

# Hierarchical Classification of Arabic Media: Employing a Two-Task BERT Model

Aisha Blfgeh<sup>\*1</sup>, Hanen Himdi<sup>2</sup>, Rehab Qarout<sup>3</sup>, Fatmah Assiri<sup>4</sup>, Amjad Qashlan<sup>5</sup>

Department of Computer Science and Artificial Intelligence,  
College of Computer Science and Engineering, University of Jeddah,  
Jeddah 23218, Saudi Arabia<sup>1,2,3</sup>

Department of Software Engineering, College of Computer Science and Engineering,  
University of Jeddah, Jeddah 23218, Saudi Arabia<sup>4</sup>

Department of Cyber Security, College of Computer Science and Engineering,  
University of Jeddah, Jeddah 23218, Saudi Arabia<sup>5</sup>

**Abstract**—Thousands of news articles are published on the web daily, causing the need to filter and extract knowledge on specific topics. With the increasing proliferation of digital content, automatically differentiating between informative news reporting and opinion-based columns has become essential to support fact-checking, content moderation, and journalistic integrity. Although previous studies have focused mainly on classifying news articles by domain in languages such as English, this research advances Arabic news classification by proposing a novel hierarchical two-task model to distinguish between news articles and columns. The proposed model leverages AraBERT, a pre-trained Arabic transformer-based language model, to first differentiate between news and columns, followed by the classification of news articles into six categories: Economy, Health, Local, Politics, Sports, and Technology. To support model development, a comprehensive dataset comprising 7,367 columns was collected from various Arabic news platforms. Experimental results demonstrate that the model achieved 95% accuracy in distinguishing news from columns and 94% accuracy in classifying news articles into their respective domains. When tested on real-world articles, the model attained 96% accuracy in the first classification task and 75% accuracy in domain classification. Additionally, an in-depth performance analysis was conducted at both the class and word levels using feature importance and attention mechanisms, respectively, to better understand the linguistic characteristics that differentiate news from columns. These findings underscore the model's effectiveness and highlight its potential for future optimizations, particularly in enhancing Arabic text classification and improving the nuanced differentiation between journalistic writing styles.

**Keywords**—Arabic text classification; AraBERT; BERT; Hierarchical classification; natural language processing; news article classification; opinion column detection; transformer models

## I. INTRODUCTION

The internet provides constant data to support news writers in producing articles and columns [1]. However, the growing volume of digital content necessitates accurate and speedy fact-checking to navigate it effectively. A significant challenge in news fact-checking relates to differentiating between news articles and opinion columns, where news refers to recent occurrences reported in newspapers or broadcasts [2]. Fuller [3] defines news as “a report of what a news organization has recently learned about matters of some significance or interest

to the specific community that news organization serves” (p. 6). However, the columns are different. Scherlen [4] defines a column as “a section of a newspaper, magazine, journal, or other serial publication where a specific writer expresses opinions, analysis, or treatment of a particular topic on a regular basis” (p. 81). Thus, a column is a place where writers and experts can express their opinions. This space is open to external specialists and in-house journalists who can contribute to various subjects. Columns typically provide in-depth analysis and opinions, often backed by arguments and supporting data. Almost any topic can be the subject of a column. News and columns may seem similar, but they have different purposes. Approximately 80% of journalists surveyed by the Media Insight Project said news organizations must clearly distinguish between news reporting and commentary [5]. Since articles and columns have different goals, distinguishing them is vital. Columns express personal viewpoints, while news articles give objective, evidence-based information. Research has shown that readers often confuse factual reporting with opinion, especially when reading online with minimal context or presentation. The Pew Research Center<sup>1</sup> found that several readers misinterpret factual and opinion statements, highlighting the difficulty of distinguishing objective reporting from subjective analysis. Such action is needed to overcome fake news while maintaining public trust. Automatically identifying opinion in news can assist fact-checking and event extraction.

Advancements in artificial intelligence (AI), particularly in natural language processing (NLP), have developed sophisticated models capable of automatically classifying texts. Veritably, there is an emerging body of work on classifying the domains of news articles (e.g. [6], [7]). Although these models can distinguish opinion pieces from reports, their consistency in making this distinction is uncertain. Indeed, advanced tools for journalists exist for verification; however, their accuracy remains questionable<sup>2</sup>. Recently, pre-trained transformer-based models have advanced language modeling, particularly in text classification [8], [9], potentially distinguishing columns from news. Even as numerous studies have explored column classification [10], [11] and information extraction [12], [13], they have been limited to the English language.

<sup>1</sup><https://www.pewresearch.org/journalism/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news>

<sup>2</sup><https://www.axios.com/2025/01/17/apple-ai-news-alerts-fake-headlines>

<sup>\*</sup>Corresponding author.

To the authors' knowledge, no prior work has focused on separating articles from opinion columns and simultaneously classifying them into their respective domains. This study introduces a novel hierarchical classification approach that distinguishes news from opinion pieces and subsequently classifies news articles into specific domains. Utilizing sequential decision-making, this approach improves the depth and precision of text classification. This study addresses a tricky issue: news and column articles often use the same formal journalistic writing style and rely solely on textual features, making textual distinctions difficult to identify.

Thus, automatic categorization of news and opinion articles improves news organization, media evaluation, information extraction, and content suggestion while maintaining transparency and reader trust in journalism. Additionally, a recent study revealed that natural language processing still struggles to distinguish news from opinion because their linguistic characteristics often overlap [14]. Hence, strong classification models become crucial. For this, the paper provides multiple contributions:

- 1) The curation of a large dataset comprising 7,367 Arabic column articles on diverse topics.
- 2) Developing an innovative Two-Task Hierarchical Classification Model, leveraging AraBERT to differentiate between news and columns, then further classify the news articles in their specific domain.
- 3) Offering a generalization through empirical experiments on an unseen dataset extracted from real-world articles.
- 4) Detailing a comprehensive model's performance through word-level and class-level analysis. Hence, offering linguistic findings that help with similar tasks.

A critical review of the literature highlights two limitations in classification model research, despite their global growth: Currently, most automated column classification, perspective identification, and opinion mining methods are limited to the English language due to linguistic and geographic biases. The Arabic NLP community has focused mainly on sentiment analysis of informal social media text or flat news classification, neglecting the structural traces of official Arabic editorial writing. Second, most Arabic classification techniques presuppose a flat taxonomy of categories, leading to stylistic neglect and homogeneity. Media text is not considered bi-dimensional, with intent/genre profiles (objective facts vs. subjective opinion analysis) and topical profiles (e.g., economy or politics). Columnists and news reporters typically use the same journalistic style and vocabulary. Therefore, traditional single-stage classifiers have poor accuracy due to their inability to differentiate fine-grained structural boundaries. The authors believe that this complex issue has not been addressed directly in Arabic NLP before. This study addresses the multi-layered challenge by presenting a dedicated Arabic dataset and a two-task hierarchical model that disentangles the editing style and subject domain.

This paper continues as follows. The Related Work section analyzes Arabic news classification and deep learning text classification studies. The Methodology section covers dataset preparation, pre-processing, and two-task hierarchical classification

model architecture. The Experimental Setup section describes the experimental design and the evaluation procedure, whereas the Model Performance Analysis section discusses classification findings on the primary and unseen datasets. The Conclusion highlights the main findings and outlines future research.

## II. RELATED WORK

### A. Methodological Paradigms in Text Classification

To systematically analyze the evolution of text classification frameworks, existing research can be broadly categorized into three distinct methodological paradigms:

- 1) Deep Learning Approaches, which utilize neural network architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to automatically extract spatial and sequential patterns;
- 2) Pre-trained Transformer Models, which leverage massive self-attention mechanisms to capture complex contextual dependencies. The following subsections critically examine these paradigms, highlighting their respective contributions, computational trade-offs, and technical limitations.

