International Journal of Natural-Applied Sciences and ENgineering
(IJNASEN)
Vol.4 No.1 (2026) pp. 6–22

Research Article

A Hybrid BERTopic–NAMAA Framework for Arabic News Topic Modeling, Classification, and Evaluation

Noor S. Dawood[1]*, Salma A. Mahmood[2]

[1]Department of Computer Science, collage of Computer Science and Information Technology, Basrah University, Basrah, Iraq.
* **Corresponding Author Email:** noor.salman@uobasrah.edu.iq - **ORCID:** 0000-0002-5247-7350

[2]Department of Computer Information Systems, collage of Computer Science and Information Technology, Basrah University, Basrah, Iraq.
Email: salma.mahmood@uobasrah.edu.iq - **ORCID:** 0000-0002-5247-7450

Abstract: Topic detection is a powerful technique for uncovering the latent semantic structures in large-scale unlabeled text. Yet, the application to Arabic is still an open challenge due to the complex morphology of the language and the absence of comprehensive solution that extends topic discovery by systematic evaluation. In this paper, a novel hybrid architecture based on BERTopic (for unsupervised topic detection) in conjunction with NAMAA (an Arabic classification technique) for topic category assignment and external evaluation is presented. The suggested architecture involves an Arabic-specific preprocessing layer, a contextual document embedding through four Arabic pre-trained transformer models (AraBERTv0.2, AraBERTv2, Asafaya, and QARiB), a topic extraction layer by utilizing the BERTopic model, an automatic topic category assignment layer via NAMAA, and an evaluation step for the entire architecture based on both intrinsic (NPMI, C_v coherence, and Topic Diversity) and extrinsic (Accuracy and Cohen's Kappa) metrics. The proposed framework was tested on two Arabic news related Economy and Sports domains: an imbalanced dataset (D1) of 7,212 documents and a balanced dataset (D2) of 3,158 documents. BERTopic yielded between 59 and 116 topics on D1 and between 62 and 125 topics on D2. On the balanced dataset Asafaya scored the best C_v coherence (0.579) and Cohen Kappa (0.937), while AraBERTv2 ranked the best NPMI (0.289) and Topic Diversity (0.702). The best classification accuracy with Asafaya on D2 was 96.8%, while 98.2 with AraBERTv2 on D1. To sum up, these results indicate that our framework delivers a robust and efficient solution for Arabic news topic discovery and evaluation, revealing also the impact of embedding model choice and class distribution on topic quality, classification accuracy and agreement with the original dataset labels.

Keywords: Arabic Natural Language Processing; BERTopic; Topic Modeling; Arabic Text Classification; Arabic News; AraBERT; NAMAA; Cohen's Kappa.

Received: *03 January 2026* **|** ***Revised:*** *22 April 2026* **|** ***Accepted:*** *27 April 2026* **|** ***DOI:*** *10.22399/ijnasen.38*

1. Introduction

The fast pace of growth in Arabic digital content and especially in online news has led to a growing need for automated techniques that can help in identifying, organizing, and making sense of large volumes of unstructured text [1]. The substantial growth in the size and heterogeneity of textual data in the last twenty years has rendered manual analysis infeasible, creating a demand for approaches that can discover meaningful thematic structures while at the same time preserving the semantic content of documents [2]. Natural Language Processing (NLP) has recently made enormous strides, allowing for better automated understanding of text such as that found in documents. Traditional ways of representing lexical semantics like BoW [3] and TF-IDF [4] are now being supplemented by probabilistic topic models such as Latent Dirichlet Allocation (LDA) [5] and more recently by Transformer-based language models led by BERT, which supplies deeper contextual representations and has

attained the best results on attested data in various NLP tasks [6], [7]. However, Arabic still suffers from being one of the hardest languages for natural language processing due to its complex root & pattern morphology, rich derivational system, optional diacritization, and being combined with Modern Standard Arabic (MSA) and a bunch of local dialects [8]. These linguistic properties constrain the straightforward adoption of techniques initially designed for other languages and necessitate Angra-specific conceptual frameworks. Topic modeling and text classification aim for similar objectives in the analysis of large text corpora. Unsupervised topic modeling can uncover latent semantic structure of a document collection without the accessibility of labeled data [9], and supervised text classification assigns a given free-text input into one of a predefined set of categories by learned semantic pattern [10]. Combining these two strategies might allow one to integrate the exploratory nature of topic modeling with the interpretability and external validation of supervised classification, thus not only discovering "good" topics, but also systematically categorizing and assessing those topics. Driven by this complementarity, in the paper we introduce a novel BERTopic based hybrid model for analyzing Arabic news articles in terms of topic [11] (a Transformer-based topic modeling method) and NAMAA [12], a pre-trained model used for Arabic text classification. BERTopic is utilized to reveal hidden topics in the Arabic news articles and NAMAA is a pre-trained classifier used to directly classify the topic representations extracted without application-specific training. Rather, this process evaluates the extent to which the contextualised document embedding, that is the representations used by BERTopic, contain enough discriminative semantic information to allow an unrelated general-purpose classifier to predict the labels of the original datasets. Hence, the proposed method provides an extrinsic evaluation for the whole topic discovery and classification process, to be viewed as complementary, not replacing the typical intrinsic measures of topic-quality. NAMAA was chosen as it is a modern Arabic classification model for topic categorization that can offer wide topical coverage with no need for extra task-based fine-tuning. Its application as a standalone classifier allows for a uniform evaluation scheme over each embedding model, and it limits the effect of supervised optimization to the training dataset on each dataset to a minimum. In this sense, the proposed framework, which is based on unsupervised topic learning and independent topic classification for a variety of tasks, aims at capturing the whole aspect of Arabic news in a more effective manner compared to approaches that are purely based on intrinsic topic models or independent document clustering.

1.1 Research Gaps and Objectives

A review of the effect of work for the Arabic topic modeling and text classification (Section 2) reveals three major challenges. First, while topic modeling and text classification have been well studied, there are very few works that integrate unsupervised topic discovery and supervised topic classification into one framework on analyzing Arabic news. Second, the evaluation of the quality of Arabic topic models at topic level has been based on intrinsic measures of topic quality (e.g. topic coherence) while some research was conducted on extrinsic evaluation with reference labels provided externally. Third, class imbalance is a common characteristic of real-world Arabic text datasets, however, its impact on the combined performance of topic modelling and topic classification using has not been sufficiently investigated, to the best of author's knowledge.This paper contributes by introducing a hybrid model which makes use of BERTopic for unsupervised topic modeling and NAMAA as a standalone pre-trained Arabic classifier (i.e., with no task-specific fine-tuning). 4 The framework uses Cohen's Kappa to measure the agreement between the topic categories assigned by the automatic methods and the labels of the original datasets and tests the class distribution effect on both topic quality and classification performance on two datasets configurations- an imbalanced (D1) and a balanced dataset (D2).

The research questions are as follows:

• To what extent can the hybrid model enable effective topic discovery, classification and evaluation in Arabic news articles?
• Which semantic topics are extracted by BERTopic from these datasets with regard to the topics?
• What is the precision of NAMAA in classifying the discovered topics into Economy and Sports?
• How well do the automatically predicted topic categories match the original labels of the dataset?
• What is the performance of the identified topics with respect to standard intrinsic quality measures such as NPMI and C_v coherence and one extrinsic measure, Topic Diversity?

