models have been a major breakthrough in Arabic Natural Language Processing (NLP), recent data indicates that conventional lexical representations continue to give a good discriminating ability in structured tasks like news classification. The given study will be a comparative analysis of classical machine learning models, transformer-based fine-tuning, and lexical contextual fusion framework as a hybrid method in Arabic multi-class news classification. A subset of the MAAD dataset (13,866 Arabic news articles) in six categories was carefully selected and cleaned and sampled to conduct experiments on a balanced and well-cleaned subset of the entire data set. To guarantee the data quality, the Arabic-character ratio filtering and length constraints were used to eliminate the corrupted and non-Arabic samples. TF-IDF features with Logistic regression and Linear Support Vector machine (SVM) were used to implement baseline models and contextual modeling was done using fine-tuning AraBERTv2. Moreover, the hybrid feature-level fusion model was suggested by adding TFIDF vectors with contextual embeddings were obtained using AraBERT. Experimental findings indicate that the traditional linear models are still very competitive whereby Linear SVM has an accuracy of 95.57 percent. The fine-tuned AraBERT obtained an accuracy of 91.6%, which puts lexical features in structured news datasets in the spotlight as still significant. The hybrid model with the highest performance in the proposed structure was 96.03% accurate and 96.02% Macro-F1 score, which shows that combining lexical statistical cues and contextual embeddings is complementary. These results show that hybrid lexical-contextual representations offer a powerful, computationally effective solution to Arabic news classification. The analysis offers reproducible experimental environments and intricate statistical investigation, to add the empirical data regarding the interaction between classical and deep methods of learning Arabic NLP.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The fast development of Arabic digital information distributed through news outlets, social networks, and online newspapers has largely boosted the need to automate Arabic text classification systems.

Correct classification makes it possible to organize large-scale content, filter it by topic, monitor sentiment, and perform downstream tasks, including information retrieval and recommendation systems. Nevertheless, Arabic Natural Language Processing (NLP) has its own linguistic peculiarities, such as rich morphology, complicated derivational models, orthographic ambiguity and dialects. These aspects make lexical sparsity and learning semantic representations more challenging, especially when performing tasks of multi-domain news classification. The current developments in large-scale Arabic language modeling have shown that both model structure and pretraining data size are significant factors in boosting performance at downstream tasks. Ghaddar, et al. [1] emphasize the importance of the data scale and data quality in pretraining Arabic language models, which have a consistent rise in the evaluation standards with the augmentation of the size and heterogeneity of corpora [1]. Similarly, the concept of Arabic-centric embedding interventions, such as ArabicMTEB, shows the additional importance of sound assessment schemes, peculiar to the Arabic language peculiarities [2]. Dialect-oriented models such as SaudiBERT also demonstrate the reason why domain- and dialect-aware representation learning should be viewed [3]. In spite of the successful models that use transformers such as AraBERT, the classical machine learning models that use the transformers of TFIDF representations with the use of linear models (e.g. Linear SVM or Logistic Regression) are also a good baseline in Arabic classification.

Lexical signals can be very discriminative in well-constructed news datasets and have been shown to compete with deep contextual models under some circumstances. Thus, it still needs a sound traditional base and only then more sophisticated architectures can be proposed. Concomitant to classification research, it has made a lot of progress in Arabic information retrieval and query expansion based on word embeddings and hybrid neural methods. It has been established that semantic matching is enhanced with word embedding based pseudo relevance feedback techniques and deep averaging networks when trying to obtain documents written in Arabic language [3] [4]. Morphology aware frameworks like MORSE-QE use root-extraction mechanisms to deal with the morphological complexity of Arabic and hence augment semantic expansion strategies [5]. These papers show the usefulness of embedding-based semantic modeling of Arabic NLP. Moreover, feature-fusion and hybrid neural -symbolic and neural -symbolic systems are also a relatively new subject of discussion as the next step in direction of improved robustness and scalability of NLP systems [6]. The combination of structured representations of lexical and contextual embeddings can both learn surface-discriminative and deep semantic relationships. The issue of structured learning and generalization mechanisms in optimizing neural network tasks is also emphasized on by optimization strategies of neural networks [7]. Based on the current advancements in Arabic language modelling [1-3] and embedding-based semantic structures [5] [7-12]. The papers discuss the multi-class classification of Arabic news using a comparative experimental design that they are involved in:

- 1 TF-IDF based baselines ('Logistic Regression and Linear SVM') which are predictable.
- 2 Fine tuning of 'AraBERT' on the basis of transformers,
- 3 A hybrid of feature level fusion lexical and relative symbol system.