### B. Deep Learning Approaches

To overcome manual feature engineering and the need for strong Arabic text classification, neural network frameworks autonomously capture latent textual information. Imad et al. [15] classified news using neural networks (CNNs). They gathered and pre-processed text using the CNN Arabic, BBC Arabic, and OSAC corpora. The study used Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BOW), and One-Hot Encoding (OHE) to extract features. The study presented a CNN model with an input layer for data vectors and various one-dimensional convolutional layers for capturing local text patterns. Max-pooling layers reduced feature map dimensionality, with fully connected layers integrating the extracted features. A dense output layer categorized texts in the model's final layer. The CNN dataset had 94.47% accuracy using One-Hot Encoding, and the OSAC dataset had 99.42% accuracy using BOW.

Similarly, Alroobaea [16] classified Arabic news items into business, entertainment, health, politics, sports, and technology using a deep learning model. Stop words, lemmatization, URLs, mentions, and hashtags were removed from 5000 tweets during pre-processing. The study applied deep learning models, including CNN, LSTM, and a hybrid CNN-LSTM framework. The CNN-LSTM model, with three convolutional layers and max-pooling layers, extracted local text features. The CNN output was fed into an LSTM layer to capture sequential dependencies. Ultimately, the hybrid CNN-LSTM model outperformed others with 93.15% classification accuracy.

### C. Pre-trained Transformer Models

The paradigm shift was the introduction of the self-attention mechanism in transformers that allowed models to pre-train on huge unlabeled datasets and fine-tune for specific

TABLE I. CRITICAL COMPARATIVE SYNTHESIS OF TEXT CLASSIFICATION PARADIGMS AND THE PROPOSED FRAMEWORK

| Methodological Paradigm                                                                                | Key Technical Strengths                                                                                                | Identified Weaknesses & Limitations                                                                                  | Proposed Framework Advantage                                                                                                           |
|--------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| <b>Traditional ML</b><br>(Selab & Guessoum; Khan et al.)                                               | Low computational complexity; fast training and inference; effective on small datasets.                                | Relies on manual feature engineering (TF-IDF, BoW); fails to capture deep semantic context or morphological nuances. | Replaces static features with end-to-end deep contextual representations using pre-trained transformer encoders.                       |
| <b>Deep Learning (CNN/LSTM)</b><br>(Imad et al.; Alrooba; Zhu; Liu)                                    | Automatically learns local n-gram patterns and temporal sequences without manual feature engineering.                  | Struggles with long-range document context; susceptible to vanishing gradients; requires training from scratch.      | Leverages pre-trained AraBERT weights optimized for Modern Standard Arabic, requiring fine-tuning rather than full training.           |
| <b>Pre-trained Transformers</b><br>(Garcia et al.; Rifat & Imran; Nugroho et al.; Escoufflaire et al.) | State-of-the-art contextual understanding, bidirectional semantic representations, and transfer learning capabilities. | Monolithic single-stage setup causes feature confusion when topics overlap heavily across different media genres.    | <b>Two-Stage Hierarchical Design:</b> Disentangles structural style/intent from subject domain prior to executing category assignment. |

tasks. To overcome the absence of several datasets, Garcia et al. [17] built a Portuguese news classification dataset from WikiNews. This dataset contains articles on politics, sports, economy and health. The authors compared Bag of Words (BoW), TF-IDF, and FastText embeddings with classifiers such as SVM, DJINN, CNNs, and BERT. BERT obtained a high F1-score of 91.8% on the new dataset.

The effectiveness of transformer models is also demonstrated in the study by Rifat and Imran [18]. Sentiment analysis and news classification in low-resourced language Khmer using transformer models. Over 100,000 Khmer news articles were used to test FastText and BERT for word embedding extraction. The fine-tuned BERT model has outperformed other methods (FastText embeddings and BERT-based feature extraction on deep neural networks) with an accuracy of 85%. This distinction shows the strength of transformer models for low-resource language challenges.

Nugroho et al. [8] used an NLP framework and deep learning models to improve scalable false-news detection for large-scale news classification. The study compared accuracy and resource utilization of pre-trained BERT models such as BERT-Tiny, BERT-Mini, BERT-Small, BERT-Medium, and BERT-Base on over 1 million news articles. The text processing was done in a distributed and scalable manner using Spark NLP. This approach could improve the efficiency of training and testing by minimizing the computational demands. BERT had higher accuracy without Spark NLP and lower training time with Spark NLP. It is more feasible for real-time management of large datasets but with a loss of accuracy.

Although Transformer models have excellent representation quality, most implementations have employed single-stage flat classification frameworks like topic labeling or sentiment assessment. They ignore the complexity of multi-layered domains and genres and assume all target classes are taxonomically homogeneous.

In the context of media genre categorization, [14] struggled to distinguish opinion pieces from news articles in the press and to classify media genres. They showed that fine-tuned contextual transformers such as CamemBERT are better than feature-based methods at detecting subtle stylistic and subjective features, particularly in genres like news and commentary, which have similar vocabulary distributions and formal journalistic styles. However, they evaluated single-stage, unified categorization frameworks on French press data.

#### D. Distinguishing the Hierarchical Framework from Existing Approaches

Arabic text classification systems typically use a flat, single-task learning architecture to categorize incoming papers into predefined themes (e.g., politics, sport, economics). These methods presume labels are in a single taxonomic plane. The Two-Task hierarchical categorization approach distinguishes the framework. Our framework classifies text into stylistic \intent and thematic \subject layers.

Recent media genre classification efforts, such as [14], highlight the utility of pre-trained transformers in distinguishing news and opinion articles, but these methods are limited to single-stage flat frameworks in non-Arabic contexts. Political news reports and political opinion columns have similar linguistic patterns, making domain vocabulary modeling difficult. Our sequential decision-making approach isolates the structural style (News vs Column) before domain classification on the specified news sample, capturing layered semantic distributions that flat models cannot resolve. In Table I, we compare the technological trade-offs between established paradigms and our suggested framework.

### III. METHODOLOGY

This research compiles a two-task hierarchical classification model leveraging an AraBERT model to classify Arabic news and columns, then news articles in their specific domains. The proposed model is textual-content-based, which focuses on the textual content of the articles without any associated metadata. The input to these models is a stream of articles, and the output is the articles' class. The following are details of the three-stage methodology of the research, as illustrated in Fig. 1.

#### A. Dataset

One of this paper's main contributions is the creation of a sizable Arabic column article corpus open to all interested researchers. This dataset possesses a journalistic theme to more authentically reflect the media available on news platforms.

1) *Columns Dataset:* A Python script utilizing an Application Programming Interface (API) was employed to extract Arabic opinion columns from multiple platforms, forming the dataset. The data collection spanned columns published from January 2020 to August 2023, resulting in a corpus of 7,500 columns sourced from news platforms in two Arab

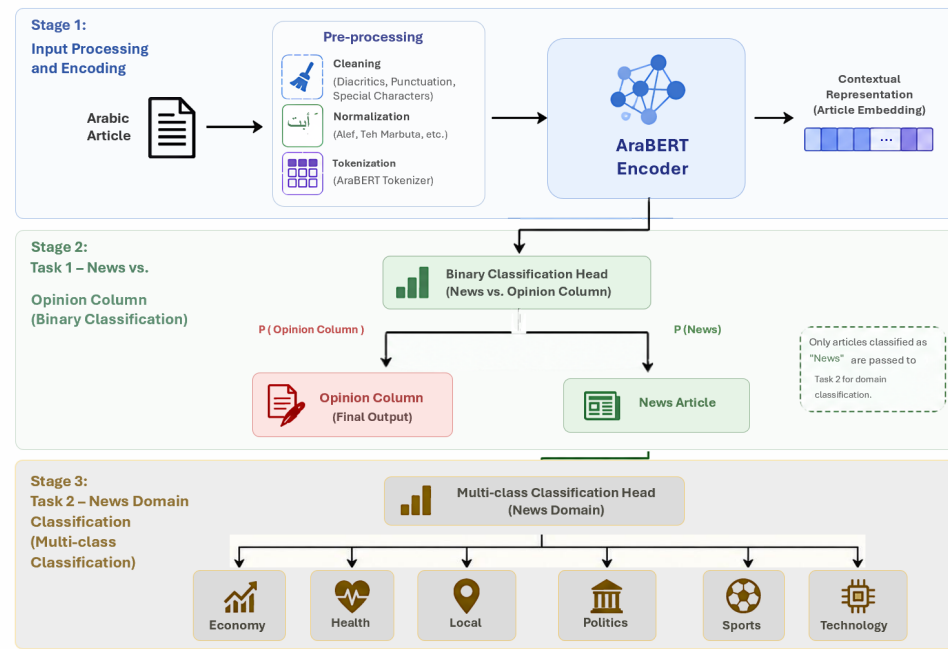


Fig. 1. Proposed two-task hierarchical AraBERT framework. The first task distinguishes news articles from opinion columns. Articles classified as news are subsequently passed to the second task, where they are categorized into one of six news domains: Economy, Health, Local, Politics, Sports, or Technology.