• What is the effect of class distribution on topic quality, classification performance, and agreement between the true and original dataset labels?

Therefore, this study has two purposes: to propose an integrated solution for Arabic news topic discovery, classification, and evaluation including BERTopic and NAMAA and to compare and evaluate the proposed solution with others in terms of solution quality and execution time. In particular, it investigates [whether] a fusion of unsupervised topic modeling and an independent pre- trained Arabic classifier can constitute a more complete evaluation framework as opposed to intrinsic topic-quality measures.

The main contributions of this work are fourfold:

1. A hybrid methodology that combines both BERTopic and NAMAA for the retrieval of topics and topic classification in the case of Arabic news.
2. An extrinsic evaluation schema based on Cohen's Kappa to measure how well the topics agree in its labels with the labels of the original dataset, which complements the intrinsic topic-quality measures.
3. A thorough experimental evaluation of the proposed framework in both balanced and imbalanced class scenarios.
4. A complete, fully documented experimental methodology that enables openness and repeatability for future comparative work in Arabic topic modeling.

2. Related Work

Although the two fields of application have grown in tandem in the result about modelling topics in texts in Arabic and classifying texts in Arabic, the two have developed as somewhat separate lines of research. Topic modeling attempts to reveal the latent semantic structure of document collections without need for labeled data, while text classification is concerned with classifying textual documents into a set of predefined categories based on the supervised learning. While both techniques are utilized in performing analysis of large-scale Arabic text, they have been barely combined within one framework, a gap that the present work aims to fill. This review subsequently provides a critical analysis of the previous work in these two areas, and highlights ac knowledge gaps filled with the current work.

2.1 Arabic Topic Modeling

Recent works show that BERTopic and contextual language models are also promising for Arabic topic discovery. Abuzayed and Al-Khalifa (2021) conducted a comparison between BERTopic and LDA and NMF on a 5-category Arabic news dataset with QARiB embeddings and reported a best NPMI score of 0.17 and concluded that monolingual Arabic embeddings are better than the multilingual versions for topic discovery [13]. Similar, Abdelrazek et al. (2022) tested 6 tm approaches using RoBERTa embeddings on a general Arabic corpus and resulted in a top C_v coherence score of 0.1147 [14]. In a more recent work, Aouichaty et al. (2024) utilized BERTopic with Moroccan Arabic legal texts where the NPMI varied from 0.11 to 0.17 [15]. These results show that contextual embeddings lead to significant improvements in the quality of Arabic topics, as was the case in English. However, the assessment of them is largely restricted to intrinsic topic-quality measures, e.g., NPMI and C_v coherence. None of these works systematically assesses whether the identified topics can be robustly linked to predefined semantic categories via an isolated classifier, or studies the impact of class imbalance on topic discovery.

2.2 Arabic Text Classification

Similar results have been attained in Arabic text classification via Transformer-based language models. Ben Ali et al. (2022) have adapted AraBERT for the task of Arabic fake-news detection and achieved an accuracy of 98.0% [16]. Nassif et al. (2022) achieved comparable high classification results with ARBERT-based contextual embeddings [17], [18]. Karajeh et al. (2023) is 97.25% and F1-score is 97.26% using E2E supervised training [19] to the best of our knowledge, the proposed approach is the finest for the seven-category SANAD dataset, based

on these two parameters. Similarly, Al- Taie (2025) presented a hybrid AraBERT–BiLSTM framework for detection of Arabic fake-news with an accuracy of 98.4% along with F1-score of 98.0% [20]. While these works confirm the efficiency of supervised Arabic text categorization, they consider classification as end- to end prediction task on the document texts. Therefore, they do not make use of the latent topic structures uncovered by unsupervised topic modeling nor consider the extent to which topic discovery may assist topic understanding or external assessment.

2.3 Critical Comparison and Research Gap

Table 1 summarizes the principal characteristics of the reviewed studies and compares them with the proposed framework.

Table 1. *Comparative summary of previous studies and the present work*

Study (Year)	Domain	Method	Model-Usage Strategy	Metrics	Gap
Abuzayed & Al-Khalifa (2021)	Arabic news (5 categories)	BERTopic + QARiB vs. LDA & NMF	Pre-trained embeddings only — no classifier	NPMI	No downstream topic classification; internal metrics only
Ben Ali et al. (2022)	Arabic fake news (social media)	AraBERT (direct fine-tuning)	End-to-end fine-tuning on task data	Accuracy	Supervised classification only; no topic modeling
Nassif et al. (2022)	Arabic fake-news detection	ARBERT + other contextual embeddings	End-to-end fine-tuning on task data	Accuracy	No topic modeling; class imbalance not addressed
Abdelrazek et al. (2022)	General Arabic text corpus	BERTopic + RoBERTa (6-method comparison)	Pre-trained embeddings only — no classifier	C_v	Internal metrics only; no evaluation against human labels
Karajeh et al. (2023)	Arabic news, SANAD (7 categories)	GCN + AraBERT (hybrid)	End-to-end fine-tuning on task data	Accuracy	Direct classification; no topic modeling stage
Aouichaty et al. (2024)	Moroccan Arabic legal texts	BERTopic + Arabic BERT (monolingual)	Pre-trained embeddings only — no classifier	NPMI	No external evaluation; no link to predefined categories
Al-Taie (2025)	Arabic fake news	AraBERT + BiLSTM (hybrid)	End-to-end fine-tuning on task data	Accuracy, F1	Direct classification; no topic modeling integration
Present study (2025)	Arabic news (Sports & Economy)	BERTopic + LLM embeddings + NAMAA (hybrid)	Pre-trained embeddings + no fine-tuning	NPMI, C_v, Accuracy, Kappa	Integrated hybrid framework with external evaluation via Cohen's Kappa

In general, the survey literature reveals that the Arabic topic modelling and the Arabic text classification have been treated independently as two separate lines of research. Existing topic-modeling research is dominated by studies which rely on intrinsic evaluation, while text classification work focuses on end-to-end supervised prediction without an explicit topic-modeling stage. In addition, evaluation of the discovered topics with respect to the ground truth dataset labels using an independent pre-trained classifier has not been considered, and the effect of class distribution on the joint performance of topic modeling and classification is under-explored. The above limitations can be handled with the integration of BERTopic and the pre-trained NAMAA classifier in one pipeline with the proposed framework. Unlike prior work, the proposed method integrates (unsupervised) topic discovery with (independent) topic classification, augment intrinsic topic-quality measures with extrinsic evaluation using Cohen's Kappa, and analyzes the effect of balanced vs. unbalanced class distributions on the framework's performance systematically.

3. Research Methodology

This is the methodology used to build and test the hybrid framework for the analysis of the topic of the Arabic news. The process comprises five consecutive steps: data gathering and preprocessing, unsupervised topic detection with BERTopic, intrinsic evaluation of topic quality, topic category assignment by the pre-trained

NAMAA model, and extrinsic evaluation of performance. These stages collectively make up a comprehensive and repeatable process for identifying, categorising and assessing topics derived from Arabic news content.