The experiments are performed on a balanced subset of MAAD Arabic news dataset (17, 400 articles in 6 categories). To carry out a sound statistical comparison of the results, accuracy, macro-averaged F1-score, and analysis of the confusion matrix are used to assess the results. This study has added the following:

- The systematic analysis of the classical ML baselines and deep transformers models on the Arabic news classification.
- Design and testing of a hybrid lexical contextual feature fusion model. .
- Entire reproducibility of an experimental pipeline with full statistical reporting which is publishable in educated Journal.

II. RELATED WORK

The classification of Arabic text has over the last ten years experienced notable developments, due to the increased functionality of feature engineering, distributed word representations, and large-scale pretrained transformer models. Despite the fact that the traditional machine learning approaches are still providing decent baselines in structured domains, recent research is putting an increasing focus on contextual embedding models and hybrid architecture to improve the linguistic complexity of Arabic. The section presents the overview of the most current literature on three main directions: conventional statistical models, transformer-based Arabic language models, and approaches based on hybrid or embedding-driven.

II.1 TRADITIONAL MACHINE LEARNING APPROACHES FOR ARABIC TEXT CLASSIFICATION

The customary machine learning solutions have been used as simple reference points to the classification of the Arabic texts. SVM and Logistic Regression with feature extraction algorithms like Bag-of-Words and TF -IDF have continued to be popular because of their interpretability and performance, as well as their ability to work in organised areas, particularly in news classification. In the recent empirical investigations, it has been confirmed that competitive performance is not necessarily unattainable by using the linear models in case there is a common method of preprocessing, and the distribution of classes is observed [13]. Recent comparative analyses show that well-configured TF-IDF + Linear SVM pipelines can either match or even surpass deep transformer architectures when there is bias in the data or high quality lexical predictors available [14]. However, lexical means are also poor in contextual semantics and in long-range dependencies especially in morphologically rich languages such as Arabic. Root-pattern and derivational complexity morphology adds the sparsity and semantic ambiguity of words, making surface-level statistical representations ineffective. All these weaknesses resulted in the transition to contextual embedding models and distributed representations.

II.2 TRANSFORMER-BASED ARABIC LANGUAGE MODELS

Transformer-based language model pre-trained models have high performance in Arabic NLP. AraBERT and its successors showed good gains in the classification, question answering and sentiment analysis tasks. Recent studies highlight how choice of data scale and quality is crucial in the Arabic language modeling.

Ghaddar, et al. [1] demonstrate that there are consistent gains on the Arabic benchmarks as the size of the pretraining corpora increases. Likewise, Arabic-centric embedding benchmarks like ArabicMTEB offer standardized large-scale assessment on various tasks which underscores the research maturity on Arabic representation learning [2]. Models that are dialect aware and region specific have also been noted. Dialect-based pretraining, presented in 2024 under the name SaudiBERT, illustrates that downstream classification can be enhanced by using dialect-based pretraining [3]. Moreover, giant Arabic-centric foundation models like JAIS are a new generation of Arabic-competent LLMs that are trained on large multilingual corpora [15]. Transformer models are also successful but have computational overhead, can overfit or be unstable when fine-tuned on more moderate-sized datasets. Further, recent research states that transformers are not necessarily always better than simpler models in cases where lexical signals are very discriminative [16]. This fact has revived the focus on hybrid modeling approaches.