TABLE II. NEWS PLATFORMS AND ARTICLE STATISTICS

| News Platform      | Country      | Total Articles | Average (words) | Length |
|--------------------|--------------|----------------|-----------------|--------|
| Okaz <sup>a</sup>  | Saudi Arabia | 2500           | 506             |        |
| Youm7 <sup>b</sup> | Egypt        | 2500           | 487             |        |
| Sabq <sup>c</sup>  | Saudi Arabia | 2500           | 499             |        |

<sup>a</sup><https://www.okaz.com.sa/>

<sup>b</sup><https://www.youm7.com/>

<sup>c</sup><https://sabq.org/>

countries, Saudi Arabia and Egypt. To reduce sample bias, neither manual article selection nor topic-based filtering was applied; all retrieved column articles were preserved. During data preparation, duplicate articles were removed by comparing their textual content with the rest of the articles. When invalid or empty cases were found, they were discarded as well. Table II provides an overview of the news platforms and the exact number of columns extracted from each source. We aimed to extract the column articles specifically published in the column section of the newspaper. This section may also be labeled under different names such as (مقالات) ‘Op-Eds’, (مدونات) ‘Columns’, (مرييات) ‘Viewpoints’, depending on the newspaper’s format.

Politics, economy, health, sports, and religion were covered in the columns. Some pieces covered specific sports events, while others covered multiple topics such as sports and politics. The dataset comprised columns authored by people from diverse backgrounds, including male and female writers, different age groups, and different Arabic nationalities. The columns were randomly selected from news sites without regard to author gender, age, or nationality to reduce bias. The dataset provides detailed metadata, including title, publication

date, article link, source, and author name, enabling descriptive analysis. Only the articles’ textual content was included in this study, and duplicate columns were deleted, leaving 7,367 columns.

2) *Sub-Categorical News Articles*: Regarding the sub-categorical news articles, the publicly available ANAD dataset [19] was employed. This dataset is a state-of-the-art, standardized Arabic news corpus compiled from articles collected over a one-year period, from January 2021 to December 2021. It comprises over 500,000 articles sourced from 12 Arabic news platforms, covering diverse topics such as sports, economics, local news, politics, technology, tourism, entertainment, automobiles, health, and art.

The researchers concentrated on six domains: economy, health, local news, politics, sports, and technology to improve model consistency and compilation. Within these domains, 7,500 ANAD articles were extracted with a 600-word limit. This limitation ensured article length consistency across news and columns categories. As illustrated in Fig. 2, the dataset mainly included articles under 600 words before standardization. A sample of articles is shown in Table III.

The final dataset had 7,500 news articles and 7,367 column articles. Every news domain—Economy, Health, Local, Politics, Sports, and Technology—had 1,250 articles, providing a balanced distribution. The diversity was handled during model training and evaluation to mitigate the potential effects of class imbalance. Throughout the study, this dataset is named the ‘main’ dataset.

## B. Data Pre-Processing

The dataset includes various articles written in modern standard Arabic, including simple diacritics, links, and non-

TABLE III. EXCERPT SAMPLES OF THE ARTICLES

|         |                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Economy | وزير الاتصالات: لا صحة لطرح حصة إضافية من وزير الاتصالات وتكنولوجيا المعلومات... في مجال الفينتك ، المدن الذكية ، و التكنولوجيا المالية ، و الخدمات الصحية.<br>Minister of Communications: There is no truth to the idea of offering an additional stake in Telecom Egypt Minister of Communications and Information Technology...in the field of fintech, smart cities, financial technology, and health services. |
| Health  | المقلبات واللحوم المقددة تزيد خطر الإصابة بأمراض القلب حيث دراسة حديثة نشرت نتائجها في الولايات المتحدة ، أن محبي .. بحسب الجمعية الأميركية لطب القلب.<br>Fried foods and cured meats increase the risk of heart disease, according to a recent study published in the United States, which showed that enthusiasts. . . according to the American Heart Association.                                               |
| Columns | مشكلة التنظيمات المؤجلة أنها لا تعي أهم أساسيات و مفاهيم... لدعوة تنظيم أصبح بنفسه يسارع إلى إعلان شهادة وفاته.<br>The problem with ideological organizations is that they are not aware of the most important basics and concepts... To invite an organization that itself became quick to announce his death certificate.                                                                                         |

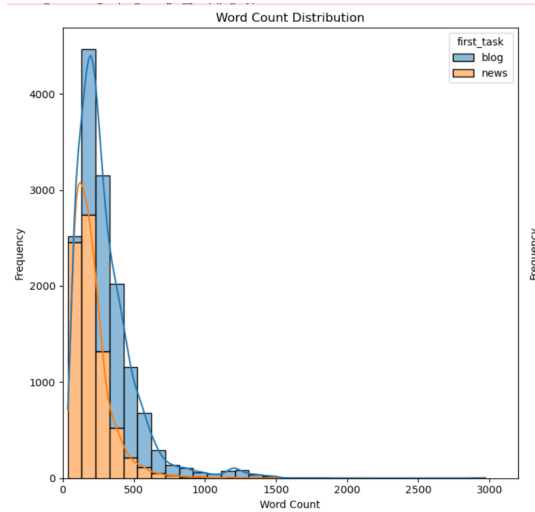


Fig. 2. Multi-categorical news articles dataset word counts.

alphabetic characters. Consequently, applying feature extraction and text classification to such data could yield poor results due to the inherent complexity of the Arabic language, including its morphological richness, dialectal variations, and the lack of labeled datasets for fine-grained classification tasks. Tasaheel [20], an Arabic NLP tool, was used to conduct all the pre-processing procedures needed to address these aberrations.

- **Cleaning:** This involves removing diacritics, punctuation marks (except periods/full stops), and non-alphabetic characters such as hashtags, emails, and web page links.
- **Normalization:** This involves elongating words by removing the repetition of three or more characters and normalizing the three Arabic letters: Alef (أ), Alef maqsoura (إ), and ta-marbouta (ة).
- **Tokenization:** This involves separating a piece of text into smaller chunks called tokens. This operation is necessary to facilitate analysis by separating words in otherwise unstructured text.

### C. Model Architectures

The AraBERT pre-trained model was chosen for this research based on a number of important criteria. Starting with its open-access accessibility makes it a widely accessible and well-documented resource for Arabic NLP research<sup>3</sup>. Second, AraBERT has shown great performance in earlier research on text categorization projects [21], [22], making it a viable option for handling complex linguistic patterns in Arabic. Given these considerations, AraBERT was found appropriate for acquiring contextual representations in Arabic text classification, especially in tasks requiring complex linguistic understanding.

1) *AraBERT*: AraBERT is a fine-tuned language model based on BERT, specifically designed for Arabic natural language processing. A model of this nature can be employed for sentiment analysis through the monitoring of social media or news platforms, as well as for named entity recognition (NER) to identify entities such as names and locations within natural language text. AraBERT is a transformer model designed to capture the meaning and structure of the Arabic language. It does this by analyzing each word in a sentence in relation to every token within a sequence, both preceding and following it. This model is likely applicable in machine translation tasks involving text-to-language conversions and vice versa.

The encoder-decoder architecture of the transformer model facilitates the self-attention mechanism. This architecture effectively employs parallel processing to encode characters for sequence-to-sequence tasks that involve distant dependencies. This architecture consists of multiple layers of the same type, utilized in parallel to process input, enabling the model to generate highly rich, contextualized embeddings for each token without sequential processing steps. The design relies heavily on self-attention mechanisms (or intra-attention), enabling each element in the sequence to identify the most relevant elements within different contextual ranges, thereby implicitly weighing these aspects based on their contextual significance. The transformer architecture in the BERT model is utilized in AraBERT, wherein the transformer layer executes self-attention across a sequence of tokens, adhering to the principles of Scaled Dot-Product Attention as delineated in Eq. (1).