3.1 Research Design and Experimental Environment

This research is a developmental and experimental study with a quantitative method of evaluation. The goal is to design and test a hybrid approach that combines unsupervised topic modeling with independent topic classification for the analysis of Arabic news. The architecture is implemented in Python 3.12, with topic modeling performed using BERTopic, loading the arabic transformer models using Hugging Face Transformers library, evaluating performance using Scikit-learn and other scientific libraries (Pandas, NumPy, Matplotlib, Gensim and NLTK) for support. All the experiments were performed in Google Colab with NVIDIA A100 GPU acceleration to speed up the generation of contextual document embeddings and run the BERTopic pipeline on big Arabic news corpora. The experimental setup provided computational efficiency and uniform implementation for all embedding models studied herein. All the experiments were performed in Google Colab with NVIDIA A100 GPU acceleration to speed up the generation of contextual document embeddings and run the BERTopic pipeline on big Arabic news corpora. The experimental setup provided computational efficiency and uniform implementation for all embedding models studied herein. The proposed framework consists of five consecutive phases:

- Data collection and preprocessing, including document loading, text normalization, and Arabic-specific preprocessing.
- Unsupervised topic modeling, where contextual document embeddings are generated, followed by dimensionality reduction and document clustering using BERTopic.
- Intrinsic topic evaluation, in which the discovered topics are assessed using NPMI, C_v coherence, and Topic Diversity.
- Topic category assignment, where the discovered topic representations are classified into predefined categories using the pre-trained NAMAA model without task-specific fine-tuning.
- Extrinsic performance evaluation, where the automatically assigned topic categories are compared with the original dataset labels using Accuracy, Cohen's Kappa, confusion matrices, and error analysis.

3.2 Dataset

The experiments are performed on two public Arabic news corpora: the Khaleej-2004 dataset, which consists of news articles in MSA collected from the Akhbar Al-Khaleej newspaper [21], and SANAD-SUB, a single-label sub-dataset of the Single-label Arabic News Articles Dataset designed for Arabic text classification research [22]. From both corpora, we select the Economy and Sports articles as these two cohesive and orthogonal news domains are well-populated in both corpora, and thus represent suitable subsets for assessing topic discovery and topic classification. The original categories in the dataset were kept during all the experiments and used as the reference labels in the extrinsic evaluation. All articles were saved as html files and were grouped according to their pre-defined category tags to aid processing and performance evaluation. In order to study the effect of class distribution on the proposed framework, two versions of the experimental dataset were formed. The first setting (D1) maintains the original class distribution in the collected corpus, and hence, it corresponds to a realistic imbalanced situation. The second setting (D2) was sampled by choosing equal number of articles from each category at random, thus resulting in a balanced dataset with same class size. D2 is created by random undersampling majority class with a fixed random seed for ensuring experimental reproducibility. Enabling both configurations allows the proposed frame- work to be tested in the wild as well as in balanced experimental settings. Table 2 shows an overview of the properties of the two dataset configurations.

3.3 Data Preprocessing

A proper pre-processing pipeline is required for the text to be augmented and to have a reliable input to the proposed hybrid scheme. The intention of the preprocessing process is to reduce the noise present in the text, to normalize different orthographic forms and to enhance the quality of the contextual document embeddings which

are subsequently employed in topic modeling and classification. All the documents were loaded using a custom multi-category data loader which iteratively walked the category paths on the file system, matched files by

Table 2. *Dataset configurations and class distribution*

Dataset	Total Articles	Sports	Economy	Imbalance Level
D1	7,212	5,633	1,579	Highly imbalanced
D2	3,158	1,579	1,579	Balanced

equating them to a recursive glob pattern, and processed them in parallel using a thread pool to maximize CPU usage and to hasten the processing on big corpora. The text of documents were discarded if it contained less than five words in order to remove texts with very little semantic information. The processed files were then passed through a series of Arabic-driven preprocessing steps. The process started with the removal of html tags and the extraction of the text content using beautifulsoup while a cleaning step based on regular expressions was applied to eliminate urls, numbers and non-textual characters. Secondly, Tashkeel (diacritics) fatha, damma, kasra and other marks were stripped away by strip_tashkeel method of pyarabic to minimize the lexical diversity while preserve the semantic contents. Third, Orthographic normalization was introduced to decrease vocabulary sparsity by unifying different forms of letters. In particular all these alif variants (□ □□ □□) were converted to the the bare form (□) and the ta marbuta (□) was converted to ha (□). Fourth, we removed Arabic stop words by means of the NLTK default list, enhanced with a very few words the most frequent in the corpus that were not contributing to the differentiation of topics (e.g., "□□□□□"، "□□□□"، "□□□", and "□□□□□"). Fifth, to enhance the coherence of text, sequences of whitespace were replaced by a single space. In the end, a quality-control step discarded documents that had been rendered empty by preprocessing, thereby guaranteeing that only the informative texts were kept for further examination. The developed preprocessing pathway was applied to all documents and that pipeline was a lexical variation reduced, irrelevant textual artifacts removed and standard document representations, with which contextual embedding models were applied, were produced to generate contextual embeddings for topic modeling through BERTopic and to assign topic category using NAMAA model.

3.4 Unsupervised Topic Detection with BERTopic

BERTopic is a Transformer-based topic modeling approach that leverages contextualized document embeddings and class-based TF-IDF to create semantically meaningful topic representations [11]. In contrast to traditional probabilistic topic models like Latent Dirichlet Allocation (LDA), BERTopic makes use of contextual embeddings, which are obtained from pre-trained Transformer models, to capture the semantic similarity among documents prior to the topic formation. The methodology includes four key phases: (1) contextual document embedding, (2) dimensionality reduction via Uniform Manifold Approximation and Projection (UMAP) [23], (3) density-based document clustering through Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [24], and (4) topic representation via c-TF-IDF and selection of the most representative keywords for each topic. To study the effect of embedding models on the quality of Arabic topic modeling, we considered four Arabic pre-trained Transformer models to serve as the embedding backbone: AraBERT v0.2 and AraBERTv2 [25], trained on Arabic news and Wikipedia useing Farasa based morphological segmentation-enabled; Asafaya (BERT-base-Arabic) [26] pre-trained on approximately 8.2 billion Arabic words from the OSCAR corpus and Arabic Wikipedia; and QARiB [27], trained on a mix of MSA and dialectal Arabic using data from Twitter, Gigaword, and OPUS, with Byte-Pair Encoding tokenization. These four models all follow the BERT-base architecture which consists of 12 Transformer layers, 12 attention heads, a hidden size of 768, with parameter counts ranging from approximately 110 million (Asafaya and QARiB) to approximately 136 million (AraBERT v0.2 and AraBERTv2), owing to differences in vocabulary size. Their major differences lie in the training corpora, vocabulary size, and preprocessing strategies, which may influence their effectiveness for Arabic news topic modeling.The normal batch size for document embeddings production was 64 for each embedding model, and they were normed to unit length prior to clustering in order to keep cosine-distance relations. GPU