II.3 HYBRID AND EMBEDDING-DRIVEN APPROACHES

Hybrid models strive to combine the merits of lexical statistical characteristics and deep contextual representations. Embedding-driven semantic enhancement has received a lot of attention in Arabic NLP in an information retrieval and query expansion context. Pseudo-relevance feedback methods based on embedding words using deep averaging networks have proven to have quantifiable improvements in the Arabic document retrieval [9]. Morphology-sensitive embedding models like MORSE-QE can deal with derivational complexity by incorporating root-extraction processes [5]. Analyses conducted by surveys also prove that automatic query expansion is an important technique of alleviating vocabulary mismatch in processing Arabic texts [8]. In addition to retrieval, recent developments in hybrid deep-symbolic structures see the need to incorporate both structured features with neural representations to enhance robustness and scalability [10]. Structured learning mechanisms have also been suggested to be used as optimization strategies in neural networks to improve on the performance of generalization [17]. Comparative studies of Arabic text classification are more recent, with feature-level fusion, using contextual embeddings but concatenated with TF feature-IDF vectors, finding it effective to eclipse standalone transformer fine-tuning on some domain-specific tasks [18] [19]. The above findings indicate that hybrid lexical-contextual integration is a viable direction to be adopted in Arabic multi-class news classification.

III. METHODOLOGY

III.1 DATASET AND LABEL SET

The experiments were run with a subset of MAAD (Multi-Label Arabic Articles Dataset) that is a publicly accessible corpus of Arabic news created to research the text classification problem [18]. We have concentrated on the source of Al Jazeera news and identified six major categories namely Economy, Health, Technology, Culture, Political, and Sport. In order to adequately evaluate every category, the number of samples per category was the same, which made 17,400 articles (2,900 in each category). The source files have structured fields (Category, Title, Summary, publish date, domain, Article) contained in each record. The last category is single-label, multi-class multi-class (6 classes).

Text construction. To maximize semantic coverage, each instance input text was constructed as a concatenation of: Title + Summary + Article, followed by normalization.

III.2 DATA PARSING AND CLEANING (PRACTICAL PIPELINE)

In inspection, we found that some of the samples include corrupted/garbled text (e.g. sequences such as H B D D ...) which is not legitimate Arabic text. Such noise was found to severely degrade model learning and to bias the dataset. Therefore, we applied a cleaning pipeline consistent with the implemented code:

1. Normalization & Noise Removal
 - URL removal.
 - Extra whitespace normalization.
 - Removal of non-Arabic / non-informative symbols.

2. Quality Filtering.

We measured text quality using:

- Length (characters).
- Arabic character ratio (fraction of Arabic Unicode characters in the text).

Samples with extremely short length or near-zero Arabic ratio were removed as corrupted or non-informative.

3. Class Balancing After Cleaning.

After filtering, class counts became uneven. To preserve fairness and comparability between models, we re-balanced the dataset by downsampling each class to the minimum available count after filtering (resulting in a perfectly balanced dataset).

III.3 TRAIN-TEST SPLIT (REPRODUCIBLE SETUP)

The balanced dataset was split into:

- 80% training
- 20% testing

using a fixed random seed (seed = 42) to ensure reproducibility.

III.4 BASELINE MODELS (FAST CLASSICAL BENCHMARKS)

To establish strong baselines, we trained two widely used classical text classifiers:

(a) TF-IDF + Logistic Regression.

- Features: TF-IDF [20].

Term weighting was computed using the standard TF-IDF formulation:

$$TF-IDF(t, d) = TF(t, d) \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

Where:

- $TF(t, d)$ denotes the frequency of term t in document d ,
- $DF(t)$ is the document frequency of term t ,
- N represents the total number of documents in the corpus.
- Classifier: Logistic Regression [21].

Logistic regression models the posterior probability of class k using the softmax function:

$$P(y = k | x) = \frac{e^{w_k^T x}}{\sum_{j=1}^K e^{w_j^T x}} \quad (2)$$

Where:

- x is the *TF-IDF* feature vector,
- w_k is the weight vector associated with class k ,
- K is the total number of classes.

The model is trained by minimizing the cross-entropy loss:

$$L = - \sum_{i=1}^n \log P(y_i | x_i)$$

- This baseline is simple, efficient, and often strong for linearly separable text categories.

(b) TF-IDF + Linear SVM

- Features: TF-IDF,
- Classifier: Linear Support Vector Machine [22].

The linear SVM aims to find a separating hyperplane that maximizes the margin between classes while minimizing classification errors. The optimization objective is defined as:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i (w^T x_i + b)) \quad (3)$$

Where:

- w is the weight vector,
- b is the bias term,
- C is the regularization parameter,
- $y_i \in -1, +1$ represents class labels,
- x_i denotes the *TF-IDF* feature vector.
- Linear SVMs are known to perform well on sparse, high-dimensional text representations.

