International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 5541-5551
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Arabic Topic Detection: A Comprehensive Review of Recent Advances

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

Article Info:

DOI: 10.22399/ijcesen.3424
Received : 21 May 2025
Accepted : 18 July 2025

Keywords

Natural Language Processing
Arabic language Processing
Topic detection
Large Language Models
Machine Learning

Abstract:

Topic detection and short-text analysis have been significantly transformed by integrating machine learning (ML) techniques and large language models (LLMs) such as BERT and GPT, particularly in platforms like Twitter. These advanced models outperform traditional rule-based and statistical approaches by leveraging transformer architectures and semantic embedding techniques (e.g., word embeddings) to uncover text's latent themes and contextual relationships. Even in low-resource language settings, LLMs can capture semantic nuances and support robust text classification and dynamic topic modeling. However, Arabic-language applications face unique challenges, primarily due to the scarcity of high-quality, task-specific annotated datasets, especially for domains like synthetic content identification and fake news detection. Successful model training in Arabic requires extensive corpora, careful linguistic preprocessing, and sensitivity to morphological complexity and dialectal variability. Additionally, LLMs are limited by computational limitations related to input length, which restricts the capacity for scaling when working with large volumes of text. In conclusion, future research should focus on establishing hybrid frameworks with contextual fine-tuning for domains, cross-lingual transfer learning, and better management of computational memory to address these obstacles and completely tap into the possibilities of ML-driven text analytics in resource-constrained settings.

1. Introduction

The rapid development of online technologies and the extensive utilization of social media platforms have resulted in unprecedented amounts of digital content and the speed with which information can be accessed and disseminated. The benefits to users of the vast amounts of online information are clear, but new challenges have arisen. Retrieving specific information and knowledge quickly and accurately from larger digital collections has become difficult in this new environment. Also, knowledge about a single topic, event, or moment is often spread across different times, places, or locations across digital platforms, making creating a coherent account or a more coherent and complete mental model challenging. This complexity arises from massive arrays of data, multiple sources, and the dispersion of information across the digital landscape. As a result, there is a need to develop complex computer systems to aggregate, arrange, and organize information and knowledge by topic or event and algorithms to discover and follow the topics that users are interested in. For this to happen effectively, we need to integrate these various technologies to improve the accessibility, clarity, and perception of our information, especially at a time of extraordinary digital growth and big data [1] .

Recent advances in natural language processing (NLP) - especially the rise of large language models (LLMs) such as BERT and GPT - have revolutionized automatic topic detection and analysis of short texts. These models rely on deep semantic embeddings and Transformer architecture to understand complex semantic relationships and subtle language patterns that traditional (rule-based or statistical) methods fail to recognize. These models can generalize across domains and adapt to resource-constrained languages, including Arabic,

making them critical tools in multilingual and diverse information systems.

Furthermore, combining large language models with unsupervised positional modeling techniques improves the accuracy of identifying underlying topics, monitoring changes in discussions over time, and filtering relevant content in real-time. This radical shift highlights the crucial role of AI frameworks in enhancing objective analysis, especially in contexts where information is abundant but fragmented.

1.1. Challenges of Topic Detection in Arabic Text

Detecting topics in Arabic texts, especially short texts posted on social media, such as Twitter, raises a number of linguistic and methodological issues that can seriously affect the performance and accuracy of models.

Understanding the contextual meaning of short texts is one of the most prominent of these challenges, as the nature of short tweets and their lack of sufficient linguistic context make it difficult to extract the main idea accurately [4]. In addition, there is a lack of effective mechanisms for pre-processing Arabic texts, which negatively affects the efficiency of analytical models and their ability to deal with the linguistic characteristics of the Arabic language, such as derivation and morphology [2].

Assessing model performance in heterogeneous data environments is yet another obstacle because it is challenging to develop precise and standardized metrics that measure the performance of multiple models. This problem is exacerbated by the fact that many platforms (e.g., Twitter) are multilingual, meaning different languages or dialects and possibly too much noise, which often makes invalid topic extraction [3].

On the other hand, the shortage of quality and quantity of Arabic training data is one of the main barriers to developing accurate models. Most large language models, such as GPT, are mainly trained on English texts, while Arabic resources are still limited in quantity and quality. The inconsistency in language representation limits the ability of these models to recognize subtle Arabic contexts, which negatively influences their performance in topic detection tasks across short and informal texts (e.g., tweets) [5].

These difficulties point to a critical need for dedicated linguistic processing environments for Arabic with high-quality annotated data sources that facilitate the creation of more precise and relevant models for practical applications.

2. Literature Review

The digital age is experiencing a massive explosion in the amount of Arabic content produced on the internet and social media platforms. This creates challenges of automatically sorting this amount of data and obtaining topics for it. With this literature review, we intend to review the research progress on topic discovery in Arabic texts, study the methods used, and analyze their effectiveness and areas of use. Some of these works are listed below:

Nahar et al. (2020) proposed a direct text analysis system designed to support researchers in maintaining topic coherence during the writing process. The system was developed using the Stand for Arabic corpus and incorporated essential preprocessing techniques, including stop-word removal and character normalization. To improve the accuracy of thematic categorization, the approach utilized N-gram models in combination with the Naïve Bayes classification algorithm [9].