acceleration were also used when running on embedding generation process to speed up computations. The BERTopic pipeline was set with the same group of hyperparameters in all the experiments. UMAP was then used for dimensionality reduction with n_neighbors = 10, n_components = 5, min_dist = 0.1 and random_state = 42 to achieve a good balance between local and global document structure and to ensure reproducibility. HDBSCAN is then run with min_cluster_size = 5, min_samples = 2, cluster_selection_epsilon = 0.2 and prediction_data = True, allowing for probability based topic assignment. Unclustered: Documents that were not assigned to any cluster with enough confidence were treated as outliers (Topic -1) and ignored in further topic-level analyses. Topic representations are created with c-TF-IDF by forming a term-frequency matrix based on CountVectorizer that uses unigrams, bigrams, and trigrams (ngram_range = (1, 3)). The pre-processed result was filtered by the customized Arabic stop-word list presented in Section 3.3 and words occurred less than two documents (min_df = 2) or occurred in more than 95% of the total number of documents (max_df = 0.95) were removed in order to eliminate noise and to enhance the interpretability of the topics. Eventually, BERTopic was trained through the fit_transform() method with the parameter calculate_probabilities = True to get the probabilities of topic assignment for further analysis. Tab. 3 shows the main features of the considered embedding models that might be at the basis of the performance differences. Although these architectural and pre-training properties give intuitive explanations to the experimental results, quantifying their effect requires specific ablation studies that are out of the scope of this work.

***Table* 3.** *Key pre-training factors that may influence the performance of the Arabic language models*

Model	Architecture	Training Data	Preprocessing Focus	Best At
AraBERTv0.2	BERT-base, 12 layers, ~110M params	Large Arabic news corpora	Limited Arabic-specific preprocessing	Stable, balanced baseline
AraBERTv2	BERT-base, 12 layers, ~110M params	~77M sentences: news, Wikipedia, Twitter	Farasa morphological segmentation	Classification accuracy & diversity
QARiB	BERT-base, 12 layers, ~110M params	Tweets, Gigaword, OPUS — MSA + dialects	Byte-Pair Encoding, dialect-aware	Topic diversity & coverage
Asafaya	BERT-base, 12 layers, ~110M params	8.2B words: Arabic Wikipedia + OSCAR	Large, clean general corpus	Coherence, Kappa & overall reliability

3.5 Intrinsic Evaluation of Topic Quality

The quality of the discovered topics was assessed using three complementary intrinsic evaluation metrics that measure different aspects of topic interpretability and distinctiveness.

Normalized Pointwise Mutual Information (NPMI) evaluates the semantic association among the top representative words within each topic by measuring their co-occurrence in the document collection. The metric is normalized to the interval [-1, 1], where higher values indicate stronger semantic coherence. NPMI was computed using a custom inverted-index implementation designed to efficiently calculate word co-occurrence statistics for large Arabic corpora [28]. Previous studies have shown that NPMI correlates well with human judgments of topic quality [29], [30].

Topic Coherence (C_v) assesses the semantic consistency of the representative words within a topic by combining indirect cosine similarity, normalized pointwise mutual information, and a sliding-window approach. The metric was computed using the Gensim CoherenceModel over a dictionary and bag-of-words corpus constructed from the tokenized documents. Röder, Both, and Hinneburg identified C_v as the coherence measure exhibiting the strongest correlation with human evaluation among more than 237,000 candidate metric combinations, making it one of the most widely adopted intrinsic evaluation measures in topic modeling research [31]. In Arabic NLP studies, C_v values above 0.50 are generally considered acceptable, whereas values exceeding 0.70 indicate highly coherent and interpretable topics.

Topic Diversity (TD) measures the distinctiveness of the discovered topics by calculating the proportion of unique words among the top 25 representative words across all topics, following the formulation proposed by Dieng, Ruiz, and Blei [32]. Values approaching 1 indicate that topics are characterized by distinct vocabularies with minimal overlap, whereas lower values suggest greater redundancy among topic representations. Together, NPMI,

C_v, and Topic Diversity provide complementary assessments of semantic coherence, interpretability, and diversity, enabling a comprehensive intrinsic evaluation of the topics generated by BERTopic.

3.6 Topic Classification Using NAMAA

Since BERTopic performs unsupervised topic modeling, the discovered topics are not associated with predefined semantic categories. To assign topic categories, this study employed NAMAA-Space/AraModernBert-Topic-Classifier [12], a pre-trained Arabic topic classification model developed within the Network for Advancing Modern Arabic NLP & AI (NAMAA) initiative. The model is built on the AraModernBERT architecture, which is an Arabic version of ModernBERT that has been designed using transtokenized embedding initialization and native long- context modeling. AraModernBERT is a 22 Transformer layers with 768 hidden units that uses an alternating attention mechanism (global attention every 3 layers, local attention with 128-token window), adopts ROtary Positional Embeddings, allows a maximum context length of 8,192 tokens and has about 149 million parameters. The model reached 94.32% accuracy and a 94.31% macro f1-score on SANAD [12]. In the above framework, NAMAA was used as an independent pre-trained classifier without task-specific fine-tuning. This protocol was chosen to determine if the topic representations created by BERTopic retain enough semantic information so that a separate classifier, learned on a different labeled dataset, can map the identified topics to the correct semantic groups. Therefore, the result is the performance of the proposed hybrid scheme, not the best classifier on the experimental datasets. Instead, the NAMAA tokenizer and classification model were loaded with the Hugging Face AutoTokenizer and AutoModelForSequenceClassification interfaces, and inference was done via the text-classification pipeline to ensure optimal batch performance. For each identified topic representative keywords (Top_Words) produced by BERTopic were batched to NAMAA in mini batches of 16 with an automatic truncation when input is beyond maximum supported sequence length. The most likely category along with the confidence score was retained for each topic.In order to be consistent with the experimental data sets, the 'native' output labels of NAMAA were mapped to the two target classes considered in this work. More concretely labels Finance, Business and Economy were merged into one tag named Economy while the Sports tag was left as is. All independent predictions for the rest of the categories were put into the Other category for further diagnostics. The original predictions, the mapped category, and the confidence score were saved in structured output files for the purpose of traceability and further evaluation.