Both baselines were evaluated on the same split using Accuracy and Macro-F1, and the results were reported with per-class precision/recall/F1.

III.5 TRANSFORMER MODEL FINE-TUNING (ARABERTV2)

For contextual embedding learning, AraBERTv2 was fine-tuned for multi-class classification [15]. AraBERT is a transformer-based model pretrained specifically on large Arabic corpora and has demonstrated strong performance across multiple downstream Arabic NLP tasks. Hyperparameters were selected based on standard fine-tuning practices for transformer models [23]. The transformer architecture relies on the scaled dot-product self-attention mechanism proposed by Vaswani, et al. [24]:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

Where:

- $Q \in R^{n \times d_k}$ denotes the query matrix,
- $K \in R^{n \times d_k}$ denotes the key matrix,
- $V \in R^{n \times d_v}$ denotes the value matrix,
- d_k is the dimensionality of the key vector,
- n represents the number of tokens in the sequence.

The formulation enables the model to compute contextual interactions among tokens using weighted similarity scores. Before training, the input texts were pre-processed and tokenized to ensure consistent formatting for the transformer model.

Tokenization configuration:

- truncation = True.
- padding = "max_length."
- max_length = 256.

Training configuration:

- batch_size_train = 8.
- batch_size_eval = 8.
- learning_rate = 2e-5.
- weight_decay = 0.01.
- epochs = 3 (and sometimes 4 during trials).
- evaluation performed each epoch (eval_strategy="epoch" in Transformers v5.x).

We used the Hugging Face Trainer API to train and evaluate the model, and we reported a full classification report and a confusion matrix.

III.6 HYBRID MODEL (LEXICAL + CONTEXTUAL FUSION)

Hybrid modeling strategies combine lexical statistical representations with contextual embeddings to exploit complementary semantic and distributional information. Feature-level fusion between sparse TF-IDF vectors and dense transformer embeddings has been shown to improve robustness and classification accuracy in multiple NLP tasks [25] [26]. We implemented a hybrid classifier combining:

- **TF-IDF lexical features.**
- **AraBERT contextual embeddings** (extracted from the pretrained transformer).

These vectors were concatenated at the feature level to form a fused representation, which was then used as input to a linear classifier. This approach is computationally efficient (relative to end-to-end deep fusion) and empirically strong on news classification tasks [14] [24]. Let:

$$X_{tfidf} \in \mathbb{R}^{n \times d_1}$$

represents lexical **TF-IDF** features,

$$X_{bert} \in \mathbb{R}^{n \times d_2}$$

Represents contextual embeddings extracted from AraBERT. The fused representation is defined as:

$$X_{Fused} = [X_{tfidf}; X_{bert}] \quad (5)$$

Where:

- n is the number of samples,
- d1 is the dimensionality of TF-IDF features,
- d2 is the dimensionality of AraBERT embeddings,
- $X_{fused} \in \mathbb{R}^{n \times (d_1 + d_2)}$

The fused vector is then passed to a linear classifier defined as:

$$\hat{y} = WX_{fused} + b$$

where:

- W is the weight matrix,
- b is the bias term.

III.7 EVALUATION METRICS

All models were evaluated using:

- **Accuracy,**
- **Macro-F1,**
- **Per-class precision/recall/F1** (classification report),
- **Confusion matrix.**

Accuracy is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

The F1-score is computed as:

$$F1 = \frac{2 \cdot Precision \cdot recall}{Precision + recall} \quad (7)$$

Macro-F1 is defined as:

$$Macro - F1 = \frac{1}{K} \sum_{k=1}^K F1_k \quad (8)$$

Where:

- TP , TN , FP , FN represent the true positives, true negatives, false positives, and false negatives,
- K is the number of classes,
- FI_k is the FI -score for class k .

Macro-F1 was emphasized because it treats all classes equally and remains informative even if class distributions change after filtering [27].

IV. EXPERIMENTAL SETUP

IV.1 FINAL DATASET CONFIGURATION

Once all the corrupted and poor-quality samples were filtered out due to Arabic-ratio filtering and length restrictions, the dataset was rebalanced to have an equal number of classes. The final data is comprised of:

- **Total samples: 13,866,**
- **Classes: 6,**
- **Samples per class: 2,311,**
- **Train size (80%): 11,092,**
- **Test size (20%): 2,774.**

Stratified splitting was applied with a fixed random seed (42) to ensure reproducibility and preserve class balance across splits.