Liu et al. (2020) introduced a dynamic topic recognition and tracking model based on the Citation-Involved Hierarchical Dirichlet Process (CIHDP), presenting a novel approach for monitoring technological evolution. The study utilized benchmark datasets such as Citeseer, Cora, and Aminer. The preprocessing pipeline included duplicate removal, stemming, stop-word elimination, and representation learning through Node2Vec, ensuring that each unique word was counted only once per document. Experimental results demonstrated that the proposed CIHDP model outperformed traditional topic modeling techniques, including Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP). A primary strength of the CIHDP method is the ability to automatically compute the optimal value for the number of topics, even if there are no pre-determined definitions—this improves the model's sophistication and flexibility to capture evolving technological trends [10].

Mottaghinia et al. (2020) explained that the systematic study and assessment of an array of Twitter topic identification techniques illustrate the importance of this research for researchers and practitioners, who can use it to identify and design improved methods of managing large datasets and better understand trends and trending topics on Twitter in real-time. Some studies utilized the Twitter API to obtain the TDT4 dataset, a collection of tweets extracted from Twitter. In efforts to detect emerging topics without relying on pre-trained datasets, researchers predominantly employed techniques such as encoding, normalization, standardization, segmentation, and clustering algorithms, including Naïve Bayes, SVM, K-means, Single-Pass, LDA, Word2Vec, and BERT. While supervised methods like SVM and Naïve Bayes

achieved high classification accuracy, their performance was constrained by the necessity of labeled training data and limited adaptability to previously unseen topics. Contextual word representation models such as Word2Vec and BERT demonstrated notable classification accuracy; however, they similarly depend on trained data and struggle to generalize to novel themes. Techniques utilizing contextual word embeddings, such as Word2Vec and BERT, outperformed traditional methods like TF-IDF by improving detection accuracy by 3–10%, especially in capturing semantic and contextual subtleties. BERT yielded the highest efficacy in tweet evaluation when integrated with conventional embedding techniques such as GloVe. The K-means clustering algorithm also proved effective in grouping short tweets when combined with preprocessing methods like Singular Value Decomposition (SVD). Furthermore, the enhanced Single-Pass algorithm (MC-TSP) improved efficiency by 10% compared to its traditional counterpart while mitigating the impact of data sequencing. Future research directions include developing real-time topic detection systems, integrating geographical and temporal contexts, adopting contextual language models such as BERT, and establishing standardized datasets for unbiased comparative analysis [3].

Over the past decade, El Kah et al. (2021) conducted a scoping review examining advancements in Arabic topic identification. The study explored various text representation methods, particularly Bag of Words (BOW) and Bag of Concepts (BOC), utilizing analytical tools such as Count Vectorizer, TF-IDF Vectorizer, and Chi-square testing. It also evaluated supervised classification algorithms, including Naïve Bayes, Decision Trees, SVM, KNN, and ANN, with particular emphasis on the effectiveness of SVM and probabilistic-based approaches. The review highlighted the need to enhance preprocessing techniques and classification algorithms, and it called on researchers to develop new, freely accessible Arabic text corpora to facilitate further research in this domain [11].

Abuzayed et al. (2021) conducted an empirical study on applying BERTopic in topic modeling in Arabic, using pre-trained Arabic language models. The study aimed to evaluate the performance of BERTopic compared to traditional methods such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Analysis (NMF) regarding topic quality, using the NPMI scale to assess topic coherence. The study used the Arabic taxonomy dataset without applying preprocessing steps due to the data's cleanliness. UMAP, HDBSCAN, and TF-IDF techniques, along with topic modeling algorithms, were employed in the methodology. The results showed that BERTopic outperformed LDA and NMF, especially in preserving topic coherence. The research suggested enhancing the criteria for evaluating topic quality, investigating more language models, broadening datasets, applying preprocessing methods, and utilizing BERTopic across different applications [12].

Grootendorst (2022) introduced BERTopic as a neural topic modeling approach designed to enhance text analysis and topic extraction by improving accuracy and consistency. The study employed a variety of datasets, including 20 Newsgroups, BBC News, Trump Tweets, and United Nations General Debate Transcripts (UN). Preprocessing techniques included lowercasing, lemmatization, stop word removal, and excluding documents containing fewer than five words. The results demonstrated the effectiveness of BERTopic in extracting coherent and meaningful topics, highlighting its superiority over traditional topic modeling methods [13]

As part of the AraProp initiative at WANLP, Singh (2022) conducted a study on identifying propaganda techniques in Arabic texts published on social media. The research utilized a dataset provided by the organizers of the WANLP 2022 competition, curated explicitly for Arabic propaganda detection. No preprocessing techniques were applied, thereby preserving the original structure of the texts. The study employed Arabic-specific pre-trained language models, including MARBERT, ARBERT, and AraBERT, as well as multilingual models such as XLM-RoBERTa and mBERT. The findings indicated that pre-trained models, particularly MARBERTv2, demonstrated promising performance, although further enhancements are necessary to improve their effectiveness in future applications [2].