3.7 Extrinsic Evaluation and Agreement Metrics

The classification stage was evaluated using four standard extrinsic metrics: Accuracy, the proportion of correct predictions among all predictions; Precision, the proportion of instances predicted positive that are truly positive; Recall, the proportion of true positive instances correctly identified; and F1-Score, the harmonic mean of Precision and Recall. These are defined respectively as

- Accuracy = (TP+TN)/(TP+TN+FP+FN),
- Precision = TP/(TP+FP),
- Recall = TP/(TP+FN), and
- F1 = 2*(Precision*Recall)/(Precision +Recall),

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives. A confusion matrix, whose diagonal entries represent correctly classified samples and whose off-diagonal entries represent misclassifications, was used to complement these aggregate scores with a class-level view of model behavior [33, 34, 35].For extrinsic evaluation, the topic-category predictions generated by NAMAA were matched with the original dataset labels associated with each topic. Performance metrics were computed using the corresponding Scikit-learn implementations. Because dataset D1 exhibits substantial class imbalance, weighted averages of Precision, Recall, and F1-score were additionally reported to provide performance estimates that appropriately reflect the underlying class distribution.To measure the extent of agreement between the automatically assigned topic categories and the labels in the original dataset after accounting for chance agreement, Cohen's Kappa (κ) was used the measure of agreement. Cohen's Kappa (κ) was also calculated through the use of the

cohen_kappa_score function of Scikit-learn, as the main extrinsic agreement measure. Unlike naive percentage of agreement which may be inflated by chance agreement, Kappa accounts for the agreement due to chance: $\kappa = (P_0 - P_e) / (1 - P_e)$, where P0 is the observed agreement and P_e is the agreement expected in the case of random allocation. Kappa is generally interpreted as follows from Landis and Koch (1977) and ranges from −1 to +1 where 0 indicates agreement at the level of chance and 1 indicates perfect agreement which can be found in Table 4 [36, 37]. In order to ensure robustness, P_0 and P_e were computed manually as a double check against the automatically calculated Kappa values and the observed agreement was also expressed in terms of how much it exceeded the chance-level agreement.

***Table* 4.** *Interpretation of Cohen's Kappa values (Landis & Koch, 1977)*

κ Value	**Strength of Agreement**
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

In addition to the numeric evaluation two diagnostic analyses were carried out. Before mapping categories, First the distribution of NAMAA's output labels was analyzed with an application focus on predictions that were placed in the Other category and at low-confidence predictions (confidence < 0.50), since these might represent ambiguous topic representations. Second, Class-specific confusion matrices (with Precision, Recall, F1-score) are analyzed to the subject of most misclassified topic categories, giving a fine-grained understanding of the main sources of disagreement in the suggested scheme.

3.8 Reproducibility, Quality Assurance, and Ethical Considerations

Several protocols were applied to make the experiments results reliable and reproducible. The random seeds (seed = 42) were fixed for Python, NumPy and PyTorch to reduce the variability among runs of the experiments. Prior to running each experiment, previously created output files were deleted to allow each run to be conducted on a level playing field. Furthermore, diagnostic reports (including the MD5 hash values over the produced topic assignments and the doc embeddings) are generated to confirm output consistency for multiple runs. All model designs, preprocessing steps, and hyper parameters are clearly described for reproducibility of the proposed framework.The research was done on publicly accessible Arabic news corpora which do not contain any personal information or sensitive content. Therefore, no human subject was involved, no data concerning individuals were gathered, and there was no need for obtaining approval from laboratory ethics committee. During this work, the ground truth labels of the original dataset were kept only for testing and were not used in training the model. In the interest of transparency and reproducibility, we describe the entire preprocessing pipeline, experiment setup and each of the evaluation steps in enough detail for the reader to replicate the setup and verify our results.

4. Results and Discussion

This part is devoted to the presentation and discussion of the results acquired using the proposed hybrid BERTopic–NAMAA with four Arabic pre-trained Transformer embedding models (AraBERTv0.2, AraBERTv2, Asafaya, and QARiB) and two dataset configurations, namely D1 which keeps the original class distribution, and D2 which is a balanced class distribution. The results are structured to examine the framework for extrinsic and intrinsic evaluations. Section 4.1 deals with the topic modeling results both in terms of the number and of the profiles of the identified topics. We evaluate intrinsic topic quality by NPMI, C_v coherence and Topic Diversity in Section 4.2. Section 4.3 presents the results on topic category classification using the pre-trained NAMAA classifier. We 4.4 report on the agreement between the automatically assigned topic categories and original dataset labels versus Cohen's κ. In this subsection, the best embedding models are compared according to each

evaluation indicator. We also conduct an error analysis based on confusion matrix in Section 4.6 to identify causes of most misclassifications and disagreements. 4.7 Our comparative analysis with the state-of-the-art work in Arabic topic modeling and text classification and the strengths and weaknesses of the presented framework. In Sect. 8, we conclude and ponder on their possible influences on Arab topic modeling and classification.

4.1 Topic Discovery Results (BERTopic)

Table 5 presents the intrinsic topic quality measures of BERTopic using the four tested Arabic Transformer embedding models on both dataset arrangements. We evaluate the quality of the topics by NPMI and C_v coherence with the Topic Diversity measure, reflecting semantic coherence, interpretability and the uniqueness of the generated topics, respectively.

***Table* 5.** *Unsupervised BERTopic results on the imbalanced dataset (D1) and the balanced dataset (D2)*

Model	NPMI (D1)	C_v (D1)	Diversity (D1)	NPMI (D2)	C_v (D2)	Diversity (D2)
AraBERTv0.2	0.255	0.545	0.704	0.283	0.553	0.687
AraBERTv2	0.264	0.528	0.709	0.289	0.574	0.702
QARiB	0.268	0.545	0.728	0.281	0.527	0.685
Asafaya	0.273	0.627	0.686	0.276	0.579	0.690

On D1 (the imbalanced dataset), Asafaya again obtained the best NPMI (0.273) and C_v (0.627) which indicates that discovered topics had the best semantic cohesion. On the other hand, QARiB achieves the best Topic Diversity (0.728), indicating that its topic representations are more lexically diverse. AraBERTv0.2 and AraBERTv2 achieved more or less comparable results on all the three metrics, yielding steady but less differentiating topic representations than Asafaya and QARiB. The performance on the balanced dataset (D2) follows a different trend. With the highest NPMI (0.289) and Topic Diversity (0.702) for AraBERTv2 and the highest C_v coherence (0.579) for Asafaya. Even QARiB survived on the imbalanced corpus, its performance dropped on D2, especially in term of coherence and diversity. This finding indicates that the relative performance of an embedding model might be related to the class prior and the nature of the corpus. The comparison between D1 and D2 shows that, overall, balancing the data led to better semantic coherence for the majority of the embedding models, but the degree of enhancement varied for each metric. These findings also show that there is not a single best embedding model for all the evaluation measures. Rather, each model represents specific strengths and weaknesses with respect to the quality measures. In general Asafaya the best topic representations across both datasets, AraBERTv2 attained the highest semantic relatedness and topic diversity on the balanced dataset and QARiB yielded higher lexical diversity on the unbalanced corpus. These results suggest that the selection of embedding model depends on the purpose of an application whether focusing on the semantic coherence, topic diversity, or the overall interpretability of topics.