IV.2 HARDWARE AND SOFTWARE ENVIRONMENT

Experiments were conducted using:

- **GPU: NVIDIA Tesla T4,**
- **Framework: PyTorch,**
- **HuggingFace Transformers (v5.x),**
- **Scikit-learn for baseline models,**
- **Python 3.12,**
- **Random seed = 42.**

IV.3 MODEL CONFIGURATIONS

Baseline Models.

TF-IDF Configuration.

- Standard term frequency–inverse document frequency weighting⁹.
- Sparse high-dimensional representation.

Logistic Regression

- Multinomial setting.
- Default regularization (L2).

Linear SVM

- Linear kernel.
- Optimized for sparse input.

AraBERT Fine-Tuning

Pretrained model:

aubmindlab/bert-base-arabertv2.

Training configuration:

- **Max sequence length: 256.**
- **Batch size: 8.**
- **Learning rate: 2e-5.**
- **Weight decay: 0.01.**
- **Epochs: 3.**
- **Evaluation per epoch.**
- **Padding: max_length.**
- **Truncation: True.**

Hybrid Model

The hybrid model combines:

- TF-IDF lexical features,
- AraBERT contextual embeddings.

Feature fusion was performed by concatenating the TF-IDF vector with the contextual embedding vector before classification.

Classifier:

- Linear classifier on fused representation.

V. RESULTS AND DISCUSSION

V.1 OVERALL MODEL PERFORMANCE

Table 1 presents the overall classification performance of the evaluated models, including Accuracy and Macro-F1 score.

Table 1: Overall Performance Comparison.

Model	Accuracy	Macro-F1
Logistic Regression	0.9483	0.9482
Linear SVM	0.9557	0.9557
Hybrid Model	0.9603	0.9602

Source: Authors, (2026).

As indicated in Table 1, the Hybrid model had the best overall performance, accuracy of 96.03 and Macro-F1 of 96.02, which is better than the traditional baselines. The Linear SVM (95.57 percent) performed well, slightly better than the Logistic Regression (94.83 percent), indicating that Linear margin based classifiers are very powerful in sparse high dimensional TFIDF representation of structured news data. Interestingly, even though it is usually assumed that transformer-based contextual models will prevail in classical approaches, our previous AraBERT-only fine-tuning performed worse (= 91.6%), meaning that the lexical cues in balanced news data sets are very discriminative.

V.2 IMPROVEMENT OVER THE STRONGEST BASELINE

In order to measure the gain of the hybrid fusion approach, Table 2 reports performance gains when compared to the best classical baseline (Linear SVM).

Table 2: Performance Improvement Over Linear SVM.

Model	Accuracy Gain	Macro-F1 Gain
Hybrid Model	+0.0046	+0.0045

Source: Authors, (2026).

Linear SVM has an error of about 0.46 percentage points higher than the hybrid model. Though numerically small, it is significant even in already high-performing systems (above 95 percent), in which marginal gains are ever more challenging to obtain. This establishes the fact that the integration of a lexical and contextual feature brings out complementary information that neither of the representation can bring out.

V.3 PER-CLASS PERFORMANCE ANALYSIS

Table 3 records the F1 score per-class of all the assessed models.

Table 3: Per-Class F1 Score Comparison.

Class	Logistic	SVM	Hybrid
Economy	0.93	0.93	0.94
Health	0.98	0.98	0.98
Technology	0.92	0.93	0.95
Culture	0.96	0.97	0.97
Political	0.92	0.93	0.92
Sport	0.99	1.00	0.99

Source: Authors, (2026).

As displayed in Table 3, the Hybrid model especially increases performance in the Technology category, which indicates that contextual embeddings are used to disambiguate semantically related technological vocabulary. Sport category was classified almost perfectly in all models suggesting that there is an excellent lexical separability in the sports related lexicon. The greatest confusion is between Political and Economics which is not surprising considering topical context and overlap in the terminology used in the news articles.