Mansy et al. (2022) aimed to develop an ensemble model for emotion recognition in Arabic Twitter texts using deep learning techniques. The study employed the SemEval-2018-Task1-Ar-Ec dataset and applied multiple preprocessing steps, including the removal of English characters, numerals, and stop words, Arabic normalization, diacritic removal, emoji substitution, elimination of repeated characters, punctuation removal, and translation of English text. The research adopted MARBERT, AraBERT, and a combination of both. The results showed that MARBERT performed better than AraBERT, achieving the best performance for emotion detection. The ensemble model also performed better than the individual models and previous work, showing promising results for finding emotions in Arabic Tweets. This framework provides a strong framework for performing

sentiment analysis of Arabic social media texts. Future work can be extended to include aspects like data balancing, experimenting with different language models, improving the ensemble process, examining Arabic dialects, and improving the processing of emojis. In addition, practical applications of the framework are suggested. These developments should help further improve the model's ability to identify emotions in Arabic text [14].

Ben Ali introduced an approach to identifying misinformation in Arabic, specifically on Tunisian Arabic Twitter and Facebook, in 2022. The study utilized pre-trained advanced language models, particularly AraBERT, to conduct classification accurately while accounting for the aggravating circumstances associated with misinformation during COVID-19. The dataset comprised over 6,000 meticulously annotated records, including tweets, Facebook posts, and comments, categorized as either false or authentic news. The preprocessing pipeline involved removing website links, user mentions, non-Arabic text, emojis, punctuation, Arabic diacritics, numerals, and hashtags. Additionally, redundant characters, excessive whitespace, and symbols were eliminated. The text was then tokenized and normalized to ensure consistency. AraBERT's performance was compared against the BERT base Arabic model, demonstrating superior classification accuracy of 95.2% versus 94.6% on the study's dataset and 98.2% versus 97.3% on an external dataset. Evaluation metrics included precision, recall, and F1-score. Overall, the results underscored AraBERT's potential and efficacy in identifying fake news in Arabic, especially in cases pertaining to local contexts, and highlighted the continued work necessary to upgrade Arabic-language datasets and produce models that can more effectively manage heterogeneous dialects [15].

George et al. (2023) suggested an incorporated clustering and BERT framework to improve topic modeling for large and unstructured text corpora. The work influences advanced techniques, combining BERT with Latent Dirichlet Allocation (LDA) to enhance topic extraction. The study uses the CORD-19 dataset and preprocessing steps such as TF-IDF, stop word removal, spelling correction, and diacritic removal. The results show that the integrated LDA-BERT approach exceeds standalone methods, while UMAP proves to be more effective for dimensionality reduction [16].

Yusung et al. (2023) conducted a study on extracting marketing insights from reviews by employing BERTopic and Deep Clustering Networks (DCN) to enhance automation in data analysis. The research utilized review data from the Naver Shopping platform and incorporated preprocessing steps such as data cleaning, stopword removal, noise reduction, data formatting, and distribution analysis. Advanced techniques were applied, including HDBSCAN, UMAP, and KcBERT—a Korean language-specific variant of BERT. The results indicate that the BERTopic and DCN models effectively identify product features, drawbacks, and relationships between products, thereby providing a robust data-driven framework for businesses to understand consumer needs better and optimize marketing strategies [17].

Rahimi et al. (2023) introduced ATEM, a novel approach to identifying new research topics by examining how the relationships between these topics change over time. When looking at the development of ideas, they identify citation networks as playing an important role. Utilizing the DBLP dataset, which contains five million computer science publications, the study employed text preprocessing techniques such as stop word removal, lowercasing, and punctuation elimination, combined with Doc2Vec, TF-IDF, and clustering algorithms including HDBSCAN, Leiden, and Cluster Aggregation. Furthermore, dynamic topic model methods, such as LDA, ANTM, and BERTopic, were added. The results indicate that ATEM outperforms classic models for accurately identifying new topics and has predictive capabilities for prospective research directions using citation data. The study presents ATEM as a robust tool for analyzing topic evolution and recommends future enhancements, such as applying it to broader datasets and incorporating additional dynamic models [18].

Al-Khalifa et al. (2023) conducted a study analyzing discussions among early Arab Twitter users regarding ChatGPT, focusing on key topics, sentiment analysis, and tone detection. Data were collected directly from Twitter using the Scrape and Tweepy packages. The preprocessing steps included diacritic removal, hashtag processing, emoji extraction, elimination of Twitter metadata and special characters, text normalization, removal of newlines and duplicates, and manual filtering of irrelevant tweets. The study employed BERTopic for topic modeling, ArabiTools for sentiment analysis, and the MARBERT Sarcasm Detector for identifying sarcastic tones. The thematic analysis revealed prominent discussion themes, including the use of ChatGPT in research, education, and customer service; its potential impact on employment and society; ethical concerns surrounding artificial intelligence, particularly bias, and misinformation; experiences of Arab users with ChatGPT, such as registration and accessibility challenges in certain Arab countries; and

controversies related to ChatGPT bans in countries like Italy due to privacy issues. Sentiment analysis classified user attitudes as neutral, positive, or negative, while sarcasm detection identified ironic or sarcastic expressions in tweets. The findings provide insight into early Arab user perceptions of ChatGPT and the broader implications of AI adoption in the region [19].