4.2 Topic Classification and Agreement Results (NAMAA)

This section analyzes phase two of the suggested hybrid framework by analyzing how well the pre-trained NAMAA classifier can categorize the topics generated by BERTopic into semantic groups. The classification's performance is evaluated by traditional predictive metrics and then an agreement analysis based on Cohen's Kappa

Table 6. *Supervised NAMAA classification results on the imbalanced dataset (D1) and the balanced dataset (D2)*

Model	Acc. (D1)	F1 (D1)	Prec. (D1)	Rec. (D1)	Acc. (D2)	F1 (D2)	Prec. (D2)	Rec. (D2)
AraBERTv0.2	0.966	0.978	0.992	0.966	0.949	0.974	1.000	0.949
AraBERTv2	0.982	0.991	1.000	0.982	0.959	0.979	1.000	0.959
QARiB	0.969	0.980	0.991	0.969	0.960	0.960	1.000	0.960
Asafaya	0.949	0.966	0.985	0.949	0.968	0.983	1.000	0.968

is presented. In the end, the top performing embedding models over all evaluation parameters are presented. The results of the classification for the proposed framework with two dataset settings using Accuracy, Precision, Recall

and F1-score are shown in Table 6. AraBERTv2 outperformed all other models on the maximum length (D1) dataset for the classification task with 98.2% accuracy and 99.1% F1-score on the positive class and it also attained perfect Precision score of (100%). QARiB secured the second place with a 96.9% accuracy and a 98.0% F1-score, closely followed by AraBERTv0.2. While Asafaya had the worst classification results on D1, it still achieved a high precision of 98.5%, suggesting false positive predictions were relatively rare. A distinctive set of results was also obtained on the balanced (D2) dataset. Asafaya surpassed all the baselines in the overall classification with 96.8% Accuracy, 98.3% F1-score, and 100% Precision. QARiB and AraBERTv2 have comparable results while AraBERTv0.2 is fairly stable across both corpora. These findings suggest that embedding models should be more or less effective based on the class distribution of the baseline corpus. Precision was consistently high (98.5–100 %) in all experiments, thus the topic categories which were automatically assigned were very trustworthy in case of positive predictions. On the other hand, there was more variability in Recall, especially on the skewed data set, indicating that the class distribution had some impact on the framework's capability to retrieve all related topic categories. The details of these errors are analyzed further in Section 4.3. To the classification results the Cohen Kappa score between the automatically predicted topic category and the label in the original dataset was added, see Table 7, to also assess the agreement.

Table 7. *Cohen's Kappa agreement scores on the imbalanced dataset (D1) and the balanced dataset (D2)*

Model	Kappa (D1)	Kappa (D2)
AraBERTv0.2	0.829	0.885
AraBERTv2	0.743	0.918
QARiB	0.864	0.923
Asafaya	0.842	0.937

All evaluated embedding models achieved substantial to almost perfect agreement ($\kappa > 0.74$) according to the Landis and Koch interpretation scale. On the imbalanced dataset (D1), QARiB obtained the highest agreement ($\kappa = 0.864$), whereas Asafaya achieved the highest agreement on the balanced dataset ($\kappa = 0.937$). Dataset balancing improved Kappa values for every evaluated embedding model, with AraBERTv2 exhibiting the largest improvement, increasing from 0.743 on D1 to 0.918 on D2. These results indicate a high level of consistency between the automatically assigned topic categories and the original dataset labels under both experimental conditions.To facilitate comparison across the different evaluation criteria, Table 8 summarizes the best-performing embedding model for each metric and dataset.

Table 8. *Summary of best-performing models by evaluation metric and dataset*

Evaluation Dimension	D1 (Imbalanced) — Best Model	D2 (Balanced) — Best Model
Topic quality (NPMI & Coherence)	Asafaya (NPMI 0.273, C_v 0.627)	AraBERTv2 (NPMI 0.289) / Asafaya (C_v 0.579)
Classification (Accuracy & F1)	AraBERTv2 (Acc. 98.2%, F1 99.1%)	Asafaya (Acc. 96.8%, F1 98.3%)
Agreement & Diversity	QARiB (Kappa 0.864, Diversity 0.728)	Asafaya (Kappa 0.937) / AraBERTv2 (Diversity 0.702)

The results demonstrate that no single embedding model consistently outperformed the others across all evaluation dimensions. Asafaya achieved the strongest overall semantic coherence and agreement, AraBERTv2 consistently produced the highest classification performance and the strongest semantic association on the balanced dataset, whereas QARiB achieved the highest topic diversity and the strongest agreement on the imbalanced dataset. These findings indicate that the optimal embedding model depends on the primary objective of the application, whether maximizing intrinsic topic quality, classification performance, or agreement with the original dataset labels. Collectively, the results demonstrate that the proposed BERTopic–NAMAA framework provides robust and consistent performance across both balanced and imbalanced Arabic news corpora while highlighting the influence of embedding-model selection and class distribution on overall system performance.

4.5 Error Analysis (Confusion Matrix)

In order to have a better insight into the classification process of the learnt model, we performed a detailed error analysis in terms of the confusion matrix and the misclassification patterns. Although the overall results reported in Section 4.2 indicate that the classification was very accurate on average, the following results give information about what kind of mistakes the proposed framework made and how these errors could have been caused.Below are the confusion matrices for the D1 and D2 results in Table 9 and Table 10, respectively. In a both datasets false positive predictions were negligible for all embedding models that is consistent with very high values of Precision in Table 6. On the D1 without equalizing the class sizes, it was AraBERTv2 which made the least number of false negative predictions which is in line with its overall best classification results on D1. In contrast, in the balanced dataset Asafaya reached the minim number of false negatives which is in line with its best results in Accuracy and F1-score for the balanced classes.Although there was considerable variation in the number of identified topics (59 topics in Asafaya, D1, to 125 topics for QARiB, D2), the classification accuracy was uniformly high for all the embedding models. This means that the suggested BERTopic–NAMAA architecture achieves a stable topic category assignment under both coarse-grained and fine-grained topic representations. In order to analyze the rest of the classification errors, Table 11 reports the discovered misclassification patters along with the unmapped labels that were initially predicted by NAMAA. Error analysis shows qualitative differences for the two dataset setups. In the skewed dataset (D1), almost all of the misclassified topics were placed in the other category, instead

Table 9. *Confusion matrix results on the imbalanced dataset (D1)*

Model	Topics	Class	TP	FP	FN	TN
Asafaya	59	Economy	10	1	0	48
		Sports	46	0	3	10
AraBERTv0.2	116	Economy	11	1	0	104
		Sports	101	0	4	11
AraBERTv2	110	Economy	3	0	0	107
		Sports	105	0	2	3
QARiB	98	Economy	11	1	0	86
		Sports	84	0	3	11

Table 10. *Confusion matrix results on the balanced dataset (D2)*

Model	Topics	Class	TP	FP	FN	TN
Asafaya	62	Economy	29	0	0	33
		Sports	31	0	2	29
AraBERTv0.2	99	Economy	26	0	3	70
		Sports	68	0	2	29
AraBERTv2	97	Economy	37	0	3	57
		Sports	56	0	1	40
QARiB	125	Economy	55	0	5	65
		Sports	65	0	0	60

Table 11. *Error patterns and unmapped labels across models on D1 and D2*

Dataset	Model	Total Errors	Sports→Other	Economy→Other	Unmapped Labels
D1 (Imbalanced)	Asafaya	3	3	0	Politics, Finance, Culture
	AraBERTv0.2	3	3	0	Politics ×2, Culture
	AraBERTv2	2	2	0	Culture, Religion
	QARiB	2	2	0	Politics, Culture
D2 (Balanced)	Asafaya	2	2	0	Culture ×2
	AraBERTv0.2	5	2	3	Politics ×2, Religion ×2, Culture
	AraBERTv2	4	1	3	Politics ×2, Religion, Culture
	QARiB	5	0	5	Politics ×4, Technology

of being misclassified as the other sense target category. Most of these mistakes were of the form Sports → Other, which indicates there were a few Sports topics that could not be confidently mapped to the target categories. When the data was made uniform (D2), the error pattern overall was different, and more Economy → Other assignments were made. This result indicates that equalizing the class distribution caused the distribution of classification errors to change, rather than suppressing the errors altogether. A further investigation for the unmapped NAMAA predictions indicates that most of the remaining errors also fall under semantically related categories with Politics, Culture, Religion, Finance and Tech forming a significant portion. These classes are lexically and contextually similar with the target domains, in particular in news articles, political, economic and sport events have many overlaps. Hence, it can be said that a large amount of the remaining misclassifications appear to be induced by semantic closeness of adjacent news class rather than random misclassification. To summarize, the confusion matrices and error analyses show that the hy- brid framework demonstrates very stable classification performance when applied to the two datasets. The consistently low false positive rates, relatively small number of false negative predictions and very modest amount of unmapped categories indicate that the combination of BERTopic with the pre-trained NAMAA classifier delivers reliable topic category assignment while being robust to different topic granularity and class distributions.