V.4 DISCUSSION

A number of important observations can be made based on the results:

1. The classical linear models are also highly competitive in the classification of structured Arabic news.
2. Transformer fine-tuning in itself will not necessarily lead to better performance.
3. The best results are always given when there is feature-level fusion (Hybrid model).
4. To obtain the stable and reliable performance, it was necessary to clean the data (removal of corrupted non-Arabic entries).
5. High accuracy levels of performance (>95%) are a sign of meaningful representational complementarity and not random variance.

The suggested hybrid model can be used as a solid framework to enhance real-world applications by solving complex classification problems. The integration of the lexical and contextual feature strategies is advantageous to the predictive modeling approaches to student specialization [28] and unique virtual keypad encoding [29]. The application of transformer-based embeddings such as AraBERT is further confirmed by methodological similarities in mobile security [30] and tendencies in generative AI [31]. Word embeddings-based for Arabic document retrieval in [32]. The broad applicability of results in areas such as e-learning requires stability by ensuring that empirical data is split [33]. The broad applicability of results in areas such as e-learning requires stability by ensuring that empirical data is split [9]. Moreover, this model is flexible, which facilitates forecasting the job market [34] and the morphosyntactic analysis of Arabic language, which addresses the linguistic complexity of Arabic language [35]. The application of explainable AI [10] allows one to find the most significant text features, and the comparison of different machine learning models guarantees a high level of accuracy in security applications [11]. Lastly, the adoption of these automated systems is an indication of a massive effect on operational efficiency within the areas of HR performance [12]. The use of deep networks in retrieving Arabic documents [36] and query expansion with deep averaging networks [37] proves the effectiveness of embeddings in information retrieval. Explainable AI [38] can be used to determine the strongest text features, and different machine learning models. These results confirm the hypothesis that lexical statistical signals and contextual embeddings reflect various ways of textual information. The combination of their integration gives a richer space of representation, hence enhancing generalization.

VI. CONCLUSION

In the present study, the authors examined the performance of classical, transformer-based and hybrid lexical-contextual methods of the Arabic multi-class news classification. The results have revealed that, even though deep learning has significantly advanced in Natural Language Processing, the traditional lexical model, including TF-IDF with Linear SVM, is still very competitive in the structured Arabic news data with the performance of 95.57% accuracy. Although the fine-tuned AraBERTv2 model found semantic contextual relations, its independent performance (91.6) suggests that contextual modeling is not the best way to fully harness the discriminatory ability of structured lexical patterns in news corpora. The hybrid feature-level fusion model, where TF-IDF vectors were combined with the AraBERT embeddings, had the best performance (96.03% accuracy and 96.02% Macro-F1) which, as one would expect, lexical statistical signals and contextual embeddings complement one another. The outcome of this test discloses the significance of representation synergy over methodological replacement in Arabic NLP tasks. In general, the paper presents empirical data that hybrid lexical-contextual learning is a robust, computationally efficient, and practical solution to Arabic news classification. Its repeatability of the preprocessing pipeline and balanced assessment framework also enhance its utility in future research and practical Arabic text analytics systems.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

Methodology: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

Investigation: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

Discussion of results: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

Writing – Original Draft: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

Writing – Review and Editing: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

Resources: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

Supervision: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

Approval of the final text: Yasir Hadi Farhan, Mustafa Tareq, Boumedyen Shannaq, Said Almaqbali, Oualid Ali.