De Leo et al. (2023) emphasize that enhancing clustering stability and the accuracy of topic detection is essential for improving topic recognition in short textual formats, such as tweets. The study compiled a dataset of tweets from 20 companies across various sectors, including food, telecommunications, and the automotive industry. Preprocessing steps involved stopword removal, lowercase conversion, elimination of mentions and hashtags, link extraction, punctuation removal, and tokenization. The analytical methods applied included HDBSCAN, Non-Negative Matrix Factorization (NMF), and Word2Vec. The semantics were built using Word2Vec while ensuring the stability and purity of the topics, particularly regarding short texts, from recursive consensus clustering. This was useful for public opinion and real-time understanding of conversation and trending topics. Also, improving the accuracy of topic detection could assist in the fight against misinformation. Analyzing tweets would help gain closer insight into consumer attitudes about brands in a digital marketing context [20].

In a recent study, Aljehani et al. (2024) proposed a BERT-based model integrated with Prototypical Networks (PN) for few-shot learning aimed at topic identification in short Arabic texts. The researchers utilized the SemEval Dataset, Arabic Social Media News Dataset (ASND), and Arabic Influencer Twitter Dataset (AITD), applying various text preprocessing techniques, including symbol and stop word removal, special character replacement, diacritic elimination, repetition handling, and tokenization. The study utilized MARBERT, a BERT-based model specially designed for Arabic and Prototypical Networks to overcome limitations due to scarcity of data. The findings indicated that the method suggested in this study provides effective and accurate topic detection for short Arabic texts and can be a helpful technique [21].

Alshammari et al. (2024) aimed to enhance the accuracy of detecting AI-generated texts (AIGT) in Arabic, tackling a notable deficiency in the literature where most prior research has concentrated on Latin-based languages, particularly English. The study proposed using a unique custom dataset alongside other specialized datasets, including Ejabah-Driven, Religious, Custom Plus, and Custom Max. Essential text preprocessing techniques were applied, such as text cleaning, number mapping, element replacement, and the removal of elongation (Tatweel). The research utilized advanced models for the detection task, including AraBERT, AraELECTRA, XLM-R, and mBERT [8].

Boutal et al. (2024) introduced a novel framework named "BERTrend," designed to detect emerging trends and weak signals within large, dynamic datasets. The framework was applied to datasets, including New York Times (NYT) articles and scientific paper abstracts from arXiv. The methodology involved segmenting documents into paragraphs, filtering out texts containing fewer than 100 Latin characters—given that short texts are often regarded as noise—and dividing the corpus into temporal slices (e.g., daily, weekly, or monthly). Text normalization through Unicode normalization was applied to ensure data consistency. The approach utilized HDBSCAN and UMAP for clustering and dimensionality reduction. The BERTrend framework, combined with BERTopic, demonstrated its ability to identify early signals related to the COVID-19 pandemic. Future work aims to incorporate additional datasets, integrate real-time data, and enhance evaluation procedures by engaging domain experts to improve weak signal detection [22].

Mu et al. (2024) evaluated the performance of Large Language Models (LLMs) compared to traditional topic modeling approaches. Unlike conventional methods, LLMs can handle unprocessed text directly, unlike traditional methods that often demand elaborate preprocessing steps like stop word removal, stemming, and tokenization. This capability removes the necessity for those intricate procedures and saves both time and computational resources. The study found that while Latent Dirichlet Allocation (LDA) performs adequately on simple texts, its accuracy diminishes when applied to complex or multi-topic documents. Conversely, large language models (LLMs) offer a more profound grasp of context and can identify topics with greater accuracy and detail without needing elaborate preprocessing. The authors also emphasize the importance of additional research to decrease the computational expenses and improve the efficiency of large language models [23].

Abdelal et al. (2024) introduced LAraBench, a standardized framework designed to evaluate the performance of large language models—including GPT-3.5, GPT-4, BLOOMZ, and Jais-13b-chat—on various Arabic language processing tasks. The study assessed the applicability of these models across multiple Natural Language Processing (NLP) tasks such as translation, sentiment analysis, named

entity recognition, automatic speech recognition (ASR), and text-to-speech (TTS). The evaluation was performed using 61 diverse datasets, incorporating procedures such as text normalization, noise removal, and prompt engineering. The results demonstrated that GPT-4 outperformed other models in numerous tasks, particularly in few-shot learning scenarios, although it remained less effective than specialized models on certain specific tasks [24].