4.6 Comparison with Previous Studies

In order to situate the proposed framework with the current state-of-the-art, the results of the experiments were compared to those of representative Arabic topic model and Arabic text classification works. As the previous works present different evaluation metrics and use different datasets and experimental settings, the comparisons can be seen as an indicator of performance and are not intended to be used for direct benchmarking. Tab. 12 compares the intrinsic topic-quality scores achieved by the proposed framework with the results from the related Arabic BERTopic work.

Table 12. *Comparison of topic-modeling quality metrics with previous studies*

Study	Year	Method / Best Model	NPMI	C_v	Diversity
Abuzayed & Al-Khalifa	2021	BERTopic + QARiB (5 classes)	0.17	N/A	N/A
Abdelrazek et al.	2022	BERTopic + RoBERTa	N/A	0.1147	N/A
Aouichaty et al.	2024	BERTopic + Arabic BERT (legal)	0.11–0.17	N/A	N/A
Present study — D1	2025	BERTopic + Asafaya	0.273	0.627	0.690
Present study — D2	2025	BERTopic + AraBERTv2	0.289	0.574	0.702

The inherent topic quality scores obtained by the proposed framework were consistently higher than the results from the literature in Table 12. Abuzayed and Al-Khalifa (2021) achieved a maximum NPMI of 0.17 with BERTopic on Arabic news articles [13], for while Aouichaty et al. (2024) obtained NPMI scores between 0.11 and 0.17 on Moroccan Arabic legal texts [15]. Similarly, Abdelrazek et al. (2022) got a C_v Coherence of 0.1147 via RoBERTa-based sentence embeddings [14]. Meanwhile, our framework achieved NPMI scores of 0.273 and 0.289 on (D1) and (D2) respectively, also the corresponding C_v coherence scores are 0.627 and 0.579. While these studies used different datasets and experimental configurations, the consistently superior intrinsic evaluation scores in this work suggest that the proposed BERTopic-NAMAA framework can produce the topic representations that are highly coherent and semantically meaningful. Moreover, previous Arabic BERTopic works that heavily depend on intrinsic topic-quality measures are complemented by an extrinsic evaluation based on Cohen's Kappa for the proposed framework which makes it well-rounded in assessing topic quality. Table 13 presents a comparison between the proposed framework and some representative Arabic Transformer-based text classification works in terms of topic category assignment. The proposed approach achieved similar classification performance to state-of-the-art Arabic Transformer-based classifiers, despite using an evaluation method that was rather different. Prior work, for example Ben Ali et al. (2022), Nassif et al. (2022), Karajeh et al. (2023), and Al-Taie (2025), has applied end-to-end supervised document classification via task-specific fine-tuning [16, 17, 19, 20]. In the current research, however latent topics are first mined by BERTopic, and then topics are classified by

the pre-trained NAMAA classifier without further fine-tuning for the experimental datasets. For the unbalanced data set (D1), the framework achieved 98.2% Accuracy

Table 13. *Comparison of classification performance metrics with previous studies*

Study	Year	Method / Best Model	Accuracy	F1	Precision	Recall
Ben Ali et al.	2022	AraBERT (fake news)	0.980	N/A	N/A	N/A
Nassif et al.	2022	ARBERT (fake news)	>0.98	N/A	N/A	N/A
Karajeh et al.	2023	GCN + AraBERT (SANAD, 7-class)	0.9725	0.9726	N/A	0.9727
Al-Taie	2025	AraBERT + BiLSTM (fake news)	0.984	0.980	N/A	N/A
Present study — D1 (AraBERTv2)	2025	BERTopic + NAMAA / AraBERTv2	0.982	0.991	1.000	0.982
Present study — D2 (Asafaya)	2025	BERTopic + NAMAA / Asafaya	0.968	0.983	1.000	0.968

and 99.1% F1-score with AraBERTv2, while Asafaya had the best performance on the balanced data set with 96.8% Accuracy and 98.3% F1-score on D2. This suggests that the hybrid framework proposed can keep up with competitive classifier performance, while also offering interpretable topic representations, which traditional end-to-end document classification methods do not provide. The bottom line is the comparison demonstrates two key contributions of the proposed framework. One, it is able to consistently obtain good intrinsic topic quality while achieving good classification performance under different embedding models and class distributions. Two, the proposed framework, by combining unsupervised topic modeling, independent topic categorical assignment and agreement evaluation in a single experimental setup, offers a more holistic evaluation approach than methods based on exclusively topic modeling or document classification.

4.7 Summary of Findings

The results of experiments show that the performance of the proposed BERTopic–NAMAA method is determined by the vector representation model and class imbalance of datasets. None of the embedding models consistently led to the best results with respect to all evaluation measures. Rather, they showed complementary strengths based on the consideration of the most important goal, such as maximizing semantic coherence, topic diversity, classification performance, or agreement with the labels in the original data set. Across both dataset configurations, the proposed framework consistently generated coherent topic representations while maintaining high classification performance and substantial-to-almost-perfect agreement with the original dataset labels. The balanced dataset generally improved semantic coherence and agreement, whereas the imbalanced dataset highlighted differences in the robustness of the evaluated embedding models. These findings confirm that intrinsic topic-quality measures and extrinsic evaluation metrics provide complementary perspectives and should be considered jointly when assessing Arabic topic-modeling systems.

5. Conclusion

This study proposed and evaluated a hybrid framework that integrates BERTopic for unsupervised topic modeling with the pre-trained NAMAA classifier for independent topic category assignment and extrinsic evaluation of Arabic news articles. The experimental results demonstrate that the proposed framework consistently produced coherent topic representations together with high classification performance across multiple Arabic Transformer embedding models and under both balanced and imbalanced class distributions.The findings further demonstrate that embedding-model selection and class distribution both influence the overall performance of Arabic topic modeling systems. While no single embedding model consistently outperformed the others across all evaluation criteria, the combination of intrinsic topic-quality measures with extrinsic agreement analysis provided a comprehensive evaluation of the proposed framework. Compared with existing Arabic topic-modeling studies, the proposed framework extends conventional evaluation by integrating unsupervised topic discovery, independent topic category assignment, and agreement analysis within a unified and reproducible experimental pipeline.