VIII. REFERENCES

- [1] A. Ghaddar, P. Langlais, M. Rezagholizadeh, and B. J. a. p. a. Chen, "On the importance of Data Scale in Pretraining Arabic Language Models," 2024.
- [2] G. Bhatia, A. El Mekki, F. Alwajih, and M. Abdul-Mageed, "Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks," in Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 4654-4670.
- [3] F. J. a. p. a. Qarah, "SaudiBERT: A large language model pretrained on saudi dialect corpora," 2024.
- [4] Y. H. Farhan, S. A. Mohd Noah, M. Mohd, and J. Atwan, "Word-embedding-based query expansion: Incorporating Deep Averaging Networks in Arabic document retrieval," *Journal of Information Science*, vol. 49, no. 5, pp. 1168-1186, 2021.
- [5] Y. Farhan, M. Tareq, B. Shannaq, O. Ali, S. Almaqbali, and V. J. I. Sriram, "MORSE-QE: A Morphology-Aware, Embedding-Driven Framework with Root Extraction for Arabic Dialectal Query Expansion," vol. 49, no. 20, 2025.
- [6] B. Shannaq, Y. H. Farhan, M. Tareq, O. Ali, and S. Almaqbali, "Overcoming Challenges in AI-Powered NLP Models: Enhancing the Capabilities of ChatGPT and DeepSeek," in 2025 3rd International Conference on Cyber Resilience (ICCR), 2025, pp. 1-6: IEEE.
- [7] Y. H. FARHAN et al., "UTILIZING WORD EMBEDDING'S FOR AUTOMATED QUERY EXPANSION IN ARABIC INFORMATION RETRIEVAL: A BLENDED METHODOLOGY," vol. 103, no. 2, 2025.
- [8] Y. H. Farhan, M. Mohd, S. A. M. J. J. o. I. S. T. Noah, and Practice, "Survey of automatic query expansion for Arabic text retrieval," vol. 8, no. 4, pp. 67-86, 2020.
- [9] B. Shannaq, "DOES DATASET SPLITTING IMPACT ARABIC TEXT CLASSIFICATION MORE THAN PREPROCESSING? AN EMPIRICAL ANALYSIS IN BIG DATA ANALYTICS," *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 3, 2025, [Online]. Available: chrome-extension://efaidnbmninnibpcapcglcfindmkaj/https://www.jatit.org/volumes/Vol103No3/19Vol103No3.pdf
- [10] B. Shannaq, "Using Explainable AI to Rank Educational Success Factors Driving the Job Market," *J. Info. Know. Mgmt.*, p. 2550064, Aug. 2025, doi: 10.1142/S0219649225500649.