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

<sup>3</sup><https://huggingface.co/aubmindlab/bert-base-arabert>

where,  $Q$ ,  $K$ , and  $V$  represent the queries, keys, and values derived from the input sequence,  $d_k$  is the dimension of the key vectors, included to scale the dot product for numerical stability. Thus, this formula computes attention by calculating the dot product of queries  $Q$  and keys  $K$ , scaling it by  $\sqrt{d_k}$ , applying the softmax function, and finally multiplying by the values  $V$ .

In a hierarchical architecture, AraBERT can serve as the primary encoder, providing contextual embeddings for the input. For both classification tasks, news vs. columns, and news domain categorization, two AraBERT models with the highest weights process the input sequences independently. The architectural details for each task are described further. In this study, we employ [23] as our primary model, which is openly accessible in the HuggingFace library [24].

**2) Two-Task Hierarchical Classification Model:** Two independently fine-tuned AraBERTv02 models form a hierarchical pipeline in the proposed framework. The initial AraBERTv02 model binary classifies news from opinion columns. Only news articles are given to the second independently fine-tuned AraBERTv02 model, which classifies them into one of six news domains. This hierarchical approach enables each model to specialize in its classification task, while improving the overall efficiency and accuracy of the framework. The dataset contains 7,500 news articles (1,250 per domain) and 7,367 opinion columns, which do not belong to any news subcategory.

*First Task: Classifying Articles as News or Columns:*

The pre-trained AraBERT tokenizer<sup>4</sup> was employed to tokenize the article text, converting it into numerical representations. This is followed by padding and truncation to maintain the same sequence length, which was set to a maximum of 350 characters. Thus, the text is now a multidimensional vector with several tokens, as Eq. (2) explains. However, since the target class is a categorical value (news or columns), a label encoding was used to encode the values 0 and 1, which represent a column and a news value, respectively, as explained in Eq. (3).

$$T = \text{Tokenizer}(X) = \{t_1, t_2, \dots, t_i\} \quad (2)$$

$$y = \text{LabelEncoder}(L) = \{0, 1\} \quad (3)$$

Consequently, the AraBERT model processes the tokenized text  $T$  to produce a pooling output  $P$ , as shown in Eq. (4). The pooling step aggregates these token outputs into a single, fixed-size vector that encapsulates the most significant information from the input text. For our binary classification of news or columns, the output is routed via a dense layer with a sigmoid activation function as explained in Eq. (5).

$$P = f_{\text{AraBERT}}(T) \quad (4)$$

$$\hat{y} = \sigma(WP + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

where,  $\sigma(z)$  is the sigmoid function, and  $W$  represents the weights,  $b$  represents the bias, and  $P$  is as shown in Eq. (4).

The model is trained utilizing the Adam optimizer, employs binary cross-entropy loss, and is trained for five epochs with a batch size of 16. Ultimately, the prediction function is shown in Eq. (6).

$$\text{Prediction} = \begin{cases} \text{news} & \text{if } \hat{y} \geq 0.5 \\ \text{columns} & \text{if } \hat{y} < 0.5 \end{cases} \quad (6)$$

**Second Task: News category classification:** If the first task categorizes the article as 'news', the following task proceeds to classify the article into its specific domains. Similar to Eq. (2), the text is encoded into vectors  $T'$ , while the target class is encoded in  $k$  domains, and  $y'$  represents the article's domain, as explained in Eq. (7).

$$y = \text{LabelEncoder}(\text{NEWS}) = \{0, 1, \dots, k-1\} \quad (7)$$

Similar to the first task, the tokenized text  $T'$  undergoes processing by another AraBERT model, resulting in a pooled output  $P'$  as shown in Eq. (8). Next, for our multi-class classification, the pooled output passes through a dense layer with a softmax activation function as shown in Eq. (9).

$$P' = f_{\text{AraBERT}}(T') \quad (8)$$

$$\hat{y}_i = \text{Softmax}(z_i) = \frac{e^{-z_i}}{\sum_{j=1}^k 1 + e^{-z_j}} \quad \forall i \in \{1, 2, \dots, k\} \quad (9)$$

The model is trained to utilize the Adam optimizer, sparse categorical cross-entropy loss, and a batch size of 16 for five epochs. Ultimately, the prediction function is shown in Eq. (10).

$$\text{Prediction} = \arg \max \hat{y}_i \quad (10)$$

The dataset included Modern Standard Arabic news sourced from various Arabic sources. Further, AraBERT consisted of 12 transformer blocks, each comprising 768 hidden units, and contained 12 self-attention heads, with a total of 110 million trainable parameters. This study employed its base version, while implementing the approach for the hierarchical two-stage classification system using two independently fine-tuned AraBERT models, where the input text is encoded with AraBERT and then directed to the second task-specific classification model, as illustrated in Fig. 3.

#### IV. EXPERIMENTAL SETUP

The proposed hierarchical framework consists of the compiled model, for which the researchers performed extensive experiments and adjusted the parameters.

Both classification models were initialized from the pre-trained AraBERTv02 (aubmindlab/bert-base-arabertv02) checkpoint. The tokenizer associated with the pre-trained model was employed to tokenize all input articles. Each input sequence was padded or truncated to a maximum length of 512 tokens. Both models were fine-tuned using the Adam optimizer with a learning rate of  $1 \times 10^{-4}$ , a batch size of 16, and five

<sup>4</sup><https://huggingface.co/aubmindlab/bert-base-arabertv02>

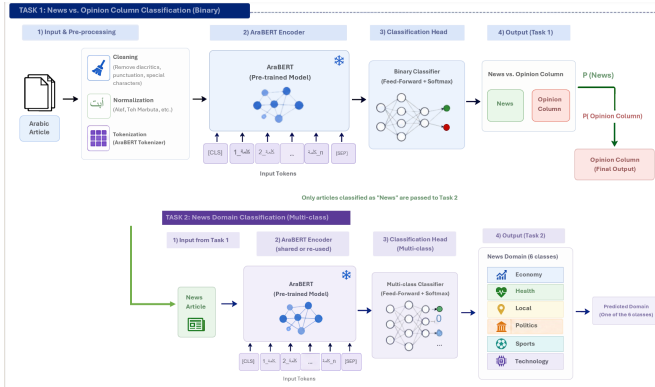


Fig. 3. Detailed architecture of the proposed two-task hierarchical AraBERT framework. Task 1 (top) is a binary classifier that distinguishes between news articles and opinion columns. Articles predicted as news are forwarded to Task 2 (bottom), a multi-class classifier that assigns the article to one of six news domains.

training epochs. For the first task (News vs. Opinion Columns), a sigmoid output layer with binary cross-entropy loss was used. For the second task (News Domain Classification), a softmax output layer with sparse categorical cross-entropy loss was employed. The dataset was randomly divided into 80% for training and 20% for testing, and the model achieving the best validation performance was retained for evaluation.

The selected hyperparameters were determined experimentally after evaluating multiple configurations. The main dataset included 7,500 news and 7,367 column articles for the first task. For the second task, each news category included 1,250 articles, ensuring a balanced distribution across categories.

In order to conduct a thorough comparison, four traditional machine learning classifiers—Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and Logistic Regression (LR)—were evaluated using the main dataset. For these baseline models, Term Frequency–Inverse Document Frequency (TF-IDF) was employed as the feature extraction method because to its efficacy in encapsulating textual data. TF-IDF allocates a weight to each term based on its occurrence in a document and its scarcity throughout the corpus, generating insightful feature vectors for text categorization. The derived TF-IDF features served as input for all machine learning classifiers, and an identical 80%/20% train-test split was employed to guarantee an equitable comparison with the proposed hierarchical model.