Alghaslan et al. (2024) present a study on fine-tuning large language models for Arabic stance detection as part of StanceEval2024. Stance detection involves identifying a writer's supportive, opposing, or neutral attitude toward a topic. The research utilizes the MAWQIF dataset and evaluates several large language models, including AraBERT, GPT-3.5-Turbo, Meta-Llama-3-8B-Instruct, and Falcon-7B-Instruct. Results reveal that GPT-3.5-Turbo achieved the highest overall performance, with an F1 score of 82.93%. The Meta-Llama-3-8B model showed moderate performance, attaining an F1 score of 74.33%, while Falcon-7B exhibited the lowest performance with an F1 score of 40.73%. GPT-3.5-Turbo demonstrated significant effectiveness in understanding and accurately classifying multilingual tweets. The study recommends further exploration of optimization parameters and adopting approaches such as Retrieval-Augmented Generation (RAG) to enhance model performance and integrate these models into practical applications for improved social media monitoring and public opinion analysis [25].

Aljehani et al. (2024) propose a topic detection model for brief Arabic texts using Few-shot Learning techniques to tackle the issue of insufficient annotated data and enhance the model's ability to classify topics in short texts with few examples. The study employs datasets including SemEval, the Arabic Social Media News Dataset (ASND), and the Arabic Influencer Twitter Dataset (AITD). Preprocessing involves tokenization, repetition handling, stop word removal, special character replacement, and symbol elimination. The model integrates Prototypical Networks (PN) with MARBERT to perform Few-shot Learning. The proposed approach achieves accuracies of 86.2% on the SemEval dataset, 80.8% on the ASND dataset, and 93.4% on the AITD dataset, outperforming traditional classifiers such as SVM, CNN, and LSTM despite utilizing only a limited number of samples. The study suggests future enhancements by incorporating additional languages, including English, and extending the model's application to classify longer documents [21].

Hossain et al. (2024) introduced AraCovTexFinder, a transformer-based language model designed to detect COVID-19-related content in Arabic texts autonomously. The study utilized datasets such as AraCov and AraEC and incorporated extensive preprocessing techniques. These techniques included extracting non-Arabic characters, normalizing whitespace using regular expressions, and removing HTML elements, hashtags, URLs, and punctuation. Additionally, hyperparameter optimization was performed to enhance the model's performance.Various models were evaluated, including RVADER, XML-RoBERTa, BERT, CT-BERT, CNN-GRU, BERT+CNN, AraBERT, and ensemble approaches. The findings demonstrated that AraCovTexFinder outperformed all baseline models, particularly in addressing the out-of-vocabulary (OOV) issue through enhanced word embedding strategies. Future research directions include expanding the model's capabilities to detect other forms of disinformation, such as politically and economically motivated content. There are also plans to integrate the model into social media platforms for real-time fake news detection and to develop open-source tools for analyzing Arabic health-related texts.

Kirilenko et al. (2024) evaluated the effectiveness of large language models (LLMs), such as ChatGPT, in conducting topic analysis of user-generated comments from social media platforms. The study specifically compared the performance of LLMs with the traditional Latent Dirichlet Allocation (LDA) approach, focusing on the challenges posed by short, unstructured content, such as user comments on tourism-related films. Data were collected from platforms including YouTube and Weibo. The results suggested that LLMs were better than LDA when classifying topics from language samples, especially with shorter texts and noise from language. Last, the topics provided by LLMs were more coherent and specific than those produced by LDAs.This approach demonstrates the potential of LLMs for extracting valuable insights from social media content, thereby enhancing the understanding of user sentiments and emerging trends [27].

Tarekegn (2024) presents a study to enhance news classification and event detection through advanced clustering techniques. The research leverages embeddings generated by large language models (LLMs) to improve the accuracy of event identification within news narratives. Utilizing the GDELT dataset, the study applied a series of preprocessing steps, including data cleaning, text normalization, noise reduction, keyword extraction, text embedding, and dimensionality reduction. Various clustering algorithms, including

Agglomerative Clustering, K-Means, HDBSCAN, and Gaussian Mixture Models (GMM), were explored alongside dimensionality reduction methods like UMAP and t-SNE. Experimental results demonstrated that embeddings based on LLMs significantly outperformed traditional embedding techniques, including TF-IDF, GloVe, and BERT, regarding clustering accuracy. The study indicates that future research should be conducted concerning the time evolution of events, exploiting more advanced artificial intelligence approaches to improve event detection and tracking capabilities [28].

Doi et al. (2024) investigate the enhancement of topic modeling for short texts using Large Language Models (LLMs), specifically GPT-3.5 and GPT-4. Short texts—such as social media posts and news headlines—pose unique challenges for traditional topic modeling techniques. The study utilized three primary datasets: StackOverflow, Tweets, and GoogleNewsT. These datasets were preprocessed through lowercasing, removal of short words, filtering of rare terms, and segmentation into subsets to comply with LLM input length limitations. The performance of GPT-3.5 and GPT-4 was compared to that of conventional models, including Latent Dirichlet Allocation (LDA), Topically-Supervised Contextual Topic Model (TSCTM), and BERTopic. The results demonstrated that GPT-4 outperformed traditional models regarding topic coherence, marking a significant advancement in short-text topic modeling. The authors suggest, and welcome research to increase topic quality, improve model efficiency, and discover new practical applications to augment the usefulness and generalizability of LLM topic modeling approaches [29].