5.1 Contributions and Practical Implications

The suggested framework constitutes a unified approach for an Arabic topic modeling that integrates unsupervised topic extraction with a separate topic category assignment as well as full evaluation. In addition to strong results in experiments, the framework defines interpretable topics, which can be used in practical applications such as organizing Arabic news, digital archiving, trend analysis, and topic-based recommender systems. In a more general sense, the proposed evaluation approach demonstrates the possibility of combining intrinsic with extrinsic evaluation in order to obtain a more thorough assessment of systems for Arabic topic-modelling.

5.2 Limitations and Threats to Validity

Understanding the results above, a few limitations should be kept in mind. First, the experiments that we have conducted are constrained to two news categories (Economy and Sports), which might limit the generalizability of the results across other categories of Arabic news. Second, the agreement analysis was conducted between the original dataset labels and not between annotations collected from humans anew. Third, the proposed model was tested with one random seed only and without running predefined statistical significance tests. Fourth, in the NAMAA representation only, with no document-classification baseline to compare against which prevents us from disentangling the contribution of that representation on the hybrid design. Lastly, all experiments were conducted only on MSA, thus dialectal Arabic was not considered in this work.

5.3 Future Work

For future work, the proposed framework can be extended by considering more Arabic news categories and more diverse news providers as well as examining its applicability on dialectal Arabic corpora. Comparative analysis against other recent neural topic-modelling techniques using larger publicly available benchmark datasets would also further attest the generalizability of the framework. Moreover, further comparisons could be made between zero-shot and task-specific fine-tuned versions of NAMAA, between temporal topic evolution, perhaps employing BERTopic's dynamic topic modeling functionality, between outlier topics in more depth, and between the scalability of the presented framework on massive collections of Arabic news in the future.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] A. M. Alayba, "Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends," Computers, vol. 14, no. 11, p. 497, Nov. 2025, doi: 10.3390/computers14110497.

[2] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," J Big Data, vol. 11, no. 1, p. 92, Jul. 2024, doi: 10.1186/s40537-024-00948-z.

[3] C. D. Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press, 2008.

[4] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, 2003. [Online]. Available: https://www.jmlr.org/papers/v3/blei03a.html
[6]A. Vaswani et al., "Attention Is All You Need," Aug. 02, 2023, arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," doi: 10.48550/arXiv.1810.04805.
[8] N. Habash, Arabic Natural Language Processing. United States: Morgan and Claypool, 2009.
[9]D. M. Blei, "Probabilistic topic models," Commun. ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.
[10] V. Keselj, "Book Review: Speech and Language Processing (second edition) by Daniel Jurafsky and James H. Martin," Computational Linguistics, vol. 35, no. 3, Sep. 2009, doi: 10.1162/coli.B09-001.
[11] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure."
[12] O. Elshehy, O. Nacar, A. Djamai, M. Ragab, K. A. Jallad, and M. Abdelazim, "AraModernBERT: Transtokenized Initialization and Long-Context Encoder Modeling for Arabic," 2026, arXiv. doi: 10.48550/ARXIV.2603.09982.
[13] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," Procedia Computer Science, vol. 189, pp. 191–194, 2021, doi: 10.1016/j.procs.2021.05.096.
[14] A. Abdelrazek, W. Medhat, E. Gawish, and A. Hassan, "Topic Modeling on Arabic Language Dataset: Comparative Study," in Advances in Model and Data Engineering in the Digitalization Era, vol. 1751, P. Fournier-Viger, A. Hassan, L. Bellatreche, A. Awad, A. Ait Wakrime, Y. Ouhammou, and I. Ait Sadoune, Eds., Communications in Computer and Information Science, vol. 1751. Cham: Springer Nature Switzerland, 2022, pp. 61–71. doi: 10.1007/978-3-031-23119-3_5.
[15] S. Aouichaty, Y. Maleh, M. T. Mohtadi, A. Hajami, and H. Allali, "Sustainable Topic Modeling for Legal Moroccan Arabic Language: A Challenging Study on BERTopic Technique," Procedia Computer Science, vol. 236, pp. 582–588, 2024, doi: 10.1016/j.procs.2024.05.069.
[16] S. Ben Ali, Z. Kechaou, and A. Wali, "Arabic fake news detection in social media Based on AraBERT," in 2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), Toronto, ON, Canada: IEEE, Dec. 2022, pp. 214–220. doi: 10.1109/ICCICC57084.2022.10101635.
[17] A. B. Nassif, A. Elnagar, O. Elgendy, and Y. Afadar, "Arabic fake news detection based on deep contextualized embedding models," Neural Comput & Applic, vol. 34, no. 18, pp. 16019–16032, Sep. 2022, doi: 10.1007/s00521-022-07206-4.
[18] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 7088–7105. doi: 10.18653/v1/2021.acl-long.551.
[19] O. Karajeh, M. N. Al-Kabi, and E. A. Fox, "Fusing AraBERT and Graph Neural Networks for Enhanced Arabic Text Classification," in 2023 24th International Arab Conference on Information Technology (ACIT), Ajman, United Arab Emirates: IEEE, Dec. 2023, pp. 1–8. doi: 10.1109/ACIT58888.2023.10453909.
[20] M. Z. Al-Taie, "Comparative Study of Machine Learning Approaches for Detecting Fake News in Arabic Text," IETI Trans. Data Anal. Forecast.vol. 3, no. 1, pp. 18–31, May 2025, doi: 10.3991/itdaf.v3i1.53575.
[21] M. Abbas and K. Smaili, "Comparison of Topic Identification Methods for Arabic Language," presented at the RANLP 2005: Recent Advances in Natural Language Processing, 2005, pp. 14–17.
[22] O. Einea, A. Elnagar, and R. Al Debsi, "SANAD: Single-label Arabic News Articles Dataset for automatic text categorization," Data in Brief, vol. 25, p. 104076, Aug. 2019, doi: 10.1016/j.dib.2019.104076.
[23] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction."
[24] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," JOSS, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.
[25] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding."
[26] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," in Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 2054–2059. doi: 10.18653/v1/2020.semeval-1.271.
[27] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations."
[28] G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction."
[29] J. H. Lau, D. Newman, and T. Baldwin, "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality," in Proceedings of the 14th Conference of the European Chapter of the Association for

Computational Linguistics, Gothenburg, Sweden: Association for Computational Linguistics, 2014, pp. 530–539. doi: 10.3115/v1/E14-1056.
[30] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic Evaluation of Topic Coherence."
[31] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China: ACM, Feb. 2015, pp. 399–408. doi: 10.1145/2684822.2685324.
[32] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," Transactions of the Association for Computational Linguistics, vol. 8, pp. 439–453, Dec. 2020, doi: 10.1162/tacl_a_00325.
[33] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
[34] C. J. Van Rijsbergen, Information Retrieval, 2nd ed. London: Butterworth, 1979.
[35]"Glossary of Terms," Machine Learning, vol. 30, no. 2–3, pp. 271–274, Feb. 1998, doi: 10.1023/A:1017181826899.
[36] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.
[37] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.