Experiments were conducted on a server with an Intel(R) Xeon(R) E5-2670 CPU, NVIDIA(R) GeForce GTX 1080 GPU with 8 GB of video memory, and 64 GB of RAM. The models were developed using Python 3 and run on Google Colab 12. The dataset and model compilation can be accessed freely on the author's GitHub<sup>5</sup>.

The metrics of precision ( $P$ ), recall ( $R$ ), F-measure ( $F$ ), and Accuracy ( $A$ ) are used to evaluate the performance of each algorithm, where TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively.

$$\text{Accuracy (A)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$\text{Precision (P)} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Recall (R)} = \frac{TP}{TP + FN} \quad (13)$$

$$\text{F-Score (F)} = \frac{2 \cdot P \cdot R}{P + R} \quad (14)$$

## V. RESULTS

The study conducted two experiments on two distinct datasets, each involving two classification tasks. In the first experiment, the aim was to distinguish between news and columns using the main dataset described in the Columns Dataset. Regarding the baseline models, the evaluation of the traditional machine learning models showed that LR produced the best overall performance, achieving 89.6% accuracy; RF ranked second, obtaining 88.0% accuracy; NB produced 86.9% accuracy, and SVM recorded the lowest performance, achieving 83.8% accuracy. Although these baseline models demonstrated satisfactory classification capability using TF-IDF features, they were consistently outperformed by the proposed model, as shown in Table IV. The model achieved an overall accuracy of 95%. Specifically, precision scores were 92% for news and 98% for columns. Both classes exhibited identical F1-scores, indicating balanced performance. To get a better perspective on the model's prediction performance, Fig. 4a displays the confusion matrix for this task's predicted instances. According to the figure, 121 columns instances were misclassified as news, and 1357 columns were accurately identified, thus underscoring the need for further refinement to reduce errors. Moreover, it confirms the previous accuracy results, showing that the model correctly classified 1468 news articles, misclassifying only 27 as columns. In support, the Area Under the Curve (AUC) in (Fig. 5) further reinforces these findings, with an AUC of 0.95, demonstrating the model's strong discriminative capability. AUC is a performance measure that evaluates the ability of a model to differentiate across classes. It denotes the area under the Receiver Operating Characteristic (ROC) curve, which demonstrates the true positive rate in relation to the false positive rate across diverse threshold settings. An AUC score approaching 1.0 signifies strong classification efficacy, whereas a value of 0.5 implies performance no superior to random chance [25].

For the second task, which was for classifying the news articles into their specific domains employing the sub-dataset as described in Sub-Categorical News Articles, the model achieved an overall accuracy of 94%, with all weighted F1-scores for all the news domains classifications reaching above 90%. Performance varied among topics: the Health class achieved a perfect recall of 98%, while Politics had the lowest recall of 83%. Technology and Politics attained the highest precision scores of 98% and 97%, respectively, while the Economy exhibited the lowest at 88%. The lowest F1-score of 89% was observed for the Politics category, which may

<sup>5</sup><https://github.com/Hanen-Tarik>

TABLE IV. CLASSIFICATION RESULTS ON THE MAIN DATASET

| First Task Classification Report  |            |           |        |          |          |
|-----------------------------------|------------|-----------|--------|----------|----------|
| Models                            | Class      | Precision | Recall | F1 Score | Accuracy |
| BERT-tuned model                  | Column     | 0.98      | 0.92   | 0.95     | 0.95     |
|                                   | News       | 0.92      | 0.98   | 0.95     |          |
| Second Task Classification Report |            |           |        |          |          |
| BERT-tuned model                  | Economy    | 0.88      | 0.92   | 0.90     | 0.94     |
|                                   | Health     | 0.93      | 0.98   | 0.95     |          |
|                                   | Local      | 0.90      | 0.96   | 0.93     |          |
|                                   | Politics   | 0.97      | 0.83   | 0.89     |          |
|                                   | Sports     | 0.95      | 0.91   | 0.93     |          |
|                                   | Technology | 0.98      | 0.93   | 0.95     |          |

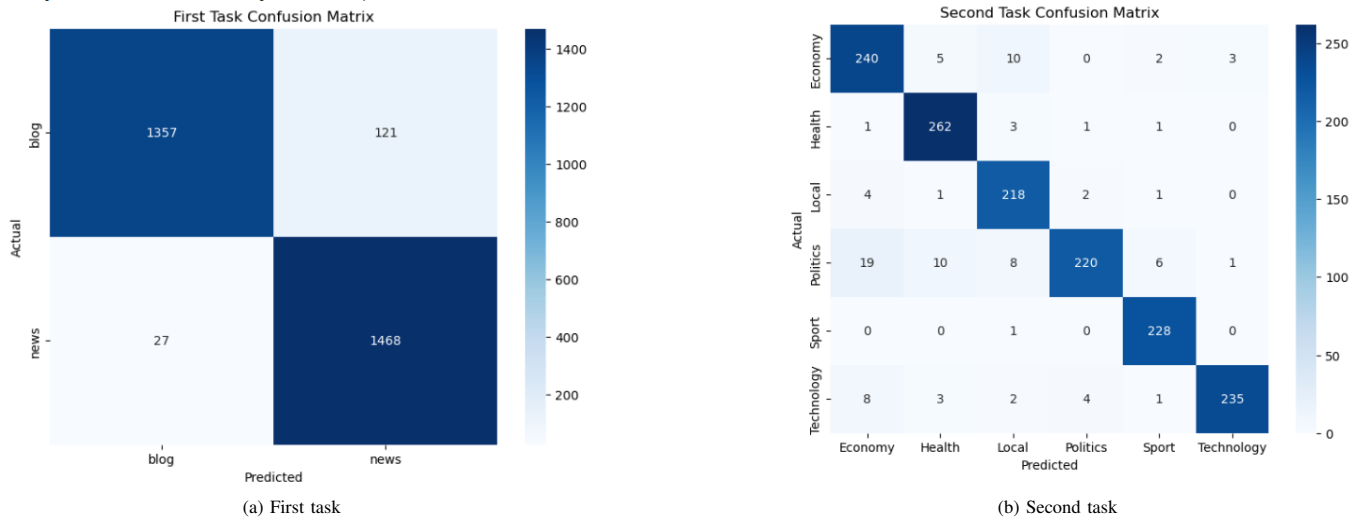


Fig. 4. Confusion matrix for both tasks on main dataset.

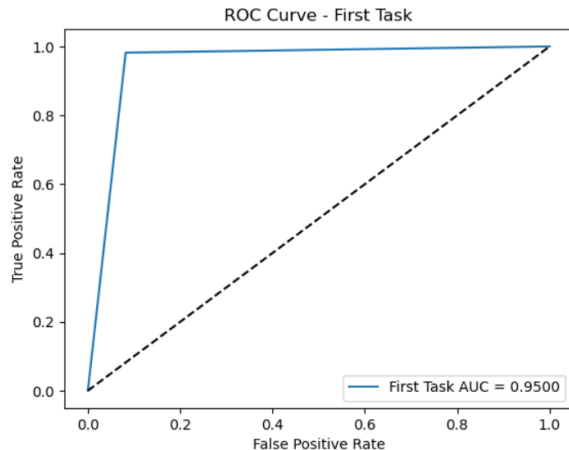


Fig. 5. ROC plot for news vs. columns classification of the main dataset.

be attributed to the diverse range of topics covered within political articles, in contrast to other domains that focus on more specific subject areas. From a more detailed perspective, the model exhibited particularly strong performance in the

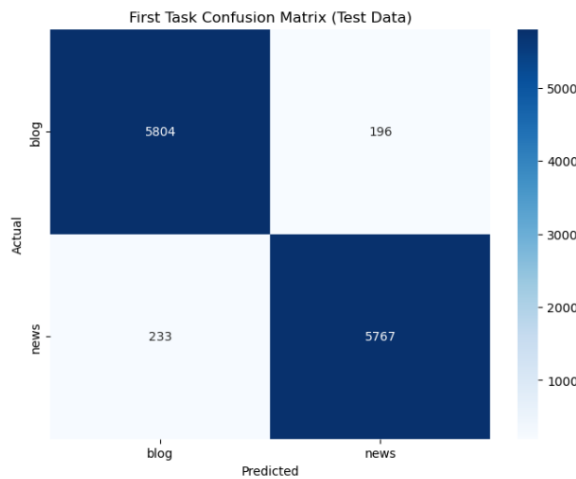
Health domain, achieving a precision of 93%, recall of 98%, and an F1-score of 95%. Similarly, the Sports category attained an F1-score of 93%, while Technology articles demonstrated high classification effectiveness with an F1-score of 95%. However, the Politics category presented challenges, 10 cases misclassified as Health, 8 as Local, and 6 as Sports. Likewise, the Economy category was frequently confused with Local (10 instances) and Health (5 instances). These findings highlight the model's classification difficulties, particularly in distinguishing Politics, Economy, and Health. Despite these challenges, the model consistently achieved high classification accuracy across all six domains, with a particularly strong performance in classifying articles in the Health, Sports, and Technology domains.