- [11] B. Shannaq, O. Ali, S. A. Maqbali, and A. Al-Zeidi, "Advancing user classification models: A comparative analysis of machine learning approaches to enhance faculty password policies at the University of Buraimi," *J. Infrass. Policy. Dev.*, vol. 8, no. 13, p. 9311, Nov. 2024, doi: 10.24294/jipd9311.
- [12] B. Shannaq and A. Alabri, "EVALUATING THE IMPACT OF CHATGPT IMPLEMENTATION ON HR PERFORMANCE SKILLS: A CASE STUDY OF ACADEMIC INSTITUTIONS IN OMAN", *BJMSR*, vol. 10, no. 1, pp. 11–21, Feb. 2025, doi: 10.46281/bjmsr.v10i1.2281.
- [13] 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*, 2020, pp. 9-15.
- [14] A. Aljamel, H. Khalil, Y. J. T. I. J. o. E. Aburawi, and I. Technology, "Comparative study of fine tuned BERT-based models and RNN-based models. case study: Arabic fake news detection," vol. 12, no. 1, pp. 56-64, 2024.
- [15] N. Sengupta et al., "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," 2023.
- [16] I. Akhadam and H. Ayyad, "Enhancing Arabic text classification: A comparative study of machine learning and deep learning approaches," in *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*, 2024, pp. 1-6: IEEE.
- [17] A. A. Jihad et al., "Optimizing Neural Networks Using a Group Learning Algorithm for Heart Disease Prediction with the Cleveland Dataset," vol. 49, no. 24, 2025.
- [18] Y. Bayu, "Multi-label emotion classification on social media comments using deep learning," 2024.
- [19] M. K. B. Melhem, L. Abualigah, R. A. Zitar, A. G. Hussien, D. J. C. a. w. d. I. Oliva, and m. I. technologies, "Comparative study on Arabic text classification: challenges and opportunities," pp. 217-224, 2022.
- [20] G. Salton, C. J. I. p. Buckley, and management, "Term-weighting approaches in automatic text retrieval," vol. 24, no. 5, pp. 513-523, 1988.
- [21] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning* (no. 4). Springer, 2006.
- [22] C. Cortes and V. J. M. I. Vapnik, "Support-vector networks," vol. 20, no. 3, pp. 273-297, 1995.
- [23] 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)*, 2019, pp. 4171-4186.
- [24] A. Vaswani et al., "Attention is all you need," vol. 30, 2017.
- [25] R. Ranjan et al., "A manifold-level hybrid deep learning approach for sentiment classification using an autoregressive model," vol. 13, no. 5, p. 3091, 2023.
- [26] J. Ye, J. Hai, J. Song, and Z. J. A. S. o. T. S. P. Wang, "Multimodal data hybrid fusion and natural language processing for clinical prediction models," vol. 2024, p. 191, 2024.
- [27] M. Sokolova, G. J. I. p. Lapalme, and management, "A systematic analysis of performance measures for classification tasks," vol. 45, no. 4, pp. 427-437, 2009.
- [28] F. J. I. Alazzawi, B. Shannaq, and A. Al-Zeidi, "Predictive Modeling for Student Specialization Selection: Analyzing Influences and Developing Tailored Recommendations," in *Advances in Information and Communication*, vol. 1285, K. Arai, Ed., in *Lecture Notes in Networks and Systems*, vol. 1285. , Cham: Springer Nature Switzerland, 2025, pp. 633–645. doi: 10.1007/978-3-031-84460-7_40.
- [29] B. Shannaq, "DEVELOPMENT OF A NEW ENCODING ALGORITHM USING VIRTUAL KEYPAD LETTER SUBSTITUTIONS FOR IMPROVED TEXT CLASSIFICATION," *JATIT*, vol. 103, no. 2, Jan. 2025, [Online]. Available: <https://www.jatit.org/volumes/Vol103No2/27Vol103No2.pdf>
- [30] B. Shannaq, "A Novel Algorithm for Differentiating Authorized Users from Fraudsters via Mobile Keypad Input Patterns During Password Updates," *TEM Journal*, pp. 768–778, Feb. 2025, doi: 10.18421/TEM141-68.
- [31] I. Alshamsi, K. F. Sadriwala, F. J. Ibrahim Alazzawi, and B. Shannaq, "Exploring the impact of generative AI technologies on education: Academic expert perspectives, trends, and implications for sustainable development goals," *J. Infrass. Policy. Dev.*, vol. 8, no. 11, p. 8532, Oct. 2024, doi: 10.24294/jipd.v8i11.8532.
- [32] Y. H. Farhan, S. A. M. Noah, M. Mohd, J. J. J. o. I. S. T. Atwan, and Practice, "Word embeddings-based pseudo relevance feedback using deep averaging networks for Arabic document retrieval," vol. 9, no. 2, pp. 1-17, 2021.
- [33] F. J. I. Alazzawi, B. Shannaq, S. Al Maqbali, and S. Al Rawahi, "The Role of Predictive Analytics for Adaptive E-Learning: A Future Pathway Framework for Personalized Education," in *Sustainable Data Management*, vol. 171, R. K. Hamdan, Ed., in *Studies in Big Data*, vol. 171. , Cham: Springer Nature Switzerland, 2025, pp. 367–376. doi: 10.1007/978-3-031-83911-5_32.
- [34] B. Shannaq, "Leveraging AI to Forecast Future Job Market Trends: Smart Solutions and Strategic Recommendations for Career Success," *J. Info. Know. Mgmt.*, p. 2550070, Aug. 2025, doi: 10.1142/S0219649225500704.
- [35] B. Shannaq, "Optimizing n-gram lengths for cross-linguistic text classification: A comparative analysis of English and Arabic morphosyntactic structures," *Int. j. adv. appl. sci.*, vol. 12, no. 4, pp. 136–145, May 2025, doi: 10.21833/ijaas.2025.04.015.
- [36] Y. H. Farhan, M. Tareq, B. Shannaq, S. Almaqbali, and O. Ali, "Enhancing Robustness and Scalability in AI-Powered Natural Language Processing: A Novel Hybrid Deep Symbolic Framework," in *2025 3rd International Conference on Cyber Resilience (ICCR)*, 2025, pp. 1-5: IEEE.
- [37] Y. H. Farhan, M. Shakir, M. Abd Tareq, B. J. J. o. I. S. T. Shannaq, and Practice, "Incorporating Deep Median Networks for Arabic Document Retrieval Using Word Embeddings-Based Query Expansion," vol. 12, no. 3, pp. 36-48, 2024.
- [38] Y. H. Farhan, S. A. Mohd Noah, M. Mohd, and J. J. J. o. I. S. T. Atwan, "Word-embedding-based query expansion: Incorporating Deep Averaging Networks in Arabic document retrieval," vol. 49, no. 5, pp. 1168-1186, 2023.