Alzaidi (2025) proposed an advanced methodology for automatically classifying Arabic news articles. The preprocessing pipeline applied to the SANAD dataset involved removing URLs, emoticons, numerals, punctuation, special characters, hashtags, and usernames. Text normalization techniques such as lowercasing, lemmatization, tokenization, and stop-word removal were also utilized. The research utilized two models, TCAODL-ANA and FastText, for analyzing classification performance in many news domains, specifically politics, technology, sports, finance, medicine, and religion. The results indicate that TCAODL-ANA successfully identified and categorized Arabic news content across various topics. In addition, the research indicates the benefits of employing a variety of feature engineering techniques, through identifying and analyzing more complex linguistic patterns which could lead to additional classification performance. Future research could explore other deep learning architectures to provide more accurate and effective clarity and detail to classification systems for Arabic documents.

Table 1. *Summary table of research on Topic detection in Arabic Text.*

Authors	method	Dataset	The result
Nahar et al. 2020	N-gram, Naïve Bayes	Stanford Arabic Corpus	Accuracy 90%
Liu et al 2020	CIHDP, LDA, HDP	Citeseer, Cora, Aminer	Perplexity (CIHDP)
Mottaghinia et al. 2020	Naïve Bayes, SVM, K-means, Single-Pass, LDA, Word2Vec, BERT	TDT4	review
El Kah et al. 2021	Naive Bayes ،Decision Trees ،SVM ،KNN ، ANN	SANAD, NADiA, OSIAN, ANT, Abu El-Khair, NADA, ASND	Naive Bayes, SVM
Abuzayed et al. 2021	BERTopic, LDA, NMF	MSA	BERTopic
Grootendorst 2022	BERTopic, LDA, NMF, CTM, Top2Vec	BBC News, Trump Tweets, UN	BERTopic
Singh 2022	Bert-base-multilingual-cased, xlm-roberta-base, bert-base-arabic, bert-base-araber, bert-base-arabertv2, ARBERT, MARBERT, MARBERTv2	tweepy library	MARBERTv2 0.61116 (micro-F1) bert-base-arabic 0.16182(macro-F1)
Mansy et al. 2022	MARBERT, AraBERT, Ensemble Model	SemEval-2018-Task1-Ar-Ec	MARBERT 0.529 (F1 Score Macro)
Ben Ali 2022	AraBERT, BERT base Arabic	scraped from Twitter and (posts/comments) from Facebook	AraBERT 0.952 Accuracy

George et al. 2023	LDA, BERT, LDA-BERT	CORD-19	BERT-LDA 0.51998 UMAP
An et al. 2023	BERTopic, DCN	Naver Shoppin	BERTopic
Al-Khalifa et al. 2023	BERTopic, MARBERT	Twitter	
Rahimi et al 2023	LDA, ANTM, BERTopic, Doc2Vec, HDBSCAN	DBLP	ATEM
De Leo et al. 2023	HDBSCAN, Word2Vec	Tweets	Word2Vec
Aljehani et al. 2024	MARBERT, Prototypical Networks (PN) for Few-shot Learning	SemEval Dataset, ASND, AITD	The accuracy of SemEval Dataset86.2%, ASND80.8%, AITD93.4%
Alshammari et al. 2024	AraBERT, AraELECTRA, XLM-R , mBERT	Discretized Custom, Ejabah-Driven, Religious, Custom Plus, Custom Max	
Boutaleb et al. 2024	BERTrend, BERTopic	(NYT) articles, Summaries of scientific papers from arXiv.	The model has demonstrated its ability to detect early signs such as the COVID-19 pandemic
Mu et al. 2024	LDA, GPT 4	News articles, Blogs	GPT 4
Abdelali et al. 2024	GPT-3.5, GPT-4, BLOOMZ, Jais-13b-chat	QADI, ADI, ArSAS, ArSarcasm, SemEval18, SANAD, ASND, AraBench, MADAR, APT.	GPT-4
Alghaslan et al. 2024	GPT-3.5-Turbo, Meta-Llama-3-8B-Instruct, Falcon-7B-Instruct	MAWQIF	GPT-3.5-Turbo
Hossain et al. 2024	RVADER, XML-RoBERTa, BERT, CT-BERT, CNN-GRU, BERT+CNN, AraBERT, Ensemble	AraCov, AraEC	The AraCov Tex Finder model outperformed
Kirilenko et al 2024	LDA , GPT-3	weibo , You Tuobe	GPT-3
Tarekegn 2024	LLM embedings, Agglomerative Clustering, K-Means, HDBSCAN, GMM, UMAP, t-SNE	GDELT	Enhanced Clustering
Doi et al. 2024	GPT-3.5, GPT-4, LDA, TSCTM, BERTopic	StackOverFlow, Tweet, GoogleNewsT	GPT-4
Alzaidi 2024	TCAODL-ANA, FastText	SANAD	effective

3. Conclusion

The cumulative analysis of the research experiences listed in the table shows a significant development in the methodologies of topic discovery for Arabic texts, which can be monitored through the following axes: First, the research field is witnessing a fundamental methodological shift from traditional statistical models to advanced artificial intelligence systems:

First, the research field is witnessing a fundamental methodological shift from traditional statistical models to advanced AI systems. In the early phase (2020-2021), the dominant models were characterized by relying on classical algorithms such as n-gram models, conditional probability models (Naïve Bayes), and automatic classification algorithms (SVM), as these methods recorded

accuracy rates ranging from 80-90% under ideal testing conditions. More recent data (2021-2024) indicates a clear dominance of deep natural language processing models, especially the BERT family of models (such as AraBERT and MARBERT) and large generative models (such as GPT-4), with accuracy rates exceeding 95% in some specialized applications.