#### A. Unseen Dataset

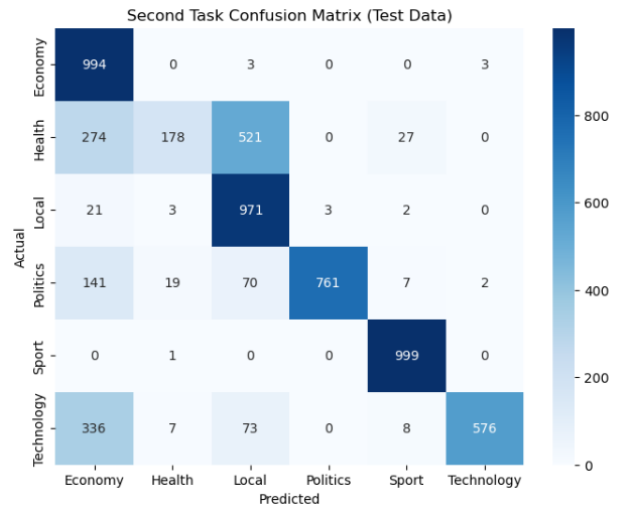
To evaluate the hierarchical classification model's generalization capability in classifying real-world instances, an experiment was conducted on an unseen dataset encompassing the two tasks. In the first task, 12,000 unseen articles were used—comprising 6,000 news articles from the ANAD dataset and 6,000 column articles manually extracted from the same news platforms as the main dataset to ensure structural

TABLE V. CLASSIFICATION RESULTS ON THE UNSEEN DATASET

| First Task Classification Report  |            |           |        |          |          |
|-----------------------------------|------------|-----------|--------|----------|----------|
| Models                            | Class      | Precision | Recall | F1 Score | Accuracy |
| BERT-tuned model                  | Column     | 0.96      | 0.97   | 0.96     | 0.96     |
|                                   | News       | 0.97      | 0.96   | 0.96     |          |
| Second Task Classification Report |            |           |        |          |          |
| BERT-tuned model                  | Economy    | 0.56      | 0.99   | 0.72     | 0.75     |
|                                   | Health     | 0.86      | 0.18   | 0.29     |          |
|                                   | Local      | 0.59      | 0.97   | 0.74     |          |
|                                   | Politics   | 1.00      | 0.76   | 0.86     |          |
|                                   | Sports     | 0.96      | 1.00   | 0.98     |          |
|                                   | Technology | 0.99      | 0.58   | 0.73     |          |



(a) First task



(b) Second task

Fig. 6. Confusion matrix for both tasks on the unseen dataset.

consistency. In the second task, 1,000 distinct news articles were selected for each of the six predefined domains from the ANAD dataset. The unseen dataset underwent the same pre-processing steps as the main dataset and was balanced across all classes, including 6,000 news and 6,000 column articles for the first task, and 1,000 articles per category for the second task. The evaluation was conducted using the hierarchical classification model, which had been fully trained on the original dataset without any retraining or fine-tuning, in order to provide a reliable assessment of its generalization performance.

As shown in Table V, the model achieved an overall accuracy of 96% in classifying news and columns, with a recall of 97% for the Column class and 96% for the news class. The confusion matrix for this task (Fig. 6a) further supports these findings, indicating that the model correctly classified 5,804 Column articles and 5,767 news articles, with minimal misclassifications in both categories. This performance is reinforced by an AUC of 96% (Fig. 7), suggesting that the model effectively differentiates between news and Column articles with high reliability.

However, in the second task, where news articles were

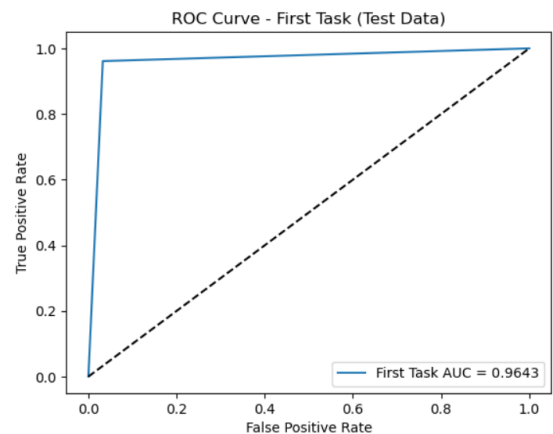


Fig. 7. ROC plot for news vs. columns classification of unseen dataset.

categorized into six domains, the model's overall accuracy dropped to 75% (Table V), indicating challenges in accurately assigning articles to their respective categories. A detailed analysis of classification performance across domains (Fig. 6b)

reveals that the Sports category achieved the highest F1-score at 98%, followed by Politics with an F1-score of 86%, correctly classifying 761 articles. Meanwhile, the Economy, Local, and Technology categories maintained moderate classification performance, each exceeding an F1-score of 70%, with consistently high numbers of correctly classified articles across all three domains.

In contrast, the Health domain exhibited poor performance, achieving an F1 score of only 29%. A significant number of Health articles were misclassified and only 178 out of 1,000 Health articles were correctly classified. Similarly, the Technology category showed a recall reduction to 58%, with an F1-score of 73%. Of the 1,000 Technology articles, 576 were correctly classified, while 336 were misclassified as Economy and 73 as Local. These results suggest that while the model performed well in distinguishing some categories, particularly Sports and Politics, further refinement is necessary to improve classification accuracy in domains such as Health and Technology.

Overall, the drop in classification accuracy to 75% can be attributed to increased confusion among semantically related domains. Examination of the confusion matrix revealed that misclassifications were most frequent in the Health and Technology categories, where overlapping terminology with domains such as Local and Economy led to reduced accuracy. These findings underscore the substantial impact of lexical similarity on the model's discriminative capability and highlight the necessity of further refining domain-specific feature representations to enhance generalization performance.

## VI. MODEL PERFORMANCE ANALYSIS

This section analyses the performance of each model, examining cases where news and columns were classified correctly (true positives) at the word level and cases where they were misclassified (false negatives) at the class level. Both analyses provided valuable insights into the problem of Arabic news classification and columns.

### A. Word-Level Analysis

To get a better perspective of the model's performance at the word-level, we further examine some of the correctly classified and misclassified articles, presented in Table VI. To further inspect the correctly classified articles, we employ an attention mechanism technique for this task, as they are components of neural networks designed to address limitations in encoder-decoder models [26]. From that, attention maps visually represent the attention process in models, particularly in transformer-based architectures. They illustrate the extent of 'attention' the model allocates to various segments of an input sequence during prediction. They mainly operate by calculating a weighted total of all elements in an input sequence, where each weight (or attention score) indicates the significance of one element relative to another. The attention map visualizes these weights in a matrix, where:

- Rows represent the amount of attention the tokens (tokenized words symbolized by #) give to other tokens.
- Columns represent how much attention the tokens receive from other tokens.

Regarding the first correctly classified article presented in the table, the column discussed an upcoming theatrical event by introducing the topic through an emotional narration of a famous story about an Arabian poet. Fig. 8 illustrates an example of an attention map for the column. The most relevant portions of the tokenized text are represented in the attention map, where the intersection between rows and columns—highlighted in yellow—indicates the attention values assigned by the token in the column to the token in the row. Notably, the token indicated by the orange arrow, corresponding to the word (البلّيج) 'the eloquent,' assigns greater attention to tokens in the row that include possessive pronouns, such as (صديقه) 'his friend' and (لهفي) 'my ambition,' as well as words associated with emotions, such as (بالحب) 'in love.'

The word (البلّيج) 'the eloquent' is commonly used in descriptive and literary contexts, emphasizing expressive writing, while (بالحب) 'in love' is intrinsically associated with emotion. The presence of these words within the column article suggests that such linguistic choices contribute to an opinion-based and subjective writing tone. This finding aligns with the nature of column articles, where writers frequently employ personal expressions and emotions, and thus use possessive pronouns to convey their perspectives. In contrast, news articles focus on factual reporting, hence enforcing an objective tone and minimizing the use of subjective linguistics.

To further substantiate this observation, we analyzed the Part of Speech (POS) distribution across all articles in the dataset using the Farasa Arabic tool [27]. The results revealed that possessive pronouns appeared in 61.7% of the column articles, compared to 43.2% in the news articles. This statistical difference might indicate that possessive pronouns serve as linguistic markers of column writing. These markers were not only identified in the attention maps but also exhibited a higher frequency of occurrence in column articles, reaffirming their significance in distinguishing columns from news articles.

On the other hand, Fig. 9 presents the attention map for the technology domain news article correctly classified, shown in Table VI. The article was a news report discussing services provided by Windows operating systems. Referring to the figure, the visual token intersections occurred in a specific region, where the token in the column, marked by the orange arrow, corresponds to the tokenized word (السريع) 'fast', while the tokens in the row include (ويندوز) 'Windows', (شاشات) 'screens', and (التطبيق) 'application', all of which are words associated with the technology domain. The model identified a semantic correlation between the term 'fast' and these technology-related words, mapping the concept of speed as a key characteristic in technological advancements. As a result, the model correctly classified the article under the 'technology' domain. This finding highlights that articles containing a high frequency of domain-specific terms are more likely to be accurately categorized.

### B. Class-Level Analysis

Regarding the classification of articles into their respective domains, the SHapley Additive exPlanations (SHAP) approach was applied to identify the most influential features contributing to the hierarchical classification model's predictions

TABLE VI. EXCERPT SAMPLES OF THE CORRECTLY AND INCORRECTLY CLASSIFIED ARTICLES.

| Case                   | Article Excerpt<br>(Translated/Truncated)                                                                                                                 | True<br>Label | Predicted<br>Label | Possible Reason for<br>Classification                                                                                              |
|------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|--------------------|------------------------------------------------------------------------------------------------------------------------------------|
| Correctly Classified   | لهفي على دعد و ما حفلت إلا بالبليغه<br>الحسناء...بالحب و الصدق...<br>The leather has adorned the beauty..her<br>love..                                    | Column        | Column             | The article presents a poetic love story characterized by highly subjective language and several non-formal linguistic expressions |
| Correctly Classified   | تواصل شركة مايكروسوفت تحسين نظام<br>التشغيل ... وشاشات اللمس الحديثة ...<br>Microsoft continues to enhance its Win-<br>dows... with touch screens that... | Technology    | Technology         | The article contains technology terminologies that fit the technology domain.                                                      |
| Incorrectly Classified | رغم العجز المالي ايران تزيد ميزانيتها ...<br>Despite the financial deficit, Iran in-<br>creases its budget...                                             | Politics      | Economy            | The article discusses political decisions affecting financial markets, resulting in semantic overlap between Politics and Economy. |

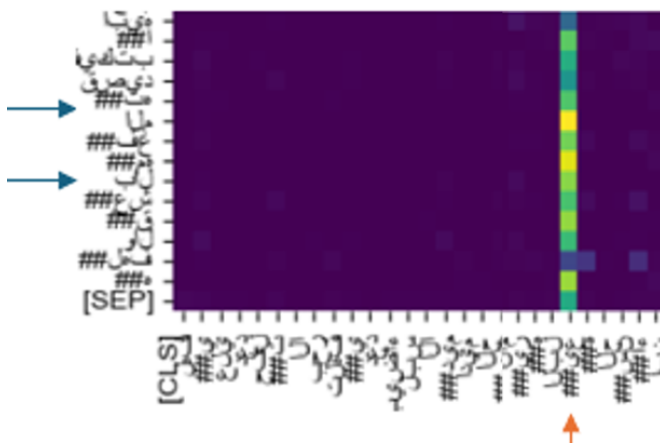


Fig. 8. Attention map of a correctly classified column article, illustrating token-level attention patterns associated with subjective writing.

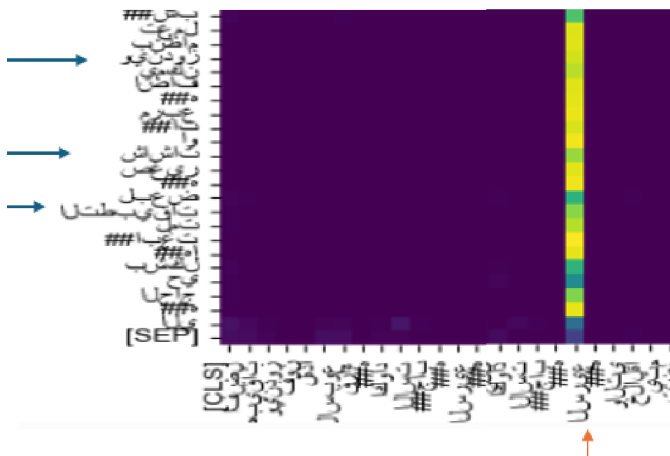


Fig. 9. Attention map of a correctly classified Technology-domain news article, illustrating contextual attention patterns among technology-related terms.

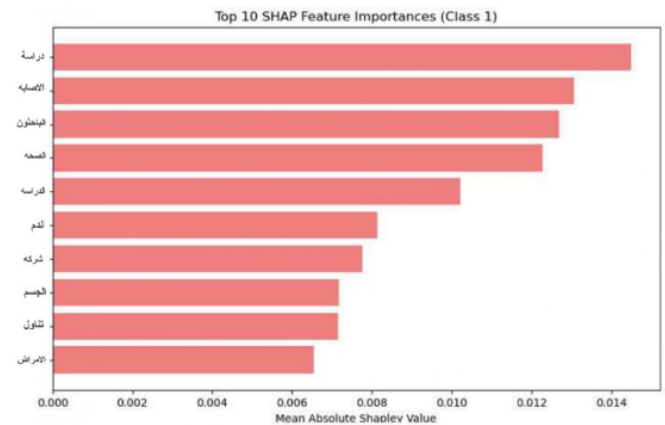


Fig. 10. Top 10 most influential features for the Health category, ranked by their mean absolute SHAP values.

across classes. Misclassifications primarily occurred when news articles contained a high proportion of terms typically not associated with other domains.

For instance, a news article categorized as political that discussed the impact of COVID-19 on government sectors in China was misclassified as health due to the high frequency of medical-related terms such as ‘COVID-19’ (كورونا), vaccine (لقاح), immune (مناعة), and disease (مرض), shown in Table VI. This finding is further supported by Fig. 10, which highlights the most influential words (features) commonly found in articles within the health domain. The figure prominently displays frequently occurring terms such as researchers (الباحثون), infected (الإصابة), health (الصحة), and diseases (الأمراض), all of which are strongly associated with the health category. Consequently, the substantial presence of domain-specific terminology plays a crucial role in determining an article’s classification, potentially leading to misclassifications when such terms appear in articles from different domains.

Moreover, for the first task, misclassified news articles were manually inspected. About 6.3% of articles misclassified as columns contained quotations, often with possessive pronouns.

For example, a sports news article reporting on a soccer match incorporated multiple quotes from fans and sports analysts as supporting details. These quotations used possessive pronouns, giving them a column-like subjective tone. These articles were likely misclassified due to linguistic overlap. See Table VI for examples of misclassified articles and an investigation of possible causes.

## VII. DISCUSSION

The results and analysis showed that the proposed model and binary and multi-class classification optimizations were effective. The first task, distinguishing news articles from opinion columns, was completed with 95% accuracy. This is due to linguistic variances between both texts. News articles are factual and objective, while opinion articles are more subjective and evaluative. These style distinctions help AraBERT grasp the contextual patterns that distinguish the two classes. A column class of 98% precision indicates the model excels at opinion-based writing, which is more subjective than factual news articles.