Second, the effectiveness of Arabic-specific language models emerged as a critical factor in improving performance. The results show a clear superiority of Arabic-specific models (such as ARABERT and MARBERTv2) compared to multilingual models, recording differences in classification accuracy of up to 45% in some comparisons. This difference emphasizes the concept of linguistic specialization when dealing with the distinct morphological and syntactic characteristics of the Arabic language, particularly in relation to derivational and inflectional morphology.

The third factor to consider is that the variety within the datasets exposes considerable methodological challenges. The sources varied between standardized colloquial texts (e.g., SANAD) and social media (e.g., Twitter), reflecting the contrast between formal and colloquial language. Also, the different evaluation measures (Accuracy, F1-Score, Perplexity) between studies prevent accurate comparisons, which calls for the development of standardized evaluation criteria.

Fourth, promising research trends emerge when combining new technologies. Algorithmic hybrids that combine deep language models with localized modeling methods (e.g., BERT combined with LDA) have shown remarkable success in improving the efficiency of overlapping topic detection. In contrast, generative models (BERT combined with LDA) have been shown to improve the efficiency of overlapping topic detection. Generative models (e.g., GPT-4) also proved to be highly efficient in handling complex tasks such as tracking the temporal evolution of topics.

In conclusion, while the advances are incredible, there are still gaps to be filled in research, namely the paucity of documentation reflecting Arabic dialects, non-existent standards on performance assessment frameworks, and the need for methodological frameworks to evaluate the progress of models evaluated in real situations other than experimental settings. The study recommends moving towards developing multi-technology hybrid systems, establishing standardized evaluation criteria, expanding training to include dialectal diversity, and enhancing the interpretability of models.

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] Riccardo Cantini and Fabrizio Marozzo, (2023). Topic Detection and Tracking in Social Media Platforms. *Springer, Cham*, https://doi.org/10.1007/978-3-031-31469-8_3

[2] G. Singh, (2022). AraProp at WANLP 2022 Shared Task: Leveraging Pre-Trained Language Models for Arabic Propaganda Detection, in *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, 496–500. doi: 10.18653/v1/2022.wanlp-1.56.

[3] Z. Mottaghinia, M.-R. Feizi-Derakhshi, L. Farzinvash, and P. Salehpour, (2021). A review of approaches for topic detection in Twitter, *J. Exp. Theor. Artif. Intell.*, vol. 33(5), 747–773, doi: 10.1080/0952813X.2020.1785019.

[4] Amani Aljehani and Syed Hamid Hasan, (2024). A BERT-based Prototypical Networks for Few-Shot Arabic Short-text Topic Detection, *J. Eng. Sci.*, vol. 20(10s), https://journal.esrgroups.org/jes/article/view/5146

[5] F. Alderazi, A. Algosaibi, M. Alabdullatif, H. F. Ahmad, A. M. Qamar, and A. Albarrak, (2024). Generative artificial intelligence in topic-sentiment classification for Arabic text: a comparative study with possible future directions, *PeerJ Comput. Sci.*, vol. 10, e2081, doi: 10.7717/peerj-cs.2081.

[6] H. Lamtougui, H. El Moubtahij, H. Fouadi, and K. Satori, (2023). An Efficient Hybrid Model for Arabic Text Recognition, *Comput. Mater. Contin.*, vol. 74(2), 2871–2888, doi: 10.32604/cmc.2023.032550.

[7] S. Aouichaty, Y. Maleh, M. T. Mohtadi, A. Hajami, and H. Allali, (2024). Sustainable Topic Modeling for Legal Moroccan Arabic Language: A Challenging Study on BERTopic Technique, *Procedia Comput. Sci.*, vol. 236, 582–588, doi: 10.1016/j.procs.2024.05.069.

[8] H. Alshammari and K. Elleithy, (2024). Toward Robust Arabic AI-Generated Text Detection: Tackling Diacritics Challenges, *Information*, vol. 15(7), 419, doi: 10.3390/info15070419.

[9] K. Nahar, R. Al-Khatib, M. Al-Shannaq, M. Daradkeh, and R. Malkawi, (2020). Direct Text Classifier for Thematic Arabic Discourse Documents, *Int. Arab J. Inf. Technol.*, vol. 17(3), 394–403, doi: 10.34028/iajit/17/3/13.

[10] H. Liu, Z. Chen, J. Tang, Y. Zhou, and S. Liu, (2020). Mapping the technology evolution path: a novel model for dynamic topic detection and tracking, *Scientometrics*, vol. 125(3), 2043–2090, doi: 10.1007/s11192-020-03700-5.

[11] A. El Kah and I. Zeroual, (2021). Arabic Topic Identification: A Decade Scoping Review, *E3S Web Conf.*, vol. 297, 01058, doi: 10.1051/e3sconf/202129701058.

[12] A. Abuzayed and H. Al-Khalifa, (2021). BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique, *Procedia Comput. Sci.*, vol. 189, 191–194, doi: 10.1016/j.procs.2021.05.096.

[13] M. Grootendorst, (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure, *arXiv*: arXiv:2203.05794. doi: 10.48550/arXiv.2203.05794.

[14] A. Mansy, S. Rady, and T. Gharib, (2022). An Ensemble Deep Learning Approach for Emotion Detection in Arabic Tweets, *Int. J. Adv. Comput. Sci. Appl.*, vol. 13(4), doi: 10.14569/IJACSA.2022.01304112.

[15] S. Ben Ali, Z. Kechaou, and A. Wali, (2022). 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, 214–220. doi: 10.1109/ICCICC57084.2022.10101635.

[16] L. George and P. Sumathy, (2023). An integrated clustering and BERT framework for improved topic modeling, *Int. J. Inf. Technol.*, vol. 15(4), 2187–2195, doi: 10.1007/s41870-023-01268-w.

Technologies in Tourism 2024, K. Berezina, L. Nixon, and A. Tuomi, Eds., in Springer

[17] Y. An, H. Oh, and J. Lee, (2023). Marketing Insights from Reviews Using Topic Modeling with BERTopic and Deep Clustering Network, *Appl. Sci.*, vol. 13(16), 9443, doi: 10.3390/app13169443.

[18] H. Rahimi, H. Naacke, C. Constantin, and B. Amann, (2023). ATEM: A Topic Evolution Model for the Detection of Emerging Topics in Scientific Archives, *arXiv*: arXiv:2306.02221. doi: 10.48550/arXiv.2306.02221.

[19] S. Al-Khalifa, F. Alhumaidhi, H. Alotaibi, and H. S. Al-Khalifa, (2023). ChatGPT across Arabic Twitter: A Study of Topics, Sentiments, and Sarcasm, *Data*, vol. 8(11), 171, doi: 10.3390/data8110171.

[20] V. De Leo, M. Puliga, M. Bardazzi, F. Capriotti, A. Filetti, and A. Chessa, (2023). Topic detection with recursive consensus clustering and semantic enrichment, *Humanit. Soc. Sci. Commun.*, vol. 10(1), 197, doi: 10.1057/s41599-023-01711-0.

[21] 2024-A BERT-based Prototypical.

[22] A. Boutaleb, J. Picault, and G. Grosjean, (2024). BERTrend: Neural Topic Modeling for Emerging Trends Detection, in *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, Miami, Florida, USA: Association for Computational Linguistics, 1–17. doi: 10.18653/v1/2024.futured-1.1.

[23] Y. Mu, C. Dong, K. Bontcheva, and X. Song, (2024). Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling, *arXiv*: arXiv:2403.16248. doi: 10.48550/arXiv.2403.16248.

[24] A. Abdelali *et al.*, (2024). LAraBench: Benchmarking Arabic AI with Large Language Models, *arXiv*: arXiv:2305.14982. doi: 10.48550/arXiv.2305.14982.

[25] M. Alghaslan and K. Almutairy, (2024). MGKM at StanceEval2024 Fine-Tuning Large Language Models for Arabic Stance Detection, in *Proceedings of The Second Arabic Natural Language Processing Conference*, Bangkok, Thailand: Association for Computational Linguistics, 816–822. doi: 10.18653/v1/2024.arabicnlp-1.95.

[26] Md. R. Hossain, M. M. Hoque, N. Siddique, and M. A. A. Dewan, (2024). AraCovTexFinder: Leveraging the transformer-based language model for Arabic COVID-19 text identification, *Eng. Appl. Artif. Intell.*, vol. 133, 107987, doi: 10.1016/j.engappai.2024.107987.

[27] A. Kirilenko and S. Stepchenkova, (2024). Automated Topic Analysis with Large Language Models, in *Information and Communication*

Proceedings in Business and Economics., Cham: Springer Nature Switzerland, 29–34. doi: 10.1007/978-3-031-58839-6_3.

[28] A. N. Tarekegn, (2024). Large Language Model Enhanced Clustering for News Event Detection, *arXiv*. doi: 10.48550/ARXIV.2406.10552.

[29] T. Doi, M. Isonuma, and H. Yanaka, (2024). Comprehensive Evaluation of Large Language Models for Topic Modeling, *arXiv*: arXiv:2406.00697. doi: 10.48550/arXiv.2406.00697.

[30] M. S. A. Alzaidi *et al.*, (2025). Enhanced automated text categorization via Aquila optimizer with deep learning for Arabic news articles, *Ain Shams Eng. J.*, vol. 16(1), 103189, doi: 10.1016/j.asej.2024.103189.