Results of the second task, the news classification into specific domains, were likewise impressive, with an accuracy of 94%. This strong performance is mostly due to the presence of domain-specific vocabulary in many news categories. Sports and technology articles use specialist phrases to assist the model in distinguishing these categories. Politics and Economics were challenging. The Politics class had a low recall rate (83%), and some articles were misclassified as belonging to other classes. The Economy class suffered similar problems, with articles being misclassified as Politics. This semantic overlap between these domains can be explained as articles on government policies, financial laws, budgets, and economic changes use similar vocabulary and contextual expressions. This blurs the distinctions between these categories, making classification tougher.

Similarly, the proposed model attained 96% accuracy on the unobserved dataset in the first task. However, its performance in the second task (classifying six domains) dropped to 75%. Performance and error analysis indicated that opinion columns were correctly identified using subjective and expressive language, while news articles were classified using strong domain-specific terms. These results show that semantic similarities between related domains caused the model's flaws rather than textual representation learning limitations. The results demonstrate that the proposed model can classify Arabic news using stylistic and semantic features.

## VIII. LIMITATIONS AND FUTURE WORK

This research demonstrated the efficacy of a two-task hierarchical model utilizing AraBERT for classifying news articles and columns, in addition to classifying news into appropriate categories. Nonetheless, some constraints must be recognized. Initially, while the collection encompasses articles from several Arabic news outlets, it predominantly features Modern Standard Arabic. Consequently, the proposed model has yet to be assessed on articles featuring several Arabic dialects, which might influence its efficacy in more heterogeneous linguistic environments. Secondly, the model was confined to six news categories; thus, its capacity to generalize to alternative news

domains remains unexamined. Thirdly, the model underwent assessment using the gathered dataset and another unseen dataset; however, its ability to generalize across other writing styles necessitates more scrutiny. Nonetheless, there exist other pathways for additional work. Initially, expanding the dataset to encompass a wider array of Arabic news articles, various Arabic dialects, and supplementary news sectors might enhance the model's generalization. Second, more studies should investigate several transformer-based models and evaluate their efficacy against AraBERT, including modern Arabic models such as RoBERTa or MARBERT. Ultimately, the integration of semantic and discourse-level analysis might enhance the model's ability to distinguish between objective reporting and subjective writing. These enhancements would help construct more contextually aware Arabic media categorization frameworks.

## IX. CONCLUSION

This research introduced a hierarchical classification framework for Arabic news articles and columns using the pre-trained AraBERT model. A two-task classification algorithm was used to identify news and column items and categorize news articles into six domains to fill a gap in Arabic computational linguistics. The approach performed well, with an accuracy of 95% for the first task and 94% for the second task on the main dataset. Curating 7,367 Arabic opinion columns from various Arabic media channels is also significant. This dataset will aid Arabic NLP research. The study includes a comprehensive model analysis incorporating word-level attention and class-level feature importance. This technique provides linguistic insights into text classification challenges and feature importance. The findings showed that classification accuracy requires domain-specific terminology. News stories using mixed-domain vocabulary and subjective phrases, similar to opinion pieces, were commonly misclassified. Finally, the model performed well with an unseen dataset, distinguishing news from columns with 96% accuracy. However, news domain classification accuracy declined to 75%, indicating difficulties in generalizing across real-world data. Despite promising results, distinguishing columns from news articles is difficult, especially when columns take on a journalistic tone. Advanced linguistic and semantic analysis techniques like discourse markers and argumentation frameworks should be investigated to improve classification. Incorporating multi-modal data sources like author metadata or writing style could boost the model's robustness.

## REFERENCES

- [1] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," *Mining text data*, pp. 163–222, 2012.
- [2] Merriam-Webster Collegiate Dictionary, "News," 2024, retrieved January 4, 2020, from <https://www.merriam-webster.com/dictionary/news>.
- [3] J. Fuller, *News Values: Ideas for an Information Age*. Chicago, IL: University of Chicago Press, 1996.
- [4] A. Scherlen, "Part i: Columns and blogs: Making sense of merging worlds," *The Serials Librarian*, vol. 54, no. 1-2, pp. 79–92, 2008.
- [5] The Media Insight Project, "How americans describe their news consumption behaviors," June 11 2018, retrieved September 9, 2024, from <https://www.americanpressinstitute.org/publications/reports/survey-research/americans-news-consumption/>.

- [6] M. Islam and S. S. Anas, "News classification using ml algorithms: An exploration of efficient information categorization," *IJCRT*, vol. 11, no. 6, 2023.
- [7] P. Saigal and V. Khanna, "Multi-category news classification using support vector machine based classifiers," *SN Applied Sciences*, vol. 2, no. 3, p. 458, 2020.
- [8] K. S. Nugroho, A. Y. Sukmadewa, and N. Yudistira, "Large-scale news classification using bert language model: Spark nlp approach," in *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology*, 2021, pp. 240–246.
- [9] A. R. Lubis, S. Prayudani, Y. Fatmi, and O. Nugroho, "Classifying news based on indonesian news using lightgbm," in *2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*. IEEE, 2022, pp. 162–166.
- [10] T. Kathirvalavakumar, "Two dimensional feature extraction and blog classification using artificial neural network," *International Journal of Applied Engineering Research*, vol. 13, no. 9, pp. 6536–6544, 2018.
- [11] R. Jannati, R. Mahendra, C. W. Wardhana, and M. Adriani, "Stance classification towards political figures on blog writing," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 96–101.
- [12] Y. Du, Y. Yi, X. Li, X. Chen, Y. Fan, and F. Su, "Extracting and tracking hot topics of micro-blogs based on improved latent dirichlet allocation," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103279, 2020.
- [13] H. Hassani, C. Beneki, S. Unger, M. Mazinani, and M. Yeganegi, "Text mining in big data analytics. big data and cognitive computing, 4 (1), 1," 2020.
- [14] L. Escoufflaire, A. Descampe, and C. Fairon, "Automated text classification of opinion vs. news french press articles. a comparison of transformer and feature-based approaches," *Language & Communication*, vol. 99, pp. 129–140, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0271530924000624>
- [15] J. Imad, A. Rachid, and B. Mohamed, "Automated arabic news classification using the convolutional neural network," *International Journal on Electrical Engineering & Informatics*, vol. 15, no. 2, 2023.
- [16] R. Alroobaea, "An empirical deep learning approach for arabic news classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.
- [17] K. Garcia, P. Shiguihara, and L. Berton, "Breaking news: Unveiling a new dataset for portuguese news classification and comparative analysis of approaches," *Plos one*, vol. 19, no. 1, p. e0296929, 2024.
- [18] M. R. I. Rifat and A. A. Imran, "Incorporating transformer models for sentiment analysis and news classification in khmer," in *Computational Data and Social Networks: 10th International Conference, CSoNet 2021, Virtual Event, November 15–17, 2021, Proceedings 10*. Springer, 2021, pp. 106–117.
- [19] M. Altamimi and A. M. Alayba, "Anad: Arabic news article dataset," *Data in Brief*, vol. 50, p. 109460, 2023.
- [20] H. T. Himdi and F. Y. Assiri, "Tasaheel: An arabic automotive textual analysis tool—all in one," *IEEE Access*, vol. 11, pp. 139 979–139 992, 2023.
- [21] F. zahra El-Alami, S. Ouatik El Alaoui, and N. En Nahnah, "Contextual semantic embeddings based on fine-tuned arabert model for arabic text multi-class categorization," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, Part A, pp. 8422–8428, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1319157821000446>
- [22] K. Alzamel and M. Alajmi, "An application of textual document classification for arabic governmental correspondence," *Kuwait Journal of Science*, vol. 52, no. 1, p. 100299, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S230741082400124X>
- [23] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.
- [24] Hugging Face, "Hugging face [computer software]," 2024, retrieved September 3, 2024, from <https://huggingface.co/>.
- [25] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [26] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [27] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for arabic," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, 2016, pp. 11–16.